Multimodal Methods for Analyzing Learning and Training Environments: A Systematic Literature Review

CLAYTON COHN, Vanderbilt University, USA
EDUARDO DAVALOS, Vanderbilt University, USA
CALEB VATRAL, Tennessee State University, USA
JOYCE HORN FONTELES, Vanderbilt University, USA
HANCHEN DAVID WANG, Vanderbilt University, USA
AUSTIN COURSEY, Vanderbilt University, USA
SURYA RAYALA, Vanderbilt University, USA
ASHWIN T S, Vanderbilt University, USA
MEIYI MA, Vanderbilt University, USA
GAUTAM BISWAS, Vanderbilt University, USA

This document serves as the appendix to our literature review on multimodal methods applied to learning and training environments. It provides supplementary material not included in the main manuscript, including a comprehensive table of all publications in the review corpus, a detailed description of the literature search and screening procedures, discussion of the limitations of the review, and extended results.

CCS Concepts: • **Applied computing** → **Education**; **Computer-assisted instruction**; **Interactive learning environments**; **Collaborative learning**; **E-learning**; **Computer-managed instruction**;

Additional Key Words and Phrases: multimodal data, data analytics, learning analytics, multimodal learning analytics, mmla, learning environments, training environments

ACM Reference Format:
Clayton Cohn, Eduardo Davalos, Caleb Vatral, Joyce Horn Fonteles, Hanchen David Wang, Austin Coursey, Surya Rayala, Ashwin T S, Meiyi Ma, and Gautam Biswas. 2025. Multimodal Methods for Analyzing Learning and Training Environments: A Systematic Literature Review. 1, 1 (December 2025), 46 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

This work is supported under National Science Foundation grants IIS-2327708, DRL-2112635, and IIS-2017000; and US Army CCDC Soldier Center Award #W912CG2220001. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or United States Government, and no official endorsement by either party should be inferred.
Authors' Contact Information: Clayton Cohn, clayton.a.cohn@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Eduardo Davalos, eduardo.davalos.anaya@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Caleb Vatral, cvatral@tnstate.edu, Tennessee State University, Nashville, TN, USA; Joyce Horn Fonteles, joyce.h.fonteles@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Hanchen David Wang, hanchen.wang.1@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Austin Coursey, austin.c.coursey@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Surya Rayala, surya.chand.rayala@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Ashwin T S, ashwindixit9@gmail.com, Vanderbilt University, Nashville, TN, USA; Meiyi Ma, meiyi.ma@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA; Gautam Biswas, gautam.biswas@vanderbilt.edu, Vanderbilt University, Nashville, TN, USA.

A Corpus Table

Table 1 enumerates the 122 papers in this literature review's corpus.

UUID	First Author	Title	Year	Publication	Corpus
2456887548 [5]	Alyuz	An Unobtrusive And Multimodal Approach For Behavioral Engagement Detection Of Students	2017	MIE	A
818492192 [7]	Andrade	Understanding Student Learning Trajectories Using Multimodal Learning Analytics Within An Embodied-Interaction Learning Environment	2017	LAK	A
425012016 [8]	Anton	The Human Condition: Modal and Interactive Advantages of Teacher over AI Feedback on Children's Mathematical Performance	2025	IDC	B
3637456466 [10]	Ashwin	Impact Of Inquiry Interventions On Students In E-Learning And Classroom Environments Using Affective Computing Framework	2020	UMUAI	A
3448122334 [12]	Aslan	Investigating The Impact Of A Real-Time, Multimodal Student Engagement Analytics Technology In Authentic Classrooms	2019	CHI	A
2668965770 [11]	Aslan	Exploring kid space in the wild: a preliminary study of multimodal and immersive collaborative play-based learning experiences	2022	ETRD	B
1886134458 [13]	Azcona	Personalizing Computer Science Education By Leveraging Multimodal Learning Analytics	2018	FIE	A
3146393211 [16]	Birt	Mobile Mixed Reality For Experiential Learning And Simulation In Medical And Health Sciences Education	2018	Information	A
1326191931 [22]	Chan	Multimodal Learning Analytics In A Laboratory Classroom	2019	MLPALA	A
4089325423 [23]	Chan	Predicting behavior change in students with special education needs using multimodal learning analytics	2023	Access	B
2936220551 [25]	Chango	Multi-Source And Multimodal Data Fusion For Predicting Academic Performance In Blended Learning University Courses	2020	CEE	A
4277812050 [26]	Chango	Improving Prediction Of Students' Performance In Intelligent Tutoring Systems Using Attribute Selection And Ensembles Of Different Multimodal Data Sources	2021	JCHE	A

1196965665 [28]	Chejara	How to build more generalizable models for collaboration quality? lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics	2023	LAK	B
1731146538 [29]	Chejara	Impact of window size on the generalizability of collaboration quality estimation models developed using Multimodal Learning Analytics	2023	LAK	B
1426267857 [31]	Chen	Affect, Support, And Personal Factors: Multimodal Causal Models Of One-On-One Coaching	2021	JEDM	A
2764645776 [32]	Chen	MindScratch: A Visual Programming Support Tool for Classroom Learning Based on Multimodal Generative AI	2025	IJHCI	B
328477558 [30]	Chen	Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs Human expert	2025	CompEdu	B
1225141845 [33]	Cheung	Exploring students' multimodal representations of ideas about epistemic reading of scientific texts in generative AI tools	2025	JSET	B
3304069824 [36]	Civit	Class integration of ChatGPT and learning analytics for higher education	2024	Expert Sys	B
3809293172 [37]	Closser	Blending Learning Analytics And Embodied Design To Model Students' Comprehension Of Measurement Using Their Actions, Speech, And Gestures	2021	IJCCI	A
570697424 [46]	Cohn	A multimodal approach to support teacher, researcher and AI collaboration in STEM+ C learning environments	2025	BJET	B
3537775194 [54]	Contero	Personalized and Timely Feedback in Online Education: Enhancing Learning with Deep Learning and Large Language Models	2025	MTI	B
4019205162 [50]	Cornide-Reyes	Introducing Low-Cost Sensors Into The Classroom Settings: Improving The Assessment In Agile Practices With Multimodal Learning Analytics	2019	Sensors	A
2846172025 [51]	Cosentino	Generative AI and multimodal data for educational feedback: Insights from embodied math learning	2025	BJET	B
1576545447 [56]	Cukurova	Artificial Intelligence And Multimodal Data In The Service Of Human Decision-Making: A Case Study In Debate Tutoring	2019	BJET	A
1609706685 [57]	Di Mitri	Learning Pulse: A Machine Learning Approach For Predicting Performance In Self-Regulated Learning Using Multimodal Data	2017	LAK	A

2070224207 [131]	Di Mitri	Detecting Medical Simulation Errors With Machine Learning And Multimodal Data	2019	CAIM	A
3009548670 [60]	Di Mitri	Real-Time Multimodal Feedback With The Cpr Tutor	2020	AIED	A
1763513559 [58]	Di Mitri	Keep Me In The Loop: Real-Time Feedback With Multimodal Data	2021	IJAIED	A
1296637108 [64]	Echeverria	Towards Collaboration Translucence: Giving Meaning To Multimodal Group Data	2019	CHI	A
1040787959 [65]	Echeverria	TeamSlides: A multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space	2024	LAK	B
1581261659 [67]	Emerson	Early Prediction Of Visitor Engagement In Science Museums With Multimodal Learning Analytics	2020	ICMI	A
1598166515 [66]	Emerson	Multimodal Learning Analytics For Game-Based Learning	2020	BJET	A
4035649049 [70]	Fernández-Nieto	Storytelling With Learner Data: Guiding Student Reflection On Multimodal Team Data	2021	TLT	A
151988148 [72]	Fernández-Nieto	Data storytelling editor: A teacher-centred tool for customising learning analytics dashboard narratives	2024	LAK	B
483140962 [76]	Fwa	Investigating Multimodal Affect Sensing In An Affective Tutoring System Using Unobtrusive Sensors	2018	PPIG	A
4278392816 [78]	Giannakos	Multimodal Data As A Means To Understand The Learning Experience	2019	IJIM	A
2243240858 [81]	Goslen	Llm-based student plan generation for adaptive scaffolding in game-based learning environments	2025	IJAIED	B
853680639 [84]	Henderson	Sensor-Based Data Fusion For Multimodal Affect Detection In Game-Based Learning Environments	2019	EDM	A
3398902089 [95]	Järvelä	What Multimodal Data Can Tell Us About The Students' Regulation Of Their Learning Process?	2019	LAI	A
86191824 [91]	Jiang	Examining How Different Modes Mediate Adolescents' Interactions During Their Collaborative Multimodal Composing Processes	2019	ILE	
141378338 [89]	Jiang	How Did the Generative Artificial Intelligence-Assisted Digital Multimodal Composing Process Facilitate the Production of Quality Digital Multimodal Compositions: Toward a Process-Genre Integrated Model	2025	TESQ	B

2166765216 [93]	Jin	Chatting with a learning analytics dashboard: The role of generative AI literacy on learner interaction with conventional and scaffolding chatbots	2025	LAK	B
2280467946 [116]	Kim	Multimodal Writing Evaluation in Digital Storytelling using Video-Based Output: Comparing performance of AI and Human Raters.	2024	ICMET	B
32184286 [100]	Kubsch	Once More With Feeling: Emotions In Multimodal Learning Analytics	2022	MMLA Handbook	A
205660768 [101]	Larmuseau	Multimodal Learning Analytics To Investigate Cognitive Load During Online Problem Solving	2020	BJET	A
1877483551 [105]	Lee-Cultura	Motion-Based Educational Games: Using Multi-Modal Data To Predict Player'S Performance	2020	COG	A
3660066725 [102]	Lee-Cultura	Children'S Play And Problem Solving In Motion-Based Educational Games: Synergies Between Human Annotations And Multi-Modal Data	2021	IDC	A
3856280479 [103]	Lee-Cultura	Children'S Play And Problem-Solving In Motion-Based Learning Technologies Using A Multi-Modal Mixed Methods Approach	2021	IJCCI	A
962997360 [107]	Lehtonen	Multimodal Communication and Peer Interaction during Equation-Solving Sessions with and without Tangible Technologies	2023	MTI	B
2429627610 [109]	Lin	Advancing self-directed learning in STEM education: integrating GPT-based learning aid with multimodal learning analytics	2025	JRTE	B
227355655 [108]	Lin	Recognitions of image and speech to improve learning diagnosis on STEM collaborative activity for precision education	2024	EIT	B
804659204 [115]	Liu	Towards Smart Educational Recommendations With Reinforcement Learning In Classroom	2018	TALE	A
3783339081 [114]	Liu	A Novel Method For The In-Depth Multimodal Analysis Of Student Learning Trajectories In Intelligent Tutoring Systems	2018	JLA	A
3796180663 [113]	Liu	Learning Linkages: Integrating Data Streams Of Multiple Modalities And Timescales	2018	JCAL	A
1161441004 [112]	Liu	Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing	2024	CompEdu	B

518268671 [118]	López	Using Multimodal Learning Analytics To Explore Collaboration In A Sustainability Co-Located Tabletop Game	2021	ECGBL	A
566043228 [21]	Ma	Automatic Student Engagement In Online Learning Environment Based On Neural Turing Machine	2021	IJIET	A
3754172825 [119]	Ma	Detecting Impasse During Collaborative Problem Solving With Multimodal Learning Analytics	2022	LAK	A
147203129 [121]	Mangaroska	Multimodal Learning Analytics To Inform Learning Design: Lessons Learned From Computing Education	2020	JLA	A
603534886 [120]	Mangaroska	Exploring students’ cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity	2022	JCAL	B
1847468084 [123]	Martin	Computationally Augmented Ethnography: Emotion Tracking And Learning In Museum Games	2019	ICQE	A
2879332689 [124]	Martinez-Maldonado	From Data To Insights: A Layered Storytelling Approach For Multimodal Learning Analytics	2020	CHI	A
549526582 [125]	Martinez-Maldonado	Lessons learnt from a multimodal learning analytics deployment in-the-wild	2023	TOCHI	B
2737776963 [128]	Milesi	“It’s Really Enjoyable to See Me Solve the Problem like a Hero”: GenAI-enhanced Data Comics as a Learning Analytics Tool	2024	CHI EA	B
1552158788 [129]	Mills	Smart glasses for 3D multimodal composition	2025	LMT	B
1278817005 [132]	Moon	Using multimodal learning analytics as a formative assessment tool: Exploring collaborative dynamics in mathematics teacher education	2024	JCAL	B
2155422499 [133]	Morell	A Multimodal Analysis Of Pair Work Engagement Episodes: Implications For Emi Lecturer Training	2022	JEAP	A
190066185 [134]	Mzwri	Bridging LMS and Generative AI: Dynamic Course Content Integration (DCCI) for Connecting LLMs to Course Content–The Ask ME Assistant	2025	JCE	B
2273914836 [135]	Nasir	Many Are The Ways To Learn Identifying Multi-Modal Behavioral Profiles Of Collaborative Learning In Constructivist Activities	2022	IJCSCL	A
1469065963 [136]	Nguyen	Examining Socially Shared Regulation And Shared Physiological Arousal Events With Multimodal Learning Analytics	2022	BJET	A

3224774131 [138]	Nguyen	Providing Automated Feedback on Formative Science Assessments: Uses of Multimodal Large Language Models	2025	LAK	B
3888330750 [71]	Nieto	Beyond the learning analytics dashboard: Alternative ways to communicate student data insights combining visualisation, narrative and storytelling	2022	LAK	B
2345021698 [140]	Noël	Exploring Collaborative Writing Of User Stories With Multimodal Learning Analytics: A Case Study On A Software Engineering Course	2018	Access	A
2609260641 [142]	Noël	Visualizing Collaboration In Teamwork: A Multimodal Learning Analytics Platform For Non-Verbal Communication	2022	DAMLE	A
2497456347 [145]	Ochoa	The Rap System: Automatic Feedback Of Oral Presentation Skills Using Multimodal Analysis And Low-Cost Sensors	2018	LAK	A
2634033325 [144]	Ochoa	Controlled Evaluation Of A Multimodal System To Improve Oral Presentation Skills In A Real Learning Setting	2020	BJET	A
3051560548 [146]	Olsen	Temporal Analysis Of Multimodal Data To Predict Collaborative Learning Outcomes	2020	BJET	A
116733479 [147]	Ouyang	Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course	2023	ETHE	B
2005607968 [203]	Ouyang	Multimodal learning analytics of collaborative patterns during pair programming in higher education	2023	ETHE	B
2995141815 [148]	Ouyang	An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective	2023	IJCSCL	B
123412197 [149]	Papamitsiou	Utilizing Multimodal Data Through Fsqca To Explain Engagement In Adaptive Learning	2020	TLT	A
85990093 [151]	Petukhova	Multimodal Markers Of Persuasive Speech : Designing A Virtual Debate Coach	2017	INTERSPEECH	A
957160695 [150]	Petukhova	Virtual Debate Coach Design: Assessing Multimodal Argumentation Performance	2017	ICMI	A
1374035721 [152]	Pham	Attentivelearner2: A Multimodal Approach For Improving Mooc Learning On Mobile Devices	2017	AIED	A

2836996318 [153]	Pham	Predicting Learners' Emotions In Mobile Mooc Learning Via A Multimodal Intelligent Tutor	2018	ITS	A
3135645357 [155]	Prieto	Multimodal Teaching Analytics: Automated Extraction Of Orchestration Graphs From Wearable Sensor Data	2018	JCAL	A
3408664396 [157]	Psaltis	Multimodal Student Engagement Recognition In Prosocial Games	2017	T-CIAIG	A
3308658121 [159]	Reilly	Exploring Collaboration Using Motion Sensors And Multi-Modal Learning Analytics	2018	EDM	A
1500258376 [161]	Sabuncuoglu	Developing a multimodal classroom engagement analysis dashboard for higher-education	2023	PACM HCI	B
1844320601 [162]	Santhosh	Gaze-Driven Adaptive Learning System with ChatGPT-Generated Summaries	2024	Access	B
3625722965 [126]	Sanusi	Table Tennis Tutor: Forehand Strokes Classification Based On Multimodal Data And Neural Networks	2021	Sensors	A
2000036002 [170]	Sharma	Predicting Learners' Effortful Behaviour In Adaptive Assessment Using Multimodal Data	2020	LAK	A
261302708 [104]	Sharma	Multimodal teacher dashboards: Challenges and opportunities of enhancing teacher insights through a case study	2023	TLT	B
780281159 [172]	Smith	Multimodal composing with generative AI: Examining preservice teachers' processes and perspectives	2025	CompComp	B
1118315889 [178]	Spikol	Using Multimodal Learning Analytics To Identify Aspects Of Collaboration In Project-Based Learning	2017	CSCL	A
3339002981 [180]	Spikol	Estimation Of Success In Collaborative Learning Based On Multimodal Learning Analytics Features	2017	ICALT	A
1637690235 [179]	Spikol	Supervised Machine Learning In Multimodal Learning Analytics For Estimating Success In Project-Based Learning	2018	JCAL	A
3796643912 [181]	Standen	An Evaluation Of An Adaptive Learning System Based On Multimodal Affect Recognition For Learners With Intellectual Disabilities	2020	BJET	A
2181637610 [182]	Starr	Toward Using Multi-Modal Learning Analytics To Support And Measure Collaboration In Co-Located Dyads	2018	ICLS	A
1315379489 [184]	Sümer	Multimodal Engagement Analysis From Facial Videos In The Classroom	2021	TAC	A

3093310941 [185]	Tanaka	Embodied Conversational Agents For Multimodal Automated Social Skills Training In People With Autism Spectrum Disorders	2017	PLOS	A
1345598079 [186]	Tancredi	Intermodality In Multimodal Learning Analytics For Cognitive Theory Development: A Case From Embodied Design For Mathematics Learning	2022	MMLA Handbook	A
1687167932 [187]	Tang	Using multimodal analytics to systemically investigate online collaborative problem-solving	2022	DistEdu	B
1285699194 [63]	Tang	A multimodal analysis of college students' collaborative problem solving in virtual experimentation activities: A perspective of cognitive load	2023	JCHE	B
433919853 [192]	Tisza	Understanding Fun In Learning To Code: A Multi-Modal Data Approach	2022	IDC	A
1770989706 [195]	Vrzakova	Focused Or Stuck Together: Multimodal Patterns Reveal Triads' Performance In Collaborative Problem Solving	2020	LAK	A
2055153191 [196]	Vujovic	Round Or Rectangular Tables For Collaborative Problem Solving? A Multimodal Learning Analytics Study	2020	BJET	A
3095923626 [200]	Worsley	A Multimodal Analysis Of Making	2017	IJAIED	A
3309250332 [199]	Worsley	(Dis)Engagement Matters: Identifying Efficacious Learning Practices With Multimodal Learning Analytics	2018	LAK	A
666050348 [201]	Worsley	Multicraft: A Multimodal Interface For Supporting And Studying Learning In Minecraft	2021	HCII	A
1441411748 [111]	Wu	Enhancing self-directed learning and Python mastery through integration of a large language model and learning analytics dashboard	2025	BJET	B
3313249608 [202]	Xu	Classroom Simulacra: Building Contextual Student Generative Agents in Online Education for Learning Behavioral Simulation	2025	CHI	B
3522635517 [204]	Yan	Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning	2024	BJET	B
1019093033 [206]	Yang	Prime: Block-Wise Missingness Handling For Multi-Modalities In Intelligent Tutoring Systems	2019	MMM	A
1436887306 [110]	Yeh	Enhancing EFL vocabulary learning with multimodal cues supported by an educational robot and an IoT-Based 3D book	2022	System	B

177743022 [207]	You	AI-Driven Intelligent Learning Companions: A Multimodal Fusion Framework for Personalized Education	2025	WOCC	B
1935812764 [208]	Yusuf	Using multimodal learning analytics to model students' learning behavior in animated programming classroom	2024	EIT	B
1675503665 [209]	Zapata	AI and peer reviews in higher education: students' multimodal views on benefits, differences and limitations	2025	TPE	B
2737977054 [14]	Zhang	Can AI-generated pedagogical agents (AIPA) replace human teacher in picture book videos? The effects of appearance and voice of AIPA on children's learning	2025	EIT	B
209328204 [213]	Zhao	METS: Multimodal learning analytics of embodied teamwork learning	2023	LAK	B
3602263061 [212]	Zhao	Towards automated transcribing and coding of embodied teamwork communication through multimodal learning analytics	2024	BJET	B

Table 1. Each of the 122 works in our corpus.

