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
MEIYI MA, Vanderbilt University, USA
GAUTAM BISWAS, Vanderbilt University, USA

Advances in learning and training technologies have increased our ability to collect and analyze rich multimodal data (e.g., speech, video, and eye gaze) from these environments and better inform participants' learning and training experiences in physical and virtual spaces. While there have been several surveys and literature reviews focusing on multimodal learning and training, they cover specific parts of the multimodal pipeline, such as conceptual models and data fusion technologies. To date, a comprehensive literature review on the *methods* informing multimodal learning and training environments has not been conducted.

This literature review provides a comprehensive analysis of research methods in multimodal learning and training environments. We propose a taxonomy and framework that encapsulates recent methodological advances in this field. We characterize the multimodal domain in terms of five modality groups: (1) Natural Language, (2) Video, (3) Sensors, (4) Human-Centered, and (5) Environment Logs. We provide descriptive statistics; conduct a thorough, qualitative thematic analysis; and discuss the current state-of-the-art, challenges, and research gaps in the application of multimodal methods. Furthermore, we recognize the need for an additional data fusion category—*mid fusion*—and introduce a novel graph-based technique for literature review corpus refinement, which we call *citation graph pruning*. Overall, our corpus suggests that leveraging multiple modalities offers a more holistic understanding of the behaviors and outcomes of learners and trainees. Even when multimodality does not enhance predictive accuracy, it often uncovers patterns that contextualize and elucidate unimodal data, revealing subtleties that a single modality may miss. However, there remains a need for further research to bridge the divide between multimodal learning and training studies and foundational AI research.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

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' addresses: Clayton Cohn, clayton.a.cohn@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235; Eduardo Davalos, eduardo.davalos.anaya@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235; Caleb Vatral, cvatral@tnstate.edu, Tennessee State University, 3500 John A Merritt Blvd, Nashville, TN, USA, 37209; Joyce Horn Fonteles, joyce.h.fonteles@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235; Hanchen David Wang, hanchen.wang.1@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235; Meiyi Ma, meiyi.ma@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235; Gautam Biswas, gautam.biswas@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235.

Manuscript submitted to ACM

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

Recent technological advancements in multimodal machine learning—including the rise of large language models (LLMs)—have improved our ability to collect, process, and analyze diverse multimodal data such as speech, video, and eye gaze in learning and training contexts. While prior reviews have addressed individual components of the multimodal pipeline (e.g., conceptual models, data fusion), a comprehensive review of empirical methods in applied multimodal environments remains notably absent. This review addresses that gap, introducing a framework and taxonomy that capture both established practices and recent innovations driven by LLMs and generative AI. We identify five *modality groups*: Natural Language, Vision, Physiological Signals, Human-Centered Evidence, and Environment Logs. Our analysis reveals that integrating modalities enables richer insights into learner and trainee behaviors, revealing latent patterns often overlooked by unimodal approaches. However, persistent challenges in multimodal data collection and integration continue to hinder the adoption of these systems in real-time classroom settings.

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

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.

Manuscript submitted to ACM

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, Meiyi Ma, and Gautam Biswas. 2024. Multimodal Methods for Analyzing Learning and Training Environments: A Systematic Literature Review. 1, 1 (August 2024), 50 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

1.1 A Brief History

Recent advances in the learning sciences, bolstered by technological progress, are driving the personalization of educational and training curricula to meet the unique needs of learners and trainees. This shift is underpinned by data-driven approaches that are integrated into the field of *learning analytics* [61]. Learning analytics focuses on gathering and evaluating data on learners' and trainees' behaviors—specifically, their approaches to learning and training tasks [94, 165]. For example, intelligent tutoring systems like Practical Algebra Tutor [78] focus on diagnosing student errors, open-ended environments like Betty's Brain [84] adaptively scaffold learning, and teacher-feedback tools (e.g., [72, 124]) assist educators in enhancing instruction through insights into student behaviors.

A central research question in learning analytics is *What types of data are necessary to gain insights into learner behaviors and performance, and enable meaningful support that advances student learning and training in different scenarios* [108, 150]*?*. Initially, the scope of data collection and analysis was constrained by available technology and computational methods in educational settings. Early learning analytics predominantly analyzed log data from computer-based environments, establishing correlations between students' behaviors and their digital interactions, thus forming the foundation for many contemporary theories and methods in the field [71, 108].

Advances in sensor and data collection technologies are extending learning analytics beyond traditional log-based analyses [108]. In physical learning spaces, log data is insufficient to capture all learner actions, affective states, and collaborative behaviors. Researchers now integrate additional data collection devices, such as video to capture physical interactions, microphones for conversations, biometric sensors for stress levels, and eye trackers for attention [150].

This enriched data collection provides a more comprehensive understanding of students' affective, cognitive, psychomotor, and metacognitive states, advancing multimodal learning analytics (MMLA) [12, 13, 157]. MMLA has matured over a decade of research, disseminated through journal special issues [52, 96, 109], conferences [60], an edited volume [64], and systematic reviews [4, 22, 39, 50, 100, 129, 157]. This review focuses on *applied research methods* in MMLA, building on this substantial foundation.

1.2 Related Work

Recent work in MMLA research, surveys, and reviews have explored the MMLA landscape through various lenses: multimodal data fusion [22], conceptual models and taxonomy [50], statistical and qualitative assessments [121, 130], virtual reality [118], technology and data engineering [26], and ethical considerations [4]. Our review focuses on applied methods supporting data collection and analysis in multimodal learning and training environments, explicitly centering on methodologies for collecting, fusing, analyzing multimodal data, and interpreting it using learning theories. We extend and modify existing taxonomies to reflect recent advances in MMLA.

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), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Recent advances in education and technology are driving the personalization of curricula to meet the unique needs of learners and trainees, underpinned by data-driven approaches integrated into the field of *learning analytics*, which focuses on gathering and evaluating data on learners' and trainees' behaviors to understand how they approach learning and training tasks [60]. For example, intelligent tutoring systems such as the Practical Algebra Tutor [65] focus on diagnosing student errors; open-ended environments like Betty's Brain [70] adaptively scaffold learning through *social constructivism*[1] [129]; and teacher-feedback tools that assist educators in enhancing instruction through insights into students' problem-solving processes [26].

As learning and training environments have become more diverse and dependent on complex forms of human activity, learning analytics has expanded beyond its initial dependence on environment log data [11] to now incorporate multimodal evidence that captures the cognitive, affective, behavioral, and psychomotor dimensions of performance [112]. Contemporary learning and training scenarios include manipulating physical objects, both verbal and non-verbal communication, coordination within teams, and high-stakes medical procedural tasks—none of which can be adequately represented solely through digital traces. The growing use of video, audio, physiological signals, eye-tracking measures, and human-centered artifacts like surveys reflects a broader methodological transition in which multimodal data collection has become essential for understanding how learners and trainees engage with instructional tasks under real-world constraints [34]. This shift has made applied multimodal learning analytics (MMLA[2]) a critical area of inquiry, requiring systematic attention to how multimodal data are obtained, transformed, and analyzed in authentic settings.

1.1 Motivation and Contributions

Applied MMLA differs fundamentally from multimodal research in core AI and machine learning. In authentic learning and training settings, data acquisition is inevitably subject to noise, partial observability, missing data, and privacy constraints. Thus, any analysis or modeling must be designed to robustly accommodate these limitations [64]. Dataset size and structure are often limited by institutional schedules, small participant pools, and infrequent opportunities for repeated measurement [34, 83]. Annotation commonly relies on human expertise, subjective judgment, and labor-intensive manual coding, imposing methodological constraints that are largely absent from large-scale AI benchmarks and corpora [46]. These realities shape every stage of the multimodal pipeline—from sensing and feature extraction to data fusion and analysis—requiring methodological adaptations that prioritize pedagogical alignment, ecological validity, and practical feasibility.

Despite sustained growth in the MMLA literature, existing surveys primarily emphasize fusion taxonomies, conceptual frameworks, and general-purpose multimodal machine learning methods, as well as domain-specific applications such as healthcare and sensing [17, 38, 49, 90, 142, 143]. While valuable, these reviews do not synthesize how methodological decisions are shaped by the constraints of real-world data collection and deployment in classrooms, clinical simulations,

[1] Social constructivism is a learning theory that posits knowledge and meaning are constructed through social interaction, language, and cultural norms.
[2] While MMLA traditionally refers to multimodal learning analytics, this review uses the term to encompass both learning and training environments.

Di Mitri et al. [50] introduced the Multimodal Learning Analytics Model (MLeAM), a conceptual framework outlining the relationship between behavior, data, machine learning, and feedback in MMLA. This framework provided a taxonomy and introduced the concept of data observability, distinguishing between quantifiable input evidence and inferred annotations (e.g., emotions, cognition). The *observability line* demarcates these domains, crucial for AI-mediated transformation from input to hypotheses in MMLA research. Chango et al. [22] surveyed fusion methods in MMLA, categorizing studies by fusion type and application stage within the multimodal pipeline. They proposed three fusion types: *early* (feature-level integration), *late* (decision-level integration), and *hybrid* (combination of both). This classification clarifies fusion approaches and their relevance to educational data mining.

Integrating insights from both surveys, we propose a classification focused on *feature observability*, distinguishing between sensory data and human-inferred annotations. This adapted scheme refines our understanding of data fusion in MMLA and creates a refined taxonomy.

1.3 Scope of This Review

For this paper, we define a *data collection medium* as a unique type of raw data stream (e.g., video, audio, photoplethysmography (PPG) sensor). A *modality* is a unique attribute derived from data from one or more streams, each conveying different information, even from the same medium [108]. *Modality groups* are distinct sets of modalities conveying similar information, derived via inductive coding (see Section 1). *Multimodal* is a combination of either multiple modalities **or** multiple data streams. For example, the same video data stream can be used for affect and pose modalities, and the affect modality could be derived from audio and video streams. Both examples are considered multimodal. We use "papers" and "works" interchangeably, including publications outside of conferences and journals (e.g., books and book chapters).

Our review includes all papers from our literature search not excluded by our criteria (see Section B.2.2). This includes multimodal learning and training analysis done "in passing." For example, a paper focused on multimodal composing environments that performs multimodal learning analysis as a byproduct is included. We are interested in the methods used for multimodal analysis, not just those where it is the primary focus. Our definitions aim to characterize the scope of our review, not to establish a "universal" definition of multimodality and multimodal analysis.

1.4 Contributions

This paper presents a systematic literature review on methodologies for multimodal learning and training environments. We examine studies that engage in data collection and analysis across various mediums and modalities, encompassing fully physical settings (e.g., physical therapy), mixed-reality contexts (e.g., manikin-based nursing simulations), and online educational platforms (e.g., computer-based physics instruction). Notably, our review excludes virtual reality environments due to their current scalability challenges in educational settings [37].

This paper makes several novel contributions:

- A **comprehensive review** of the research methods used in multimodal learning and training environments, the challenges encountered, and relevant results that have been reported in the literature. Simultaneously, we also identify the research gaps in the data collection and analysis methodologies;
- A **congruent framework and taxonomy** that reflects the recent advances in multimodal learning and training methodologies;

workplace training, and other applied learning and training contexts. As a result, the empirical practices that govern the operationalization of multimodal methods in situ (i.e., the natural environment) remain fragmented and underexplored.

This gap has become more consequential with recent advances in generative AI (GenAI) and multimodal large language models (LLMs), which have transformed natural language processing, computer vision, and cross-modal representation learning [76]. The emergence of LLM-mediated systems has introduced a clear inflection point in MMLA practice, with post-LLM studies operating under different assumptions about model capability, automation, and the role of human interpretation than earlier work. Yet prior reviews do not explicitly distinguish between pre-LLM and post-LLM methodological paradigms, limiting their ability to contextualize current practices within a rapidly evolving landscape.

In response, this review provides an applied methodological synthesis of multimodal learning and training research grounded in empirical studies conducted in real-world environments. We examine how multimodal data are collected, transformed, fused, and analyzed under authentic constraints, and we explicitly contrast methodological patterns before and after the emergence of LLMs. Through this analysis, we pursue the following objectives and make the following contributions, summarized in the table below.

Focus	Objective	Contribution
Methodological Landscape	To characterize how multimodal data are collected, fused, and analyzed in applied learning and training contexts.	We systematically map the methodological space across data collection media, modality types, fusion strategies, and analysis approaches, offering a comprehensive account of empirical practices in the field.
Unified Framework and Taxonomy	To organize and make sense of the methodological diversity in applied MMLA.	We introduce a unified, empirically-grounded framework and taxonomy that reveal how methodological choices are shaped by environmental characteristics, data observability, and analytic goals.
Methodological Challenges	To identify and structure the key challenges encountered in real-world multimodal learning and training research.	We identify persistent methodological and practical issues and examine their implications for system design, deployment, and evaluation.
Methodological Archetypes	To identify how multimodal methods are employed to address different classes of research problems.[3]	We identify recurring methodological archetypes and illustrate their application through representative case studies, demonstrating what multimodal learning and training methodology looks like in practice.

1.2 Literature Review Structure

Section 2 situates this review within prior multimodal literature reviews. Section 3 describes the methodology for constructing and distilling our review corpus, including the scope definition and analysis procedures. Section 4 presents an empirically derived theoretical framework and taxonomy of multimodal methods for learning and training environments, with findings aligned to each framework component and illustrated through practical examples. Section 5 introduces three research archetypes that capture how MMLA research is conducted in practice, each presented with a case study. Section 6 discusses persistent challenges, future research directions, and the implications of our findings.

[3] Although archetypes emerged inductively during analysis, we formalized them as an objective once their explanatory value became clear.

- An **additional data fusion classification** that we call *mid fusion* (i.e., it is between *early fusion* and *late fusion*) that allows for differentiating processed features relative to the observability line.
- A graph-based **corpus reduction procedure** using a citation graph, which we refer to as *citation graph pruning*, that allows for programmatically pruning literature review corpora. This is described in detail in Section 3.2.1.

1.5 Structure of our Literature Review

The remainder of this literature review is structured as follows. Section 2 presents the theoretical framing and our taxonomy for multimodal methods in learning and training environments. Section 3 details the procedures for our literature search, study selection, feature extraction, and analysis. Section 4 presents our findings for each component of our framework (each subsection corresponds to a box in Figure 1), including an analysis of each of the 5 modality groups (Section 4.2). Section 5 presents three research categories ("archetypes") that best characterize the multimodal learning and training field. Section 6 highlights current trends, state-of-the-art, results, challenges, and research gaps, addressing limitations and future research directions. Section 7 concludes with a recap of this work's contributions.

2 FRAMEWORK AND TAXONOMY

In this section, we provide a detailed description of the multimodal learning and training analytics process, outlining both the overarching framework and the specific features that constitute our taxonomy.

2.1 Framework

We constructed our theoretical framework by integrating established multimodal learning analytics frameworks and through inductive analysis of the papers in our review corpus. The framework decomposes the multimodal learning and training analytics process into four primary components depicted in Figure 1: (1) the learning or training environment, (2) multimodal data, (3) learning analytics methods, and (4) feedback.

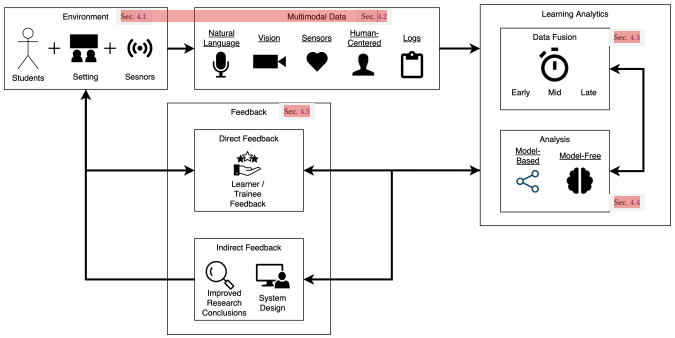


Fig. 1. Multimodal Learning and Training Environments Literature Review Framework

Table 1. Comparison of existing multimodal review studies and contributions of this review.

Review	V	A	T	S	H	Fus.	App. Chall.	X-Mod. Chall.	Trans.	MMLA Focus	L/T Env Focus	Sim
Education-Focused MMLA Reviews												
Di Mitri et al. (2018) [38]	✓	✓	✓	✓	✓	✓	×	×	×	✓	✓	×
Chango et al. (2022) [17]	✓	✓	✓	✓	×	✓	×	×	×	✓	✓	×
Alwahaby et al. (2022) [1]	✓	✓	✓	✓	✓	×	×	×	×	✓	✓	×
Shankar et al. (2018) [111]	✓	✓	✓	✓	✓	✓	×	×	×	✓	✓	×
Crescenzi et al. (2020) [32]	✓	✓	✓	✓	✓	×	✓	×	×	✓	✓	×
Mu et al. (2020) [90]	✓	✓	✓	✓	✓	✓	×	×	×	✓	✓	×
General Multimodal Deep Fusion Surveys												
Zhao et al. (2024) [142]	✓	✓	✓	✓	×	✓	×	✓	✓	×	×	×
Hussain et al. (2024) [53]	✓	✓	✓	✓	×	✓	×	✓	✓	×	×	×
Gaw et al. (2022) [48]	✓	✓	×	✓	×	✓	✓	✓	×	×	×	×
Teoh et al. (2024) [124]	✓	✓	✓	✓	×	✓	✓	✓	×	×	×	✓
Cengiz et al. (2025) [14]	×	×	✓	✓	×	✓	×	✓	✓	×	×	×
Mondal et al. (2025) [87]	✓	✓	✓	✓	×	✓	×	✓	✓	×	×	×
Healthcare / Sensing Reviews												
Shaik et al. (2024) [110]	✓	✓	✓	✓	×	✓	✓	✓	×	×	×	✓
Khoo et al. (2024) [61]	×	✓	✓	✓	✓	✓	✓	✓	×	×	×	×
This Review (2025)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Legend: V = Vision, A = Audio, T = Text, S = Sensors, H = Human-Centered; Fus. = discusses any multimodal fusion strategy; App. Chall. = applied, real-world data collection challenges and constraints; X-Mod. Chall. = cross-modal interaction challenges; Trans. = transformer-era multimodal methods covered; MMLA Focus = directly aligned with multimodal learning analytics; L/T Env = learning or training environment relevance; Sim = simulation-based or training simulation environments.