B Corpus Distillation Procedure

This appendix contains a detailed account of the steps we took to gather relevant works for our literature review and how we distilled the initial search results to the 73 and 49 papers for Corpora A and B, respectively.

B.1 Literature Search

The literature search for both corpora was based on search strings collaboratively defined and agreed upon by the authors as representative of the target research space. Rather than conducting queries manually, we used SerpAPI [165], a third-party Google Scholar scraping API selected for its ability to return organic search results—unlike alternatives such as `scholarly` [35] and `gscholar` [194], whose outputs differ from browser-based queries.

For Corpus A, we queried Google Scholar via API for papers published between January 2017 and October 2022. The 2017 cutoff was chosen to capture developments from the past five years while excluding earlier foundational work, which is discussed in Sections 1 and 2 of the main manuscript but not included in the corpus. The Corpus B search was conducted in August 2025 and backdated to begin in November 2022, covering the period following the release of ChatGPT. We timed the search to follow major conference publication cycles (LAK, AIED, EDM, and L@S) to ensure comprehensive coverage.

The Corpus A search included 14 distinct phrases, each queried three times using variations of the word *multimodal* (*multimodal*, *multi-modal*, and *multi modal*) as prefixes.[1] For Corpus B, we used 12 updated queries reflecting recent developments in GenAI and LLMs, employing only the standard spelling of *multimodal* after confirming that alternative spellings had no impact on results. We also omitted broad terms such as "multimodal survey" and "multimodal literature review," which surfaced naturally in other targeted searches. The complete list of search phrases is shown in Table 2.

Table 2. Full Corpus Search Terms

education technology
explainable artificial intelligence
learning analytics
learning environments
learning environments literature review
learning environments survey
literature review
simulation environments
survey
training environments
training environments literature review
training environments survey
tutoring systems
xai

(a) Corpus A Search Terms

education technology
learning analytics
learning environments
training environments
simulation environments
llm learning environments
llm training environments
llm learning analytics
pedagogical agents
llm pedagogical agents
ChatGPT in education
generative AI in education

(b) Corpus B Search Terms

For each search string, we collected the top five pages (100 publications) returned by Google Scholar. This top-5 cutoff was imposed for practical and financial reasons related to the subsequent construction of a citation graph (see

[1] The term "xai" was included to identify works on explainable AI in learning and training contexts; however, no relevant results were returned during the initial search.

Appendix B.2.1). SerpAPI limits citation queries to 20 citations per API call, requiring multiple calls for highly cited papers (e.g., five calls for a paper with 100 citations). Without a cutoff, the number of API calls would become intractable.

The initial search yielded 4,200 papers for Corpus A (14 search terms ×3 multimodal spelling variants ×100 results) and 1,200 papers for Corpus B (12 search terms ×1 multimodal spelling variant). The full corpus reduction procedure is detailed in Table 3 and discussed in the following subappendices. Each step is referenced using the corresponding Step ID in Table 3.

Step	Procedure	Removed A	Remain A	Removed B	Remain B
0	Literature search	0	4200	0	1200
1	Remove duplicates	2079	2121	355	845
2	Remove non-English	1	2120	0	845
3	Remove degree-0 nodes/disconnected components	589	1531	33	812
4	Iteratively remove degree-1 nodes	468	1063	253	559
5	Title reads	675	388	305	254
6	Abstract reads	261	127	110	144
7	Full paper reads	54	**73**	95	**49**

Table 3. Corpus reduction procedure.

We removed 2,079 duplicates from Corpus A and 355 from Corpus B by hashing paper titles (Table 3, Step 1), retaining the official published version when multiple copies existed. We then excluded one non-English paper from Corpus A (Step 2), identified using spaCy FastLang [191] and verified through manual inspection. After these steps, the combined search yielded 2,120 unique English-language papers for Corpus A and 855 for Corpus B.

B.2 Study Selection

To reduce the corpora to a reviewable set, we applied both quantitative and qualitative methods. First, we performed citation graph pruning (CGP) to distill the corpus algorithmically (Appendix B.2.1). This was followed by qualitative screening, detailed in Appendix B.2.2.

B.2.1 Citation Graph Pruning (Quantitative Corpus Reduction). For visualization, analysis, and corpus distillation, we used NetworkX [82] to construct a directed citation graph for all remaining papers. Each node corresponds to a paper identified by its Google Scholar UUID, and each directed edge denotes a citation from one corpus paper to another. Following SerpAPI's "cited by" results, only inbound citation queries were required; citations from papers outside the list of remaining papers were ignored.

We first removed all 0-degree nodes and disconnected components (Step 3)—papers that neither cited nor were cited by any other paper in the corpus and components with no edges to or from the primary (i.e., largest by number of nodes) component. Because incoming and outgoing citations jointly determine degree, this approach balances early papers (with few outgoing edges) and recent papers (with few incoming edges). Step 3 removed 589 papers from Corpus A and 33 from Corpus B, resulting in connected citation graphs of 1,531 and 812 papers, respectively.

We then applied iterative degree-1 pruning (Step 4), removing nodes with only one citation edge and repeating the process until none remained. Corpus A required four iterations, removing 468 papers and yielding 1,063; Corpus B required two, removing 253 and yielding 559. This approach allowed us to eliminate loosely connected papers unlikely to be central to the field. Given that multimodal learning and training research spans multiple disciplines (e.g., computer science, education, psychology), the authors agreed that papers with minimal citation connectivity were unlikely to meet the scope of this review. The CGP algorithm is detailed in Algorithm 1.

Algorithm 1 Citation Graph Pruning Algorithm

Require: Acyclic directed graph $G = (V, E)$

procedure DEGREE TRIMMING(G, n)
    $S, D \leftarrow \{\}, \{\}$
    for all $v \in V$ **do**
        if $\deg(v) <= n$ **then** $S = S \cup \{v\}$
    for all $v \in S$ **do**
        for all $e \in E$ **do**
            if $v \in e \land e \notin D$ **then** $D = D \cup \{e\}$
    return $(V \setminus S, E \setminus D)$
procedure SUBCONNECTED GRAPH TRIMMING(G)
    $[S_1, S_2, S_3, ..., S_n] = \text{ConnectedComponent}(G)$, where each $S_i = (V_i, E_i)$
    $j = \arg\max\{|V_1|, |V_2|, |V_3|, ..., |V_n|\}$
    return (V_j, E_j)
procedure ITERATIVE TRIMMING(G)
    while True **do**
        $G' = \text{DegreeTrimming}(G, 1)$, where $G' = (V', E')$
        if $|V| == |V'|$ **then**
            break
    return (V', E')
$G' = \text{DegreeTrimming}(G, 0)$ ▷ Remove 0-deg vertices
$G' = \text{SubconnectedGraphTrimming}(G')$ ▷ Keep largest connected subgraph
$G' = \text{IterativeTrimming}(G')$ ▷ Iteratively remove 1-deg vertices until equilibrium
return G'

At this point, we concluded the quantitative pruning procedure. The resulting citation graphs served as the basis for subsequent qualitative screening.

B.2.2 Quality Control (Qualitative Corpus Reduction). Following quantitative pruning, qualitative screening further reduced each corpus according to the procedures summarized in Table 3. For Corpus A, the remaining 1,063 papers proceeded through title, abstract, and full-paper review. For Corpus B, due to time constraints, we used an LLM-as-a-Judge workflow [214] for title, abstract, and full-paper decisions, with human verification on the final distilled set.

Title Screening. For Corpus A, four reviewers independently evaluated all 1,063 titles for relevance to multimodal learning or training. Inclusion and exclusion were determined by majority vote, with ties resolved by a fifth reviewer. This resulted in 388 retained titles and 675 exclusions (Table 3, Step 5 for A). For Corpus B, title decisions were made jointly by GPT-4o and Gemini 2.5; agreement between both models determined inclusion or exclusion, and disagreements were adjudicated by a human reviewer. Title screening retained 254 papers, excluding 305 (Step 5 for B).

Abstract Screening. For abstract screening (Step 6), each Corpus A abstract was reviewed by two reviewers using the exclusion criteria in Table 4. Papers without unanimous reviewer agreement underwent a second round of review using majority voting. This yielded 127 retained abstracts and 261 exclusions. For Corpus B, both LLM judges independently evaluated all abstracts under the same criteria. Agreement resulted in automatic inclusion or exclusion; disagreements were resolved by a human reviewer. A total of 144 abstracts were retained.

Full-Paper Screening. Full-paper review followed the same exclusion framework with two additional criteria introduced during reading (Table 5). For Corpus A, 127 papers were divided among five reviewers. Papers were labeled "immediate accept," "immediate exclude," or "borderline." Exclusion required unanimous agreement across all reviewers. After this stage, 73 papers remained (Table 3, Step 7). For Corpus B, both LLM judges evaluated all 144 papers end-to-end using the cumulative exclusion criteria, selecting 79 papers (including a human tie-breaker) for inclusion. Two human reviewers then manually reviewed and discussed each of these papers to assess their alignment with the scope of this review. Based on consensus coding [34], 30 papers were excluded, resulting in a final set of 49 papers. This human-in-the-loop validation ensured that all retained papers met the inclusion criteria and were within the scope of this literature review.

1. Paper does not involve a learning or training environment
2. Environment is VR-only
3. No multimodal data are analyzed
4. No multimodal analysis methods are applied
5. Paper is not original applied research

Table 4. Exclusion criteria for abstract screening.

Across both corpora, qualitative screening ensured that only papers presenting original multimodal methods applied to learning or training environments advanced to the final analysis set: 73 papers for Corpus A and 49 for Corpus B.

1. Results are not informative about learning or training
2. Analysis methods cannot be determined from the manuscript

Table 5. Additional exclusion criteria for full-paper screening.

B.3 Feature Extraction

Feature extraction was performed after the full paper review stages (Table 3, Steps 7) and was conducted manually by two human reviewers for all 73 papers in Corpus A and all 49 papers in Corpus B. Extracted features included identifying information (e.g., title, author, year) and methodological descriptors (e.g., data collection media, modalities, fusion strategies, and analysis methods). Table 6 lists the initial feature set.

To ensure consistency, feature categories were initially discretized through inductive coding [190]. Four reviewers each coded a portion of the papers in Corpus A to define discrete feature sets. For example, "video camera," "webcam," and "Kinect" were consolidated under the medium "video." Reviewers then re-extracted features into these discrete sets. The resulting circumscribing features are shown in Table 7 (Cohen's $\kappa = 0.87$).

A second Corpus A feature extraction round gathered additional features supporting later analysis. These circumscribing features—environment setting, domain, participant interaction structure, didactic nature, level of instruction or training, analysis approach, and analysis results—are listed in Table 8. All were discretized except analysis results, which were recorded in free form for thematic analysis [20]. As with the first round, feature extraction for the second

Feature	Description
UUID	Universally unique identifier on Google Scholar
Title	Publication title
First Author	Publication's first author
Year	Year first publicly available
Environment Type	Type of environment analyzed
Data Collection Media	Types of data collected
Modalities	Modalities used during analysis
Analysis Methods	Methods applied in the analysis
Fusion Type	Data fusion strategies used
Publication Source	Journal, conference, workshop, etc.

Table 6. Initial features extracted from each paper.

Feature	Feature Set
Environment Type	learning, training
Data Collection Media	video, audio, screen recording, eye tracking, logs, physiological sensor, interview, survey, participant produced artifacts, researcher produced artifacts, motion, text
Modalities	affect, pose, gesture, activity, prosodic speech, transcribed speech, qualitative observation, logs, gaze, interview notes, survey, pulse, EDA, body temperature, blood pressure, EEG, fatigue, EMG, participant artifacts, researcher artifacts, audio spectrogram, text, pixel
Analysis Methods	classification, regression, clustering, qualitative, statistical methods, network analysis, pattern extraction
Fusion Type	early, mid, late, hybrid, other

Table 7. First set of circumscribing features and their feature sets.

feature set of Corpus A involved independent coding by two reviewers followed by consensus. For this round, Cohen's κ prior to consensus was 0.71.

Once the feature sets were finalized, this process was applied to Corpus B using two human reviewers for consensus coding (Cohen's $\kappa = 0.68$ prior to consensus). For each corpus, final feature sets represent agreement between the reviewers who coded each paper.

C Extended Results

Through our inductive analysis of the review corpus, we developed a theoretical framework that captures the core components of multimodal learning and training pipelines along with their interrelations. As illustrated in Figure 1, the framework decomposes the MMLA process into four primary, sequential component processes: (1) the learning or training environment from which student data are collected through sensors, (2) multimodal data and the modalities derived from them, (3) learning analytics for making sense of that data, and (4) feedback for stakeholders like students,

Feature	Feature Set
Environment Setting	physical, virtual, blended, unspecified
Domain of Study	STEM, humanities, psychomotor skills, other, unspecified
Participant Interaction Structure	individual, multi-person
Didactic Nature	instructional, training, informal, unspecified
Level of Instruction or Training	K-12, university, professional development, unspecified
Analysis Approach	model-free, model-based
Feedback	direct, indirect

Table 8. Second set of circumscribing features and their feature sets.

teachers, and researchers. We provide an overview of each component below, followed by subsections presenting taxonomies and findings corresponding to each.

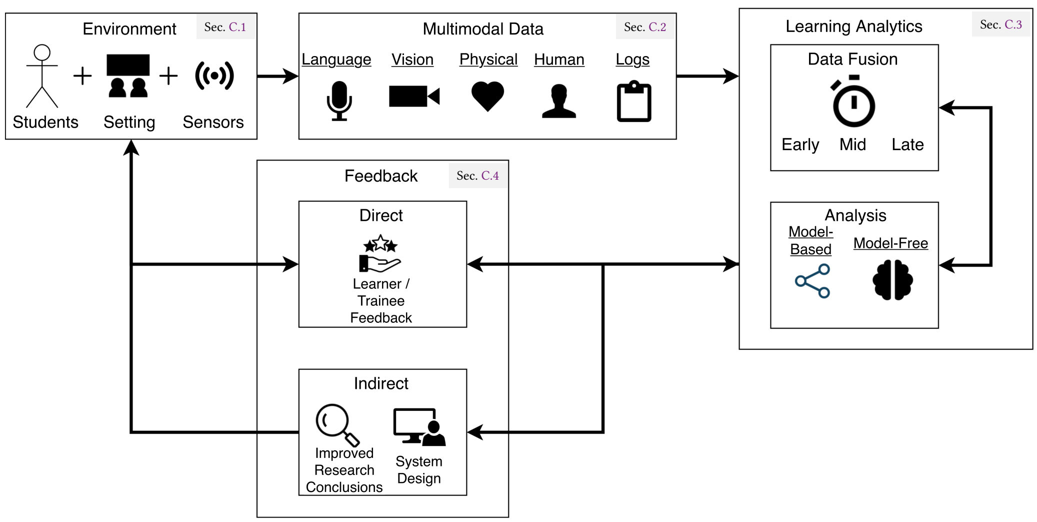

Fig. 1. Multimodal Learning and Training Environments Literature Review Framework

In the following subsections, each framework component is presented via: (1) its significance within the context of multimodal learning and training methodology; (2) a taxonomy derived from data extracted from the reviewed studies; (3) relevant findings, including a comparison of methodologies from the pre-LLM and post-LLM eras and their challenges; and (4) examples of how each component is put into practice.

C.1 Environments

Our paper explores a spectrum of environments on a learning-training continuum (Figure 2). The environments span from traditional classrooms to online courses and are categorized along two dimensions: the learning-training axis [126, 140, 151, 196] and the physical-virtual space continuum [22, 50, 153]. Systems such as nurse training simulations with manikins and embodied learning environments used in K-12 education (where students actively move around the classroom as part of the learning experience [74]) combine physical and virtual elements and are referred to as *mixed-reality* environments [97, 98].

Multimodal methods in learning environments aim to enhance educational outcomes by analyzing student engagement and learning patterns. In contrast, training environments focus on skill acquisition and task proficiency, serving individuals from personal development to professional enhancement across fields such as healthcare [60], athletics [126], the workplace [3, 97], and the military [84]. These settings range from fully virtual simulations to physical training drills, with mixed reality bridging the gap. MMLA objectives differ between learning and training, requiring context-specific strategies. While the distinction between learning and training can be ambiguous, as seen in game-based platforms [123, 201], our review spans this spectrum. We employ a fuzzy qualitative categorization to place each study on this continuum, acknowledging the approach's complexity and utility for analyzing MMLA research sub-communities.

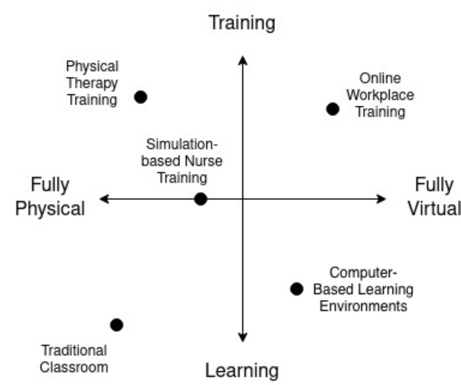

Fig. 2. Learning-Training Continuum

In the following subsections, we present findings for the three components specified in our framework for environment: **learners/trainees** (students), **setting**, and **data collection media** (sensors).

C.1.1 Learners/Trainees ("Students" in Figure 1). Learners and trainees are central to the design, deployment, and evaluation of multimodal learning and training analytics systems. The identity of the participants, the subjects they study, their methods of interaction, and the instructional settings in which they are situated influence the multimodal data that can be collected, the models that can be developed, and how the resulting analytics should be implemented. Therefore, clearly defining these learner characteristics is essential to our framework and provides a consistent perspective for understanding how studies are situated within authentic learning and training environments.

Across both Corpora A and B, we describe the learner context along four dimensions: (1) **domain of study**, (2) **participant interaction structure**, (3) **didactic nature of the environment**, and (4) **level of instruction or training**. The same taxonomy is applied throughout, and individual studies can receive multiple labels along each dimension.