2 Background and Related Work

Across more than a decade of research, MMLA has been shaped by conceptual frameworks, sensing technologies, and advances in machine learning that seek to understand how multimodal evidence informs the study of learning processes. Foundational contributions include work on data observability [112], the distinction between directly measurable signals and inferred constructs [38], and the role of feature-level and decision-level fusion in multimodal pipelines [142]. These developments establish MMLA as a field focused on how various types of evidence are collected, transformed, integrated, and interpreted to explain learner behavior and design instructional support. As the collection of multimodal data becomes more prevalent in real-world learning and training environments, it is essential to systematically synthesize the methodological foundations of MMLA.

Prior surveys and literature reviews of multimodal learning analytics span several domains and research traditions, including general multimodal machine learning, healthcare sensing, and multimodal systems engineering [12, 49, 61, 110]. Table 1 summarizes recent reviews across twelve methodological dimensions central to applied multimodal learning and training analytics. Education-focused reviews [1, 17, 38] established influential conceptual models and taxonomies but offered limited analysis of how multimodal methods operate under real-world constraints such as noise, privacy

The environment, as the context for learner activities, is categorized as either *learning* or *training*, with the former supporting knowledge acquisition and the latter focusing on skill proficiency (Section 2.2.1). Learning environments range from physical classrooms, tutoring centers, online learning centers (e.g., Khan Academy), and individual or group-based computer learning environments. Skill-based training happens through practice and repetition and can include military training, nursing training, physical training, workplace training, etc. We exclude virtual reality environments due to scalability issues [37]. We further dissect the environment into sub-components: *human participants*, the *setting* (which includes physical, virtual, or blended spaces), and the *sensors* for data collection (Sections 2.2.6 to 2.2.2). The framework's second component is *multimodal data*, which comprises the data streams generated by the environmental sensors. We classify them into five modality groups (Section 2.2.3): (1) natural language, (2) vision, (3) sensors, (4) human-centered, and (5) environment logs, detailed in Sections 4.2.1, 4.2.2, 4.2.3, 4.2.4, and 4.2.5. The next block in our framework (see Figure 1) is *Learning Analytics*, which involves the methods for analyzing multimodal data, and is divided into *data fusion* (early, mid, late, and hybrid) and *analysis* approaches. Analysis approaches can be model-based or model-free, further detailed Section 2.2.11). Finally, *feedback* is the output of MMLA, differentiated into (1) *direct* feedback for students and instructors, and (2) *indirect* feedback for researchers and system designers (Section 4.5).

2.2 Taxonomy

In this section, we delve deeper into each component of our framework, exploring features extracted from our corpus.

2.2.1 Environment Type. Our paper explores a spectrum of environments on a learning-training continuum (Figure 2), from traditional classrooms to online courses, categorized along two dimensions: the learning-training axis [95, 104, 115, 154] and the physical-virtual space continuum [19, 38, 117].

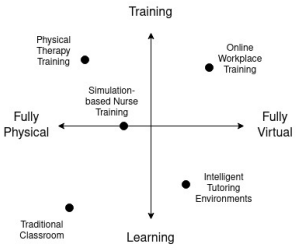


Fig. 2. Learning-Training Continuum

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 in fields like healthcare [51], athletics [95], and the military [69]. These settings may range from fully virtual simulations to physical training drills, with augmented and mixed realities bridging the gap. MMLA objectives differ between learning and training, necessitating context-specific strategies. While the distinction between learning and training can be ambiguous, as seen in game-based platforms [92, 158], our review spans this spectrum. We employ a fuzzy qualitative categorization to place each study within this continuum, acknowledging the complexity yet utility of this approach for analyzing MMLA research sub-communities.

2.2.2 Data Collection Mediums. Current learning and training environments use several computational measures of performance and behaviors, such as learning gains, establishing and progressing toward desired objectives, and employing effective plans of action to achieve these objectives. Multimodal data can provide the basis for computing these measures, ranging from logs and surveys to analyses of student artifacts. A diverse array of *data collection*

restrictions, and small datasets. These reviews also preceded the widespread use of transformer-based multimodal modeling and do not incorporate recent advances in GenAI and LLMs.

General multimodal fusion surveys [48, 53, 87, 142] provide comprehensive treatments of multimodal integration techniques, but are geared toward large-scale machine learning applications rather than learning or training. They assume controlled data-collection conditions, high-volume datasets, and domain-general modeling objectives, limiting their applicability to the methodological challenges faced in applied MMLA. Surveys in healthcare and sensing [61, 110] examine multimodal data collection under conditions of measurement noise and human variability, but they do not articulate the pedagogical, behavioral, or feedback-oriented dimensions characteristic of learning and training research.

Our literature review shows that no existing review synthesizes applied methodological practices across real-world learning and training environments, including the implications of human-subject constraints, ecological validity, task structure, and heterogeneous modality integration under real-world conditions. No evidence-derived taxonomy exists that is grounded in how multimodal data are actually collected, transformed, and analyzed in applied MMLA research. These gaps underscore the need for a review that connects methodology to the characteristics of learning and training environments while incorporating recent advances in multimodal modeling and heterogeneous data integration.

3 Methods

This section details the procedures used to construct the literature corpus, filter it with quantitative and qualitative techniques, extract methodological features, and prepare the data for analysis. Our objective is to curate a representative set of studies on multimodal data collection and analysis in learning and training contexts. Each stage of the pipeline was designed to support the unified framework and taxonomy presented in Section 4. As noted in Section 1, the emergence of LLMs has reshaped the MMLA landscape. To examine this shift, we partition the review into two corpora: Corpus A (2017-October 2022, pre-ChatGPT) and Corpus B (November 2022-2025, post-ChatGPT). Independent literature searches were conducted for each corpus using comparable search strategies and inclusion criteria, unless otherwise specified. The comparative findings in the two corpora are presented in Section 4, and full search and distillation details are provided in our Supplementary Materials.

3.1 Scope

This review focuses on multimodal learning and training analytics in empirical studies conducted in authentic instructional and training settings. We exclude studies centered exclusively on virtual reality environments, as they face scalability constraints for applied educational deployment. These boundaries informed both the inclusion criteria and corpus filtering, ensuring methodological alignment with an applied synthesis.

Given inconsistent terminology across the literature, we define key concepts to clarify our analytical focus and justify the scope of this review. *Learning* environments emphasize knowledge acquisition and conceptual understanding through didactic instruction, exploration, and problem-solving, whereas *training* environments focus on skill development through structured practice and repetition. A *data collection medium* is a raw data stream produced by a sensing or logging device (e.g., video, audio, physiological sensors). A *modality* is a derived attribute conveying a specific type of information (e.g., affect, pose, gaze, transcribed speech), and may be computed from one or more data streams. The mapping between media and modalities is many-to-many: a single video stream can yield both affect and pose, while affect can also be derived from fused audio and video. *Modality groups* are inductively identified clusters of related modalities used to structure our taxonomy. We adopt the term *multimodal* to refer to studies involving multiple modalities or synchronized data streams, following established usage in learning analytics [34].

mediums plays a pivotal role in gaining a comprehensive understanding of learners' progress, interactions, strategies, and struggles within these environments. The mediums listed in Table 1 (and all definitions in Section 2.2) were identified through our qualitative analysis of the corpus.

In the context of video data, we distinguish between depth cameras and traditional cameras. Though both fall under the video medium, depth cameras are typically employed with the motion modality to emphasize skeletal features. Furthermore, the scope of the motion medium extends beyond general video data, encompassing technologies such as real-time location systems (e.g., accelerometers, gyroscopes, or magnetometers). These technologies offer diverse approaches to capturing raw motion data, providing granularity in understanding participants' physical movements.

Medium	Definition
Video	Sequence of video frames from a camera source [27, 55, 117].
Audio	Audio signals captured by a microphone [114, 115, 142].
Screen Recording	Sequence of video frames of the contents displayed on a device screen [5, 74, 86].
Eye	Eye movement data and gaze points captured by tracking devices [21, 112, 143].
Logs	Participant's actions within the system and its state data [10, 120, 135].
Sensor	Specialized sensors used to gather participants' physiological data [69, 75, 87].
Interview	Structured or unstructured conversations between researchers and participants [11, 95, 105].
Survey	Standardized sets of questions administered to participants [38, 43, 116].
Participant Produced Artifacts	Materials produced by research participants using various mediums, including physical objects created for a task or written responses to formative assessment questions [8, 20, 106].
Researcher Produced Artifacts	Materials produced by the researchers that contribute to analysis and findings, such as observational notes [69, 93, 138].
Motion	Raw motion data collected by different devices/technologies [51, 95, 154].
Text	Raw textual input [158].

Table 1. Definitions for data collection mediums.

Researcher-produced artifacts can range from detailed field observation notes capturing contextual nuances to the labeling of data. This often requires manual coding that enhances data interpretability and contributes to more nuanced analyses and findings. Similarly, participant-produced artifacts constitute a valuable dimension in capturing participants' engagement and comprehension. These artifacts include materials, such as physical objects crafted by participants or pre/post-test results. We constrain participant-produced artifacts to include artifacts collected during the learning and training experiences, which excludes post hoc artifact collection.

2.2.3 Modalities. We previously defined *modalities* as unique attributes characterized by one or more data streams, where each modality conveys different information. Table 2 shows several modalities that are used for analyzing and understanding participants' interactions with and within learning and training environments.

In the context of a modality as a unique attribute defined by data from one or more data streams, it is important to note that multimodality can arise from a combination of multiple modalities and multiple data streams. For example, the same video data stream could be used to derive both the AFFECT and POSE modalities. Similarly, the AFFECT modality can be derived from separate audio and video data streams.

3.2 Corpus Construction and Distillation Procedure

The literature search was conducted programmatically using Google Scholar via SerpAPI, which was chosen for its ability to reliably return organic search results across different queries. Twenty-one search queries were used to perform our search (see Supplementary Materials). These queries were informed by the authors' expertise in multimodal learning analytics and aligned with the objectives of an applied methodological review focused on the collection, transformation, and analysis of multimodal data in various learning and training contexts. For each query, the top 100 results were retrieved (five pages from Google Scholar with 20 records per page). Duplicate records were removed using hash-based matching on the UUIDs provided by Google Scholar, and non-English papers were excluded from the dataset. This process yielded 2,120 papers for Corpus A and 845 for Corpus B, the latter reflecting the more recent emergence of LLMs relative to the longer history of MMLA.

Inclusion and exclusion criteria were guided by the need to focus on applied multimodal methods in authentic learning and training environments. A study was included if it (1) collected or analyzed data from at least two modalities or two data collection media; and (2) examined a learning or training setting involving human participants. These settings include fully physical environments (e.g., traditional classroom instruction), mixed-reality contexts (e.g., manikin-based nursing simulations), and technology-enhanced or online instructional environments (e.g., Khan Academy). Studies were excluded if they involved purely synthetic datasets, controlled laboratory tasks lacking instructional relevance, medical imaging applications, or multimodal architectures designed only for model training.

Following the initial literature search, we applied a programmatic corpus reduction procedure using a pragmatic filtering heuristic[4], resulting in 1,063 remaining papers for Corpus A and 559 for Corpus B. Each paper in both corpora was then qualitatively screened by at least two of the authors of this paper. We used a multi-stage review process based on Kitchenham's systematic review methodology [63]—involving sequential filtering of titles, abstracts, and full text—which we adapted to the goals of an applied multimodal synthesis based on our inclusion and exclusion criteria. Altogether, this process resulted in 73 papers for Corpus A and 49 for Corpus B, totaling 122 papers in the final corpus used for analysis. Figure 1 shows the distribution of included papers by year, with a dividing line when ChatGPT was released at the end of 2022.[5]

3.3 Data Extraction

All of the 122 papers were coded to capture methodological characteristics related to multimodal learning and training analytics. The features extracted for analysis included data-collection media, derived modalities, fusion strategies, analysis methods, types of environment, participant interaction structures, didactic characteristics, levels of instruction, domains of study, student and system feedback, and publication metadata. These features provide the empirical foundation for the unified framework and taxonomy discussed in Section 4. Notably, several categorical distinctions, such as modality groups and

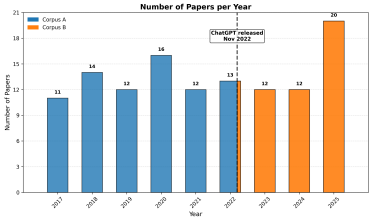


Fig. 1. Distribution of the *Number of Papers* selected for this review by year. Blue bars represent works published prior to the release of ChatGPT in November 2022 (Corpus A); orange bars represent those published afterward (Corpus B).

[4] We refer to this process as *citation graph pruning* and detail it in the Supplementary Materials.
[5] Five papers included in Corpus B were published in 2022.

Modality	Description	Modality Group
Affect	Participant's emotional or affective state [48, 120, 142].	NLP, Vision, Sensor
Pose	Participant's physical position, location, or body posture [5, 136, 139].	Vision, Sensor
Gesture	Participant's gestures and body language [6, 115, 157].	Vision
Action	Participant's observable actions or activities [62, 86, 119].	Vision, Sensor
Prosodic Speech	Speech information, e.g. volume, pauses, and intonation [104, 135, 137].	NLP
Transcribed Speech	Textual speech obtained through transcriptions, i.e., speech-to-text [11, 38, 85].	NLP
Qualitative Observations	Researcher observations about the participant and the study task [75, 92, 156].	Human-centered
Logs	Participant's actions and system state data [10, 65, 98].	Sensor
Gaze	Participant's data on the direction and focus of the eye gaze [54, 55, 161].	Vision, Sensor
Interview	Notes from interviews between researchers and participants [9, 53, 74].	Human-centered
Survey	Participant's responses to surveys/questionnaires [112, 114, 116].	Human-centered
Pulse	The participant's pulse, indicating their heart rate [81, 82, 147].	Sensor
EDA	Participant's electrodermal activity [80, 91, 131].	Sensor
Temperature	Participant's body temperature [83, 112, 131].	Sensor
Blood Pressure	Participant's blood pressure [82, 112, 147].	Sensor
EEG	Participant's electroencephalography activity [65, 112, 131].	Sensor
Fatigue	The level of fatigue experienced during the activity [81, 82].	Vision, Sensor
EMG	Participant's electromyography activity [49, 51].	Sensor
Participant Produced Artifacts	Variety of artifacts produced by the participant during the study, e.g. tests [21, 99, 106].	Human-centered
Researcher Produced Artifacts	Variety of artifacts produced by the researcher about the study and participants [27, 57, 105].	Human-centered
Spectrogram	Representation of audio frequencies in the form of a spectrogram [90].	NLP
Text	Participant's raw text data generated in the study environment [158].	NLP
Pixel	RGB pixel values from cameras or sensors [119].	Vision

Table 2. This table defines various modalities and provides exemplary references to papers. The first column lists the modality in the corpus, the second column describes each modality, and the third column categorizes the modality into relevant modality groups, as further discussed in Sec. 4.2.

2.2.4 Analysis Methods. In this literature review, the term *analysis method* refers to specific techniques for deriving insights from multimodal data in learning and training contexts. These methods are tailored to the research goals and the data characteristics. A classification scheme shown in Table 3 categorizes the analysis methods employed within the multimodal learning and training domain. The methods range from supervised techniques like classification to unsupervised methods such as clustering, and qualitative analyses. More recently, deep learning algorithms have been developed for analysis of multiple data streams [63, 64]. Similarly, reinforcement learning techniques are being developed for educational recommendations [87]. Evaluating these methods is essential for understanding current

participant interaction structures, were not predetermined. Instead, these categories emerged inductively during the extraction process as recurring patterns observed across studies. This approach aligns with creating a bottom-up taxonomy based on empirical evidence rather than conceptual assumptions. Complete definitions of all extracted features and their possible values are presented in Section 4, with further elaboration in Supplementary Materials.

3.4 Analysis Procedure

The extracted features were analyzed using qualitative thematic analysis informed by the framework presented in Figure 2. Studies were grouped by modality, environment, and analytic approach to identify methodological trends and patterns. Five modality groups—(1) natural language, (2) vision, (3) physiological signals, (4) human-centered evidence, and (5) logs—were derived through iterative coding and used to structure the analysis presented in the following section. Within each group, we examined common fusion strategies and analytic methods, along with challenges and constraints unique to applied settings. Our analysis also identified three archetypes of multimodal learning and training research, which are presented in Section 5. These archetypes highlight recurring methodological configurations and illustrate how multimodal methods are aligned with different research aims. Together, these procedures support the development of a unified, empirically grounded understanding of multimodal methods in learning and training environments.

4 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 2, 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, teachers, and researchers.

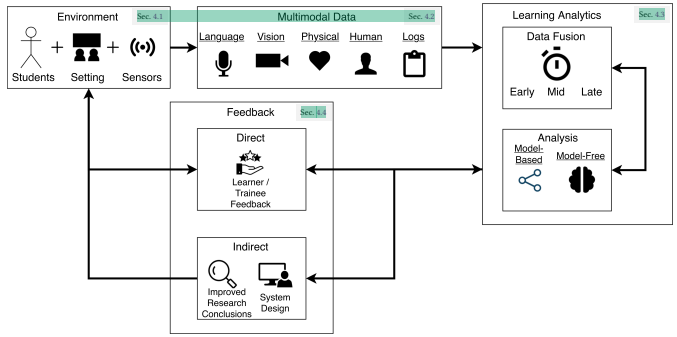


Fig. 2. Multimodal learning and training environments literature review framework.

trends in data analysis and informing future research. This review concentrates on the examination and interpretation of the data through these methods, rather than an in-depth critique of the analytical techniques themselves, unless such meta-analysis yields further valuable insights.

Method	Definition
Classification	Assigning pre-defined labels to input data based on feature analysis through supervised learning (often via deep learning approaches) [5, 120, 137].
Regression	Predicting continuous numerical values through supervised learning to understand input-output relationships [48, 117, 135].
Clustering	Grouping similar data points based on patterns or similarities using unsupervised learning [6, 19, 27].
Qualitative	Examining and interpreting non-numerical data to uncover patterns, themes, or meanings [74, 75, 92].
Statistical	Using descriptive and inferential techniques to analyze data and draw conclusions [85, 89, 106].
Network analysis	Studying relationships and interactions using graph-based approaches [23, 38, 104].
Pattern Extraction	Identifying meaningful patterns or structures within data, including techniques like Markov analysis and sequence mining [102, 112, 143].

Table 3. Analysis methods and their definitions.

2.2.5 Data Fusion. In multimodal learning and training, data fusion is essential for leveraging multiple data sources to enhance understanding of learning processes. Data fusion integrates information from diverse sources into a unified dataset, enabling enhanced analysis and understanding over single-modality studies. Such integration facilitates deeper insights into learners' cognitive states, emotions, and behaviors, informing personalized educational interventions and the use of adaptive pedagogical strategies.

The conventional classification of data fusion methods in MMLA, as reviewed by Chango et al. [22], includes early, late, and hybrid fusion. *Early fusion* merges raw data from different sources at the initial processing stage, capturing inter-modal interactions but faces challenges with data heterogeneity and model complexity. *Late fusion* involves separate analyses of each modality with outcomes integrated later, allowing for detailed, modality-specific insights but potentially missing inter-modal dynamics. *Hybrid fusion* combines these approaches, integrating data at various processing stages to harness both inter-modal relationships and in-depth, single-modality analysis, though it increases complexity and necessitates strategic feature selection.

We contend that the traditional three-state categorization inadequately captures the nuances of multimodal analysis. Our qualitative review reveals difficulties in classifying data fusion practices due to ambiguities in defining *raw* versus *processed* features. For example, some researchers might classify the joint position data measured by a Microsoft Kinect camera as a raw feature, and thus permissible in early fusion, since it is available from the camera without any additional processing. However, others might classify this as a processed feature, and thus a part of hybrid or late fusion, since the Kinect camera is computing this data from the raw depth data, regardless of whether this computation is obfuscated to the end user. Thus we've introduced a new category, *mid fusion*, which involves moderately processed data integration, as conceptualized by Di Mitri et al. [50] using the observability line. To elaborate, Di Mitri et al. state "*The distinction between observable/unobservable is conceptual and can vary in practice.*" [50]. Here, *early fusion* combines unprocessed,

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.

4.1 Environments

Our paper explores a spectrum of environments on a learning-training continuum (Figure 3). The environments span from traditional classrooms to online courses and are categorized along two dimensions: the learning-training axis [84, 94, 100] and the physical-virtual space axis [15, 30, 101]. Environments like manikin-based nurse training simulation combine physical and virtual elements and are referred to as *mixed-reality* [136].

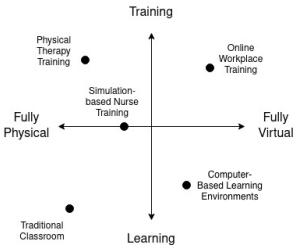


Fig. 3. Learning-training continuum.

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, while also promoting teamwork skills across fields such as healthcare [39], athletics [84], the workplace [88], and the military [52]. 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 [81, 134], 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. 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).

4.1.1 Learners/Trainees ("Students" in Figure 2). 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, how they work together in groups and teams, their methods of interaction with their teams and the environment, 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 characterize the learner context along four dimensions, detailed in Table 2. This taxonomy is applied consistently throughout the review, and individual studies may receive multiple labels within each dimension.

Across learner and trainee characteristics, a key distinction concerns the **structure** of learning contexts and its implications for multimodal design and analysis. Structured domains of study (e.g., STEM+C) involve constrained tasks and clear evaluation criteria, supporting quantitative, model-based approaches such as rule-based assessment and

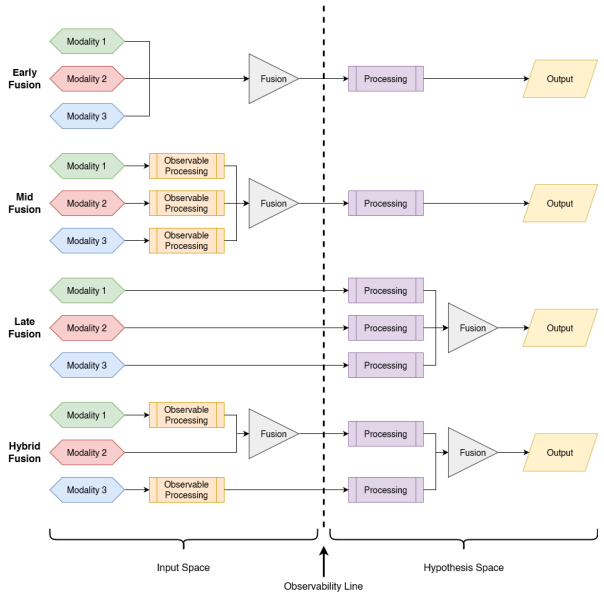


Fig. 3. Multimodal data fusion scheme according to when fusion is performed relative to the observability line.

observable features; *mid fusion* combines observable features that have undergone some processing; and *late fusion* combines processed features that cross into the hypothesis space, becoming inferences rather than direct observations.

For example, Kinect sensor's raw pixel or depth data are suitable for early fusion, while joint position data, processed but observable, fit mid-fusion. In contrast, inferred constructs like motivation, derived from joint data, align with late fusion. The *mid fusion* category, while interpretatively flexible, clarifies ambiguities and aids in identifying MMLA sub-communities by their fusion methods. For a detailed definition of observable modalities, see section 2.2.3. Following Chango et al.'s methodology [22], we also introduce an *other* category for studies not conforming to the four primary groups or lacking specified fusion points. These data fusion categories are summarized in Table 4.

2.2.6 Environment Setting. Analyzing the contextual settings in which these studies happen, we categorize these environments based on the nature of the setting. In a **Virtual** setting, activities occur entirely within a virtual space [5, 137, 142]. A **Physical** setting is where activities take place in a real-world environment without the use of digital technologies [115, 135, 157]. The **Blended** setting combines elements of both virtual and physical environments

Feature	Values	A	B	Comparison
Domain of study	• **STEM+C**: STEM + computing and healthcare [3, 118] • **Humanities**: literature, debate, writing [100, 103] • **Psychomotor**: motor coordination (e.g., CPR, woodworking, games) [50, 86]	• STEM+C dominant (55/73, 75%) [18, 134] • Humanities limited (11/73, 15%) • Psychomotor rare (5/73, 7%)	• STEM+C dominant (37/49, 76%) • Humanities increased (13/49, 27%) via LLM-supported open-ended tasks [73] • Psychomotor minimal (1/49, 2%)	• STEM+C stable across corpora • Humanities growth with LLM-enabled analysis [73]
Particip. interact. structure	• **Individual**: single learner interacting with a system (e.g., educational games [68], intelligent tutors [16], open-ended environments [70], creative platforms [33]) • **Multi-learner**: pairs, small groups, or full classes (e.g., paired programming [135], game-based competitions [102], collaborative play [9])	• Individual settings common • Multi-learner studies: 42% (31/73) [40, 97]	• Increased focus on collaboration • Multi-learner studies: 51% (25/49) [19]	• Shift toward collaborative learning • Expanded use of multimodal data to analyze group processes (e.g., collaboration, coordination, reflection)
Didactic nature	• **Instruction**: formal activities with defined objectives (e.g., courses, labs) [15, 30] • **Training**: skill development through practice (e.g., clinical simulations, vocational drills) [83, 84] • **Informal**: loosely structured contexts without fixed goals (e.g., games, exploratory play) [22, 42]	• Instruction dominant (45/73, 62%) • Training moderate (15/73, 21%) • Informal present (12/73, 16%)	• Instruction strongly dominant (40/49, 82%) • Training stable (11/49, 23%) • Informal rare (1/49, 2%)	• Shift toward formal instruction • Training remains steady across eras • Informal settings decline as LLMs concentrate in instructional contexts
Level of instruction	• **K–12**: primary and secondary education [43, 97] • **University**: undergraduate and graduate education [30, 86, 91] • **Professional**: workplace learning and continuing education [33, 39, 102]	• University dominant (36/73, 49%) • K–12 substantial (30/73, 41%) • Professional limited (5/73, 7%)	• University dominant (29/49, 59%) • K–12 stable (19/49, 39%) • Professional minimal (2/49, 4%)	• Continued emphasis on university learners • K–12 representation remains substantial • Professional development remains underrepresented

Table 2. Taxonomy of learners and trainees, with trends and comparisons across Corpora A and B.

supervised learning [139]. Unstructured domains (e.g., humanities and informal learning) emphasize open-ended tasks with subjective outcomes that are poorly captured by simple metrics [140] and instead require qualitative or model-free analyses such as thematic or interaction coding [54]. Recent advances in LLMs are beginning to narrow this gap by enabling rubric-based evaluation of unstructured work, including narrative coherence and content organization [62].

Participant interaction structure further shapes analytic possibilities. Multi-learner settings enable analyses of dialogue, coordination, and socially shared regulation using multimodal audio-visual data [58, 71], but introduce challenges such as smaller effective sample sizes and degraded data quality due to noise and overlapping speech [13]. Individual settings, by contrast, more readily support scalable personalization and fine-grained learner modeling.

Category	Description
Early Fusion	Draws inferences and computes analytics from multiple sources of raw data at the earliest stage of processing before any modality-specific analysis. [80, 141, 157].
Mid Fusion	Represents a compromise that mixes early and late fusion for analysis. Combines processed, observable features generated from individual sources with analysis using other sources of data within the input space [43, 54, 55].
Late Fusion	Analysis is performed on individual modalities, and the inferences generated are combined to generate outcomes at a later stage, i.e., in the hypothesis space [107, 117, 120].
Hybrid Fusion	Combines the strengths of both early and late fusion methods. Data from various sources are combined at multiple stages of processing [5, 6, 119].
Other	Studies that do not fit into the early, mid, late, or hybrid categories, or where the fusion point was not specified or fusion was not performed [74, 75, 92].

Table 4. Categories of data fusion approaches.

[6, 48, 120]. Finally, the **Unspecified** setting refers to environments that are not clearly described in the paper [43, 87]. We aim to unveil the contextual relevance of multimodal learning and training, discerning how these approaches manifest in computer-based spaces, traditional classrooms, and blended scenarios combining virtual and physical elements. Additionally, acknowledging instances where sufficient information is not provided, directs our attention to research gaps and unexplored areas within the literature.

2.2.7 Domain of Study. We recognized the importance of identifying the subject matter domain that study participants engage in, thus defining five domain categories. **STEM+C** includes participants engaged in Science, Technology, Engineering, Mathematics, and Computing disciplines, encompassing healthcare and medicine [6, 135, 137]. **Humanities** focuses on activities related to literature, debate, and oral presentation [115, 120, 142]. **Psychomotor Skills** emphasizes activities that develop motor skills and coordination [51, 65, 98]. The **Other** category covers subjects outside the previously mentioned categories [99, 138]. Finally, the **Unspecified** category includes papers that do not provide sufficient information about the subject matter [8, 18, 87]. This categorization helps us better contextualize the use of multimodal analytics, exploring how these apply across a diverse domains, and is intentionally broad. We realized that additional granularity may hinder our ability to analyze and interpret current trends in the multimodal design of subject-related environments. Importantly, papers reporting results from multiple studies have labels corresponding to the domain of each separate study [138, 154].

2.2.8 Participant Interaction Structure. We categorized papers by how they enabled interactions with participants, i.e., the participant interaction structure as being **Individual** [18, 21, 82] or **Multi-Person**, which often emphasized collaborative or group dynamics [119, 123, 156]. It is noteworthy that some papers analyzed both the individual learner and groups of learners, reflecting the diversity in studies even within individual publications [8, 19].

2.2.9 Didactic Nature. This refers to the approach used for delivering the learning and training. This results in yet another lens through which we can understand, analyze, and differentiate learning and training environments, and resulted in four categories. **Formal** instruction occurs in traditional classrooms, online courses, or other structured environments with clear objectives [19, 38, 75]. **Informal** learning takes place in unstructured environments without set goals, such as using Minecraft to support diverse learners [27, 54, 112, 158]. **Training** focuses on skill development.

Feature	Values	A	B	Comparison
Environment function	• **Learning**: knowledge acquisition and conceptual understanding [31, 55] • **Training**: skill development through practice [39, 40]	• Learning dominant (57/73, 78%) [30, 42, 55] • Training rarer (16/73, 22%)	• Learning dominant (40/49, 82%) [123] • Training stable (12/49, 24%) [83]	• Learning contexts dominate across eras • Training remains a consistent secondary focus
Interaction setting	• **Virtual**: fully online or simulated, no physical co-presence [72, 120] • **Physical**: in-person classrooms, labs, clinics [118, 133] • **Mixed**: combined physical and virtual interaction [69]	• Virtual dominant (51/73, 70%) [3, 120] • Physical less common	• Physical dominant (34/49, 69%) [139] • Virtual reduced	• Shift from virtual to physical environments • Reflects post-COVID return to in-person learning

Table 3. Setting taxonomy, with trends and comparisons across Corpora A and B.

Didactic nature also influences methodological choices. Instructional environments are typically structured, facilitating controlled data capture and model-based analysis [141], while training emphasizes repetition and performance [84]. Informal learning settings are more open-ended and heterogeneous, often producing noisy or incomplete data that require qualitative, human-in-the-loop approaches [89]. Correspondingly, instructional goals emphasize conceptual understanding, whereas informal learning prioritizes creativity, exploration, and inquiry [44].

Finally, the level of instruction introduces distinct ethical and logistical constraints. K–12 contexts impose heightened restrictions due to the involvement of minors, with multimodal data collection—especially video and physiological sensing—raising privacy, consent, and health-data concerns [25]. GenAI further complicates deployment due to the risks of misuse and unintended behavior [56]. Adult learning environments, particularly in higher education, offer greater flexibility and fewer institutional barriers, helping explain their dominance in the literature. Expanding multimodal learning analytics in K–12 settings will require coordinated engagement among educators, researchers, parents, and policymakers to ensure ethical and effective adoption [114].

4.1.2 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 mixed scenarios—clarifying how multimodal learning analytics systems are deployed across different contexts. With both corpora, we characterize setting along two dimensions, presented in Table 3.

Across environment settings, a key distinction concerns the level of **physical engagement**, ranging from largely **stationary** to highly **active**. Learning environments are typically stationary, involving seated interaction with digital systems or peers, whereas training environments more often require physical movement and object manipulation. This distinction strongly shapes multimodal data capture and analysis. Active settings generate complex, high-dimensional data from motion, video, and physiological sensors, often requiring model-based approaches such as tree-based or deep learning models [83], but they also introduce practical constraints related to device standardization and scalability, and may require qualitative analyses to supplement findings.

practical training, and professional development in specific fields [57, 95, 99]. The **Unspecified** category includes papers that lack sufficient information about the didactic nature of the study [48].

2.2.10 Level of Instruction or Training. We sought to delineate the level of instruction or training for participants, defining four categories to provide valuable insights into the educational contexts targeted by the analyses in our corpus. **K-12** participants are those in kindergarten through 12th grade[55, 110, 154]. **University** participants include undergraduate and graduate students[38, 65, 98]. **Professional Development** participants are involved in professional development training[49, 51, 119]. The **Unspecified** category refers to papers that lack information about the participants' level of instruction or training[19, 43, 92]. It is important to note that studies featuring multiple groups of participants, or those reporting results across various studies, may have been assigned multiple labels.