Domain of Study. Our corpus revealed three primary domains of study. **STEM+C** includes Science, Technology, Engineering, Mathematics, and Computing, as well as healthcare and medicine [7, 71, 178]. **Humanities** spans literature, debate, oral presentation, and writing [151, 157, 162]. **Psychomotor Skills** refers to domains emphasizing motor coordination, such as CPR training [131], woodworking [29], and video games like PAC-MAN [78].

Both Corpora A and B primarily focused on STEM+C learning (A: 55/73, 75%; B: 37/49, 76%), covering topics from programming [25] to nursing [70] to geometry and chemistry [28]. Psychomotor skills were less represented (A: 5/73, 7%; B: 1/49, 2%). Corpus B showed increased attention to the humanities (A: 11/73, 15%; B: 13/49, 27%), where LLMs enabled support for open-ended tasks such as multimodal composition and essay writing [112].

A key distinction between domains lies in their level of structure: **structured** domains (e.g., STEM+C) are characterized by constrained problems and clear evaluation metrics. In contrast, **unstructured** domains (e.g., humanities) involve open-ended tasks with subjective or variable outcomes. For instance, algebra problems typically follow well-defined procedures. They can be assessed using rule-based systems such as decision trees [189], whereas creative writing tasks [210] resist formulaic evaluation and are often poorly served by metrics like word count or sentence length [77]. However, with the advent of LLMs, analytical rubrics are being developed that break down writing into specific criteria, such as uniqueness of storyline, logical sequencing of ideas and content (e.g., a connected beginning, middle, and end of the narrative), style of writing, and engaging vocabulary [80, 99].

Structural differences affect how multimodal learning environments are designed and how data are captured and analyzed. Structured domains often afford quantitative, model-based analysis, while unstructured domains typically require model-free, qualitative methods such as thematic or interaction coding [90]. However, the line between the two is becoming increasingly blurred with the integration of LLMs, which can interpret and evaluate multimodal data even in unstructured, dynamic contexts such as embodied learning.

Participant Interaction Structure. Participant interaction structure describes whether learners engage individually (**individual**) or with others (**multi-learner**). Individual settings typically involve a single learner interacting with systems such as educational games [103], intelligent tutoring systems [26], open-ended learning environments [106], and creative platforms [54]. Multi-learner environments include pairs, small groups, or full-class activities such as paired programming [159], game-based competitions [155], and collaborative play [11].

Our corpus reveals a growing emphasis on collaborative learning, with multi-learner studies increasing from 42% (31/73) in Corpus A [64, 146] to 51% (25/49) in Corpus B [148, 204]. These studies leverage multimodal data to examine activities such as collaborative experimentation [63], clinical simulations [213], and group reflections using dashboards [65, 71]. In contrast, individual-focused studies emphasize personalization [207] and self-regulation, often through AI-driven tutors [93], gaze-adaptive systems [162], or LLM-integrated dashboards [111].

Multi-learner settings introduce social interaction as a central dimension, enabling insights not possible in individual contexts. Learners often externalize their thinking, allowing researchers to analyze dialogue [213], coordination [108], and socially shared regulation of learning (SSRL) via audio and video [88]. However, such contexts present analytic and logistical challenges, including smaller effective sample sizes (n) [87] and reduced transcription quality due to classroom noise [17].

Didactic Nature of the Environment. Didactic nature refers to how learning or training is presented to participants. **Instruction** involves formal activities with defined objectives (e.g., courses, labs) [22, 50, 89]. **Training** emphasizes skill development through practice (e.g., clinical simulations, vocational drills) [70, 125, 126]. **Informal** settings lack fixed goals and occur in loosely structured contexts (e.g., game-based learning, exploratory play) [37, 66, 149, 201].

Both corpora are dominated by instruction, with an even stronger emphasis in Corpus B (A: 45/73; 62%; B: 40/49; 82%). For example, Liu et al. [114] analyzed student interactions in a chemistry virtual lab, integrating logs with audio and video to uncover learning difficulties not evident from system data alone. Training environments accounted for 20-25% of both corpora (A: 15/73; 21%; B: 11/49; 23%), such as a simulated social skills trainer for youth with autism using audiovisual cues like head pose and smiling ratio. Informal learning settings declined notably from Corpus A to B (A: 12/73; 16%; B: 1/49; 2%), as LLMs were primarily applied in traditional instructional contexts. For instance, Santhosh et al. [162] combined real-time gaze-based engagement detection with ChatGPT-generated summaries to support reading comprehension while studying.

Training prioritizes repetition and performance [126], while instructional and informal settings differ in design, data, and analysis. Instructional tasks are structured, allowing controlled data capture and model-based analysis [211]. Informal settings are open-ended, generating noisy, varied, and often incomplete data requiring qualitative, model-free, and human-in-the-loop decision-making approaches [6, 177]. Goals also vary: instruction targets conceptual understanding [94], while informal learning fosters creativity, exploration, and inquiry [62].

Level of Instruction or Training. Our framework defines three levels of instruction: (1) **K–12** (primary and secondary education) [67, 89, 146], (2) **University** (undergraduate and graduate) [50, 131, 134], and (3) **Professional Development** (e.g., workplace learning, continuing education) [54, 60, 155].

Both corpora show similar trends across instructional level, skewed heavily toward university learners (A: 36/73; 49%; B: 29/49; 59%). For example, Civit et al. [36] demonstrates that ChatGPT's classroom integration, paired with physiological signals (e.g., galvanic skin response) supports emotional state detection and structured AI-tutoring interventions with college students, improving both engagement and learning outcomes. K–12 settings followed (A: 30/73; 41%; B: 19/49; 39%), while professional development was least represented (A: 5/73; 7%; B: 2/49; 4%).

K-12 and adult learning contexts pose distinct challenges. Research in K-12 settings faces significant ethical and logistical constraints due to the involvement of minors. Multimodal data capture, especially video and physiological sensing, raises privacy and health data concerns, often requiring approval from parents, teachers, administrators, and district officials [45]. The emergence of GenAI raises additional concerns, including student misuse [130] and unintended LLM behaviors [92], making school-based research difficult and often requiring on-site presence.

In contrast, adult learning environments offer greater flexibility, with fewer institutional hurdles and support for both in-person and remote studies. Despite the formative importance of K–12 education—where students acquire foundational knowledge, social skills, problem-solving strategies, and good study habits [19, 68]—the dominance of university-focused research is unsurprising. Expanding the reach of multimodal systems in K–12 contexts will require coordinated efforts from educators, researchers, parents, and policymakers to ensure these technologies are deployed ethically and effectively [171].

C.1.2 Setting. Settings describe where and how multimodal learning and training activities unfold. Whether learners are on virtual platforms, in physical classrooms, in clinical simulations, or in play spaces constrains which traces can be captured and how analytics and AI-based tools can be meaningfully embedded. Setting links sensing choices, models, and interpretations to the realities of computer-based, in-person, and blended scenarios, and clarifies how multimodal learning analytics systems are deployed across different contexts. With both corpora, we characterize setting along two dimensions: (1) **environment function** and (2) **environment interaction setting**.

Environment Function. We distinguish environments by their primary function, in line with this review's dual focus on **learning** [50, 51, 91] and **training** [60, 64, 71] contexts. Some research examines both (e.g., language learning and woodworking in the same study [29]). Learning environments dominate both corpora (A: 57/73; 78% [50, 66, 91]; B: 40/49; 82% [147, 148]), with training making up a smaller portion (A: 16/73; 22%; B: 12/49; 24% [71, 125]).

A key distinction between environments is the level of physical engagement: **active** versus **stationary**. Although exceptions exist—such as embodied learning contexts where students move around the classroom and stationary training tasks like oral presentations [144]—most learning environments involve seated participants interacting with a computer, classroom teacher, or each other. In contrast, training typically entails physical activity, such as movements and interactions with objects, to accomplish a task. This distinction, in turn, shapes both data capture and analysis. In active environments, motion capture, video, and physiological sensors generate complex, high-dimensional data that generally require model-based methods (e.g., deep learning). For example, Vatral et al. [193] employed gradient-boosted regression trees on eye-gaze and speech features to predict nursing trainees' self-efficacy. Extending this multimodal approach, Martinez-Maldonado et al. [125] integrated smartwatches, microphones, and positioning sensors. However, the authors noted the need for standardized, researcher-provided devices to ensure data quality—highlighting current limitations of "bring-your-own-device" strategies for scalable deployment.

Environment Interaction Setting. **Virtual** environments occur entirely online or in simulated spaces without physical co-presence [111, 180, 185]. **Physical** environments involve in-person activity in real-world spaces such as classrooms, labs, training facilities, and clinics [28, 178, 200]. **Blended** settings combine both, as in robotics courses where students program physical robots using online interfaces [104]. A shift is evident across corpora: virtual environments dominated Corpus A (51/73; 70%) [7, 180, 185], while Corpus B was led by physical settings (34/49; 69%) [208, 213], reflecting the post-COVID return to in-person learning.

Virtual environments are easier to scale and monitor, primarily because students tend to focus on their computer screens, which limits their movement and enables streamlined data collection through computer-based logs, screen recordings, and webcam feeds [175]. For example, researchers have combined heart rate data from Fitbits with system logs to investigate PhD students' self-regulated learning [61]. These conditions facilitate large-scale, quantitative, and model-based analyses.

However, virtual learning lacks the authenticity of physical settings, especially for K-12 education [79, 168]. Studying learners *in situ* (i.e., in their natural environment) offers critical insight into real-world learning processes. It enables access to affordances like social interactions using sensors that are not replicable online. Yet physical environments require on-site researcher presence, heightened IRB oversight, and face logistical barriers to scale [4]. Noise, technical failures, and unstructured dynamics often produce incomplete or messy datasets [55]. Consequently, physical settings frequently rely on model-free methods, such as qualitative coding or statistical correlations between observed behaviors and outcomes [90].

C.1.3 Data Collection Media ("Sensors" in Figure 1). Data collection media determine which aspects of learning and training can be observed, modeled, and ultimately supported. Their selection shapes the granularity of multimodal traces, the feasibility of signal fusion, and the types of constructs that can be inferred (e.g., performance, collaboration, reflection; see the following subsections). Media can also be combined to form multimodal signals. For instance, prosodic and semantic features extracted from audio can be fused with visual cues to predict affect [154]. Across both corpora, we identified a common set of data collection media, summarized in Table 9.

Figures 3 and 4 compare the distribution of data modalities across Corpora A and B. Video and audio data collection were prevalent across both corpora, indicating the richness and usefulness of integrating these modalities in multimodal settings. Audio data can yield prosodic information, such as tone, pitch, pauses, and volume, alongside the semantic meaning of the spoken words, which can be fused with other data streams to derive modalities downstream [27, 169]. Video data can be used to derive visual modalities like activity, gesture, pose, gaze, and affect [18, 141].

In the post-COVID era, multimodal learning and training studies experienced notable shifts. As research moved from virtual to physical settings, the use of participant-produced artifacts increased while reliance on environment log data declined. This transition, along with the rise of LLMs, enabled richer forms of textual feedback not dependent on rule-based systems derived from logs. Surveys and interviews also became more prominent, reflecting a growing emphasis on stakeholder agency in system design and validation rather than purely technological advancement [43, 44].

Medium	Definition
Video	Sequences of image frames captured from a camera source [37, 67, 153].
Audio	Audio signals captured by a microphone [150, 151, 185].
Screen Recording (Screen)	Sequences of image frames displaying a device's screen contents [5, 91, 114].
Eye	Eye movement data and gaze points captured by tracking devices [26, 149, 186].
Logs	Participant's actions within the system and its state data [13, 157, 178].
Physiological Sensors (Physical)	Specialized sensors used to gather participants' physiological data [84, 95, 115].
Interview	Structured or unstructured conversations between researchers and participants [16, 126, 142].
Survey	Standardized sets of questions administered to participants [50, 56, 152].
Participant-Produced Artifacts (PPA)	Materials produced by study participants using various mediums, including physical objects created for a task or written responses to formative assessment questions [10, 25, 144].
Researcher-Produced Artifacts (RPA)	Materials produced by the researchers that contribute to analysis and findings, such as observational notes [84, 124, 181].
Motion	Raw motion data collected via various different devices/technologies [60, 126, 196].
Text	Raw textual input [89, 134, 201].

Table 9. Data collection media.

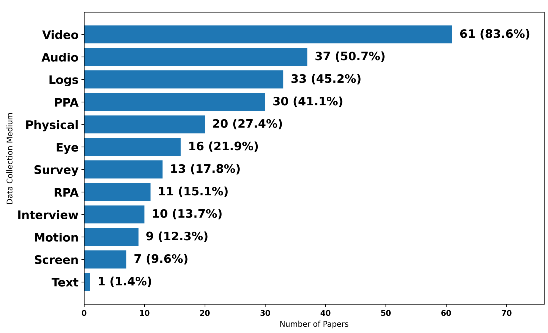

Fig. 3. Corpus A data collection media distribution.

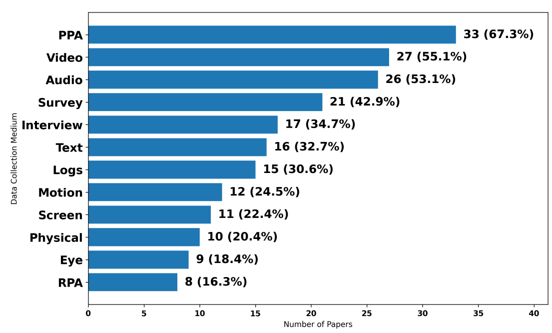

Fig. 4. Corpus B data collection media distribution.

The most striking shift was in textual input: Corpus A included only one study using raw text [201], while nearly one-third of Corpus B papers captured text as a primary data source [54, 109, 111, 134, 202]—almost entirely due to LLM-mediated interactions.

Research goals, target modalities, and participant interaction structures shape data collection methods. For example, studying socially shared regulation of learning requires both environment log data to capture cognitive activity and discourse data to analyze metacognitive and social processes [47, 174, 175]. In contrast, CPR tutor training relies on motion data and physiological signals (e.g., EMG, accelerometers, gyroscopes) to evaluate chest compression performance [60].

However, some modalities pose challenges to adoption. Video and sensor-based methods raise privacy concerns and are often perceived as invasive [122, 156], while LLM-based systems may raise skepticism due to risks of hallucination, toxicity, and misuse [83, 130]. Self-reported measures, such as interviews and surveys, are valuable but must typically be triangulated with other modalities to ensure reliability [143]. Synchronizing data from multiple data streams is challenging and requires standardization, time alignment, and feature engineering. This process is often time-consuming and requires both domain knowledge and technical expertise, creating additional barriers to adoption.

C.2 Multimodal Data

Multimodal data form the foundation of MMLA systems: choices about what to capture, how to represent it, and which streams to combine determine the learner states and processes that can be modeled, the inferences that can be made about their behaviors and performance, and how analytics can inform action. In our framework, multimodal data sit at the intersection of learner behavior and analytics, linking observable activity across multiple modalities to higher-level constructs (e.g., collaboration quality, regulation of self- and group-learning, knowledge and skill acquisition) [96].

These modality choices shape both traditional MMLA pipelines and GenAI-enabled systems by defining system architecture that determines interpretability, robustness, and analytic scope. We adopt a unified taxonomy of five **modality groups**, which partitions modalities based on how they are derived and the information they convey: (1) natural language, (2) vision, (3) physiological signals, (4) human-centered evidence, and (5) logs. Table 10 presents each modality alongside its corresponding modality group(s), setting the stage for the subsequent sections that examine the constructs and analytic methods enabled by each category.

Figures 5 and 6 show the modality distributions for Corpora A and B, largely reflecting trends in data collection media. Prior to the development of LLMs, COVID-era studies (Corpus A) focused on individual modalities such as pose, logs, affect, gaze, and prosodic speech. These modalities were often collected in virtual settings using tools like microphones, webcams, and trace data.

In contrast, Corpus B reflects a shift toward in-person, multi-party, human-centered studies integrating LLMs and sensor-rich physical environments. This newer corpus emphasizes participant-produced artifacts, transcribed speech, physical activity, surveys, and raw textual inputs, captured through 3-D video, lapel microphones, and student-created materials. Several modalities present in Corpus A—such as prosodic speech, spectrogram, EMG, and blood pressure—are notably absent in Corpus B, underscoring a broader shift from physiological signal-based approaches toward artifact- and LLM-centric methods. Importantly, many post-LLM systems employ large language models not as end-to-end multimodal architectures, but as analytic or interpretive layers operating on representations extracted from other modalities.

Approximately two-thirds of papers in both corpora use 3–5 distinct modalities to guide their research (A: 49/73 [5, 57]; 67%, B: 34/49 [30, 134]; 69%). This reflects a methodological balance: the number is sufficient to enable triangulated inferences across heterogeneous signals (e.g., aligning logs with gaze, or artifacts with surveys), while remaining tractable in terms of data collection, synchronization, and model complexity. This norm becomes especially relevant when incorporating more sophisticated analytic systems, such as LLM-based agents, which must operate within practical limits on data richness and annotation effort. The following subsections detail how the modalities within each modality group are operationalized in practice.

Modality	Description	Modality Group
Affect	Participant's facial expression, or emotional or affective state [57, 157, 185].	NLP, Vision, Physical
Pose	Participant's physical position, location, or body posture [5, 179, 182].	Vision, Physical
Gesture	Participant's gestures and body language [7, 151, 200].	Vision
Activity	Participant's observable actions or activities [76, 114, 155].	Vision, Physical
Prosodic Speech (Pros. Speech)	Elements of speech beyond word meaning, e.g. volume, pauses, and intonation [140, 178, 180].	NLP
Transcribed Speech (Trans. Speech)	Textual speech transcribed from audio [16, 50, 113].	NLP
Qualitative Observations (Qual. Obs.)	Researcher observations about the participant and study task [95, 123, 199].	Human-centered
Logs	Participant's environment actions and system state data [13, 78, 131].	Logs
Gaze	Participant's eye gaze, e.g., movement, direction and focus [66, 67, 206].	Vision, Physical
Interview	Notes from interviews between researchers and participants [12, 64, 91].	Human-centered
Survey	Participant's responses to surveys/questionnaires [149, 150, 152].	Human-centered
Pulse	The participant's pulse, indicating their heart rate [102, 103, 192].	Physical
EDA	Participant's electrodermal activity [101, 121, 170].	Physical
Temperature (Temp.)	Participant's body temperature [105, 149, 170].	Physical
Blood Pressure (BP)	Participant's blood pressure [103, 149, 192].	Physical
EEG	Participant's electroencephalography activity [78, 149, 170].	Physical
Fatigue	The level of fatigue experienced during the activity [102, 103].	Vision, Physical
EMG	Participant's electromyography activity [58, 60].	Physical
Participant Produced Artifacts (PPA)	Artifacts produced by the participant during the study, e.g., pre/post-tests [26, 133, 144].	Human-centered
Researcher Produced Artifacts (RPA)	Artifacts produced by the researcher about the study and participants, e.g., field notes [37, 70, 142].	Human-centered
Spectrogram (Spect.)	Representation of audio frequencies in the form of a spectrogram [119].	NLP
Text	Participant's raw text data generated in the study environment [201].	NLP
Pixel	RGB pixel values from cameras or sensors [155].	Vision

Table 10. Modalities, their definitions, and the modality groups they fall into (detailed in Section C.2).

C.2.1 Natural Language. Natural language captures how learners and trainees speak, write, and interact with peers, instructors—and increasingly, LLM-based systems—across modalities such as prosodic and transcribed speech, raw text, and affect derived from language or speech (see Table 10). Because much teaching, collaboration, feedback, and assessment are inherently language-based, NLP signals often encode rich information about learners' metacognition (e.g.,

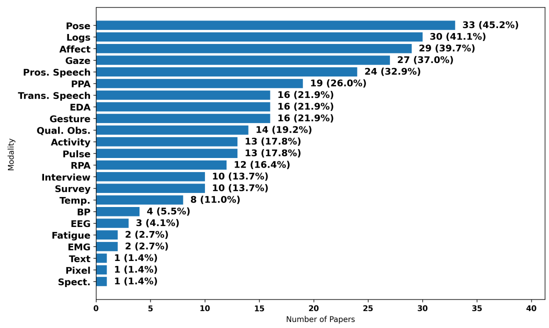

Fig. 5. Corpus A modalities distribution.

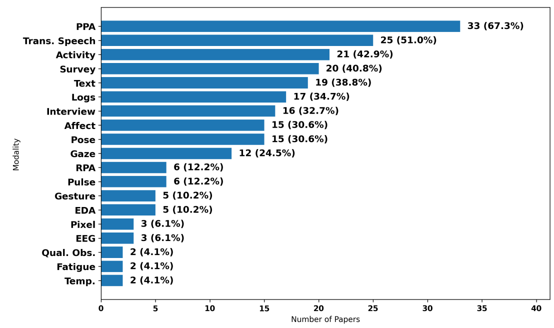

Fig. 6. Corpus B modalities distribution.

goal setting [45], planning [73], and reflective behaviors [175]) as well as collaborative processes such as information pooling and consensus building [176].

Natural language is also frequently used to contextualize other modalities, including gaze, posture, and interaction logs. For example, Snyder et al. [174] employed Markov modeling to infer students' metacognitive states (planning, enacting, monitoring, and reflecting) during collaborative problem solving by integrating environment log data and collaborative discourse, enabling ChatGPT-generated summaries of collaboration to be grounded in students' actions within the learning environment. Zhou et al. [215] used video and conversation data to automatically detect gaze, nonverbal speech, and resource-management behaviors during group learning. Analysis revealed distinctive interaction patterns, including loops between gaze-linked referring/following and resource-management behaviors. These patterns differentiated groups based on their shared understanding and collaborative learning outcomes.

The use of NLP has skyrocketed in recent years, increasing from 35/73 papers (48%) in Corpus A—where prosodic speech was the primary NLP modality—to 40/49 (82%) in Corpus B [107, 112]—where text and transcribed speech were the primary modalities. This shift is closely tied to studies that deploy LLM-enabled systems, such as GPT-based learning aids [112], GenAI-supported multimodal composing [172], and multimodal LLMs for assessment [138]. Natural language features such as raw text, word embeddings, term frequency, loudness, and pitch are consistently reported as informative, with strong associations with predictive outcomes such as learner productivity, performance, and collaborative confusion or conflict. Collaborative settings, in particular, highlight these features as frequently among the most predictive, especially when combined with other multimodal signals [44].

While NLP outputs across Corpora A and B are broadly similar, their methodological approaches differ substantially. Corpus A predominantly uses traditional machine learning techniques such as SVMs [119, 151, 179] and logistic regression [56, 113, 150], often supplemented by qualitative methods (e.g., transcript coding, case studies) to interpret learner discourse and interaction [64, 123, 145]. In contrast, Corpus B centers on LLM-enabled pipelines, particularly for analyzing dialogue and delivering feedback [54, 138, 202]. These systems enable free-flowing conversations that were not feasible prior to the LLM era. In some cases, multimodal LLMs are also used for assessment [138]. Here, LLMs and GenAI tools serve dual roles: they act as interactive components within the learning environment and as generators of textual data for downstream multimodal analysis, with other modality groups (e.g., vision, logs, physiological signals) offering contextual signals around these language-based interactions.

The natural language challenges faced by MMLA researchers also differ considerably between corpora. Corpus A, with its reliance on traditional machine learning, frequently cites issues such as small, imbalanced datasets [38, 40, 41, 113, 114, 150], which hinder the training and adaptation of deep learning and transformer models [12, 39, 42, 150, 181]. The corpus's emphasis on audio data introduces additional challenges, particularly the time-intensive nature of feature extraction. Tools such as NLTK [117], openSMILE [69], and TAACO [52, 53] can aid this process, but often yield large and difficult-to-interpret feature spaces [155], while manual preprocessing and feature engineering further constrain scalability [100, 113]. Conversational agents are rare in Corpus A; when used, they typically deliver static messages or summative feedback rather than engaging in multi-turn interaction [56, 114, 185]. While using raw text with LLMs (like in Corpus B) mitigates many of these issues, others arise such as transcription errors from automatic speech recognition (ASR) in noisy classrooms [100, 178, 201] and concerns surrounding LLMs like adverse interactions with students and how to effectively evaluate LLM output [24, 92].

C.2.2 Vision. Vision-based modalities offer continuous insight into how learners move, attend, react, and interact in learning and training environments. Cameras and eye trackers capture signals such as affect, pose, gesture, activity, gaze, and fatigue[2], which are often inaccessible through logs or language alone. These visual signals are critical for modeling non-verbal behavior and engagement. In MMLA, they support both model-based approaches (e.g., convolutional neural networks [179]) and qualitative methods (e.g., interaction analysis [172]), and are frequently triangulated with other modalities such as logs and speech [172].

The rise of multimodal LLMs such as GPT [1] and Gemini [188] has broadened the role of vision-language models (VLMs) in MMLA, shifting their use from traditional tasks like classification and coding to more complex applications in interpretation and sense-making. For example, Yan et al. [205] use GPT-4V to interpret screenshots of nursing students' learning analytics dashboards, enabling the system to "see" and reason about visual elements such as charts and graphs. This visual understanding is integrated with retrieval-augmented generation (RAG), enabling the chatbot to produce explanations grounded in both the dashboard's visual structure and its educational data context. Unlike traditional machine learning pipelines, VLM-based multimodal integration requires minimal feature engineering. While neural networks often require pixel data to be normalized or transformed into visual embeddings aligned with textual representations [85], VLMs can directly accept raw text and image inputs from the user.

There has been a marked shift away from vision modalities in recent years. Whereas three of Corpus A's top four modalities were vision-based (pose [180], affect [123], and gaze [145]), none appears among the top seven in Corpus B. Except for the activity modality, the percentage of papers using each vision modality declined[3] (see Figures 5 and 6). Overall, vision-focused papers fell from 59/73 (81%) in Corpus A to 27/49 (55%) in Corpus B. We hypothesize that the rise in activity papers (A: 13/73; 18% [64, 102], B: 21/49; 43% [8, 172]) reflects a return to in-person research after COVID.

Vision modality group implementations were broadly consistent across corpora, favoring quantitative, model-based analyses to classify attributes such as pose [109, 178], gaze [180, 202], and affect [152, 162]. These modalities were typically used in traditional machine learning or shallow deep learning pipelines to infer learner states such as engagement, collaboration quality, or skill level, often supplemented by qualitative interpretation of predicted classes [161, 203]. Vision data functioned both as features (e.g., using gaze to predict engagement [208]) and as prediction targets (e.g., deriving affect from image data [161]). In Corpus B, vision was often one component of larger multimodal pipelines involving sensors and logs [104], providing contextual grounding for downstream analytics or dashboards [46]. While LLM usage increased overall, vision integration with multimodal LLMs was rare; instead, vision inputs like gaze were used to inform or condition LLM-based interactions (e.g., as context for ChatGPT [162]).

[2]We define **pixel** data in Table 10, but omit further discussion due to its limited use across both corpora.
[3]We ignore the fatigue modality due to underrepresentation: one paper in Corpus A and two in Corpus B.

As with natural language, vision-based modalities often contributed significant predictive power to multimodal pipelines. For instance, Acosta et al. [2] demonstrated that integrating trace log data with vision features such as facial action units, head pose, and gaze—extracted using OpenFace [15]—led to more accurate predictions of collaborative satisfaction than any individual modality or subset. Student survey responses served as ground truth, and two specific facial action units emerged as the most predictive features across both high- and low-performing modality combinations. Similarly, Ma et al. [119] used early fusion to combine video features (e.g., facial expressions, body movements, inter-learner distance), linguistic features (text embeddings), and audio signals (e.g., speaking time, pitch) to predict *impasse*, i.e., moments of stalled progress during collaborative problem solving due to conflicting ideas. Their results highlighted facial muscle movements as particularly strong predictors of impasse, underscoring the importance of visual signals in capturing nuanced social dynamics.

However, vision-based multimodal analytics face several practical and methodological challenges. Many learning and training environments lack controlled lighting, fixed camera setups, or specialized hardware (e.g., eye trackers), limiting the feasibility of fine-grained gaze or pose analysis [9, 49, 198]. Additionally, small and noisy datasets often lead researchers to rely on pre-extracted features rather than raw pixel data, which can obscure model assumptions and reduce adaptability across tasks or domains. For instance, commercial tools like iMotions provide real-time emotion tracking from facial muscle movements. Yet, the inferred states (e.g., joy, anger, fear) are typically assumed as ground truth without independent validation. Synchronizing and fusing visual data with other modalities, such as natural language, logs, or physiological signals, remains complex and time-consuming, with missing data and differing temporal resolutions further complicating joint modeling. Additionally, there is growing concern that opaque vision components may introduce bias or misinterpret learner behaviors, particularly for underrepresented populations or non-standard learning contexts [9].

C.2.3 Physiological Signals. Physiological signal-based modalities capture learners' physiological and motion-related traces, including affect, pose, activity, gaze, pulse, electrodermal activity (EDA), body temperature, blood pressure, electroencephalography (EEG), fatigue, and electromyography (EMG). These modalities link learners' observable actions with their internal states, enabling the interpretation of engagement, cognitive load, stress levels, and coordination in both learning and training contexts. Unlike vision-based data, physiological signal modalities are typically used as primary features in predictive models rather than as target outputs.

Physiological signal modalities appear more frequently in Corpus B (23/49; 47% [23, 93]) than Corpus A (A: 20/73; 27% [57, 101]), with notable differences in how they are deployed. In Corpus A, physiological signals are primarily tied to the EDA modality (16/20; 80% [78, 200]), and are disproportionately used in training contexts. Although training constitutes 22% (16/73) of Corpus A studies, it accounts for 40% of those using physiological signals (8/20) [57, 124]). In contrast, physiological signal use in Corpus B shifts toward motion-oriented modalities like activity and pose—each of which appears in 12/23 (52%) of physiological signal-enabled studies [104, 125].

In practice, this leads to different methodological approaches across the corpora. In Corpus A, physiological signals are commonly used for predictive modeling and to examine their relationships with learning behaviors and outcomes. Signal processing converts raw data streams (such as EDA, pulse, and accelerometer readings) into interpretable metrics, including learning gains, team dynamics, and shared arousal. These physiological signals support offline analyses, such as identifying patterns associated with performance, and in-time feedback during training, enabling timely interventions. Additionally, these features are often integrated with other modalities (such as visual data, logs, and language) to provide context for engagement, stress, and coordination.

Corpus B emphasizes interaction-rich environments, such as simulations and collaborative tasks, helping to capture arousal and cognitive load during the learning process. While the methods employed may vary, the physiological signals in Corpus B more clearly bridge the gap between "*vision-like*" behavioral traces that focus on students' observable actions and their cognitive and emotional states, thereby reinforcing their integrative role in connecting what learners are doing, how they move, and how their bodies respond [137].

The wide range of modalities derived from physiological and motion-based sensors has yielded diverse and insightful findings across multimodal learning and training pipelines. For example, Que et al. [158] combined gaze data from EyeLink 1000 Plus eye trackers with heart rate, inter-beat intervals, and electrodermal activity from an Empatica E4 wristband to predict three types of cognitive load during an English as a Second Language (ESL) reading task: *extraneous load* (avoiding irrelevant information while learning), *intrinsic load* (reflecting the complexity of learning material), and *germane load* (resources available for processing intrinsic load, e.g., comprehension). They found that extraneous load was predicted by increased fixation count and lower mean heart rate; intrinsic load by increased fixation count and mean saccade amplitude; and germane load by increased fixation count and heart rate variability.

While the above example is illustrative, many other studies demonstrate that physiological signals do not yield generalizable findings across contexts: different sensors work best in different environments, for different tasks, and with different populations of learners and trainees. Moreover, integrating and interpreting heterogeneous sensor streams remains technically challenging, and reliance on specialized hardware, such as eye trackers and wristbands, raises practical concerns about cost, scalability, and privacy. Teachers and students have also emphasized the importance of understanding how machine learning models generate predictions [48], highlighting the need for interpretable, human-centered, explainable AI (XAI) approaches when using physiological signals. Without these, stakeholders may struggle to trust or act on insights derived from these modalities [160, 164].

C.2.4 Human-Centered. Human-centered modalities include qualitative observations [136, 148], interviews [54, 142], surveys [63, 118], and artifacts produced by participants [28, 144] or researchers [84, 128]. These modalities anchor multimodal learning and training analytics in the lived experiences of learners and the perspectives of educators and researchers, offering insights into how participants perceive, interpret, and reflect on tasks—insights that are often inaccessible through sensor or log-based data streams alone. They are frequently used to complement quantitative findings with rich detail (e.g., via case studies or error analyses), and are often treated as ground truth in predictive modeling or for correlating learning behaviors with outcomes. Human-centered data are crucial for validating inferences from other modalities, improving the interpretability of model outputs for stakeholders, and understanding how multimodal systems are experienced and perceived in practice.

While both corpora incorporated human-centered data, its prevalence rose sharply from Corpus A to Corpus B (A: 45/73; 61% [66, 159], B: 46/49; 94% [93, 125]), reflecting the community's growing emphasis on stakeholder agency in the design and evaluation of intelligent systems. In earlier work, methodological studies in educational AI often lacked input from teachers, students, or learning scientists [41]. With the emergence of LLMs, however, user-centered approaches, such as participatory design and co-design [86, 163], have become critical, as trust in GenAI systems hinges on their perceived safety, effectiveness, and alignment with stakeholder needs [45]. Across both corpora, participant-produced artifacts were the most frequent human-centered modality, with a stronger emphasis in Corpus B (33/46; 72%) [89, 202] versus A (19/45; 42%) [13, 101]).

Human-centered data are most often used for model-free, qualitative analysis, but it can also be annotated for quantitative purposes. One noteworthy approach in multimodal learning and training research is *quantitative ethnography* [166], which involves applying qualitative coding to data and then extracting quantitative features for analysis. This enables the study of complex human behavior through techniques such as network analysis. For example, Sung et al. [183] employed multimodal *epistemic network analysis*[4] (ENA) [167] during a guided reading study in a college biology course to examine how sequences of students' self-regulated learning behaviors differed between mastery and non-mastery groups. Think-aloud data was analyzed to identify self-regulation strategies (e.g., monitoring, assessing, summarizing), while environment log data was used to differentiate between in-class and out-of-class engagement with guided reading questions. Quiz scores served as measures of learning outcomes. In both groups, monitoring-related verbalizations frequently co-occurred with other self-regulated learning codes; however, the co-occurring codes differed by group: in the mastery group, monitoring was more often paired with domain-specific strategies, whereas in the non-mastery group, it was more often paired with domain-general strategies. This discrepancy led the authors to conduct follow-up qualitative analyses to better understand the contextual nuances of the students' learning processes.

The challenges surrounding human-centered modalities stem largely from the inherent subjectivity of qualitative data and analysis. Observations, participant-produced artifacts, and self-reported measures (e.g., surveys and interviews) can introduce coder bias as well as cultural and linguistic biases [127, 139], which may propagate into downstream models and compromise generalizability. Furthermore, the manual processes involved in data collection, coding, and interpretation are labor-intensive and difficult to scale. Compounding these challenges is the limited standardization of coding schemes across studies, particularly for artifacts, interviews, and observational data, which hinders replication and cross-study comparison.

C.2.5 Logs. Environment logs capture learners' and trainees' interactions with digital tools, platforms, and learning environments. In MMLA, these time-stamped traces (e.g., clicks, navigation, tool use) provide a behavioral record that can be aligned with other modalities to infer cognitive strategies, engagement, and progress in problem-solving. While modalities like language or vision may reveal what participants are thinking or feeling, log data indicate what they are actually doing. These streams are highly integrative, often providing context for interpreting focal modalities such as natural language and vision. Log data are used similarly across both corpora in terms of frequency and methodology (A: 30/73; 40% [5, 149], B: 17/49; 35% [46, 134]).

Studies often use log data in supervised machine learning contexts, extracting features such as mouse clicks, click frequencies, click sequences, and inactivity (i.e., no clicks) as inputs for models like logistic regression, support vector machines (SVM), and random forests to predict outcomes like task performance and engagement [76, 121, 206]. Statistical analyses are also common, linking log-derived metrics to learning gains and behavioral patterns, for example, by mining clickstream sequences to infer cognitive strategies such as constructing, debugging, and assessing [173].

LLMs have broadened the interpretive scope of log data by enabling its direct integration into prompts as contextualized natural language. For example, Fonteles et al. [74] used late fusion in an embodied learning setting: gaze, speech, and log data were first classified with unimodal deep learning models, then combined as text-based input to an LLM to interpret students' socially shared regulation. Similarly, Cohn et al. [44] translated students' block-based programming actions into natural language to contextualize collaborative discourse during RAG with an LLM-based pedagogical agent. Incorporating log data to situate discourse within the learning environment improved semantic alignment and

[4]ENA transforms coded qualitative data into visual networks that reveal how coded concepts co-occur over time. Data segments (e.g., collaborative discourse turns) are coded according to a theoretical framework; nodes represent codes, and edges reflect their co-occurrence within a pre-defined window. Edge thickness encodes co-occurrence frequency, enabling temporal comparisons across groups and assessment of learning processes.

retrieval performance relative to using discourse alone. Students also reported positive interactions with the agent, indicating its potential to enhance engagement and support in collaborative learning.

The main drawback of environmental log data is that it is usually only applicable in digital settings, such as virtual or blended environments, but not in fully physical ones. Time alignment and temporal granularity present challenges for multimodal fusion. Researchers often need to reconcile different sampling rates, synchronize events across various modalities, and manage high-dimensional time series. Additionally, log-based models frequently struggle to generalize across different systems, partly due to the limited adoption of interoperability standards, such as xAPI and LMS-based[5] logging. The engineering costs can also be quite high, as building robust logging infrastructures and analysis pipelines demands significant software development effort, which can impede both reuse and scalability.

C.3 Learning Analytics

Learning analytics involves transforming data into actionable insights to better understand how students learn and train. This module connects the diverse data streams generated during learning and training activities with the inferences researchers draw to understand learner behavior and deliver more effective feedback. It consists of two main components: **data fusion**, which focuses on integrating diverse data streams, and **analysis**, which centers on interpreting this data. In the following subsections, we will explore both components in detail.