2.2.11 Analysis Approach. Our systematic categorization of analysis methodologies identified two principal approaches: **Model-based** and **Model-free**. Model-based analysis employs a formal model to reveal the data's intrinsic structure and the interrelationships between variables. This approach involves hypothesizing about data structure and variable connections, often using mathematical functions to delineate the relationships in machine learning, or computational models to simulate system dynamics in cyber-physical systems. Conversely, model-free analysis eschews assumptions about data distribution, relying instead on empirical statistics, like correlations, to discern patterns and relationships directly from the data. It is important to note that these categorizations are not exclusive; a study may be classified as both model-based and model-free if it incorporates both types of approaches.

3 METHODS

This section outlines the methodology we employed to compile our literature corpus and ensure comprehensive coverage of pertinent research. We utilized a combination of quantitative (graph-based) and qualitative (quality control) techniques to refine our corpus to a representative yet manageable size. We introduce a novel graph-based method for literature corpus reduction, termed *citation graph pruning* (CGP) that is detailed in Section 3.2.1. CGP employs a directed citation graph, considering each paper's citation network, to identify and exclude outlier papers with minimal connections to the corpus, thus deemed beyond the review's scope. This graph-based pruning method is a unique contribution to literature review methodologies and has not been previously reported. Additionally, our quality control process, elaborated in Section 3.2.2, is derived from Kitchenham's systematic review procedures [77]. For an exhaustive description of our search strategy, corpus distillation, and feature extraction methods, refer to Appendix B.

3.1 Literature Search

Our literature search employed 42 search strings, collaboratively developed by the authors to encapsulate the relevant work for this review. We generated 14 search phrases, each queried thrice with variations of *multimodal* (multimodal, multi-modal, multi modal), detailed in Appendix B. Searches were conducted programmatically using Google Scholar via SerpAPI [128], chosen for its accurate retrieval of organic search results. For each search string, we selected the top five pages (100 publications) as ranked by Google Scholar, resulting in 4,200 papers. After removing 2,079 duplicates through hashing, and excluding 1 non-English paper, we obtained 2,120 unique papers.

3.2 Study Selection

After the initial search, we distilled the corpus quantitatively via citation graph pruning, which we discuss in Section 3.2.1. Subsequent distillation, performed qualitatively, is discussed in Section 3.2.2.

Virtual environments are comparatively easy to scale and instrument, as constrained movement supports streamlined data collection through logs, screen recordings, webcams, and wearables [117]. These conditions enable large-scale quantitative analyses, such as integrating discourse with system logs to study self-regulated learning [116]. Physical environments offer greater ecological validity—particularly in K–12 contexts [51]—and support social and embodied interactions not replicable online. However, they require on-site data collection, heightened ethical oversight, and contend with noise, technical failures, and unstructured dynamics that often yield incomplete data, leading studies to rely more frequently on qualitative or model-free methods such as observational coding and correlational analysis [1, 34].

4.1.3 Data Collection Media ("Sensors" in Figure 2). Data collection media (see Table 4) 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). Figures 4 and 5 compare data modality distributions across Corpora A and B. Video and audio were prevalent across both—indicating the richness and usefulness of integrating these modalities.

Medium	Definition
Video	Sequences of image frames captured from a camera source [43, 101].
Audio	Audio signals captured by a microphone [100, 122].
Screen Recording (Screen)	Sequences of image frames displaying a device's screen contents [55, 75].
Eye	Eye movement data and gaze points captured by tracking devices [16, 98].
Logs	Participant's actions within the system and its state data [10, 103].
Physiological Sensors (Physical)	Specialized sensors used to gather participants' physiological data [52, 59].
Interview	Structured or unstructured conversations between researchers and participants [84, 95].
Survey	Standardized sets of questions administered to participants [30, 35].
Participant-Produced Artifacts (PPA)	Materials produced by study participants using various media, including physical objects created for a task or written responses to formative assessment questions [8, 123].
Researcher-Produced Artifacts (RPA)	Materials produced by the researchers that contribute to analysis and findings, such as observational notes [52, 82].
Motion	Raw motion data collected via various different devices/technologies [39, 84].
Text	Raw textual input [91, 134].

Table 4. Data collection media taxonomy.

Post-COVID, 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 [23, 24]. The most

3.2.1 Citation Graph Pruning (Quantitative Corpus Reduction). For visualization and analysis, we used NetworkX to construct a *citation graph* from the initial 2,120 papers. This graph, a directed acyclic graph (DAG), features nodes representing papers identified by their Google Scholar UUID and directed edges denoting citations, i.e., paper A cites paper B. The degree of a node (paper) p is defined as the sum of incoming and outgoing edges, representing papers citing and cited by p, respectively. SerpAPI was utilized to retrieve the citation lists.

We first eliminated all 0-degree nodes, assuming their irrelevance to the field or lack of influence on subsequent research. Further analysis of the DAG's structure revealed one major component with 1,531 papers and 44 smaller, disconnected components (sizes 2-5), detailed in Appendix B.2.1. The disconnected components were then removed. Subsequent pruning involved iteratively removing 1-degree nodes until no new 1-degree nodes emerged, a process we term *citation graph pruning*, outlined in Algorithm 1. This pruning reduced the corpus to 1,063 papers.

Algorithm 1 Citation Graph Pruning Algorithm

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

procedure DegreeTrimming(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 $v \in e \wedge e \notin D$ **then** $D = D \cup \{e\}$
  return $(V \setminus S, E \setminus D)$
procedure SubconnectedGraphTrimming(G)
  $[S_1, S_2, S_3, ..., S_n]$ = 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 IterativeTrimming(G)
  while True **do**
    $G' =$ DegreeTrimming($G, 1$), where $G' = (V', E')$
    if $|V| == |V'|$ **then**
      break
  return (V', E')
$G' =$ DegreeTrimming($G, 0$) ▷ Remove 0-deg vertices
$G' =$ SubconnectedGraphTrimming(G') ▷ Keep largest connected subgraph
$G' =$ IterativeTrimming(G') ▷ Iteratively remove 1-deg vertices until equilibrium
return G'

3.2.2 Quality Control (Qualitative Corpus Reduction). Upon manually reviewing the 1,063 titles post-pruning, we found many papers irrelevant to our review's focus, such as those on training multimodal neural networks and applying multimodal methods in medical imaging. Using regex keyword searches (specified in Appendix B.2.2), we identified 217 titles for potential exclusion. After careful consideration, we removed 204 papers, retaining 13 for further evaluation, thus narrowing our corpus to 859 works. Consistent with Kitchenham's guidelines [77], we refined our corpus by sequentially reviewing titles, abstracts, and full texts, applying majority voting for exclusions, as detailed in Appendix B.2.2. This process reduced our corpus to 388 from title evaluation, 127 from abstracts, and 75 from full-text

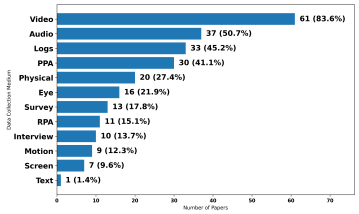


Fig. 4. Corpus A data collection media distribution.

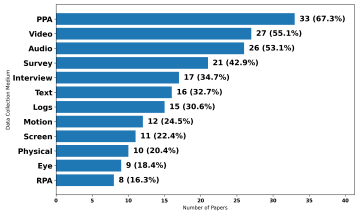


Fig. 5. Corpus B data collection media distribution.

striking shift was in textual input: Corpus A included only one study using raw text [134], while nearly one-third of Corpus B papers captured text as a primary data source [33, 72, 91]—almost entirely due to LLM-mediated interactions.

4.2 Multimodal Data

Multimodal data underpin MMLA systems by linking observable learner activity to analytic processes that support inference and action. Decisions about what data to capture, which modalities to derive, and how those modalities are combined, determine which learner states can be modeled and how analytics connect behavior to higher-level constructs such as collaboration quality, self- and group-regulation, and knowledge or skill acquisition. In our framework, multimodal data occupy the boundary between environment and analytics, mediating how evidence from different sources is interpreted and operationalized.

Research goals, target constructs, and participant interaction structures drive these decisions. For instance, analyses of socially shared regulation of learning typically integrate environment logs with discourse data to capture both cognitive activity and metacognitive or social processes [27, 116, 117]. In contrast, procedural skill training (e.g., CPR) prioritizes motion and physiological signals to evaluate performance quality [39]. Such choices shape both traditional and GenAI-enabled pipelines by constraining system architecture, interpretability, and the scope of analytics.

Approximately two-thirds of papers in both corpora use 3–5 distinct modalities to guide their research (A: 49/73 [2, 36]; 67%, B: 34/49 [20, 91]; 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. Table 5 presents the modalities, their descriptions, and the modality groups to which they may belong, derived through inductive coding based on the types of information they convey.

We present approaches and trends across modality groups, summarized in Table 6. Figures 6 and 7 show the modality distributions for Corpora A and B, largely reflecting trends in data collection media and 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. The following subsections detail how modalities within each modality group are deployed and examine challenges specific to each group.

assessments. Subsequent feature extraction led to the exclusion of two additional papers deemed outside our review's scope, culminating in a final corpus of 73 papers.

3.3 Feature Extraction

Once the corpus was finalized, we extracted several features from each of the 73 papers. This included identifying information (e.g., title, first author, publication year), and information related to the paper's methods (e.g., data collection medium, modalities, and analysis methods). Specifically, we extracted the following features from each paper (as outlined in Section 2.2): UUID, title, authors, publication year, environment type, data collection mediums, modalities, analysis methods, fusion types, publication, environment settings, domains of study, participant interaction structures, didactic natures, levels of instruction, and analysis approaches. We detail our feature extraction scheme and each feature's set of values in Appendix B.3.

3.4 Analysis Procedure

Leveraging our Figure 1 framework, we conducted a qualitative thematic analysis on the extracted features from our corpus. This yielded descriptive statistics and identified dominant trends for each framework component. We classified multimodal data into five comprehensive modality groups: (1) *natural language*, (2) *vision*, (3) *sensors*, (4) *human-centered*, and (5) *logs*. For each group and the entire corpus, we explored the state-of-the-art, challenges, research gaps, and outcomes of multimodal learning and training analyses. Furthermore, we distilled multimodal learning and training research into three distinct research types, termed *archetypes*. Our thematic findings for each framework component are detailed in Section 4, the archetypes in Section 5, and a comprehensive discussion of the corpus and field in Section 6.

4 FRAMEWORK INSIGHTS

We present our findings for the individual components in the Figure 1 framework, i.e., Environment, Multimodal Data, Data Fusion, Analysis, and Feedback. The results for each component are presented in the subsections that follow. For reference, terminology definitions are enumerated in Section 2.2.

4.1 Environments

We investigate learning and training environments for the three components, specified in our framework, i.e., setting, learners/trainers, and data. Setting refers to the environment where the learning and training occur, learners and trainers refer to the environment participants, and sensors refers to the data collection mediums used in the environment.

4.1.1 Setting. In Section 2.2.6, we categorized environments into four types: virtual, physical, blended, and unspecified. Our corpus revealed that virtual environments were predominant. This trend may be attributed to the increasing reliance on online platforms for educational engagement, a phenomenon that the COVID-19 pandemic may have accelerated (evidenced by a spike in our corpus's use of virtual environments in 2020). We initially hypothesized that recent technological advances may have engendered a rise in virtual multimodal learning and training; however, a temporal analysis of our corpus' use of environment settings did not support this. 51/73 papers (70%) incorporated at least some virtual component (i.e., used either virtual or blended environments), which suggests most multimodal learning and training research relies, at least in part, on virtual environments to collect and analyze data [6, 137, 142]. In addition, we consider the distribution of learning versus training environments, as described in Section 2.2.1. There were more than three times as many learning environments papers (57/73; 78%) [38, 54, 74] relative to training environments

Modality	Description	Modality Group
Affect	Facial expression, or emotional or affective state [36, 103].	N, V, P
Pose	Participant's physical position, location, or body posture [2, 119].	V, P
Gesture	Participant's gestures and body language [3, 100].	V
Activity	Participant's observable actions or activities [47, 102].	V, P
Prosodic Speech	Elements of speech beyond word meaning, e.g., volume, pauses, and intonation [94, 118].	N
Transcribed Speech	Textual speech transcribed from audio [30, 74].	N
Qualitative Observations	Researchers' observations about participant and study task [59, 81].	H
Logs	Participant's environment actions and system state data [10, 50].	L
Gaze	Participant's eye gaze, e.g., movement, direction, and focus [42, 138].	V, P
Interview	Interview notes between researchers and participants [40, 55].	H
Survey	Participant's responses to surveys/questionnaires [98, 99].	H
Pulse	The participant's pulse, indicating their heart rate [68, 127].	P
EDA	Participant's electrodermal activity [79, 113].	P
Temperature (Temp.)	Participant's body temperature [98, 113].	P
Blood Pressure (BP)	Participant's blood pressure [98, 127].	P
EEG	Participant's electroencephalography activity [50, 113].	P
Fatigue	The level of fatigue experienced during the activity [68, 69].	V, P
EMG	Participant's electromyography activity [37, 39].	P
Participant Produced Artifacts (PPA)	Artifacts produced by the participant during the study, e.g., pre/post-tests [16, 88].	H
Researcher Produced Artifacts (RPA)	Artifacts produced by the researcher about the study and participants, e.g., field notes [22, 95].	H
Spectrogram (Spect.)	Representation of audio frequencies in the form of a spectrogram [78].	N
Text	Participant's raw text data generated in the study environment [134].	N
Pixel	RGB pixel values from cameras or sensors [102].	V

Table 5. Modalities and their modality groups. N = Natural Language, V = Vision, P = Phys. Signals, H = Human-Centered, L = Logs.

4.2.1 Natural Language. Natural language captures how learners and trainees speak, write, and interact with peers, instructors—and increasingly, LLM-based systems. Because much teaching, collaboration, feedback, and assessment are inherently language-based, NLP signals like prosody and text embeddings often encode rich information about learners' metacognition (e.g., goal setting, planning, and reflective behaviors) as well as collaborative processes such as information pooling and consensus building [25, 46]. For example, Snyder et al. [117] 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.

papers (16/73; 22%) [51, 53, 95]. This imbalance underscores the focus of educational literature on knowledge acquisition. In contrast, the lower frequency of training settings may reflect a narrower scope centered on skill enhancement and professional development. Notably, environments emphasizing physical activity were largely absent from our corpus. This includes environments focusing on activities like rehabilitative therapy and athletic training, as well as *embodied learning* [140] environments that require students to physically engage in the learning activity.

4.1.2 Learners/Trainees. This review examines key elements of the learner's domain, including the domain of study, participant interaction structure, didactic nature, and the level of instruction or training. These elements collectively contribute to a comprehensive understanding of the learner's experience and the educational context. Our corpus predominantly focuses on STEM+C domains of study (55/73; 75%) [20, 57], with humanities (11/73; 15%) [107, 115] and psychomotor skills (5/73; 7%) [65, 98] being less represented. Four papers did not specify the domain of study, and two addressed domains outside of STEM+C, humanities, and psychomotor skills. This distribution suggests a significant emphasis on STEM+C education, reflecting global trends toward these disciplines' importance in technology-driven societies and their relevance to the job market and societal advancement.

Individual-focused learning and training environments are the most prevalent participant interaction structure (45/73; 62%) [9, 138], compared to multi-person environments that are present in 31 (42%) papers [53, 110]. This indicates most studies focus on individual learning and training experiences, which may allow for personalized and self-paced progress. However, the notable presence of multi-person settings underscores the importance of collaborative and social learning environments in educational research. The didactic nature of environments is predominantly formal and pedagogical (45/73; 62%) [86, 91, 143], followed by training (15/73; 21%) [53, 93, 106] and informal learning (12/73; 16%) [27, 65, 158]. This suggests that formal instruction is the predominant mode, with a smaller yet significant focus on training and informal learning, which may include more interactive, practical, or workplace-based scenarios. University-level instruction dominates (36/73; 49%) [21, 105], followed closely by K-12 environments (30/73; 41%) [82, 101]. Professional-level learning is less frequent (5/73; 7%) [51, 105]. The prominence of university-level participants may reflect the research emphasis and academic focus of higher education, while the strong representation of K-12 indicates ongoing interest in foundational education practices. The underrepresentation of professional settings suggests a research gap in lifelong learning and continuing education.

The data on learner characteristics in our corpus highlights a landscape where STEM education is prioritized, individual learning experiences are valued, formal instruction is the standard, and university and K-12 education levels are emphasized. However, the presence of other educational levels and informal learning contexts indicates that there exist a diverse range of learning experiences and instructional approaches. This diversity presents both challenges and opportunities for educators and researchers, emphasizing the need to tailor educational strategies to various learning environments and address the unique requirements of different learner demographics.

4.1.3 Data Collection Mediums. Figure 4 presents the distribution of the various data collection mediums used by the papers in our literature corpus. As depicted, the current state-of-the-art in data collection mediums reflects a diverse array of technologies and methodologies, with video leading (61/73; 84%) [117, 156], followed by audio (37/73; 51%) [23, 153]. These two mediums indicate a preference for rich multimedia data that can capture the complexities of learning and training, as

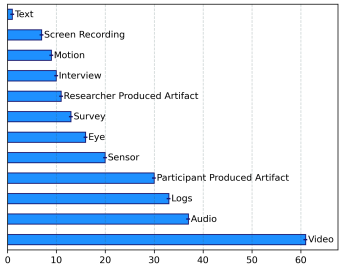


Fig. 4. Data collection mediums distribution.

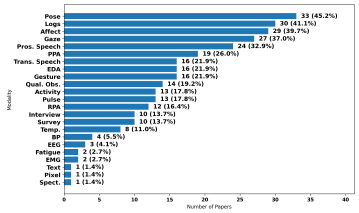


Fig. 6. Corpus A modalities distribution.

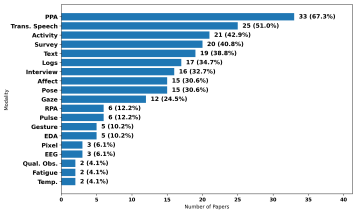


Fig. 7. Corpus B modalities distribution.

The use of natural language modalities in MMLA—particularly through text and transcribed speech—has increased precipitously with the integration of LLMs, revealing distinct challenges for both corpora. Corpus A, with its reliance on traditional machine learning, frequently cites issues such as small, imbalanced datasets [75, 99] that hinder the training and adaptation of deep learning and transformer models [99]. The corpus's emphasis on audio data introduces additional challenges, particularly the time-intensive nature of feature extraction and complexity [102], while manual preprocessing and feature engineering further constrain scalability [67, 74]. Conversational agents are rare in Corpus A; when used, they typically deliver static messages or summative feedback rather than engaging in multi-turn interaction [35, 75, 122]. 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 [67, 118, 134] and concerns surrounding LLMs like adverse interactions with students and how to effectively evaluate LLM output [56]. Additionally, LLM-based systems raise skepticism due to risks of hallucination, toxicity, and misuse [28].

4.2.2 Vision. Vision-based modalities offer continuous insight into how learners move, attend, react, and interact in learning and training environments. 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 [119]) and qualitative methods (e.g., interaction analysis), and are frequently triangulated with other modalities such as logs and speech [81].

The rise of multimodal LLMs such as GPT, Gemini, and Claude 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. [136] 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.

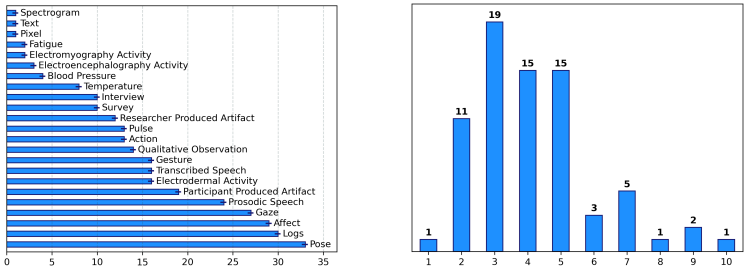
As with natural language, vision-based modalities often contribute significant predictive power to multimodal pipelines. For instance, Ma et al. [78] used early fusion to combine video features (e.g., facial expressions, body movements, inter-learner distance), linguistic features (text embeddings), 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 robust predictors of impasse, underscoring the importance of visual signals in capturing nuanced social dynamics.

well as interactions within the environments. Logs (33/73; 45%) [65, 110] and participant-produced artifacts (30/73; 41%) [8, 80] are also popular, suggesting a strong inclination toward capturing learner behaviors and outputs directly from both the environments and the participants themselves.

Despite these advances, the field faces challenges in integrating data from disparate sources and ensuring data quality and privacy. For instance, sensor data (20/73; 27%) [87, 147] presents challenges in standardization and interpretation. Although less prevalent, eye-tracking and motion capture data raise concerns about intrusiveness and the need for sophisticated analysis techniques. There is also a notable gap in text-based data collection (only one paper [158] in the corpus), as learning and training environment research currently relies primarily on transcribed speech.

4.2 Multimodal Data

Figure 5 breaks down of the different modalities used in our corpus. "Pose" is the most prevalent modality, appearing in



(a) Frequency counts for the number of papers in our corpus containing each modality.

(b) The number of modalities used per paper, i.e., how many papers (y-axis) used n modalities (x-axis).

Fig. 5. A breakdown of the individual modalities used in our corpus both in terms of frequency count (left) and the number of modalities used per paper (right).

some form in roughly 45% of papers (33/73) [51, 93, 107]. Logs, affect, gaze, and prosodic speech modalities are also common. At least one of the top five modalities appears in all but eight papers in our corpus (65/73; 89%). The remaining modalities appear less frequently, with raw text, raw pixel value, and audio spectrogram only appearing in one paper each. A large majority of papers (60/73; 82%) use 2-5 modalities in their multimodal analyses. One paper used only a single modality[1], and one paper used 10 modalities. We hypothesize that researchers typically choose between 2-5 as a compromise between overhead and informativeness, but more research is required to evaluate this quantitatively.

Diving deeper into the multimodal data, we identified five modality groups that best characterize the types of data driving multimodal learning and training methods: natural language, vision, sensors, human-centered, and logs. The

[1]By our definition of "multimodal" in Section 1.3, we consider a paper to be multimodal if multiple modalities are used during analysis *or* multiple data collection mediums are used. One paper [27] collected both video and audio data, from which the authors derived a single modality: researcher-produced artifacts. For this reason, there is one paper in our corpus that uses only one modality in its analysis pipeline, still adhering to our definition.

Modality	Description / Features	A	B	Comparison
Natural language	• Text- and speech-derived features (e.g., raw text, embeddings, term frequency, prosody) • Strong predictors of productivity, performance, and collaboration states • Central in collaborative and dialogue-rich settings	• Moderate use (35/73, 48%) • Prosodic speech dominant • Traditional ML + qualitative discourse analysis [78, 99]	• Dominant modality group (40/49, 82%) [41, 73] • Text and transcribed speech dominant • LLM-centered dialogue, feedback, and assessment [93]	• Sharp shift toward language-centric pipelines • LLMs act as both analytic tools and interaction partners
Vision	• Video-derived pose, gaze, affect, activity • Used as features or prediction targets • Quantitative, model-based analysis	• Very common (59/73, 81%) • Pose, gaze, affect among top modalities [81, 120]	• Reduced prevalence (27/49, 55%) • Activity increases post-COVID [6] • Rare integration with multimodal LLMs	• Decline in vision dominance • Vision increasingly contextual rather than central
Physio. signals	• Signals such as EDA, heart rate, motion • Capture arousal, stress, and cognitive load • Often fused with vision, language, and logs	• Less common (20/73, 27%) • Dominated by EDA (80%) • Concentrated in training contexts [133]	• Increased use (23/49, 47%) • Shift toward activity and motion signals [83] • Used in interaction-rich settings	• Broader deployment and integration • Bridge between observable behavior and internal state
Human-centered	• Participant- and researcher-produced artifacts • Surveys, interviews, reflections, annotations • Support stakeholder agency and interpretation	• Common (45/73, 61%) • Participant artifacts primary [42] • Mostly qualitative analysis	• Nearly universal (46/49, 94%) • Strong emphasis on design and trust [57] • Increased quantitative ethnography [108]	• Major growth in user-centered approaches • Human data central to GenAI validation and safety
Logs	• Interaction traces (clicks, sequences, inactivity) • Support behavioral and strategy inference • Traditionally rule-based and supervised ML	• Widely used • Predict performance and engagement [47, 79]	• Declining relative reliance • Logs integrated into LLM prompts as text	• Shift from rule-based to language-mediated interpretation • Logs increasingly contextualized via LLMs

Table 6. Multimodal data: modality groups across Corpora A and B.

However, vision modalities face several practical and methodological challenges and are often perceived as invasive [80]. 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 [7, 29, 132]. 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. Missing data,

following subsections present our findings with respect to each modality group. For each modality group, we identify the individual modalities it comprises and discuss our findings with respect to its prevalence in the corpus, current state-of-the-art, challenges faced, research gaps, and results achieved.

4.2.1 Natural Language. 35 out of the 73 (48%) corpus papers collected and analyzed some form of natural language data. The natural language modality group comprises prosodic speech (24/73; 33%), transcribed speech (16/73; 22%), raw text (1/73), audio spectogram (1/73), and affect (when derived from text or audio; 2/73). All but three natural language papers included prosodic or transcribed speech, but only eight papers incorporated both. Because prosodic speech is devoid of semantic meaning, and transcribed speech lacks important prosodic information, combining the two provides a more holistic language representation. However, research combining the two modalities was not well-represented in our corpus and represents a notable research gap.

Traditional machine learning methods were the most prevalent quantitative approaches in the natural language modality group. In particular, support vector machines [90, 115, 136] and logistic regression models [43, 85, 114] were often used with natural language features. Other approaches like random forest [101], linear regression [135], and naive Bayes [136] were also used, typically to predict outcomes such as learning or training gains. There was a noticeable lack of deep learning approaches for natural language processing (NLP) in our corpus. While some papers incorporated recurrent neural networks (e.g., LSTM models [70]), these were a relative rarity. Very few used transformer [149] models like BERT, which was surprising given their prevalence in more recent NLP work at large. This indicates that the multimodal methods for learning and training environments using natural language lag behind the current state-of-the-art in NLP [148]. However, this is likely in large part due to the small sample sizes and noisy data innate to learning and training environments that are insufficient to train many deep learning models.

Education- and training-specific datasets are often small, imbalanced, and contain domain-specific terminology that language models may not have encountered frequently during training [28, 30, 85, 86, 114]. These issues complicate the effective training of deep learning models [9, 29, 32, 114, 138]. Additional challenges include the complexity and time cost of cleaning, processing, and labeling data. Software packages like NLTK [88], openSMILE [56], and TAACO [40, 41] facilitate the programmatic extraction of audio- and text-based features, yet this can result in large, opaque feature sets [119]. Conversely, manual preprocessing and feature engineering can be time-intensive, potentially limiting the data researchers are willing to collect and analyze [79, 85]. This may explain why qualitative analysis of smaller sample sizes is common in natural language studies.

Qualitative analyses using natural language primarily involve presenting descriptive statistics, case studies, and researchers' observations, and conducting various forms of qualitative coding [53, 92, 107]. Many natural language studies focus on collaborative learning and training [89, 110, 135], favoring multi-person environments to leverage the richness of collaborative discourse. However, analyzing transcribed speech poses challenges. Several studies noted that automatic speech recognition (ASR) is a bottleneck in multimodal pipelines using transcribed speech [79, 135, 158]. Learning environments often consist of multiple groups participating simultaneously, creating noisy conditions that hinder ASR accuracy, particularly in K-12 settings [79] and among non-English speakers [158].

Only one paper in the corpus used raw text as input [158], which is surprising given the prevalence of text-based transformer models [17, 47, 122]. Considering the capabilities of large language models (LLMs), text-based features could significantly enhance multimodal learning and training pipelines, as raw text quality does not depend on ASR. One potential avenue for leveraging textual features is through conversational agents, which were notably absent in our corpus. While several works addressed multimodal agents or tutors [43, 86, 142], these agents typically

differing temporal resolutions, and a lack of standardization further complicate 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 [7].

4.2.3 Physiological Signals. Physiological signal-based modalities capture learners' physiological and motion-related traces. 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. The wide range of modalities derived from physiological and motion-based sensors (see Table 5) has yielded diverse and insightful findings across multimodal learning and training pipelines. For example, Que et al. [104] combined gaze data from 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. They found that *extraneous load* (avoiding irrelevant information while learning) was predicted by increased fixation count and lower mean heart rate; *intrinsic load* (reflecting the complexity of learning material) by increased fixation count and mean saccade amplitude; and *germane load* (resources available for processing intrinsic load, e.g., comprehension) 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 [80]. Teachers and students have also emphasized the importance of understanding how machine learning models generate predictions [28], 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 [105].

4.2.4 Human-Centered. Human-centered modalities ground MMLA in learners' lived experiences and in educators' and researchers' perspectives, providing insight into how participants perceive, interpret, and reflect on tasks—insight that is typically inaccessible through sensor- or log-based data alone. Such data are commonly used to enrich quantitative findings with contextual detail (e.g., through case studies or error analyses) and are often treated as ground truth in predictive modeling or in correlating learning behaviors with outcomes. Human-centered modalities play a critical role in validating inferences from other data sources, improving the interpretability of model outputs for stakeholders, and elucidating how multimodal systems are experienced and perceived in practice.

For example, Sung et al. [121] applied multimodal *epistemic network analysis* (ENA[6]) [109] in a guided reading study in a college biology course to examine differences in self-regulated learning between mastery and non-mastery groups. Think-aloud data were coded for self-regulation strategies, environment logs distinguished in- and out-of-class engagement, and quiz scores indexed learning outcomes. While monitoring behaviors co-occurred frequently with other strategies in both groups, mastery students paired monitoring with domain-specific strategies, while the rest paired it with domain-general strategies—prompting follow-up qualitative analyses to contextualize these differences.

[6]ENA transforms coded qualitative data into visual networks that represent the co-occurrence of coded concepts over time. Data segments are coded according to a theoretical framework; nodes denote codes, and edges encode their co-occurrence within a defined window, enabling temporal and group-level comparisons of learning processes.

provided summative performance metrics or canned responses. No studies addressed conversational agents that engage dynamically with learners as peers, mentors, or collaborators.

Despite these gaps and challenges, natural language features consistently produced positive outcomes. Researchers successfully correlated and predicted various learning outcomes using these features. This was especially evident in studies focusing on collaborative learning and training, where the collaborative environments provided discourse rich in natural language features. Collaboration was examined both as an independent and dependent variable [135, 136, 157]. In these collaborative settings, natural language features frequently were the most informative among all modalities [85]. Additionally, natural language features were usually the most predictive when combined with features derived from other modalities. This reinforces previous work, where multimodal data harnessed more predictive power than any individual modality [130]. Researchers often reported that including natural language features in the multimodal pipeline led to improved predictive performance [110]. Overall, the results reported in our corpus clearly indicate that natural language features have: 1) high correlations with performance outcomes, and 2) provide enhanced predictive capabilities when combined with features derived from other modalities.

4.2.2 Vision. Among the five groups of modalities analyzed, vision-based modalities were the most utilized, appearing in 59 out of 73 papers (81%). The vision modality group includes papers that collected data using cameras or eye-tracking devices and analyzed it for pose recognition, affect detection, gesture recognition, activity recognition, fatigue estimation, participant gaze, or raw image pixel data. Pose, affect, and gaze were the most common, present in 33 (56%), 25 (42%), and 27 (46%) of the 59 papers, respectively. Gesture recognition appeared in 16 papers (27%), activity recognition in 11 papers (19%), and fatigue estimation and raw pixel data in 2 and 1 papers, respectively.

This distribution is expected. Pose recognition was the most frequent due to the availability of off-the-shelf deep learning models and the use of Microsoft Kinect cameras, which facilitate pose data collection. Gaze tracking was common with specialized hardware like eye-tracking glasses. Affect recognition was also prevalent, supported by off-the-shelf models. Notably, raw pixel data was the least used, appearing in only one paper. Researchers typically processed raw images using other models before analysis, highlighting the importance of mid-fusion techniques. This pattern reveals a mismatch between core and applied computer vision research, with the latter relying on pre-trained models due to smaller datasets.

In terms of analysis methods, there was a slight preference for quantitative techniques in the vision subset, with 69% of papers using model-based methods compared to 63% in the full corpus. Despite this, many papers combining qualitative and quantitative analysis also used vision data. Only 24% of the vision papers employed mixed-methods analysis, often combining classification with qualitative analysis of classes.

4.2.3 Sensors. We identified 20 papers (27%) in sensor-based learning and training research, covering various physiological and behavioral data modalities. These papers focused on affective responses (11), body pose analysis (7), electrodermal activity (16), pulse rate (11), activity (5), blood pressure (4), temperature (8), electroencephalography (3), electromyography (2), fatigue (2), and gaze tracking (8).

Of these 20 papers, 12 are learning-based, and 8 are training-based. This suggests sensors are more frequently used in training-based research, which represents only (16,73, 22%) of the full corpus but 40% of papers using sensors. Within MMLA, wearable sensors monitor learners' emotional and physiological states, predict behavior and performance, provide real-time feedback, and enable multimodal data integration [53, 69, 147]. Sensor use ranges from classroom

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 [85], which may propagate into downstream models, compromise generalizability, and often necessitate triangulation with other modalities to ensure reliability [96]. 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.

4.2.5 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% [2, 98], B: 17/49; 35% [26, 91]). For example, Cohn et al. [24] 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 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 mixed 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[7] 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.

4.3 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. We explore both components in detail in the following subsections.

4.3.1 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 integrated 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.

[7] Learning Management System

environments to specialized training scenarios (e.g., CPR instruction [51]), serving as assessment tools and mechanisms for real-time educational interventions. However, integrating and interpreting sensor data presents challenges, particularly for accurate and practical real-time applications [49, 131].

The state-of-the-art in sensor-driven multimodal learning and training analytics features advanced predictive modeling, real-time feedback systems, and multimodal data fusion. However, there is a need for more granular data analysis to identify subtle patterns and correlations not apparent through traditional methods. Contextual and behavioral analytics link physiological responses to specific learning activities in real-time. Signal processing methods aggregate sensory information into physical or learning characteristics, such as relative learning gains [147], team dynamics [53], and shared physiological arousal [102]. The field also requires robust, interactive visualizations that convey complex sensory data intuitively, and Explainable AI (XAI) methods to clarify how sensor data contributes to predictive models, enhancing interpretability [125, 126].