C.3.1 Data Fusion. Data fusion is essential for leveraging multiple data sources to enhance our understanding of learning and training. Only through fusion can we construct unified representations of learners and trainees that surpass the explanatory power of unimodal approaches. Just as humans rely on multiple senses to understand the world, data fusion allows researchers to integrate diverse modalities to better capture the conditions under which learners struggle, improve, and progress.

The conventional classification of fusion methods in MMLA, as defined by Chango et al. [27], includes three types: *early, late*, and *hybrid* fusion. Early (feature-level) fusion merges raw data from different sources at the initial processing stage. While it captures inter-modal interactions effectively, it faces challenges related to data heterogeneity and model complexity. Late (decision-level) fusion processes each modality independently before integrating results, enabling modality-specific insights but often overlooking inter-modal dynamics. Hybrid fusion blends these approaches, fusing data at multiple stages to exploit both inter-modal synergies and unimodal depth. However, this increases pipeline complexity and requires careful feature selection and synchronization.

We argue that this three-way classification does not adequately reflect the complexity of modern multimodal analysis. Our review revealed difficulties in categorizing fusion practices due to inconsistencies in defining "*raw*" versus "*processed*" features. For example, skeletal joint position data from a Microsoft Kinect may be considered raw by some since it is directly available from the device, but processed by others, since it is computed internally by the Kinect system from raw depth data.

To resolve such ambiguity, we adopt and formalize the notion of *mid fusion*, drawing from the concept of the *observability line* proposed by Di Mitri et al. [59] that separates the *input space* (i.e., observable evidence) from the *hypothesis space* (i.e., inferred constructs). While the authors note that the boundary between observable and unobservable features is conceptual and context-dependent, we use this distinction to define four fusion categories that are summarized in Table 11 and illustrated in Figure 7.

[5] Learning Management System

Category	Description
Early Fusion	Draws inferences and computes analytics from multiple sources of raw, directly observable data at the earliest stage of processing before any modality-specific analysis [101, 184, 200].
Mid Fusion	Represents a compromise that mixes early and late fusion for analysis by combining processed, observable features generated from individual sources with analysis using other sources of data within the input space [56, 66, 67].
Late Fusion	Analysis is performed on individual modalities, and the inferences (abstracted and unobservable) are combined to generate outcomes at a later stage, i.e., in the hypothesis space [145, 153, 157].
Hybrid Fusion	Combines the strengths of both early and late fusion methods. Data from various sources are combined at multiple stages of processing [5, 7, 155].
Other	Studies that do not fit into the early, mid, late, or hybrid categories, or where the fusion point was not specified, fusion was not performed, or fusion was performed qualitatively through observation [91, 95, 123].

Table 11. Data fusion approaches.

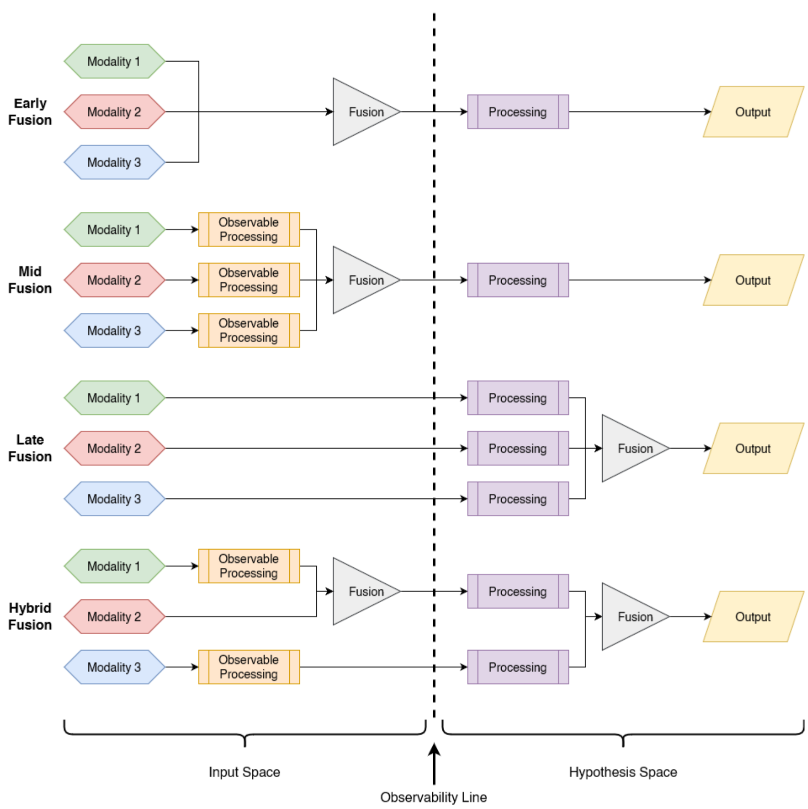

Fig. 7. Multimodal data fusion scheme according to when fusion is performed relative to the observability line.

For instance, Kinect raw pixel or depth data fit early fusion; skeletal joint position data—processed but still observable—fit mid fusion. We do not consider standard feature engineering practices, such as normalization, standardization, binning, or one-hot encoding, as constituting "processed" data for the purposes of our taxonomy. Data subjected only to these transformations would still fall under early fusion, and inferred constructs that are not observable, like planning and motivation, align with late fusion. Although the boundary between categories remains flexible, introducing mid fusion helps clarify methodological ambiguity and highlights distinctions among MMLA sub-communities in terms of their fusion practices.

Fusion strategies demonstrated a notable consistency across both corpora, with mid fusion being the most commonly used approach. In Corpus A, 27 out of 73 papers (37%) utilized mid fusion [76, 192], while in Corpus B, 19 out of 49 papers (39%) followed suit [65, 108]. This indicates a tendency to process features before integrating modalities. Hybrid fusion was also present in both corpora, though it appeared less frequently in Corpus B (Corpus A: 19/73; 26% [31, 206]; Corpus B: 8/49; 16% [11, 128]). In contrast, standalone instances of early or late fusion—outside of hybrid contexts—were rare in both datasets. It is important to note that a significant portion of the papers either did not perform fusion, did not disclose their fusion methodology, or employed alternative non-canonical strategies (e.g., "qualitative" fusion, which involves considering multiple modalities simultaneously during qualitative analysis). These instances accounted for 27% of papers in Corpus A (20/73), and 31% in Corpus B (15/49) [124, 186].

The ways in which fusion was implemented varied widely across studies. Mid fusion was particularly common in settings that used devices such as the Microsoft Kinect. Hybrid pipelines were often preferred in studies that incorporated three or more modalities, likely because of their flexibility in handling complex, heterogeneous data. No discernible pattern emerged between fusion type and the nature of input or output variables: both mid and hybrid fusion approaches were used to combine input features such as discourse embeddings, prosody, affect, behavior traces from physiological sensors, and log data. These combinations were frequently used to predict collaboration and learning quality, or to support students and teachers through real-time feedback and multimodal dashboards [29, 72, 132].

While less common, fusion with multimodal LLMs was explored in Corpus B to enable end-to-end interpretation of complex, multimodal artifacts. For example, Whitehead et al. [197] used GPT-4o to annotate students' posture during collaborative physics tasks by fusing cropped video frames with expert-defined textual prompts and a coding scheme. Fusion occurred at inference time within the model, which produced categorical posture annotations (e.g., sitting, leaning) as tabular outputs for downstream analysis. Results showed high test-retest reliability and strong agreement with human raters for simpler behaviors, though accuracy declined for more context-dependent postures. Additionally, the authors noted that performance in this context is heavily reliant on data quality, "careful prompt engineering," and human validation.

Several non-LLM challenges related to fusion emerge as well—perhaps none more significant than the alignment, integration, and deployment of heterogeneous data sources in real-world settings (i.e., cross-modal interaction). These challenges include reconciling disparate sampling rates, addressing inconsistent data quality, and managing missing values, all of which complicate synchronization and modeling. Fusion pipelines often demand extensive preprocessing, manual calibration, and domain expertise to ensure that signals are both temporally aligned and semantically coherent—requirements that are especially difficult to fulfill in real-time or online learning environments. Consequently, despite methodological advancements, the practical barriers to achieving robust, generalizable fusion remain a central bottleneck for MMLA research and its broader implementation.

C.3.2 Analysis. Analysis is how researchers transform multimodal traces into evidence about learning and training. The research questions determine which forms of analysis are appropriate (e.g., supervised vs. unsupervised, qualitative vs. quantitative, temporal vs. static), depending on the types of insights researchers hope to gain. We classify **analysis approaches** as either **model-based** or **model-free**. Model-based analysis relies on formal models to uncover the underlying structure of the data and the interrelationships between variables. These models often involve mathematical formulations, such as machine learning functions, or computational simulations that encode theoretical assumptions about learning processes. In contrast, model-free approaches avoid such assumptions, instead using empirical statistics (e.g., correlations) or qualitative analyses to identify patterns and relationships directly from the data.

Similarly, we use the term **analysis method** to refer to the specific techniques employed to derive insights from multimodal data in learning and training contexts. These methods, which are summarized in Table 12, range from supervised and unsupervised machine learning (e.g., classification, clustering) to qualitative approaches and network-based analyses. It is important to note that there is no one-to-one mapping between analysis approaches and methods, as both model-based and model-free approaches can employ a variety of methods.

Method	**Definition**
Classification	Assigning pre-defined labels to input data based on feature analysis through supervised learning (often via deep learning approaches) [5, 157, 180].
Regression	Predicting continuous numerical values through supervised learning to understand input-output relationships [57, 153, 178].
Clustering	Grouping data based on patterns or similarities using unsupervised learning [7, 22, 37].
Qualitative	Manually examining and interpreting data to uncover patterns or themes [91, 95, 123].
Statistical	Using statistical methods (e.g., correlation) to analyze data and draw conclusions [113, 118, 144].
Network analysis	Studying relationships and interactions using graph-based approaches [31, 50, 140].
Pattern Extraction	Identifying meaningful patterns or structures within data, including techniques like Markov analysis and sequence mining [136, 149, 186].

Table 12. Analysis methods.

There is a notable shift from model-based analysis in Corpus A (57/73; 78% [12, 78]) to model-free approaches in Corpus B (33/49; 67% [116, 203]). Model-based analyses in both corpora primarily involve supervised learning methods, such as classification and regression, often supplemented by statistical techniques (e.g., correlation analysis). These studies typically use input features derived from speech, video, log, and sensor data to predict outcome variables such as performance or engagement [5, 157, 178]. They focus primarily on individual learners, reflecting the difficulty of capturing complex social dynamics within formal, parameterized models.

In contrast, model-free approaches take a more exploratory stance, employing qualitative, clustering, statistical, and pattern-extraction techniques. Qualitative methods (e.g., interaction analysis) draw on theory and observation to interpret multimodal traces [123], while statistical and pattern-based approaches highlight relationships between behavior and outcomes (e.g., correlations between strategies and learning gains). These methods are especially prevalent in collaborative learning settings, where they are used to unpack social signals and discourse [50, 140].

For example, Xu et al. [203] used k-means clustering to identify collaborative patterns in undergraduate pair programming using standardized ratings of process quality (9 dimensions) and programming outcomes (4 dimensions). The resulting clusters differed meaningfully in both collaboration quality and performance. High-performing pairs engaged in knowledge construction, consensus-building discourse, positive affect, and iterative loops between talk and code adjustment. In contrast, lower-performing clusters were marked by self-talk, fragmented regulation, excessive debugging, and weaker coordination between discourse and behavior. Clusters were labeled consensus-achieved, argumentation-driven, individual-oriented, and trial-and-error, with the consensus-achieved group showing the strongest outcomes. Here, clustering functioned as a mid-level segmentation step, enabling data-driven insights into how multimodal interactions help explain relationships between collaboration and learning outcomes without relying on prior assumptions.

While both model-based and model-free methods are valuable across both corpora, each comes with inherent trade-offs. A persistent tension exists between the predictive strength and structure offered by model-based approaches—allowing researchers to leverage domain knowledge to define variable relationships that effectively guide analysis—and the interpretive richness and flexibility of model-free analyses that allows for the discovery of unanticipated insights. Choosing between them is not always straightforward. A balanced and often beneficial strategy is to employ both approaches in tandem: model-based analysis to test hypothesized relationships, while model-free methods to reveal latent patterns.

C.4 Feedback

In multimodal learning and training environments, feedback emerges when systems are deployed in real-world contexts (e.g., classrooms), typically taking one of two forms. **Direct feedback** refers to feedback explicitly provided to the user by the system—such as a pedagogical agent assisting a student— to improve performance or other learning metrics. **Indirect feedback**, conversely, is not intended for the end user but is derived from analysis of system use or learner behavior. It informs researchers and developers on how to refine their systems. Such feedback may arise from observing user-system interactions or analyzing outcomes across learner populations, ultimately leading to deeper insights that can be used to improve systems. Both types of feedback are essential for advancing MMLA and helping close the loop between methodological innovation and applied practice.

Every paper in Corpora A and B incorporated indirect feedback in some capacity [120, 179], highlighting the importance of using authentic studies with human subjects to refine system behavior. By contrast, the extent to which direct feedback was employed varied considerably across the two corpora. In the pre-LLM era, only 41 of 73 papers (43.8%) provided direct feedback to users, compared to 30 of 49 papers (61.2%) in the post-ChatGPT era [56, 124]. The LLM era has also enabled significantly more dynamic forms of direct feedback: learners and trainees can now engage in *dialogic* interactions, receiving feedback through rich exchanges with LLM systems that retain conversation history and support stateful interaction [30, 36].

The way multimodality is employed to deliver direct feedback differs substantially across the two corpora. For example, before LLMs, Petukhova et al. [151] introduced the *Virtual Debate Coach*, which monitors trainees' speech, prosody, posture, and gestures through multimodal sensing and analysis. The system extracts features such as filled pauses, speech pitch, and gestures derived from 3-D video coordinates to train an SVM classifier that estimates debaters' confidence levels. Feedback is then generated using predefined rules and expert-informed strategies.

Researchers used indirect feedback in this case to extend the system's machine learning capabilities by enabling automatic detection and interpretation of behavioral variation, as well as assessment of debater proficiency. Direct feedback was provided both formatively and summatively to help students improve their performance and confidence;

however, student-agent interactions are stateless, lacking dialogue-state tracking, turn-level language modeling, or access to conversational history. This represents a canonical pre-LLM feedback paradigm in which multimodal features are manually engineered and combined with rule-based or heuristic logic to produce feedback within a discrete response space.

By contrast, multimodal LLMs operate in a continuous space and can process heterogeneous data directly, without requiring engineered features. Nguyen and Park [138] employed GPT-4o to automatically score and generate explanatory feedback on students' multimodal science assessments. The authors demonstrated that LLMs can ingest handwritten student assessments—including textual and visual content as a single input—with over 90% transcription accuracy, achieving grading alignment comparable to human raters (Cohen's $k = 0.84$). Feedback quality was further enhanced through prompt engineering with few-shot exemplars, yielding responses that were more accurate and better aligned with teacher feedback.

While their system provided direct feedback to students in the form of a score and an accompanying explanation, they also used indirect feedback to improve system design through thematic error analysis. As the authors themselves note, the findings "present opportunities for designing learning analytics systems that allow for iterative evaluation and modification [of LLMs'] assessment output" [138]. Their analysis revealed that the LLM (1) failed to evaluate the depth of students' responses accurately, (2) hallucinated information not present in the prompt, and (3) exhibited inaccurate numerical reasoning. These were identified as the most critical issues to address in future iterations.

Additionally, the ease of deploying LLM-based feedback systems at scale (e.g., via API calls to OpenAI) has contributed to the emergence of multimodal dashboards and tools that serve as feedback layers for teachers and students, supporting guided reflection and debriefing rather than functioning solely as research instruments [65, 104]. In parallel, the multimodal capabilities of enterprise LLMs such as ChatGPT, Claude, and Gemini have facilitated the integration of GenAI-based systems with logs, artifacts, and other multimodal traces to generate personalized, data-driven feedback. These systems are designed to support self-directed learning, enhance engagement, and improve learner performance [54, 109, 111].

However, the rise of LLM-based feedback systems has introduced several challenges for multimodal learning and training. Human feedback often outperforms or qualitatively differs from AI-generated feedback, particularly in complex tasks [74]. This gap highlights ongoing design tensions around trust, interpretability, and the roles of human and AI actors in direct feedback ecosystems [8, 51].

In addition, the innate fusion capabilities afforded by contemporary multimodal LLMs often come at the expense of user control and transparency. While feature engineering is time-intensive, it enables researchers to evaluate which features contribute to model performance. In contrast, LLMs typically accept a single multimodal input [138], internally extracting features that are neither observable nor modifiable by users, and whose influence can only be evaluated indirectly through techniques such as perturbation analysis.

Recent work has shown promising results using multimodal late fusion with LLMs for direct feedback by first distilling each modality into text and then leveraging the LLM to perform textual fusion [74, 75] before feedback. However, this approach relies heavily on prompt engineering. Most studies employ ad hoc prompting strategies, with limited attention to systematically aligning generated feedback with established pedagogical principles [45]. This gap is often attributed to the absence of established learning frameworks in software engineering pipelines [48].

C.5 Summary

The four framework components—Environment, Multimodal Data, Learning Analytics, and Feedback—collectively illustrate how multimodality is used in learning and training environments. The environment determines which modalities can be captured and in what context, setting the stage for meaningful data collection. These interactions yield rich multimodal data streams, each offering unique windows into learning and training processes. Learning analytics fuses the heterogeneous data for analysis to extract insights, uncover patterns, and make inferences about learning and performance. These insights are used to generate feedback, either directly to learners and trainees or indirectly to researchers, engineers, and system designers to inform theory and improve educational tools. Across all four components, multimodality is the connective tissue that enables holistic, context-aware, and actionable understandings of learning and training in complex environments.

However, the approach to multimodal learning and training research differs markedly between Corpus A and Corpus B. Table 13 outlines key methodological shifts from pre-LLM multimodal learning analytics to more recent GenAI-enabled practices, highlighting how large transformer-based models have redefined data requirements, fusion strategies, and analytic workflows. Although researchers in Corpus B continue to apply and refine traditional methods established in Corpus A, the rapid adoption of LLMs and GenAI signals a clear and ongoing paradigm shift in the field.

D Literature Review Limitations

This review has three primary limitations: the use of Google Scholar for the literature search, the application of citation graph pruning for corpus reduction, and inconsistencies in versioning across published papers. Each is discussed below.

D.1 Google Scholar

Google Scholar, while widely used in academia and industry, presents challenges for reproducibility. Its proprietary ranking algorithm is opaque and presumably variable, with results influenced by factors such as user location, search history, the time of query, and possible A/B testing. Because the algorithm is constantly evolving and individualized, exact reproduction of our search results cannot be guaranteed.

To mitigate this, we used SerpAPI—a web scraping service that queries Google Scholar via randomized headers and proxies without user account data (Figure 8). SerpAPI's documentation confirms that no personal information is attached to API requests, and the API supports reproducibility through standardized, non-personalized queries. Additionally, SerpAPI recommends validating results using the included search URLs in browser incognito mode.