There is a noticeable gap in longitudinal studies to assess the sustained impacts of sensor-based technologies. Expanding sensor research to diverse learning contexts and demographics will help understand the broader applications. Sensor research often occurs in controlled environments, so scaling for widespread use and ensuring generalizability across diverse settings remains challenging. One example is Echeverria et al.'s study [53] using accelerometer data in nurse training simulations, which could benefit from integrating additional sensory inputs like gyroscope and magnetometer data for a multidimensional analysis. Investigating user experience and acceptance of wearable technologies in education, particularly regarding comfort, usability, perceived effectiveness, and privacy, is also needed.

4.2.4 Human-Centered. The human-centered approach offers insights into participants' experiences, perceptions, and behaviors, often identifying nuances that may be missed in quantitative analyses. Out of 73 papers, 45 (62%) incorporate at least one human-centered modality (qualitative observation, interview, survey, researcher-produced artifact, or participant-produced artifact), indicating a strong focus on human experiences. Participant-produced artifacts are the most common (19/73, 26%), followed by qualitative observation (14/73, 19%), researcher-produced artifacts (14/73, 16%), and both interview notes and survey responses (10/73, 14%). Participant artifacts often include diverse materials and standardized tests, with pre- and post-tests being the most prevalent for calculating learning gains [54, 123, 156]. The considerable use of qualitative observations highlights the importance of insights from direct human interpretation of behaviors. Common combinations include qualitative observations and participant artifacts [85, 86, 156, 157], participant artifacts and researcher artifacts [38, 123, 136, 139], and interview notes and qualitative observations [9, 74, 105, 157]. One study applied clustering, NLP, and linear modeling to researcher artifacts detailing student behaviors [27].

A predominant strategy involves transforming human-centered modalities into quantifiable data for statistical analysis. Examples include López et al. using survey data [89], Ochoa and Dominguez using participant-produced artifacts [106], and Bert et al. using both participant-produced artifacts and interview transcriptions [11]. This shows a preference for quantifiable insights from human-centered modalities. Fourteen papers focus on qualitative analysis, emphasizing rich, qualitative insights. Most papers adopt multiple analysis methods, with only 16/45 using one method exclusively, 15/45 integrating two methods, and 13/45 using three. Worsley and Blikstein [157] employ four analysis methods to identify correlations between multimodal data, experimental condition, design quality, and learning, using both human-annotated and automatically annotated data.

Human-centered approaches pose challenges related to subjectivity, scalability, resource intensiveness, and generalizability. The subjectivity of human-centered modalities may introduce bias [97, 103]. These approaches are resource-intensive, requiring trained researchers for data collection, coding, and analysis. Manual collection and analysis can

The conventional classification of fusion methods in MMLA, as defined by Chango et al. [17], 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 fails to capture the complexity of contemporary MMLA. Our review revealed persistent challenges in categorizing fusion practices, driven largely by inconsistent definitions of "*raw*" versus "*processed*" features. For example, skeletal joint position data from a Microsoft Kinect may be considered raw because it is directly provided by the device, yet processed because it is internally computed from depth data. Interpretations depend on how sensors are conceptualized: when treating the Kinect as an integrated sensor, skeletons may be viewed as raw observations; conversely, when using the Kinect as a video camera, skeletons are inherently derived observations.

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. [38] 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 primary fusion categories that are summarized in Table 7 and illustrated in Figure 8.

While less common, Corpus B explored fusion with multimodal LLMs to enable end-to-end interpretation of complex, multimodal artifacts. For example, Whitehead et al. [131] 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.

4.3.2 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

be time-consuming and may not scale well, especially in large-scale educational settings. Despite these challenges, human-centered approaches offer transparent and interpretable insights. These insights highlight gaps in integrating qualitative and quantitative methods. Developing methodologies that combine qualitative nuance with quantitative rigor is essential. The lack of standardized coding practices for human-centered modalities hampers replicability and comparability. Establishing standardized coding frameworks is crucial to enhance the reliability and credibility of machine learning analyses. Additionally, automating human-coding processes is a significant research need.

4.2.5 Logs. Thirty papers (40%) in the corpus incorporated log data (log-analysis papers). Logs, often from computer-based environments, link complementary modalities to learning outcomes and behaviors. Logs are frequently combined with video (25/30, 83%), eye-tracking (12/30, 40%), audio (12/30, 40%), participant-produced artifacts (11/30, 36%), survey responses (6/30, 27%), sensors (8/30, 26%), and motion (3/30, 10%). This highlights the diverse ways environmental logs are contextualized. Human-centered artifacts were less commonly combined with log data. Overall, log-analysis papers focus on computer-based learning environments and individualized instructional or informal activities.

The state-of-the-art in log-analysis features various approaches. Nearly all classification and regression papers used machine learning algorithms, such as support vector machine, random forest, naive Bayes, and logistic regression [62, 91, 161], to predict students' achievement, engagement, or emotional state. Deep learning approaches like CNNs [136] and LSTMs [98, 110] were used in only three papers. Statistical methods were used to correlate learning variables (e.g., perceived student emotion) to outcome variables (e.g., learning gains).

Analyzing logs presents hurdles, including time-cost, data scarcity, generalizability, and engineering expenses. Temporal aspects introduce difficulties, such as aligning time frames, handling different sampling rates, and managing time-series data. These complexities often result in smaller datasets, limiting scope and scalability. Data scarcity exacerbates the challenge of producing generalizable findings, while high software development and engineering costs hinder integrating modern features like real-time collaboration tools.

These challenges create gaps in log-analysis research. There is a deficiency in applying methods and findings from one educational setting to another, likely due to diverse educational contexts. Embracing standardized log formats and consistent practices may overcome this barrier, leading to more unified research approaches and broader applicability of insights. The low adoption rate of industry standards like xAPI [134], LTI [1], and Learning Management Systems (LMS) in educational technology research reflects a broader issue of aligning with best practices. Addressing these gaps and embracing these standards could enhance interoperability, scalability, and more robust analysis of educational data, paving the way for more impactful and transformative educational research and practices.

4.3 Data Fusion

In our analysis corpus of 73 papers, we observed multiple approaches to data fusion in multimodal learning and training. The choice between different types of fusion depends on the characteristics of the data, the nature of the environmental task, and the desired level of integration. Each fusion strategy has strengths and limitations, and researchers often select the most suitable approach based on the specific requirements of their study and research goals. One noteworthy observation in this corpus is that several papers do not explicitly explain or justify their fusion choices.

Figure 6 shows the distribution of fusion types across the 73 papers in the corpus. 54 (74%) perform early, mid, late, or hybrid fusion. The distribution of fusion types reveals that mid fusion is the most prevalent (27/73; 37%), showcasing its popularity in integrating information from different modalities by combining derived, observable features. Hybrid fusion follows closely with 19 papers (26%), utilizing a combination of early, mid and/or late fusion strategies. Early

Type	Description / Examples	A	B	Comparison
Early	• Fusion on raw, directly observable data • No modality-specific inference prior to fusion • E.g., raw Kinect depth data	• Rare outside hybrid pipelines • Appears mainly in sensor-heavy studies [133]	• Similarly rare • Often subsumed within hybrid approaches	• Standalone early fusion uncommon • Raw-data fusion limited by noise and scalability
Mid	• Fusion on processed but observable features • Clarifies boundary between raw data and inference • E.g., skeletal joint positions from video	• Most common approach • 27/73 papers (37%) [47, 127]	• Remains dominant • 19/49 papers (39%) [41, 71]	• Stable preference across eras • Reflects balance between interpretability and flexibility
Late	• Fusion on inferred, unobservable constructs • Operates in hypothesis space • E.g., planning, motivation, collaboration quality	• Rare as a standalone strategy • Mostly paired with mid fusion [103]	• Similarly rare • Used selectively for high-level inference	• Less used: limited inter-modal interaction, potential information loss • Often embedded w/in hybrid pipelines
Hybrid	• Fusion at multiple stages of processing • Combines strengths of early, mid, and late fusion • Common with 3+ modalities	• Moderate prevalence • 19/73 papers (26%) [21, 138]	• Less frequent • 8/49 papers (16%) [9, 115]	• Decline in post-LLM era • Possibly supplanted by language-centric pipelines
Other	• No fusion, unspecified fusion, or qualitative fusion • Modalities integrated interpretively by researchers	• Substantial portion • 20/73 papers (27%) [82]	• Slightly increased • 15/49 papers (31%) [72]	• Persistent ambiguity in fusion reporting • Highlights need for clearer methodological detail

Table 7. Data fusion approaches across Corpora A and B.

about learning processes. In contrast, model-free approaches avoid such assumptions, instead using empirical statistics 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 8, 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.

There is a notable shift from model-based analysis in Corpus A (57/73; 78% [8, 50]) to model-free approaches in Corpus B (33/49; 67% [135]). 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 physiological sensor data to predict outcome variables such as performance or engagement [2, 118]. They focus primarily on individual learners, reflecting the difficulty of capturing complex social dynamics within formal, parameterized models.

fusion is observed only in 3 papers , while late fusion is employed in 8 papers. 20 papers (27%) adopt other types of fusion strategies, no fusion, or do not explicitly mention data fusion.

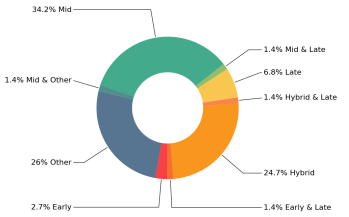


Fig. 6. Distribution of Fusion Types

4.3.1 Early Fusion. In early fusion, the joint feature representation incorporates information from all fused modalities, enabling the model to learn relationships and patterns directly from the raw, integrated features. This approach is advantageous when the modalities offer complementary information. In our corpus, early fusion was utilized in less than 5% of the papers. Early fusion may not always be the most suitable because it may not be clear what features are the most important until after processing and analyzing them. Further, early fusion may be computationally prohibitive, as the dimensionality of raw data is typically higher than that of its processed output.

4.3.2 Mid Fusion. Mid fusion combines features derived after prior processing but within the input space of observable features across modalities. It is advantageous when individual modalities require unique processing, and combining feature-level decisions is more effective than integrating raw features. 27/73 papers (37%) used mid-fusion as opposed to early fusion (3/73). This suggests that mid fusion is deemed more suitable for addressing the challenges and objectives in multimodal learning and training within the corpus.

4.3.3 Late Fusion. In late fusion, the models for each modality operate independently until the final decision or inference stage, where their outputs are aggregated to make overall inferences. This approach is suitable when modalities are semantically more independent, and their contributions are better understood when combined at a later stage. In our corpus, 8 papers (11%) employed late fusion, with 3 of them also employing other types of fusion [20, 21, 141]. Except for 1 paper that used regression [117], the majority used late fusion for classification purposes.

4.3.4 Hybrid Fusion. Hybrid fusion integrates information at different stages of the analysis pipeline and its design varies based on the learning or training task under consideration, the goals of the analysis, and the characteristics of the data. Late fusion was employed in 19 out of the 54 (35%) papers that performed fusion, highlighting the significance of this approach. Most papers (14/19; 74%) incorporated at least 4 modalities. Classification was the predominant analysis method (15/19; 79%).

4.4 Analysis

We defined our corpus's analysis approaches as model-based and model-free (see Section 2.2.11). The choice depended on the data and research questions. Model-based methods rely on assumptions about system operations, while model-free methods demand careful attention to data quality and reliability. Their methodologies are different and they create a separation of research communities. However, they are best used together to complement each other's strengths and weaknesses.

As shown in Fig. 7, 46 papers (63%) in our corpus use model-based methods, 16 papers (22%) employ model-free methods, and 11 papers (15%) use both. This distribution, with 78% (57/73) of papers employing model-based analysis, indicates a strong preference for developing models to inform analyses processes. Conversely, model-free approaches.

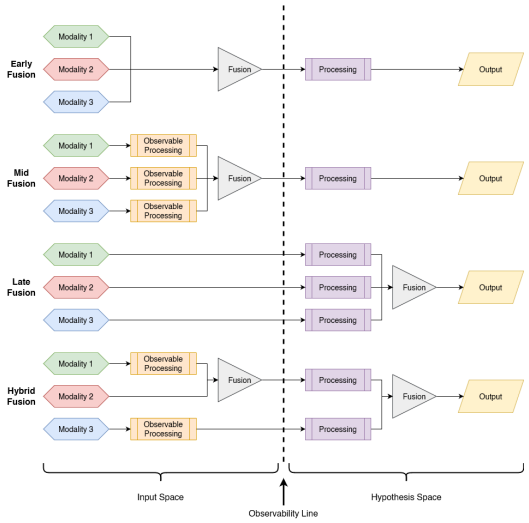


Fig. 8. Multimodal data fusion scheme according to when fusion is performed relative to the observability line.

Method	Definition
Classification	Assigning pre-defined labels to input data based on feature analysis through supervised learning (often via deep learning approaches) [103, 120].
Regression	Predicting continuous numerical values through supervised learning to understand input-output relationships [36, 101].
Clustering	Grouping data based on patterns or similarities using unsupervised learning [3, 22].
Qualitative	Manually examining and interpreting data to uncover patterns or themes [59, 81].
Statistical	Using statistical methods (e.g., correlation) to analyze data and draw conclusions [74, 77].
Network analysis	Studying relationships and interactions using graph-based approaches [21, 94].
Pattern extraction	Identifying meaningful patterns or structures within data, including techniques like Markov analysis and sequence mining [92, 98].

Table 8. Analysis methods taxonomy.

which make up 37% (27/73) of the papers, offer a valuable alternative for investigating learning and training outcomes in a more exploratory manner.

4.4.1 Model-Based. Model-based methodologies, such as machine learning models, employ mathematical frameworks to generate results from given inputs. Among papers using only model-based approaches, common analysis methods include classification (34/46; 74%), statistical analysis (17/46; 37%), regression (8/46; 17%), and clustering (7/46; 15%). These methods train models using data samples to predict factors like learning outcomes. When qualitative and pattern recognition techniques use model outputs to guide their analysis, they are also considered model-based. A notable aspect of model-based approaches is their focus on individual experiences (31/46; 67%) over collaborative ones (17/46; 37%). This trend likely arises from the complexities of mathematically representing intricate social interactions in group settings. Modeling an individual's cognitive, behavioral, and emotional states is challenging; thus, accurately reflecting collaborative dynamics in models is mostly confined to a niche within MMLA and social network analysis.

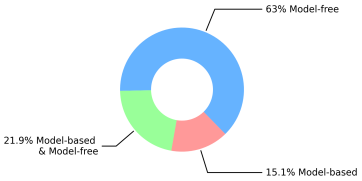


Fig. 7. Analysis approaches percentage distribution.

4.4.2 Model-Free. Model-free methods adopt a comprehensive, exploratory strategy, focusing on relationships between variables without assuming a specific link between input and output. Predominantly, these involve qualitative (11/16; 69%), statistical (9/16; 56%), and pattern recognition (3/16; 19%) methods. Qualitative methods are used in scenarios like use case and interaction analysis, where observations and learning theories guide the understanding of the learning process. Statistical and pattern recognition methods provide descriptions and correlation metrics between learning activities (e.g., behaviors and strategies) and outcome metrics. Serving as a counterbalance to the limitations of model-based methods, model-free approaches are widely used in collaborative settings. They are instrumental in dissecting social signals and providing insights into the dynamics of collaboration, including group health and communication.

4.5 Feedback

This review focuses on MMLA analysis methods, with feedback being a significant yet secondary aspect of the MMLA framework, and merits some discussion. Feedback in multimodal learning analytics is a bidirectional process essential for completing the analysis cycle, categorized as either *direct* or *indirect*. Direct feedback involves learners or system users and aims to enhance user performance or other metrics. Indirect feedback represents feedback not intended for the end user (e.g., feedback that improves system design).

Direct feedback can take two forms. One form is the prototypical feedback in the context of a learning or training environment for improving the user's performance. Although an exhaustive review of direct feedback literature is outside this paper's scope, seminal works by Hattie & Timperley [68] and Adarkwah [3] provide foundational insights. Users also contribute to MMLA in many forms by offering feedback, integral to user-centered design [2]. Conversely, indirect feedback does not involve the end user but informs system improvement or research findings. It arises from observing user-system interactions or studying learner behavior, leading to enhanced system design or theoretical understanding. Improved research conclusions occur when the study of learners and trainees in these environments

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 [81]. At the same time, 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 [30, 94].

For example, Xu et al. [135] used k-means clustering to analyze collaborative patterns in undergraduate pair programming based on process quality (9 dimensions) and programming outcomes (4 dimensions). The clusters revealed significant differences in collaboration quality and performance. High-performing pairs engaged in knowledge construction and positive discourse, while lower-performing groups exhibited fragmented regulation and excessive debugging. Clusters were categorized as consensus-achieved, argumentation-driven, individual-oriented, and trial-and-error, with the consensus-achieved group showing the best outcomes. This clustering approach provided insights into the relationship between collaboration and learning outcomes without 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 allow for 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.

4.4 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 [119], highlighting the importance of using authentic human-subjects studies 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%) post-ChatGPT [35, 82]. The LLM era has also enabled significantly more dynamic forms of direct feedback: learners and trainees now engage in *dialogic* interactions, receiving feedback through rich exchanges with LLM systems that retain conversation history and support stateful interaction [20].

The way multimodality is employed to deliver direct feedback differs substantially across the two corpora. For example, before LLMs, Petukhova et al. [100] 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

leads to new understandings of the subjects and their populations. Such feedback is vital for advancing research in multimodal learning environments.