We also contacted SerpAPI directly, who confirmed: "No, we don't add any personal information," and noted that others can reproduce results by using the same search parameters. While we agree this may be optimistic given Google's opacity, we are reasonably confident that our initial search was free from personal bias due to the nature of the SerpAPI interface. For reference, all searches were conducted in Nashville, TN, USA.

D.2 Citation Graph Pruning

As detailed in Appendix Section B.2.1, we used citation graph pruning to programmatically reduce our corpus. This approach may have excluded some relevant works that had few citation connections within the corpus.

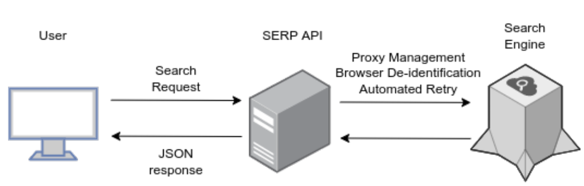

Fig. 8. Searching Google Scholar via SerpAPI.

Dimension	Pre-LLM MMLA Methods (2017–2022)	Post-LLM / GenAI-Enabled MMLA Methods (Late 2022–Present)
Feature Engineering	Predominantly manual and domain-specific feature extraction (e.g., handcrafted gaze metrics, prosodic features, rule-based textual features).	Reduced reliance on manual feature engineering through pretrained representations and prompt-based abstraction, though handcrafted features remain common in applied settings.
Model Architectures	Classical machine learning (e.g., SVMs, random forests) and task-specific deep learning models (e.g., CNNs, LSTMs).	Increasing use of transformer-based foundation models (e.g., LLMs, VLMs, multimodal transformers—particularly GPT-series models), often combined with task-specific components.
Fusion Strategies	Explicit early, mid, or late fusion pipelines designed and tuned per task.	Hybrid fusion approaches combining explicit fusion pipelines with implicit cross-modal reasoning enabled by pretrained models.
Data Requirements	Substantial labeled datasets are required for model training and validation.	Support for reduced annotation through transfer learning and zero- or few-shot inference, depending on task and context.
Adaptability Across Tasks	Limited generalization; models are typically trained for a single task or environment.	Improved cross-task and cross-domain transferability enabled by pretrained models, though adaptation remains context-dependent in applied environments and can require substantial prompt engineering.
Handling of Unstructured Data	Limited support for open-ended or qualitative data (e.g., discourse, reflection, embodied activity).	Improved capacity to process unstructured and open-ended multimodal data, particularly in language-rich and mixed-modality tasks.
Human-in-the-Loop Interaction	Primarily offline analysis and post-hoc interpretation of multimodal data.	Emerging support for interactive and human-in-the-loop analytics, including AI-assisted feedback and sense-making in certain contexts (e.g., assessment).
Interpretability and Transparency	Relatively interpretable pipelines with explicit features and model logic.	Foundation models introduce new interpretability challenges, alongside emerging practices for prompting, validation, and human oversight.
Scalability and Deployment	Deployment constrained by sensing setups, preprocessing pipelines, and model retraining requirements.	Easier prototyping and deployment via APIs and pretrained models, coupled with new constraints related to cost, latency, privacy, and governance.
Methodological Constraints	Strong dependence on controlled data collection, domain expertise, and context-specific sensing infrastructures.	Shift toward software-centric constraints, including model access, computational cost, data privacy, and alignment with institutional policies.

Table 13. Comparison of Pre-LLM (Corpus A) and Post-LLM (Corpus B) Methodological Affordances in Applied Multimodal Learning and Training Analytics

However, our goal was to identify core contributions in the field—papers that either built upon, or were built upon by, other relevant works. We reasoned that isolated papers with few citation ties were less likely to represent foundational or widely used methods.

Importantly, even after CGP, over half the remaining papers were ultimately excluded during qualitative screening. This suggests that CGP still retained many irrelevant works, reinforcing our confidence that the method did not omit significant in-scope works.

D.3 Versioning

Many papers in our corpus appeared in multiple forms across preprint servers and publication venues, often with inconsistent metadata. We used the earliest available public release date when possible. However, discrepancies may remain, particularly for papers from 2022-2023, where preprint and publication dates may differ by months.

As a result, it is possible that some Corpus B papers were written before the public release of ChatGPT (November 2022). Nonetheless, the sharp rise in publications in 2025 (see Figure 1 in the main manuscript) strongly suggests that generative AI was the driving force behind these works, which our own analysis reinforces. Any misclassification in dating is expected to be minor and unlikely to affect our overall findings or conclusions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Halim Acosta, Seung Lee, Bradford Mott, Haesol Bae, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. 2024. Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning. (2024).

[3] Mohammed AM AlGerafi, Yueliang Zhou, Mohamed Oubibi, and Tommy Tanu Wijaya. 2023. Unlocking the potential: A comprehensive evaluation of augmented reality and virtual reality in education. *Electronics* 12, 18 (2023), 3953.

[4] Haifa Alwahaby, Mutlu Cukurova, Zacharoula Papamitsiou, and Michail Giannakos. 2022. The evidence of impact and ethical considerations of multimodal learning analytics: A systematic literature review. *The multimodal learning analytics handbook* (2022), 289–325.

[5] Nese Alyuz, Eda Okur, Utku Genc, Sinem Aslan, Cagri Tanriover, and Asli Arslan Esme. 2017. An unobtrusive and multimodal approach for behavioral engagement detection of students. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*. ACM, Glasgow UK, 26–32. https://doi.org/10.1145/3139513.3139521

[6] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.

[7] Alejandro Andrade. 2017. Understanding student learning trajectories using multimodal learning analytics within an embodied-interaction learning environment. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, Vancouver British Columbia Canada, 70–79. https://doi.org/10.1145/3027385.3027429

[8] Jacqueline Anton, Giulia Cosentino, Kshitij Sharma, Mirko Gelsomini, Micah Mok, Michail Giannakos, and Dor Abrahamson. 2025. *The Human Condition: Modal and Interactive Advantages of Teacher over AI Feedback on Children's Mathematical Performance*. Association for Computing Machinery, New York, NY, USA, 183–203. https://doi.org/10.1145/3713043.3728863

[9] TS Ashwin, Nihar Sanda, Umesh Timalsina, and Gautam Biswas. 2025. Challenges of Applying Computer Vision for Emotion Detection in Educational Settings: A Study on Bias. In *International Conference on Artificial Intelligence in Education*. Springer, 388–395.

[10] T. S. Ashwin and Ram Mohana Reddy Guddeti. 2020. Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. *User Modeling and User-Adapted Interaction* 30, 5 (Nov. 2020), 759–801. https://doi.org/10.1007/s11257-019-09254-3

[11] Sinem Aslan, Ankur Agrawal, Nese Alyuz, Rebecca Chierichetti, Lenitra M Durham, Ramesh Manuvinakurike, Eda Okur, Saurav Sahay, Sangita Sharma, John Sherry, et al. 2022. Exploring kid space in the wild: a preliminary study of multimodal and immersive collaborative play-based learning experiences. *Educational technology research and development* 70, 1 (2022), 205–230.

[12] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E. Mete, Eda Okur, Sidney K. D'Mello, and Asli Arslan Esme. 2019. Investigating the Impact of a Real-time, Multimodal Student Engagement Analytics Technology in Authentic Classrooms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300534

[13] David Azcona, I-Han Hsiao, and Alan F. Smeaton. 2018. Personalizing Computer Science Education by Leveraging Multimodal Learning Analytics. In *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, San Jose, CA, USA, 1–9. https://doi.org/10.1109/FIE.2018.8658596

[14] Jie Bai, Xiulan Cheng, Hui Zhang, Yihang Qin, Tao Xu, and Yun Zhou. 2025. Can AI-generated pedagogical agents (AIPA) replace human teacher in picture book videos? The effects of appearance and voice of AIPA on children's learning. *Education and Information Technologies* (2025), 1–21.

[15] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[16] James Birt, Zane Stromberga, Michael Cowling, and Christian Moro. 2018. Mobile Mixed Reality for Experiential Learning and Simulation in Medical and Health Sciences Education. *Information* 9, 2 (Jan. 2018), 31. https://doi.org/10.3390/info9020031

[17] Nathaniel Blanchard, Michael Brady, Andrew M Olney, Marci Glaus, Xiaoyi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney D'Mello. 2015. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *International conference on artificial intelligence in education*. Springer, 23–33.

[18] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics* 3, 2 (2016), 220–238.

[19] John D Bransford, Ann L Brown, Rodney R Cocking, et al. 2000. *How people learn*. Vol. 11. Washington, DC: National academy press.

[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77.

[21] Capital Normal University, Beijing, China, Xiaoyang Ma, Min Xu, Yao Dong, and Zhong Sun. 2021. Automatic Student Engagement in Online Learning Environment Based on Neural Turing Machine. *International Journal of Information and Education Technology* 11, 3 (2021), 107–111. https://doi.org/10.18178/ijiet.2021.11.3.1497

[22] Man Ching Esther Chan, Xavier Ochoa, and David Clarke. 2020. Multimodal Learning Analytics in a Laboratory Classroom. In *Machine Learning Paradigms*, Maria Virvou, Efthimios Alepis, George A. Tsihrintzis, and Lakhmi C. Jain (Eds.). Vol. 158. Springer International Publishing, Cham, 131–156. http://link.springer.com/10.1007/978-3-030-13743-4_8

[23] Rosanna Yuen-Yan Chan, Chun Man Victor Wong, and Yen Na Yum. 2023. Predicting behavior change in students with special education needs using multimodal learning analytics. *IEEE Access* 11 (2023), 63238–63251.

[24] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[25] Wilson Chango, Rebeca Cerezo, and Cristóbal Romero. 2021. Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Computers & Electrical Engineering* 89 (Jan. 2021), 106908. https://doi.org/10.1016/j.compeleceng.2020.106908

[26] Wilson Chango, Rebeca Cerezo, Miguel Sanchez-Santillan, Roger Azevedo, and Cristóbal Romero. 2021. Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources. *Journal of Computing in Higher Education* 33, 3 (Dec. 2021), 614–634. https://doi.org/10.1007/s12528-021-09298-8

[27] Wilson Chango, Juan A Lara, Rebeca Cerezo, and Cristóbal Romero. 2022. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 4 (2022), e1458.

[28] Pankaj Chejara, Luis P Prieto, Maria Jesus Rodriguez-Triana, Reet Kasepalu, Adolfo Ruiz-Calleja, and Shashi Kant Shankar. 2023. How to build more generalizable models for collaboration quality? lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 111–121.

[29] Pankaj Chejara, Luis P Prieto, Maria Jesus Rodriguez-Triana, Adolfo Ruiz-Calleja, and Mohammad Khalil. 2023. Impact of window size on the generalizability of collaboration quality estimation models developed using Multimodal Learning Analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 559–565.

[30] Angxuan Chen, Mengtong Xiang, Junyi Zhou, Jiyou Jia, Junjie Shang, Xinyu Li, Dragan Gašević, and Yizhou Fan. 2025. Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs Human expert. *Computers & Education* 226 (2025), 105198.

[31] Lujie Karen Chen. 2021. Affect, Support, and Personal Factors: Multimodal Causal Models of One-on-one Coaching. *Journal of Educational Data Mining* 13, 3 (2021), 36–68.

[32] Yunnong Chen, Shuhong Xiao, Yaxuan Song, Zejian Li, Lingyun Sun, and Liuqing Chen. 2025. MindScratch: A Visual Programming Support Tool for Classroom Learning Based on Multimodal Generative AI. *International Journal of Human–Computer Interaction* (2025), 1–19.

[33] Kason Ka Ching Cheung, Jack Pun, Wangyin Kenneth-Li, and Jiayi Mai. 2025. Exploring students' multimodal representations of ideas about epistemic reading of scientific texts in generative AI tools. *Journal of Science Education and Technology* 34, 2 (2025), 284–297.

[34] Bonnie Chinh, Himanshu Zade, Abbas Ganji, and Cecilia Aragon. 2019. Ways of Qualitative Coding: A Case Study of Four Strategies for Resolving Disagreements. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312879

[35] Steven A. Cholewiak, Panos Ipeirotis, Victor Silva, and Arun Kannawadi. 2021. *SCHOLARLY: Simple access to Google Scholar authors and citation using Python*. N/A. https://doi.org/10.5281/zenodo.5764801

[36] Miguel Civit, María José Escalona, Francisco Cuadrado, and Salvador Reyes-de Cozar. 2024. Class integration of ChatGPT and learning analytics for higher education. *Expert Systems* 41, 12 (2024), e13703.

[37] Avery H. Closser, John A. Erickson, Hannah Smith, Ashvini Varatharaj, and Anthony F. Botelho. 2022. Blending learning analytics and embodied design to model students' comprehension of measurement using their actions, speech, and gestures. *International Journal of Child-Computer Interaction* 32 (June 2022), 100391. https://doi.org/10.1016/j.ijcci.2021.100391

[38] Keith Cochran, Clayton Cohn, and Peter Hastings. 2023. Improving NLP model performance on small educational data sets using self-augmentation. In *Proceedings of the 15th International Conference on Computer Supported Education, CSEDU 2023, Prague, Czech Republic, April 21-23, 2023, Volume 1*. scitepress. https://doi.org/10.5220/0011857200003470

[39] Keith Cochran, Clayton Cohn, Peter Hastings, Noriko Tomuro, and Simon Hughes. 2023. Using BERT to Identify Causal Structure in Students' Scientific Explanations. *International Journal of Artificial Intelligence in Education* N/A, N/A (2023), 1–39.

[40] Keith Cochran, Clayton Cohn, Nicole Hutchins, Gautam Biswas, and Peter Hastings. 2022. Improving automated evaluation of formative assessments with text data augmentation. In *International Conference on Artificial Intelligence in Education*. Springer, N/A, N/A, 390–401.

[41] Keith Cochran, Clayton Cohn, Jean Francois Rouet, and Peter Hastings. 2023. Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation. In *International Conference on Artificial Intelligence in Education*. Springer, N/A, N/A, 217–228.

[42] Clayton Cohn. 2020. *BERT efficacy on scientific and medical datasets: a systematic literature review*. DePaul University, N/A.

[43] Clayton Cohn, Joyce Horn Fonteles, Caitlin Snyder, Namrata Srivastava, Desmond Campbell, Justin Montenegro, Gautam Biswas, et al. 2025. Exploring the design of pedagogical agent roles in collaborative stem+ c learning. In *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning-CSCL 2025, pp. 330-334*. International Society of the Learning Sciences.

[44] Clayton Cohn, Surya Rayala, Caitlin Snyder, Joyce Fonteles, Shruti Jain, Naveeduddin Mohammed, Umesh Timalsina, Sarah K Burriss, Namrata Srivastava, Menton Deweese, et al. 2025. Personalizing Student-Agent Interactions Using Log-Contextualized Retrieval Augmented Generation (RAG). *arXiv preprint arXiv:2505.17238* (2025).

[45] Clayton Cohn, Surya Rayala, Namrata Srivastava, Joyce Horn Fonteles, Shruti Jain, Xinying Luo, Divya Mereddy, Naveeduddin Mohammed, and Gautam Biswas. 2025. A theory of adaptive scaffolding for LLM-based pedagogical agents. *arXiv preprint arXiv:2508.01503* (2025).

[46] Clayton Cohn, Caitlin Snyder, Joyce Horn Fonteles, Ashwin TS, Justin Montenegro, and Gautam Biswas. 2025. A multimodal approach to support teacher, researcher and AI collaboration in STEM+ C learning environments. *British Journal of Educational Technology* 56, 2 (2025), 595–620.

[47] Clayton Cohn, Caitlin Snyder, Justin Montenegro, and Gautam Biswas. 2024. Towards a human-in-the-loop LLM approach to collaborative discourse analysis. In *International Conference on Artificial Intelligence in Education*. Springer, 11–19.

[48] Clayton Cohn, Ashwin T S, Naveeduddin Mohammed, and Gautam Biswas. 2025. CoTAL: Human-in-the-Loop Prompt Engineering for Generalizable Formative Assessment Scoring. (2025). https://arxiv.org/abs/2504.02323 Submitted to the International Journal of Artificial Intelligence in Education (IJAIED). Currently under review.

[49] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo. 2018. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports medicine-open* 4, 1 (2018), 24.

[50] Hector Cornide-Reyes, René Noël, Fabián Riquelme, Matías Gajardo, Cristian Cechinel, Roberto Mac Lean, Carlos Becerra, Rodolfo Villarroel, and Roberto Munoz. 2019. Introducing Low-Cost Sensors into the Classroom Settings: Improving the Assessment in Agile Practices with Multimodal Learning Analytics. *Sensors* 19, 15 (July 2019), 3291. https://doi.org/10.3390/s19153291

[51] Giulia Cosentino, Jacqueline Anton, Kshitij Sharma, Mirko Gelsomini, Michail Giannakos, and Dor Abrahamson. 2025. Generative AI and multimodal data for educational feedback: Insights from embodied math learning. *British Journal of Educational Technology* (2025).

[52] Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods* 51 (2019), 14–27.

[53] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods* 48 (2016), 1227–1237.

[54] Óscar Cuéllar, Manuel Contero, and Mauricio Hincapié. 2025. Personalized and Timely Feedback in Online Education: Enhancing Learning with Deep Learning and Large Language Models. *Multimodal Technologies and Interaction* 9, 5 (2025), 45.

[55] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 2020. The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology* 51, 5 (2020), 1441–1449.

[56] Mutlu Cukurova, Carmel Kent, and Rosemary Luckin. 2019. Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology* 50, 6 (Nov. 2019), 3032–3046. https://doi.org/10.1111/bjet.12829

[57] Daniele Di Mitri, Maren Scheffel, Hendrik Drachsler, Dirk Börner, Stefaan Ternier, and Marcus Specht. 2017. Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, Vancouver British Columbia Canada, 188–197. https://doi.org/10.1145/3027385.3027447

[58] Daniele Di Mitri, Jan Schneider, and Hendrik Drachsler. 2022. Keep Me in the Loop: Real-Time Feedback with Multimodal Data. *International Journal of Artificial Intelligence in Education* 32, 4 (Dec. 2022), 1093–1118. https://doi.org/10.1007/s40593-021-00281-z

[59] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. 2018. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning* 34, 4 (2018), 338–349.

[60] Daniele Di Mitri, Jan Schneider, Kevin Trebing, Sasa Sopka, Marcus Specht, and Hendrik Drachsler. 2020. Real-time multimodal feedback with the CPR tutor. In *International conference on artificial intelligence in education*. Springer, 141–152.

[61] Muhterem Dindar, Jonna Malmberg, Sanna Järvelä, Eetu Haataja, and Paul A Kirschner. 2020. Matching self-reports with electrodermal activity data: Investigating temporal changes in self-regulated learning. *Education and Information Technologies* 25, 3 (2020), 1785–1802.

[62] Aaron Doering and Jeni Henrickson. 2015. Fostering creativity through inquiry and adventure in informal learning environment design. *Journal of Technology and Teacher Education* 23, 3 (2015), 387–410.

[63] Xu Du, Miao Dai, Hengtao Tang, Jui-Long Hung, Hao Li, and Jinqiu Zheng. 2023. A multimodal analysis of college students' collaborative problem solving in virtual experimentation activities: A perspective of cognitive load. *Journal of Computing in Higher Education* 35, 2 (2023), 272–295.