5 ARCHETYPES

Following the analysis in Section 4, we re-examined our corpus to classify prevailing research objectives in applying multimodal methods to learning and training environments. We identified three primary research objectives, termed *archetypes*: *Designing and Developing Methods*, *Analyzing Outcomes*, and *Exploring Behaviors*. These archetypes, detailed in subsequent subsections, often overlap within studies; for example, method development research may also yield insights into participant behaviors and outcomes. While these archetypes broadly define the field, they are not exhaustive, and some studies may not align precisely with these categories.

5.1 Designing and Developing Methods

The Designing and Developing Methods archetype encompasses studies that focus on designing, presenting, and evaluating multimodal research methods that can be applied to learning and training environments. These studies prioritize methodological innovation over the derivation of generalizable findings about a population. Although the developed methods often aim to predict outcomes (Section 5.3) and discern behaviors (Section 5.2), the primary focus remains on the method itself, not the implications of its findings on the study participants. These methods are typically quantitative, utilizing supervised learning techniques such as classification [54, 98, 131] and regression [55, 110, 117], and their efficacy is reported through performance metrics like F1-score [5, 10, 161]. Data collection often involves video, audio, and log data [85, 89, 138], targeting modalities such as affect, pose, prosodic speech, and logs [120, 138, 142], with data fusion techniques like mid or hybrid fusion being common [18, 48, 95]. The research predominantly employs model-based approaches [55, 98, 117].

Our corpus reveals a broad spectrum of tasks addressed by *Designing and Developing Methods* research, ranging from personalized feedback in CPR training [98] to engagement detection in educational games [120], and skill classification in sports [95]. The versatility of multimodal methods is evident in the diverse settings, domains, instructional levels, and didactic approaches, without a dominant trend in any specific area.

However, a notable gap in the corpus is the limited focus on evaluating the impact of these methods on end users (stakeholders) and the lack of stakeholder involvement in the method development process. While methods for tasks like feedback generation [49, 51, 107] and engagement detection [5, 18, 120] are presented, their practical effectiveness in enhancing learning outcomes and engagement is seldom empirically validated. Furthermore, the integration of stakeholder feedback into the development of methods is rare, which can lead to a disconnect between the objectives of researchers and the needs of practitioners [14]. This aspect will be further discussed in Section 6. Although some studies in our corpus do consider stakeholder impact [9, 107, 142], such instances are infrequent and not representative of the corpus as a whole.

5.2 Analyzing Outcomes

The *Analyzing Outcomes* archetype focuses on identifying specific outcome metrics, such as learning gains, engagement levels, and accuracy rates. The goal is to uncover findings that apply to broader populations, distinguishing it from the *Designing and Developing Methods* archetype, which focuses on refining analytical techniques. Outcome analysis typically employs supervised learning methods like classification [19, 83, 98] and regression [55, 117, 142], along with insights from model behaviors, statistical patterns, and unsupervised methods [65, 85, 131].

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 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.

Alternatively, multimodal LLMs operate in a continuous space and can process heterogeneous data directly, without requiring engineered features. Nguyen and Park [93] employed GPT-4o to score and generate explanatory feedback on students' multimodal science assessments, demonstrating that LLMs can ingest handwritten student responses—including textual and visual content as a single input—with over 90% transcription accuracy and 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 more accurate responses 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" [93]. 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 [41, 69]. The multimodal capabilities of enterprise LLMs such as ChatGPT, Claude, and Gemini have also 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 [33, 72].

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 [46]. This gap highlights ongoing design tensions around trust, interpretability, and the roles of human and AI actors in direct feedback ecosystems [6, 31].

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 [93], 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 [46] 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 [25]. This gap is often attributed to the absence of established learning frameworks in software engineering pipelines [28].

4.5 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. Learning and training

Outcome analysis has been applied across various learning and training contexts, focusing on constructs like attention and engagement [9, 55, 141], task performance and accuracy [10, 43, 95], learning outcomes [21, 49, 147], and collaborative outcomes [90, 136, 156]. Despite diverse environments, common outcome variables provide generalizable insights. However, this archetype has limitations. Focusing on outcome variables may overlook the complexities of learning processes, risking interventions tailored to high-performing learners and neglecting individual differences [59]. Additionally, like the *Designing and Developing Methods* archetype, these studies often exclude stakeholder perspectives, potentially leading to biased conclusions. The importance of stakeholder feedback is discussed in Section 6.

5.3 Exploring Behaviors

The *Exploring Behaviors* archetype investigates human behavior and experiences in learning and training contexts, using an exploratory approach to uncover influencing factors. This research examines a variety of human signals that vary temporally, socially, and spatially, tailored to specific learning objectives. Unlike other archetypes, it often incorporates qualitative observations [27, 74, 81, 82], and employs data exploration techniques like correlation analysis [89, 104] and pattern recognition [6, 38, 102, 123]. Data fusion in this context is typically qualitative [11, 75, 158], involving the manual integration of multimodal data sources. This approach enables triangulation of student and trainee behaviors, providing richer context to researchers, statistical analyses, or data visualizations, thereby facilitating deeper insights into the behaviors under study.

Exploring behaviors research aims to fill knowledge gaps in learning theory and technological applications by investigating human behavior in educational contexts. Reilly et al. [123] applied a Markov transition model to assess how students' physical behaviors during a collaborative programming task correlate with collaboration quality, task performance, and learning gains. Noel et al. [104] utilized correlation analysis alongside social network metrics and annotated behaviors to distinguish collaborative dynamics in a software engineering course. Closser et al. [27] conducted a qualitative study, using a coding scheme to analyze students' actions, speech, and gestures in embodied learning activities to understand their conceptualization of measurement. These studies, often grounded in learning theory, employ multimodal learning analytics to dissect the components of effective collaboration, showcasing the nuanced insights that multimodal methods can provide into collaborative learning processes. The research spans various mediums, modalities, and settings, with a discernible focus on collaboration.

6 DISCUSSION

Sections 4 and 5 reveal several trends in multimodal learning and training, including key results, challenges, research gaps, and future research directions. In the following subsections, we discuss each of these and address the limitations of our literature review. Overall, we characterize the current state of the field by presenting several key insights:

- **Environments**: Learning environments outnumber training environments 7:2, mostly focusing on STEM with at least one virtual component (virtual or blended).
- **Participants**: Participants are primarily university or K-12 students, with multi-person environments slightly more common than individual ones (3:2).
- **Data and Modalities**:
  - Common data collection mediums: video, audio, environment logs, participant-produced artifacts.
  - Popular modalities: pose, logs, affect, gaze, prosodic speech.

activities yield rich multimodal data streams, each offering unique windows into learning and training processes. Learning analytics fuses heterogeneous data for analysis to extract insights, uncover patterns, and infer 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 9 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.

Dimension	Pre-LLM (2017–2022)	Post-LLM / GenAI-Enabled (Late 2022–Present)
Feature Engineering	Predominantly manual and domain-specific feature extraction (e.g., hand-crafted 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 pre-trained 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 9. Comparison of pre-LLM (Corpus A) and post-LLM (Corpus B) methodological affordances in applied MMLA.

- Most papers use 2-5 modalities, focusing on vision analysis and human-centered modalities (artifacts, surveys, interviews).

- **Analysis Methods**:
  - Common methods: classification (for predicting outcomes), statistical analysis (for feature selection and correlation), qualitative analysis (case studies, coding, thematic analysis).
  - Model-based papers outnumber model-free ones 3:1.
- **Data Fusion**:
  - 75% of papers use early, mid, late, or hybrid fusion.
  - Mid fusion is most prevalent, followed by hybrid fusion.
  - Fused modalities often yield better results than unimodal ones, suggesting researchers should explore data fusion for a holistic understanding of behaviors and outcomes.
- **Publication Mediums**:
  - Most popular venues: *British Journal of Educational Technology* (BJET) and *International Conference on Learning Analytics & Knowledge* (LAK).

6.1 Results

The results of our corpus's papers illustrate that multimodal methods are often successful at predicting learning and training outcomes. The review also identifies the most important features that have been used for predicting those outcomes [86, 136, 137]. Vrzakova et al. point out that even when multimodality does not improve a model's predictive capabilities, patterns in the multimodal data can be informative. Often, multimodal patterns help contextualize, and add interpretability to the unimodal primitives by revealing nuances that cannot be identified by one modality alone [153]. These same patterns can also highlight performance differences among students and trainees:

> Our results demonstrate how NLP and ML techniques allow us to use different modalities of the same data, voice and transcript, and different modalities of different data sources, voice data from interviews, answers to a goal orientation questionnaire, and answers to open ended questions about energy, in order to better understand individual differences in students' performances. [79]

Human-centered approaches allow researchers to dive deeper and gain a more holistic understanding into the learning and training processes. The richness innate to human-centered data — e.g., contextual qualitative observations, tangible artifacts produced by participants and researchers, participant perspectives gleaned from interviews and surveys, etc. — allows researchers to gain unique insights into participants' experiences and behaviors by identifying subtleties that more opaque (often quantitative) approaches may miss.

Our corpus's results also establish that multimodal methods are generally better-performing and more informative relative to unimodal approaches. This is largely due to different modalities conveying markedly different types of information, which helps create more holistic representations of learners that are much richer than is possible with only a single modality. Ma et al. [90] demonstrate this via several key findings:

> The results showed that Linguistic + Audio + Video (F1 Score = 0.65) yielded the best impasse detection performance...
>
> We found that the semantics and speaker information in the linguistic modality, the pitch variation in the audio modality, and the facial muscle movements in the video modality are the most significant

In the following section, we outline three **archetypes** of multimodal learning and training research that exemplify how this research is conducted in practice. Rather than introducing new categories, the archetypes synthesize recurring configurations of methods, environments, and analytic goals observed across the corpus. Organizing the literature around these categories helps elucidate the field's diverse goals and recurring methodological approaches. To anchor each archetype, we introduce a representative case study that highlights its core characteristics in context.

5 Archetypes

Following the analysis in Section 4, we reexamined our corpus to classify prevailing research objectives in the application of multimodal methods to learning and training environments. We identified three primary research objectives, termed archetypes: **Designing and Developing Methods**, **Analyzing Outcomes**, and **Exploring Behaviors**. These archetypes, detailed in subsequent subsections, often overlap within studies, i.e., they are not mutually exclusive, and individual studies may instantiate multiple archetypal patterns. For example, method development research may also yield insights into participant behaviors and outcomes. While these archetypes broadly define the field, they are not exhaustive, and some studies may not align precisely with these categories.

5.1 Designing and Developing Methods

The Designing and Developing Methods archetype comprises studies centered on the design, construction, and evaluation of multimodal research methods for learning and training environments, emphasizing methodological innovation rather than population-level inference. Although these methods often aim to predict outcomes (Section 5.2) and identify behaviors (Section 5.3), the primary contribution lies in the methods themselves rather than in the substantive interpretation of results for participants. The approaches are predominantly quantitative, employing supervised learning techniques such as classification [86, 113] and regression [43, 101], with performance commonly reported using metrics such as F1-scores [2, 10]. Data collection typically involves video, audio, and log data [74, 77], targeting modalities including affect, pose, prosodic speech, and logs [103, 122], and applying mid-level or hybrid fusion strategies [36, 84] through model-based approaches [43, 101].

5.1.1 Case Study: Designing and Developing Methods in Psychomotor Skill Training. A representative example of the Designing and Developing Methods archetype is the *Table Tennis Tutor (T3)* system, which supports automated classification of forehand strokes in psychomotor skill training [84]. The environment is a novice-oriented table tennis practice setting, with a single trainee and an expert coach in a casual physical sports setting. Data are collected using smartphone inertial sensors (accelerometer and gyroscope) worn by the trainee and a Microsoft Kinect V2 depth camera positioned at the table center. The multimodal data comprise physiological and vision modalities, including inertial motion signals and 3-D skeletal joint trajectories, synchronized via a Multimodal Learning Hub [107]. Across 33 training sessions, 510 forehand strokes were recorded and manually annotated using time-aligned video and sensor data. An expert interview with the coach informed system design and acceptability, though no controlled intervention or learning outcome study with trainees was conducted.

Within the learning analytics component of our framework, T3 employs mid-level fusion by temporally aligning and integrating features from smartphone and Kinect sensors after initial processing. The analysis is model-based, using an LSTM network to perform binary classification of correct versus incorrect strokes. Three configurations were evaluated—smartphone-only, Kinect-only, and fused multimodal sensing—with performance assessed using accuracy, precision, and recall. The multimodal configuration achieved the highest precision (0.73), demonstrating the methodological benefits of

unimodal indicators of impasse.

...all of our multi-modal models outperformed their unimodal models...

These results underscore the considerable advantages of employing multimodal methods to understand learning and training experiences, behaviors, and outcomes. By integrating diverse modalities, researchers can uncover patterns that combine to create rich, holistic depictions of students' learning and training. This comprehensive perspective is crucial for capturing the complexities of learner and trainee experiences and behaviors, and suggests that multimodal approaches are not merely additive, but synergistic, offering opportunities for more informative and in depth analyses that are invaluable for advancing educational practice and research.

6.2 Challenges, Limitations, and Research Gaps

In Worsley and Blikstein [157], a primary "takeaway" is that various strategies for employing multimodal learning analytics offer a "meaningful glimpse" into complex datasets that traditional approaches may miss. However, multimodal data complexity presents challenges. Liu et al. [86] note that "data from different sources are often difficult to integrate." Temporal data alignment and sampling rate issues frequently arise, making data collection and labeling time-consuming and requiring "significant human time and effort" [85].

A major challenge is the lack of data. Most studies analyze small groups, making it difficult to use quantitative algorithms, explaining the limited use of deep learning. Kubsch et al. cite data scarcity as a "major challenge for building robust and reliable multimodal models" [79]. Small datasets hinder the development of scalable approaches. Researchers noted:

...the design and sample size of the focus group do not allow us to generalize the results. [105]

The limited number of pair work EEs does not allow us to make any strong claims in terms of the framework's reliability. [99]

...the size of the dataset used is relatively small, and the subject pool is not overly diverse, limiting our ability to explore culture or ethics-related factors in the model reliably. [23]

...training a model on a reduced dataset introduces a bias to the model, affecting the validity of the model's predictions when the data inputs come from a different distribution than the training set. [79]

There is a lack of large, open-source datasets that are curated for researchers in multimodal learning and training environments. This represents a major research gap. Despite several papers mentioning data scarcity as a noteworthy challenge, few papers focus on compiling such datasets or developing methods for smaller datasets. Current methods are often one-off and not designed to generalize. Researchers rely on derived features (e.g., affect and pose) rather than raw inputs (e.g., pixel values). Researchers in our corpus also relied on derived, observable features (particularly in computer vision, e.g., affect and pose) as model input. This differs from core computer vision approaches and creates useful space for exploring end-to-end model training using raw inputs in the future.

The field lags behind core AI and ML, where methods often generalize across tasks and domains. For example, GPT-4 was tested on several benchmarks and exams [111]. Resource and access limitations, along with privacy concerns, hinder the application of advanced AI methods in learning and training environments. Similarly, conversational agents

combining heterogeneous sensor streams. Feedback in this study is primarily indirect, informing researchers and system designers about sensor placement, fusion strategies, and model performance, while learner-facing feedback remains conceptual. As such, T3 exemplifies the Designing and Developing Methods archetype by prioritizing multimodal infrastructure, fusion techniques, and classification performance over evaluation of downstream learning impact.

5.2 Analyzing Outcomes

The Analyzing Outcomes archetype focuses on identifying relationships between multimodal signals and outcome metrics (e.g., learning gains), prioritizing assessments of system impact and findings intended to generalize across populations. Studies commonly employ supervised learning techniques, including classification [15, 86] and regression [43, 122], alongside insights from statistical trends and unsupervised analyses [50, 74]. This archetype has been applied across diverse learning and training contexts, examining constructs such as attention and engagement [43], task accuracy and performance [10, 35], learning outcomes [16, 37], and collaborative dynamics [78, 119]. While the consistent use of common outcome variables has enabled cross-context relevance, an emphasis on outcomes can obscure the underlying complexity of learning processes, potentially biasing interventions toward high-achieving learners and overlooking individual variability in needs and trajectories [45].

5.2.1 Case Study: Analyzing Outcomes in Embodied Math Learning with GenAI Feedback. A representative example of the Analyzing Outcomes archetype is the study by Cosentino et al. [31], which investigates the impact of LLM-generated formative feedback on learning and cognitive engagement in an embodied mathematics environment. Conducted in a physical classroom in Norway, the study involved middle-school students (ages 11-13) solving integer arithmetic problems using full-body movement on a body-scale number line within a multisensory setting featuring floor and wall projections. Multimodal sensing included mobile eye-tracking glasses, multiple video cameras, and system-level motion tracking, producing data spanning vision (e.g., gaze and fixation patterns), physiological signals (e.g., pupil dilation and body movement), and interaction logs (e.g., task attempts, correctness, and AI-generated feedback). Using a between-groups design, the study compared teacher-provided feedback with feedback generated by GPT-4, aligning multimodal measurements with interpretable learning and engagement outcomes.