[64] Vanessa Echeverria, Roberto Martinez-Maldonado, and Simon Buckingham Shum. 2019. Towards Collaboration Translucence: Giving Meaning to Multimodal Group Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–16. https://doi.org/10.1145/3290605.3300269

[65] Vanessa Echeverria, Lixiang Yan, Linxuan Zhao, Sophie Abel, Riordan Alfredo, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, Abra Osborne, Simon Buckingham Shum, et al. 2024. TeamSlides: A multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space. In *Proceedings of the 14th learning analytics and knowledge conference*. 112–122.

[66] Andrew Emerson, Elizabeth B. Cloude, Roger Azevedo, and James Lester. 2020. Multimodal learning analytics for game-based learning. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1505–1526. https://doi.org/10.1111/bjet.12992

[67] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. ACM, Virtual Event Netherlands, 107–116. https://doi.org/10.1145/3382507.3418890

[68] Mary C English and Anastasia Kitsantas. 2013. Supporting student self-regulated learning in problem-and project-based learning. *Interdisciplinary journal of problem-based learning* 7, 2 (2013), 6.

[69] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. N/A, N/A, 1459–1462.

[70] Gloria Milena Fernandez-Nieto, Vanessa Echeverria, Simon Buckingham Shum, Katerina Mangaroska, Kirsty Kitto, Evelyn Palominos, Carmen Axisa, and Roberto Martinez-Maldonado. 2021. Storytelling With Learner Data: Guiding Student Reflection on Multimodal Team Data. *IEEE Transactions on Learning Technologies* 14, 5 (Oct. 2021), 695–708. https://doi.org/10.1109/TLT.2021.3131842

[71] Gloria Milena Fernandez Nieto, Kirsty Kitto, Simon Buckingham Shum, and Roberto Martinez-Maldonado. 2022. Beyond the learning analytics dashboard: Alternative ways to communicate student data insights combining visualisation, narrative and storytelling. In *LAK22: 12th international learning analytics and knowledge conference*. 219–229.

[72] Gloria Milena Fernandez-Nieto, Roberto Martinez-Maldonado, Vanessa Echeverria, Kirsty Kitto, Dragan Gašević, and Simon Buckingham Shum. 2024. Data storytelling editor: A teacher-centred tool for customising learning analytics dashboard narratives. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 678–689.

[73] Joyce Fonteles, Eduardo Davalos, T. S. Ashwin, Yike Zhang, Mengxi Zhou, Efrat Ayalon, Alicia Lane, Selena Steinberg, Gabriella Anton, Joshua Danish, Noel Enyedy, and Gautam Biswas. 2024. A First Step in Using Machine Learning Methods to Enhance Interaction Analysis for Embodied Learning Environments. In *Artificial Intelligence in Education*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 3–16.

[74] Joyce-Horn Fonteles, Clayton Cohn, ..., and Gautam Biswas. 2026. Analyzing Embodied Learning in Classroom Settings: A Human-in-the-Loop AI Approach for Multimodal Learning Analytics. *Journal of Learning and Instruction* (2026). in press, special issue on Implementing Multimodal Learning Analytics (MMLA) in Ecological Settings for Generating Actionable Insights.

[75] Joyce Horn Fonteles, Clayton Cohn, Divya Mereddy, TS Ashwin, and Gautam Biswas. 2025. Exploring Agentic Multimodal Late Fusion With LLMs for Embodied Learning. *Planning* 7, 15 (2025), 46–7.

[76] Fwa, Hua Leong and Lindsay Marshall. 2018. Investigating multimodal affect sensing in an Affective Tutoring System using unobtrusive sensors. *Psychology of Programming Interest Group* 29 (Oct. 2018), 78–85.

[77] Kristin A Gansle, George H Noell, Amanda M VanDerHeyden, Gale M Naquin, and Natalie J Slider. 2002. Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review* 31, 4 (2002), 477–497.

[78] Michail N. Giannakos, Kshitij Sharma, Ilias O. Pappas, Vassilis Kostakos, and Eduardo Velloso. 2019. Multimodal data as a means to understand the learning experience. *International Journal of Information Management* 48 (Oct. 2019), 108–119. https://doi.org/10.1016/j.ijinfomgt.2019.02.003

[79] Sahin Gökçearslan, Cansel Tosun, and Zeynep Gizem Erdemir. 2024. Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: A systematic literature review. *International Journal of Technology in Education* 7, 1 (2024), 19–39.

[80] Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv preprint arXiv:2310.08433* (2023).

[81] Alex Goslen, Yeo Jin Kim, Jonathan Rowe, and James Lester. 2025. Llm-based student plan generation for adaptive scaffolding in game-based learning environments. *International journal of artificial intelligence in education* 35, 2 (2025), 533–558.

[82] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. 2008. Exploring network structure, dynamics, and function using NetworkX. *N/A* N/A, N/A (1 2008), N/A. https://www.osti.gov/biblio/960616

[83] Emma Harvey, Allison Koenecke, and Rene F Kizilcec. 2025. " Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[84] Nathan L Henderson, Jonathan P Rowe, Bradford W Mott, and James C Lester. 2019. Sensor-based Data Fusion for Multimodal Affect Detection in Game-based Learning Environments. In *Proceedings of the EDM and Games Workshop at the 12th International Conference on Educational Data Mining*, Vol. 2592. International Educational Data Mining Society, Montreal, CA, 1–7.

[85] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

[86] Nicole M Hutchins and Gautam Biswas. 2024. Co-designing teacher support technology for problem-based learning in middle school science. *British Journal of Educational Technology* 55, 3 (2024), 802–822.

[87] Jeroen Janssen, Ulrike Cress, Gijsbert Erkens, and Paul A Kirschner. 2013. Multilevel analysis for the analysis of collaborative learning. In *The international handbook of collaborative learning*. Routledge, 112–125.

[88] Sanna Järvelä, Andy Nguyen, Eija Vuorenmaa, Jonna Malmberg, and Hanna Järvenoja. 2023. Predicting regulatory activities for socially shared regulation to optimize collaborative learning. *Computers in Human Behavior* 144 (2023), 107737.

[89] Lianjiang Jiang and Chun Lai. 2025. How Did the Generative Artificial Intelligence-Assisted Digital Multimodal Composing Process Facilitate the Production of Quality Digital Multimodal Compositions: Toward a Process-Genre Integrated Model. *TESOL Quarterly* (2025).

[90] Shiyan Jiang, Amato Nocera, Cansu Tatar, Michael Miller Yoder, Jie Chao, Kenia Wiedemann, William Finzer, and Carolyn P Rosé. 2022. An empirical analysis of high school students' practices of modelling with unstructured data. *British Journal of Educational Technology* 53, 5 (2022).

[91] Shiyan Jiang, Blaine E. Smith, and Ji Shen. 2021. Examining how different modes mediate adolescents' interactions during their collaborative multimodal composing processes. *Interactive Learning Environments* 29, 5 (July 2021), 807–820. https://doi.org/10.1080/10494820.2019.1612450

[92] Junfeng Jiao, Saleh Afroogh, Kevin Chen, Abhejay Murali, David Atkinson, and Amit Dhurandhar. 2025. LLMs and Childhood Safety: Identifying Risks and Proposing a Protection Framework for Safe Child-LLM Interaction. *arXiv preprint arXiv:2502.11242* (2025).

[93] Yueqiao Jin, Kaixun Yang, Lixiang Yan, Vanessa Echeverria, Linxuan Zhao, Riordan Alfredo, Mikaela Milesi, Jie Xiang Fan, Xinyu Li, Dragan Gasevic, et al. 2025. Chatting with a learning analytics dashboard: The role of generative AI literacy on learner interaction with conventional and scaffolding chatbots. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. 579–590.

[94] David H Jonassen. 2006. On the role of concepts in learning and instructional design. *Educational Technology Research and Development* 54, 2 (2006), 177–196.

[95] Sanna Järvelä, Jonna Malmberg, Eetu Haataja, Marta Sobocinski, and Paul A. Kirschner. 2021. What multimodal data can tell us about the students' regulation of their learning process? *Learning and Instruction* 72 (April 2021), 101203. https://doi.org/10.1016/j.learninstruc.2019.04.004

[96] Sanna Järvelä, Jonna Malmberg, Eetu Haataja, Marta Sobocinski, and Paul A. Kirschner. 2021. What multimodal data can tell us about the students' regulation of their learning process? *Learning and Instruction* 72 (April 2021), 101203. https://doi.org/10.1016/j.learninstruc.2019.04.004

[97] Alexandra D Kaplan, Jessica Cruit, Mica Endsley, Suzanne M Beers, Ben D Sawyer, and Peter A Hancock. 2021. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Human factors* 63, 4 (2021), 706–726.

[98] Fengfeng Ke, Sungwoong Lee, and Xinhao Xu. 2016. Teaching training in a mixed-reality integrated learning environment. *Computers in Human Behavior* 62 (2016), 212–220.

[99] Sungeun Kim and Dongsuk Oh. 2025. Evaluating Creativity: Can LLMs Be Good Evaluators in Creative Writing Tasks? *Applied Sciences* 15, 6 (2025), 2971.

[100] Marcus Kubsch, Daniela Caballero, and Pablo Uribe. 2022. Once More with Feeling: Emotions in Multimodal Learning Analytics. In *The Multimodal Learning Analytics Handbook*, Michail Giannakos, Daniel Spikol, Daniele Di Mitri, Kshitij Sharma, Xavier Ochoa, and Rawad Hammad (Eds.). Springer International Publishing, Cham, 261–285. https://link.springer.com/10.1007/978-3-031-08076-0_11

[101] Charlotte Larmuseau, Jan Cornelis, Luigi Lancieri, Piet Desmet, and Fien Depaepe. 2020. Multimodal learning analytics to investigate cognitive load during online problem solving. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1548–1562. https://doi.org/10.1111/bjet.12958

[102] Serena Lee-Cultura, Kshitij Sharma, Giulia Cosentino, Sofia Papavlasopoulou, and Michail Giannakos. 2021. Children's Play and Problem Solving in Motion-Based Educational Games: Synergies between Human Annotations and Multi-Modal Data. In *Interaction Design and Children*. ACM, Athens Greece, 408–420. https://doi.org/10.1145/3459990.3460702

[103] Serena Lee-Cultura, Kshitij Sharma, and Michail Giannakos. 2022. Children's play and problem-solving in motion-based learning technologies using a multi-modal mixed methods approach. *International Journal of Child-Computer Interaction* 31 (March 2022), 100355. https://doi.org/10.1016/j.ijcci.2021.100355

[104] Serena Lee-Cultura, Kshitij Sharma, and Michail N Giannakos. 2023. Multimodal teacher dashboards: Challenges and opportunities of enhancing teacher insights through a case study. *IEEE Transactions on Learning Technologies* 17 (2023), 181–201.

[105] Serena Lee-Cultura, Kshitij Sharma, Sofia Papavlasopoulou, and Michail Giannakos. 2020. Motion-Based Educational Games: Using Multi-Modal Data to Predict Player's Performance. In *2020 IEEE Conference on Games*. IEEE, Osaka, Japan, 17–24. https://doi.org/10.1109/CoG47356.2020.9231892

[106] Krittaya Leelawong and Gautam Biswas. 2008. Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education* 18, 3 (2008), 181–208.

[107] Daranee Lehtonen, Jorma Joutsenlahti, and Päivi Perkkilä. 2023. Multimodal Communication and Peer Interaction during Equation-Solving Sessions with and without Tangible Technologies. *Multimodal Technologies and Interaction* 7, 1 (2023), 6.

[108] Chia-Ju Lin, Wei-Sheng Wang, Hsin-Yu Lee, Yueh-Min Huang, and Ting-Ting Wu. 2024. Recognitions of image and speech to improve learning diagnosis on STEM collaborative activity for precision education. *Education and Information Technologies* 29, 11 (2024), 13859–13884.

[109] Chia-Ju Lin, Wei-Sheng Wang, Hsin-Yu Lee, Pin-Hui Li, Yueh-Min Huang, and Ting-Ting Wu. 2025. Advancing self-directed learning in STEM education: integrating GPT-based learning aid with multimodal learning analytics. *Journal of Research on Technology in Education* (2025), 1–19.

[110] Vivien Lin, Hui-Chin Yeh, Huai-Hsuan Huang, and Nian-Shing Chen. 2022. Enhancing EFL vocabulary learning with multimodal cues supported by an educational robot and an IoT-Based 3D book. *System* 104 (2022), 102691.

[111] Ming Liu, Zhongming Wu, Haimin Dai, Yifei Su, Laiba Malik, Jian Liao, Wei Zhang, Shuo Guo, Li Liu, and Junqiang Zhao. 2025. Enhancing self-directed learning and Python mastery through integration of a large language model and learning analytics dashboard. *British Journal of Educational Technology* (2025).

[112] Meilu Liu, Lawrence Jun Zhang, and Christine Biebricher. 2024. Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education* 211 (2024), 104977.

[113] Ran Liu, John Stamper, Jodi Davenport, Scott Crossley, Danielle McNamara, Kalonji Nzinga, and Bruce Sherin. 2019. Learning linkages: Integrating data streams of multiple modalities and timescales. *Journal of Computer Assisted Learning* 35, 1 (Feb. 2019), 99–109. https://doi.org/10.1111/jcal.12315

[114] Ran Liu, John C Stamper, and Jodi Davenport. 2018. A Novel Method for the In-Depth Multimodal Analysis of Student Learning Trajectories in Intelligent Tutoring Systems. *Journal of Learning Analytics* 5, 1 (April 2018), 41–54. https://doi.org/10.18608/jla.2018.51.4

[115] Su Liu, Ye Chen, Hui Huang, Liang Xiao, and Xiaojun Hei. 2018. Towards Smart Educational Recommendations with Reinforcement Learning in Classroom. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. IEEE, Wollongong, NSW, 1079–1084. https://doi.org/10.1109/TALE.2018.8615217

[116] Sichen Liu and Eunyoung Kim. 2024. Multimodal Writing Evaluation in Digital Storytelling using Video-Based Output: Comparing performance of AI and Human Raters.. In *Proceedings of the 6th International Conference on Modern Educational Technology*. 117–123.

[117] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. https://doi.org/10.48550/ARXIV.CS/0205028

[118] María Ximena López, Francesco Strada, Andrea Bottino, and Carlo Fabricatore. 2021. Using Multimodal Learning Analytics to Explore Collaboration in a Sustainability Co-located Tabletop Game. In *15th European Conference on Game-Based Learning*. Academic Conferences LTD, Brighton, UK.

[119] Yingbo Ma, Mehmet Celepkolu, and Kristy Elizabeth Boyer. 2022. Detecting Impasse During Collaborative Problem Solving with Multimodal Learning Analytics. In *12th International Learning Analytics and Knowledge*. ACM, Online USA, 45–55. https://doi.org/10.1145/3506860.3506865

[120] Katerina Mangaroska, Kshitij Sharma, Dragan Gašević, and Michail Giannakos. 2022. Exploring students' cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity. *Journal of Computer Assisted Learning* 38, 1 (2022), 40–59.

[121] Katerina Mangaroska, Kshitij Sharma, Dragan Gašević, and Michalis Giannakos. 2020. Multimodal Learning Analytics to Inform Learning Design: Lessons Learned from Computing Education. *Journal of Learning Analytics* 7, 3 (Dec. 2020), 79–97. https://doi.org/10.18608/jla.2020.73.7

[122] Victoria I Marín, Jeffrey P Carpenter, Gemma Tur, and Sandra Williamson-Leadley. 2023. Social media and data privacy in education: An international comparative study of perceptions among pre-service teachers. *Journal of Computers in Education* 10, 4 (2023), 769–795.

[123] Kit Martin, Emily Q. Wang, Connor Bain, and Marcelo Worsley. 2019. Computationally Augmented Ethnography: Emotion Tracking and Learning in Museum Games. In *Advances in Quantitative Ethnography*, Brendan Eagan, Morten Misfeldt, and Amanda Siebert-Evenstone (Eds.). Vol. 1112. Springer International Publishing, Cham, 141–153. http://link.springer.com/10.1007/978-3-030-33232-7_12

[124] Roberto Martinez-Maldonado, Vanessa Echeverria, Gloria Fernandez Nieto, and Simon Buckingham Shum. 2020. From Data to Insights: A Layered Storytelling Approach for Multimodal Learning Analytics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. https://doi.org/10.1145/3313831.3376148

[125] Roberto Martinez-Maldonado, Vanessa Echeverria, Gloria Fernandez-Nieto, Lixiang Yan, Linxuan Zhao, Riordan Alfredo, Xinyu Li, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, et al. 2023. Lessons learnt from a multimodal learning analytics deployment in-the-wild. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–41.

[126] Khaleel Asyraaf Mat Sanusi, Daniele Di Mitri, Bibeg Limbu, and Roland Klemke. 2021. Table Tennis Tutor: Forehand Strokes Classification Based on Multimodal Data and Neural Networks. *Sensors* 21, 9 (April 2021), 3121. https://doi.org/10.3390/s21093121

[127] Beloo Mehra. 2015. Bias in Qualitative Research: Voices from an Online Classroom. *The Qualitative Report* N/A, N/A (Jan 2015), N/A pages. https://doi.org/10.46743/2160-3715/2002.1986

[128] Mikaela E Milesi, Riordan Alfredo, Vanessa Echeverria, Lixiang Yan, Linxuan Zhao, Yi-Shan Tsai, and Roberto Martinez-Maldonado. 2024. "It's Really Enjoyable to See Me Solve the Problem like a Hero"': GenAI-enhanced Data Comics as a Learning Analytics Tool. In *Extended abstracts of the CHI conference on human factors in computing systems*. 1–7.

[129] Kathy A Mills and Alinta Brown. 2025. Smart glasses for 3D multimodal composition. *Learning, Media and Technology* 50, 2 (2025), 156–177.

[130] Joseph Mintz, Wayne Holmes, Leping Liu, and Maria Perez-Ortiz. 2023. Artificial Intelligence and K-12 education: Possibilities, pedagogies and risks. *Computers in the Schools* 40, 4 (2023).

[131] Daniele Di Mitri. 2019. Detecting Medical Simulation Errors with Machine learning and Multimodal Data. In *17th Conference on Artificial Intelligence in Medicine*. Springer International Publishing, Poznan, Poland, 1–6.

[132] Jewoong Moon, Sheunghyun Yeo, Seyyed Kazem Banihashem, and Omid Noroozi. 2024. Using multimodal learning analytics as a formative assessment tool: Exploring collaborative dynamics in mathematics teacher education. *Journal of Computer Assisted Learning* 40, 6 (2024), 2753–2771.

[133] Teresa Morell, Vicent Beltrán-Palanques, and Natalia Norte. 2022. A multimodal analysis of pair work engagement episodes: Implications for EMI lecturer training. *Journal of English for Academic Purposes* 58 (July 2022), 101124. https://doi.org/10.1016/j.jeap.2022.101124

[134] Kovan Mzwri and Márta Turcsányi-Szabo. 2025. Bridging LMS and generative AI: dynamic course content integration (DCCI) for enhancing student satisfaction and engagement via the ask ME assistant. *Journal of Computers in Education* (2025), 1–38.

[135] Jauwairia Nasir, Aditi Kothiyal, Barbara Bruno, and Pierre Dillenbourg. 2021. Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer-Supported Collaborative Learning* 16, 4 (Dec. 2021), 485–523. https://doi.org/10.1007/s11412-021-09358-2

[136] Andy Nguyen, Sanna Järvelä, Carolyn Rosé, Hanna Järvenoja, and Jonna Malmberg. 2023. Examining socially shared regulation and shared physiological arousal events with multimodal learning analytics. *British Journal of Educational Technology* 54, 1 (Jan. 2023), 293–312. https://doi.org/10.1111/bjet.13280

[137] David J Nguyen and Jay B Larson. 2015. Don't forget about the body: Exploring the curricular possibilities of embodied pedagogy. *Innovative Higher Education* 40, 4 (2015), 331–344.

[138] Ha Nguyen and Saerok Park. 2025. Providing Automated Feedback on Formative Science Assessments: Uses of Multimodal Large Language Models. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. 803–809.

[139] Helen Noble and Joanna Smith. 2015. Issues of validity and reliability in qualitative research. *Evidence Based Nursing* 18, 2 (Feb. 2015), 34–35. https://doi.org/10.1136/eb-2015-102054

[140] Rene Noel, Fabian Riquelme, Roberto Mac Lean, Erick Merino, Cristian Cechinel, Thiago S. Barcelos, Rodolfo Villarroel, and Roberto Munoz. 2018. Exploring Collaborative Writing of User Stories With Multimodal Learning Analytics: A Case Study on a Software Engineering Course. *IEEE Access* 6 (2018), 67783–67798. https://doi.org/10.1109/ACCESS.2018.2876801

[141] Omid Noroozi, Iman Alikhani, Sanna Järvelä, Paul A Kirschner, Ilkka Juuso, and Tapio Seppänen. 2019. Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior* 100 (2019), 298–304.

[142] René Noël, Diego Miranda, Cristian Cechinel, Fabián Riquelme, Tiago Thompsen Primo, and Roberto Munoz. 2022. Visualizing Collaboration in Teamwork: A Multimodal Learning Analytics Platform for Non-Verbal Communication. *Applied Sciences* 12, 15 (July 2022), 7499. https://doi.org/10.3390/app12157499

[143] Teresa M Ober, Maxwell R Hong, Daniella A Rebouças-Ju, Matthew F Carter, Cheng Liu, and Ying Cheng. 2021. Linking self-report and process data to performance as measured by different assessment types. *Computers & Education* 167 (2021), 104188.

[144] Xavier Ochoa and Federico Dominguez. 2020. Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1615–1630. https://doi.org/10.1111/bjet.12987

[145] Xavier Ochoa, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. 2018. The RAP system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, Sydney New South Wales Australia, 360–364. https://doi.org/10.1145/3170358.3170406

[146] Jennifer K. Olsen, Kshitij Sharma, Nikol Rummel, and Vincent Aleven. 2020. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1527–1547. https://doi.org/10.1111/bjet.12982

[147] Fan Ouyang, Mian Wu, Luyi Zheng, Liyin Zhang, and Pengcheng Jiao. 2023. Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 4.

[148] Fan Ouyang, Weiqi Xu, and Mutlu Cukurova. 2023. An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective. *International Journal of Computer-Supported Collaborative Learning* 18, 1 (2023), 39–66.

[149] Zacharoula Papamitsiou, Ilias O. Pappas, Kshitij Sharma, and Michail N. Giannakos. 2020. Utilizing Multimodal Data Through fsQCA to Explain Engagement in Adaptive Learning. *IEEE Transactions on Learning Technologies* 13, 4 (Oct. 2020), 689–703. https://doi.org/10.1109/TLT.2020.3020499

[150] Volha Petukhova, Tobias Mayer, Andrei Malchanau, and Harry Bunt. 2017. Virtual debate coach design: assessing multimodal argumentation performance. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, Glasgow UK, 41–50. https://doi.org/10.1145/3136755.3136775

[151] Volha Petukhova, Manoj Raju, and Harry Bunt. 2017. Multimodal Markers of Persuasive Speech: Designing a Virtual Debate Coach. In *Interspeech 2017*. ISCA, Stockholm, Sweden, 142–146. https://doi.org/10.21437/Interspeech.2017-98

[152] Phuong Pham and Jingtao Wang. 2017. AttentiveLearner2: A Multimodal Approach for Improving MOOC Learning on Mobile Devices. In *Artificial Intelligence in Education*, Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict Du Boulay (Eds.). Vol. 10331. Springer International Publishing, Cham, 561–564. http://link.springer.com/10.1007/978-3-319-61425-0_64

[153] Phuong Pham and Jingtao Wang. 2018. Predicting Learners' Emotions in Mobile MOOC Learning via a Multimodal Intelligent Tutor. In *Intelligent Tutoring Systems*, Roger Nkambou, Roger Azevedo, and Julita Vassileva (Eds.). Vol. 10858. Springer International Publishing, Cham, 150–159. http://link.springer.com/10.1007/978-3-319-91464-0_15

[154] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[155] L.P. Prieto, K. Sharma, Ł. Kidzinski, M.J. Rodríguez-Triana, and P. Dillenbourg. 2018. Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning* 34, 2 (April 2018), 193–203. https://doi.org/10.1111/jcal.12232

[156] Paul Prinsloo, Sharon Slade, and Mohammad Khalil. 2022. The answer is (not only) technological: Considering student data privacy in learning analytics. *British Journal of Educational Technology* 53, 4 (2022), 876–893.

[157] Athanasios Psaltis, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, and Petros Daras. 2018. Multimodal Student Engagement Recognition in Prosocial Games. *IEEE Transactions on Games* 10, 3 (Sept. 2018), 292–303. https://doi.org/10.1109/TCIAIG.2017.2743341

[158] Ying Que, Yueyuan Zheng, Janet H Hsiao, and Xiao Hu. 2025. Using eye movements, electrodermal activities, and heart rates to predict different types of cognitive load during reading with background music. *Scientific Reports* 15, 1 (2025), 32635.

[159] Joseph M Reilly, Milan Ravenell, and Bertrand Schneider. 2018. Exploring Collaboration Using Motion Sensors and Multi- Modal Learning Analytics. In *Proceedings of the 11th International Conference on Educational Data Mining*. International Educational Data Mining Society, Buffalo, NY, USA.

[160] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. 2021. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950* N/A, N/A (2021), N/A.

[161] Alpay Sabuncuoglu and T Metin Sezgin. 2023. Developing a multimodal classroom engagement analysis dashboard for higher-education. *Proceedings of the ACM on Human-Computer Interaction* 7, EICS (2023), 1–23.

[162] Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. 2024. Gaze-Driven Adaptive Learning System with ChatGPT-Generated Summaries. *IEEE Access* 12 (2024), 173714–173733.

[163] Juan Pablo Sarmiento and Alyssa Friend Wise. 2022. Participatory and Co-Design of Learning Analytics: An Initial Review of the Literature. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 535–541. https://doi.org/10.1145/3506860.3506910

[164] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies* 154 (2021), 102684.

[165] SerpApi. N/A. Google Scholar API. https://serpapi.com/google-scholar-api. [Accessed 08-02-2024].

[166] DW Shaffer. 2017. *Quantitative ethnography*. Cathcart Press.

[167] David Williamson Shaffer, Wesley Collier, and Andrew R Ruis. 2016. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of learning analytics* 3, 3 (2016), 9–45.

[168] Tamar Shamir-Inbal and Ina Blau. 2021. Facilitating emergency remote K-12 teaching in computing-enhanced virtual learning environments during COVID-19 pandemic-blessing or curse? *Journal of Educational Computing Research* 59, 7 (2021), 1243–1271.

[169] Kshitij Sharma and Michail Giannakos. 2020. Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology* 51, 5 (2020), 1450–1484.

[170] Kshitij Sharma, Zacharoula Papamitsiou, Jennifer K. Olsen, and Michail Giannakos. 2020. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. ACM, Frankfurt Germany, 480–489. https://doi.org/10.1145/3375462.3375498

[171] Qi Si, Tracey S Hodges, and Julianne M Coleman. 2022. Multimodal literacies classroom instruction for K-12 students: a review of research. *Literacy Research and Instruction* 61, 3 (2022), 276–297.

[172] Blaine E Smith, Amanda Yoshiko Shimizu, Sarah K Burriss, Melanie Hundley, and Emily Pendergrass. 2025. Multimodal composing with generative AI: Examining preservice teachers' processes and perspectives. *Computers and Composition* 75 (2025), 102896.

[173] Caitlin Snyder. 2024. *Understanding Students' Collaborative Problem Solving during STEM+ C Learning using Multimodal Analysis*. Ph. D. Dissertation. Vanderbilt University.

[174] Caitlin Snyder, Clayton Cohn, Joyce Horn Fonteles, and Gautam Biswas. 2025. Using collaborative interactivity metrics to analyze students' problem-solving behaviors during STEM+ C computational modeling tasks. *Learning and Individual Differences* 121 (2025), 102724.

[175] Caitlin Snyder, Nicole M Hutchins, Clayton Cohn, Joyce Horn Fonteles, and Gautam Biswas. 2024. Analyzing students collaborative problem-solving behaviors in synergistic STEM+ C learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 540–550.

[176] Caitlin Snyder, Cai-Ting Wen, Nicole M Hutchins, Caleb Vatral, Chen-Chung Liu, and Gautam Biswas. 2024. Investigating collaborative problem solving behaviors during stem+ c learning in groups with different prior knowledge distributions. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning-CSCL 2024, pp. 107-114*. International Society of the Learning Sciences.

[177] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems* 34, 11 (2022), 8135–8153.

[178] Daniel Spikol, Emanuele Ruffaldi, and Mutlu Cukurova. 2017. Using Multimodal Learning Analytics to Identify Aspects of Collaboration in Project-Based Learning. In *Making a Difference: Prioritizing Equity and Access in CSCL*, Vol. 1. International Society of the Learning Sciences, Philadelphia, PA USA, 263–270.

[179] Daniel Spikol, Emanuele Ruffaldi, Giacomo Dabisias, and Mutlu Cukurova. 2018. Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning* 34, 4 (Aug. 2018), 366–377. https://doi.org/10.1111/jcal.12263

[180] Daniel Spikol, Emanuele Ruffaldi, Lorenzo Landolfi, and Mutlu Cukurova. 2017. Estimation of Success in Collaborative Learning Based on Multimodal Learning Analytics Features. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, Timisoara, Romania, 269–273. https://doi.org/10.1109/ICALT.2017.122

[181] Penelope J. Standen, David J. Brown, Mohammad Taheri, Maria J. Galvez Trigo, Helen Boulton, Andrew Burton, Madeline J. Hallewell, James G. Lathe, Nicholas Shopland, Maria A. Blanco Gonzalez, Gosia M. Kwiatkowska, Elena Milli, Stefano Cobello, Annaleda Mazzucato, Marco Traversi, and Enrique Hortal. 2020. An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1748–1765. https://doi.org/10.1111/bjet.13010

[182] Emma L Starr, Joseph M Reilly, and Bertrand Schneider. 2018. Toward Using Multi-Modal Learning Analytics to Support and Measure Collaboration in Co-Located Dyads. In *ICLS 2018*. International Society of the Learning Sciences, London, UK, 448–455.

[183] Hanall Sung, Matthew L Bernacki, Jeffrey A Greene, Linyu Yu, and Robert D Plumley. 2024. Beyond frequency: Using epistemic network analysis and multimodal traces to understand temporal dynamics of self-regulated learning. *Journal of Science Education and Technology* (2024), 1–18.

[184] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2023. Multimodal Engagement Analysis From Facial Videos in the Classroom. *IEEE Transactions on Affective Computing* 14, 2 (April 2023). https://doi.org/10.1109/TAFFC.2021.3127692

[185] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE* 12, 8 (Aug. 2017), e0182151. https://doi.org/10.1371/journal.pone.0182151

[186] Sofia Tancredi, Rotem Abdu, Ramesh Balasubramaniam, and Dor Abrahamson. 2022. Intermodality in Multimodal Learning Analytics for Cognitive Theory Development: A Case from Embodied Design for Mathematics Learning. In *The Multimodal Learning Analytics Handbook*, Michail Giannakos, Daniel Spikol, Daniele Di Mitri, Kshitij Sharma, Xavier Ochoa, and Rawad Hammad (Eds.). Springer International Publishing, Cham, 133–158. https://link.springer.com/10.1007/978-3-031-08076-0_6

[187] Hengtao Tang, Miao Dai, Shuoqiu Yang, Xu Du, Jui-Long Hung, and Hao Li. 2022. Using multimodal analytics to systemically investigate online collaborative problem-solving. *Distance Education* 43, 2 (2022), 290–317.

[188] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* N/A, N/A (2023), N/A.

[189] Thanyalak Thiannoi, Putcharee Junpeng, and Thanapong Intharah. 2024. Efficiency of Decision Tree Depth to Diagnose Mathematical Procedures in Number and Algebra for Seventh-grade Students. In *2024 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*. 1–6. https://doi.org/10.1109/ITC-CSCC62988.2024.10628321

[190] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. https://doi.org/10.1177/1098214005283748 arXiv:https://doi.org/10.1177/1098214005283748

[191] Thomas Thiebaud". 2020. Spacy FastLang. https://spacy.io/universe/project/spacy_fastlang. [Accessed 08-02-2024].

[192] Gabriella Tisza, Kshitij Sharma, Sofia Papavlasopoulou, Panos Markopoulos, and Michail Giannakos. 2022. Understanding Fun in Learning to Code: A Multi-Modal Data approach. In *Interaction Design and Children*. ACM, Braga Portugal, 274–287. https://doi.org/10.1145/3501712.3529716

[193] Caleb Vatral, Madison Lee, Clayton Cohn, Eduardo Davalos, Daniel Levin, and Gautam Biswas. 2023. Prediction of Students' Self-confidence Using Multimodal Features in an Experiential Nurse Training Environment. In *International Conference on Artificial Intelligence in Education*. Springer.

[194] Bastian Venthur. 2010. GitHub - venthur/gscholar: Query Google Scholar with Python. https://github.com/venthur/gscholar. [Accessed 08-02-2024].

[195] Hana Vrzakova, Mary Jean Amon, Angela Stewart, Nicholas D. Duran, and Sidney K. D'Mello. 2020. Focused or stuck together: multimodal patterns reveal triads' performance in collaborative problem solving. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. ACM, Frankfurt Germany, 295–304. https://doi.org/10.1145/3375462.3375467

[196] Milica Vujovic, Davinia Hernández-Leo, Simone Tassani, and Daniel Spikol. 2020. Round or rectangular tables for collaborative problem solving? A multimodal learning analytics study. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1597–1614. https://doi.org/10.1111/bjet.12988

[197] Ridwan Whitehead, Andy Nguyen, and Sanna Järvelä. 2025. Utilizing multimodal large language models for video analysis of posture in studying collaborative learning: A case study. *Journal of Learning Analytics* 12, 1 (2025), 186–200.

[198] Marcelo Worsley. 2012. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 353–356.

[199] Marcelo Worsley. 2018. (Dis)engagement matters: identifying efficacious learning practices with multimodal learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, Sydney, Australia, 365–369. https://doi.org/10.1145/3170358.3170420

[200] Marcelo Worsley and Paulo Blikstein. 2018. A Multimodal Analysis of Making. *International Journal of Artificial Intelligence in Education* 28, 3 (Sept. 2018), 385–419. https://doi.org/10.1007/s40593-017-0160-1

[201] Marcelo Worsley, Kevin Mendoza Tudares, Timothy Mwiti, Mitchell Zhen, and Marc Jiang. 2021. Multicraft: A Multimodal Interface for Supporting and Studying Learning in Minecraft. In *HCI in Games: Serious and Immersive Games*, Xiaowen Fang (Ed.). Vol. 12790. Springer International Publishing, Cham, 113–131. https://link.springer.com/10.1007/978-3-030-77414-1_10

[202] Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025. Classroom Simulacra: Building Contextual Student Generative Agents in Online Education for Learning Behavioral Simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[203] Weiqi Xu, Yajuan Wu, and Fan Ouyang. 2023. Multimodal learning analytics of collaborative patterns during pair programming in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 8.

[204] Lixiang Yan, Vanessa Echeverria, Yueqiao Jin, Gloria Fernandez-Nieto, Linxuan Zhao, Xinyu Li, Riordan Alfredo, Zachari Swiecki, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning. *British Journal of Educational Technology* 55, 5 (2024), 1900–1925.

[205] Lixiang Yan, Linxuan Zhao, Vanessa Echeverria, Yueqiao Jin, Riordan Alfredo, Xinyu Li, Dragan Gaševi'c, and Roberto Martinez-Maldonado. 2024. VizChat: enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots. In *International conference on artificial intelligence in education*. Springer, 180–193.

[206] Xi Yang, Yeo-Jin Kim, Michelle Taub, Roger Azevedo, and Min Chi. 2020. PRIME: Block-Wise Missingness Handling for Multi-modalities in Intelligent Tutoring Systems. In *MultiMedia Modeling*, Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve (Eds.). Vol. 11962. Springer International Publishing, Cham, 63–75. http://link.springer.com/10.1007/978-3-030-37734-2_6

[207] Cunqian You, Huijuan Lu, Ping Li, Xiaoyu Zhao, and Yudong Yao. 2025. AI-Driven Intelligent Learning Companions: A Multimodal Fusion Framework for Personalized Education. In *2025 IEEE 34th Wireless and Optical Communications Conference (WOCC)*. IEEE, 424–428.

[208] Abdullahi Yusuf, Norah Md Noor, and Shamsudeen Bello. 2024. Using multimodal learning analytics to model students' learning behavior in animated programming classroom. *Education and Information Technologies* 29, 6 (2024), 6947–6990.

[209] Gabriela C Zapata, Bill Cope, Mary Kalantzis, Anastasia Olga Tzirides, Akash K Saini, Duane Searsmith, Jennifer Whiting, Nikoleta Polyxeni Kastania, Vania Castro, Theodora Kourkoulou, et al. 2025. AI and peer reviews in higher education: students' multimodal views on benefits, differences and limitations. *Technology, Pedagogy and Education* (2025), 1–19.

[210] Claire M Zedelius, Caitlin Mills, and Jonathan W Schooler. 2019. Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior research methods* 51, 2 (2019), 879–894.

[211] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *Comput. Surveys* 57, 5 (2025), 1–42.

[212] Linxuan Zhao, Dragan Gašević, Zachari Swiecki, Yuheng Li, Jionghao Lin, Lele Sha, Lixiang Yan, Riordan Alfredo, Xinyu Li, and Roberto Martinez-Maldonado. 2024. Towards automated transcribing and coding of embodied teamwork communication through multimodal learning analytics. *British Journal of Educational Technology* 55, 4 (2024), 1673–1702.

[213] Linxuan Zhao, Zachari Swiecki, Dragan Gasevic, Lixiang Yan, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, Abra Osborne, Xinyu Li, Riordan Alfredo, et al. 2023. METS: Multimodal learning analytics of embodied teamwork learning. In *LAK23: 13th International learning analytics and knowledge conference*. 186–196.

[214] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

[215] Qi Zhou, Wannapon Suraworachet, and Mutlu Cukurova. 2024. Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and information technologies* 29, 1 (2024), 1071–1098.