The analysis used hybrid fusion to synchronize gaze, pupil, and log data, creating indicators of learning and engagement via model-free statistical testing (t-tests with normalization and Welch's correction). Key outcome variables included task performance, cognitive load as measured by pupillary oscillations, time spent in areas of interest, and transitions between them. While no significant performance differences were found, students receiving GenAI feedback showed lower cognitive load and more balanced visual processing compared to those receiving teacher feedback. The feedback affected learning directly through interactions or hints and indirectly through multimodal analyses of attention and cognitive effort. The study's main contribution is its assessment of feedback's impact on learning outcomes, aligning with the Analyzing Outcomes archetype.

5.3 Exploring Behaviors

The Exploring Behaviors archetype examines human behavior and experience in learning and training contexts through exploratory analyses aimed at identifying influencing factors. These studies analyze temporally, socially, and spatially varying human signals aligned with specific learning objectives and frequently incorporate qualitative observations [22, 68]. Common analytical approaches include correlation analysis [77, 94] and pattern recognition [3, 92]. Data fusion is typically qualitative [59, 134], relying on manual integration of multimodal sources to support triangulation.

are underrepresented, with few papers discussing agents and none employing interactive, multi-turn agents (one paper [142] did mention exploring this in the future). The rise of generative AI may have a substantial impact on multimodal analytics, in terms of multi-turn agents and otherwise. The lack of standardized coding practices and protocols is another gap. Most papers use domain-specific coding schemes, making replication difficult. Developing reliable methods for automating coding and creating standardized log formats would benefit the field.

Training literature is sparse compared to learning literature. Physical training environments are underrepresented, and sensor data is rarely used in learning environments. Most papers use quantitative or qualitative analysis, with few employing mixed-methods approaches. Professional development environments and longitudinal analyses are also underrepresented.

Finally, little work focuses on the direct impact of methods on learners or trainees, or considers their input during development. Recently, particularly in education, researchers have adopted a more stakeholder-centric approach to method development [33, 35] by incorporating *user-centered design* [2], i.e., focusing on users and their needs throughout the design process. Other stakeholder-centric approaches like *participatory design* [127] and *co-design* [113] are prevalent in learning sciences but not well-represented in our corpus.

While significant strides have been made in the field, numerous challenges and research gaps remain. The complexity of integrating multimodal data, scarcity of large and diverse datasets, and limitations in data alignment continue to hinder the development of robust and scalable models. The underrepresentation of more advanced AI methods, standardized coding practices, and stakeholder-centric approaches further limits the field's progress. Addressing these challenges will not only advance the state of multimodal learning and training research, but also enhance the utility and impact of educational technologies in diverse learning and training environments.

6.3 Future Research Directions

The results demonstrate that multimodal methods can be powerful in learning and training settings. However, persisting challenges and limitations highlight several research directions requiring further exploration. In the following subsections, we discuss directions that would provide the greatest benefit to the field.

6.3.1 LLMs. The recent boom in generative AI and multimodal LLMs creates tremendous opportunities for multimodal learning and training research. State-of-the-art models like GPT-4x [111] and Gemini [144] now offer multimodal capabilities and allow for prompt engineering approaches that can bypass the need for traditional model training (i.e., parameter updates) and large datasets [34]. Smaller, open-source models can also be trained via parameter-efficient methods to ease the computational overhead endemic to large transformer models [46]. We see both prompt engineering and multimodal conversational agents as two promising research directions.

Advances in multimodal transformers (especially those combining vision and text) have demonstrated these models' ability to perform multiple multimodal tasks. Examples include video-moment retrieval with step-captioning [162] and diagram generation via LLM planning [163]. Other work has built multimodal pipelines around LLMs by performing log-based discourse segmentation and using students' environment actions to contextualize students' discourse in the prompt before inference [36, 132, 133]. Given the recent proliferation of multimodal LLMs in core AI research, we expect to see an increase in LLM integration with multimodal learning and training environments.

6.3.2 Data Scarcity Mitigation. Data scarcity is a major issue, causing multimodal learning and training methods to lag behind core AI approaches. Compiling large learning corpora could help, but challenges exist. Collecting multimodal data for large studies is more difficult than for unimodal ones, with a negative correlation between the number of modalities

Collectively, this work leverages multimodal learning analytics to decompose learning and collaboration processes, yielding nuanced insights into the behaviors under study.

5.3.1 Case Study: Exploring Behaviors in Online Collaborative Problem-Solving. A canonical example of the Exploring Behaviors archetype is the study by Tang et al. [123], which examines college students' attention during online collaborative problem solving (CPS) using a multimodal approach. In synchronous online sessions for a computer networking course, triads of students tackled subnetting problems via a video conferencing platform (Tencent Meeting). The study used EEG devices, video recordings, and performance assessments to collect data on physiological signals (EEG-derived attention), behavioral interactions, and performance logs, highlighting different collaborative patterns, including centralized, distributed, and individual participation.

Learning analytics in this study focuses on exploratory interpretation rather than prediction, using correlation analysis, ANOVA, and a Hidden Markov Model (HMM) to reveal latent attention states and transitions. Researchers manually triangulate hand-coded behavioral sequences from video with EEG-derived attention measures to build a systemic account of collaborative problem-solving (CPS) processes. The findings indicate that attention varies with collaborative structures: centralized groups show synchronized attention, distributed groups exhibit fluctuating patterns, and individual groups maintain persistent inattention. Feedback is primarily indirect, aiding theory-building by illustrating how different collaborative configurations influence attentional engagement, exemplifying the Exploring Behaviors archetype's emphasis on understanding human behavior through multimodal analytics.

6 Discussion and Conclusions

In this review, we examined how multimodal data are collected, fused, and analyzed in applied learning and training contexts, with particular attention to how the field has evolved in response to emerging technologies like GenAI and LLMs. We mapped the methodological landscape across data sources, modality types, fusion strategies, and analysis approaches, highlighting both enduring practices and newer techniques enabled by foundation models. We presented a unified, empirically grounded framework and taxonomy that explain how methodological choices are shaped by learning environments, data, and analytic goals. This allowed us to synthesize persistent methodological challenges that hinder scalability and real-world deployment. We concluded by identifying three methodological archetypes that illustrate how multimodal components are combined in practice to design intelligent systems, assess learning outcomes, and explore learner behaviors.

Together, these contributions offer a coherent account of the current methodological landscape in multimodal learning and training research. In the following subsections, we reflect on the insights our framework affords, identify research gaps, examine persistent methodological and practical challenges, and outline promising directions for future research. We conclude by presenting broader implications for the design, deployment, and interpretation of multimodal learning and training systems.

6.1 Framework Insights and Research Gaps

Multimodality is not simply an exercise in data aggregation: when designed effectively, these systems capture patterns no single modality reveals alone, thus producing insights that are greater than the sum of their parts [133]. Multimodal methods have consistently demonstrated effectiveness in predicting learning outcomes and identifying salient features that drive those outcomes [75, 120]. Each modality encodes distinct information (e.g., metacognition in collaborative discourse [116, 117]), and when combined thoughtfully, yields more comprehensive and context-rich representations of

analyzed and sample size [130]. Ethical concerns, particularly regarding privacy and surveillance in educational datasets involving children, complicate data collection [42]. One solution is designing generalizable methods requiring limited data, such as zero and few-shot learning approaches, which have advanced in core AI domains [76].

6.3.3 Standardization. Blanco et al. [45] emphasize the need for uniform coding standards for multimodal, temporal, and human-focused data. Current e-learning norms like xAPI [134] and LTI [1] are used in platforms like Canvas and Moodle but mainly for unimodal data and are limited by proprietary licenses. Adapting these frameworks for multimodal data is challenging, leading to minimal use in research.

Multimodal learning and training research merges AI, multimodal data, and educational contexts, requiring novel software. This has led to disparate approaches by different teams [160]. Creating uniform standards is crucial for the reliability of machine learning techniques and improving human-centric data analysis [44]. Adopting a unified log format for multimodal data could reduce reliance on context-specific methods and improve generalizability. Researchers and engineers should comply with existing standards and methodologies.

6.3.4 Active Environments. Environments, where study participants are physically active, provide an opportunity for researchers to accommodate motion-based modalities into their multimodal pipelines, e.g., via inertial measurement unit (IMU) sensors. This type of research was largely absent from our corpus, and we envision it being particularly useful for embodied learning and physical training research.

Embodied learning scenarios, where learners explore concepts through body movement, involve extensive multimodal data, capturing sensory inputs essential for movements, gestures, speech, gaze, interactions, and coordination [6]. Interaction analysis is common but challenging due to human analysts' cognitive limits and the fast-changing nature of embodied contexts [164]. Leveraging multimodal methods to support human analysts in such scenarios is promising. MMLA must address the complexities of 1) multimodal data collection from heterogeneous sensors, 2) alignment, and 3) analysis to derive meaningful insights into learners' behaviors, providing educators with a comprehensive understanding of engagement and problem-solving [58].

Physical training environments, like rehabilitation therapy, weight lifting, running, and cycling, often use IMU sensors for human activity recognition (HAR). However, this is not typically done using multimodal data. Combining spatial modalities (like pose and gesture) with physiological modalities (such as blood pressure, body temperature, and electrodermal activity) could provide a more holistic interpretation of trainees' actions. Multimodality can decompose activities into sub-activities too nuanced to identify unimodally and add interpretability that IMU data alone cannot provide. For example, Xia et al. co-trained deep learning models using activities' images and IMU data, improving HAR generalizability [159]. While some physical training works in our corpus leveraged multimodality [51, 95], this was rare, and further research is needed to better inform physical training environments.

6.3.5 Explainability. Many AI and ML approaches use black-box algorithms with outputs that lack explainability, hindering teachers' and trainers' ability to guide students and fostering distrust in AI systems. Prior work has aimed to create more explainable systems using data visualization tools to make learning processes transparent [73, 151]. LLMs have potential for enhancing explainability through *Chain-of-Thought* prompting, which elicits reasoning chains from the model [34, 155]. Feedback from teachers and students shows they see potential for LLMs to improve learning outcomes, but explainability is crucial for their acceptance [33].

6.3.6 Longitudinal Analyses. The vast majority of studies in our corpus focus on using multimodality to either predict overall learning and training outcomes or identify features correlating with those outcomes; however, these approaches

learning and training. Even when predictive accuracy does not improve, multimodality can surface meaningful patterns, offering greater interpretability and opportunities to design more responsive and effective educational systems [128].

Recent advances in LLMs further extend the affordances of multimodal systems by enabling late-fusion classification and prediction through the distillation of individual modalities into semantically meaningful labels, accompanied by human-readable justifications that articulate model reasoning [46]. This interpretive capacity introduces new opportunities for transparency and dialogue in learning analytics. When paired with structured prompt engineering and iterative refinement via active learning [28], LLMs can generate time-aligned behavioral explanations across modalities, allowing researchers and educators to examine outputs through natural language contextualized by other modalities. These capabilities offer a dual benefit: enhancing interpretability and supporting sense-making within human-in-the-loop workflows across diverse environments and tasks (e.g., embodied learning, assessment, feedback).

While our findings highlight the distinct advantages of multimodal approaches, the literature reveals both areas of maturity and critical, underexplored gaps across the components of our framework. Within the environment component, most research remains concentrated in small-scale classroom settings, typically within K-12 or postsecondary contexts. These studies are often anchored to short-term outcomes or single-session observations, with limited attention to training environments, younger learners in primary education, or adult learners outside the university system. Longitudinal perspectives, which are essential for understanding how learners and trainees evolve over time, are notably scarce.

Challenges surrounding multimodal data collection, synchronization, and alignment continue to limit scalability. Although widely acknowledged, these issues remain unresolved in most studies. Recent efforts have targeted the cross-modal integration bottleneck through open-source tools like SyncFlow [126], which enable automated collection and alignment of webcam, screen, and audio data via web browsers or standalone software. However, such tools are rarely adopted in practice and remain at an early stage of development. The scarcity of large, open-source datasets tailored for multimodal learning further constrains replication, benchmarking, and broader methodological progress.

Within learning analytics, there is growing use of LLMs and machine learning pipelines to process multimodal signals; however, the interpretability of system decisions remains a significant challenge. Explanations for model decision-making are often lacking or inaccessible to practitioners. Furthermore, the integration of LLMs in this space is limited, with most studies relying on proprietary GPT-based APIs used out of the box. This leaves a gap in the development of more advanced pipelines and little exploration of open-source alternatives that could promote transparency and trust.

The feedback component reflects early momentum toward LLM-generated formative and explanatory feedback. However, few studies rigorously examine whether such feedback aligns with curricular goals, teachers' expectations, or broader pedagogical intent. Trust and adoption challenges, particularly from educators, remain an impediment to adoption, and little is known about how learners engage with or respond to LLM feedback in authentic settings.

These findings illuminate the current state of multimodal learning and training research and reveal a set of persistent, interrelated gaps and challenges. The following subsections build on this analysis by categorizing the core challenges that hinder progress and outlining future directions to guide the design of more impactful MMLA systems.

6.2 Challenges and Limitations

We highlight three categories of persistent challenges, which we supplement with a discussion of this review's limitations.

Data. Data collection, labeling, and scarcity issues were widely reported in both corpora. Studies with human subjects often have small samples because data collection is time-consuming, resource-intensive, and constrained by privacy requirements. These constraints hinder the deployment of AI systems in classroom settings and limit large-scale data

do not consider how students and trainees evolve over time. Conducting longitudinal studies and analyses would provide insight into how participants' behaviors and abilities develop as they progress in their learning or training. Longitudinal investigations have been successfully executed using unimodal and digital trace data [15], but less frequently within multimodal studies. The challenges of scalability and standardization of multimodal logs have restricted longitudinal MMLA research [160], affecting both research and software development in multimodal learning and training. There exists a void in the literature concerning longitudinal multimodal learner models encompassing a comprehensive view of learners' and trainees' evolution over time, making this an area ripe for further research exploration.

6.3.7 Stakeholder Input and Impact. Section 5 revealed a disconnect between researchers designing multimodal learning and training methods and the stakeholders these methods were intended to benefit. Few efforts incorporated user input in their method development pipelines or evaluated the impact of their methods on stakeholders' real-world experiences. A larger emphasis on *design-based research* [7], i.e., iteratively designing and refining methods based on real-world research, would help bridge this gap. Additionally, employing *participatory design* (i.e., actively involving stakeholders throughout the method design process) and *co-design* (i.e., involving contributions from all stakeholders throughout the design and development processes) [73] would help researchers develop multimodal methods better aligned with stakeholder experiences and outcomes.

6.4 Literature Review Limitations

We acknowledge the limitations of our literature review. While Google Scholar is widely used, it poses reproducibility challenges due to its opaqueness, non-determinism, and user-specific results. Although reconstructing our initial corpus *in its exact form* is unlikely, the authors are confident that the variability in Google Scholar searches does not prohibit the *overall* reproducibility of the corpus. This is because SerpAPI does not use individual user data when conducting web scrapes, as API calls are made via proxy and random headers.

Initially distilling our literature search corpus using citation graph pruning (see Section 3.2.1) is another potential limitation, as relevant papers may have been excluded due to minimal citations. However, since this paper reviews prominent methods in multimodal learning and training, the authors agreed that works not significantly citing other related papers (outgoing citations) or significantly cited by related papers (incoming citations), were outside our review's scope. For a detailed account of this review's limitations, see Appendix C.

7 CONCLUSIONS

In this paper, we conducted a comprehensive literature review of research methods in multimodal learning and training environments. We developed a novel approach, *citation graph pruning*, to distill our literature corpus. We presented a taxonomy and framework reflecting current advances, identifying and analyzing five modality groups (Natural Language, Vision, Sensors, Human-Centered, and Logs) through descriptive statistics, qualitative thematic analysis, and discussions on state-of-the-art findings, challenges, and research gaps. We derived three archetypes characterizing current research and identified the need for a new type of data fusion, *mid fusion*, which combines derived, observable features. We concluded with promising research directions and the limitations of our work. As multimodal learning and training analytics expand with generative AI, this review aims to inspire new methods and research.

REFERENCES

[1] 1EdTech. 2019. Learning Tools Interoperability Core Specification 1.3 | IMS Global Learning Consortium — imsglobal.org. https://www.imsglobal.org/spec/lti/v1p3/. [Accessed 25-01-2024].

collection, which is rarely released publicly, especially when participants are minors. Consequently, many studies rely on limited or demographically narrow datasets, reducing the generalizability and reliability of findings [95]. Models trained on such data are susceptible to distributional shifts, leading to bias and compromised predictive validity [67].

Cross-Modal Interaction. Integrating multimodal data is complex and resource-intensive, requiring reconciliation of heterogeneous representations that differ in structure, granularity, and temporal dynamics [75]. Even minor misalignments can cascade through the pipeline, degrading performance and interpretability. Real-time alignment is harder still, as modalities arrive asynchronously with variable latency and noise, disrupting temporal and semantic coherence. These challenges are amplified by the field's reliance on mid-fusion, where modalities are preprocessed into intermediate features and integrated using enterprise software with unverifiable predictions. Such predictions are frequently treated as ground truth for downstream modeling, introducing inaccuracies that are rarely subjected to human validation.

LLMs. LLMs can reduce annotation burdens and accelerate data pipelines, but their integration introduces new challenges. The scale and opacity of enterprise systems (e.g., ChatGPT) complicate transparency, trustworthiness, and pedagogical alignment, prompting several states and school districts to restrict classroom use over unresolved ethical, pedagogical, and governance concerns [106]. Training smaller, open models may mitigate some issues, yet is often infeasible due to data scarcity, compute costs, and the technical complexity of self-hosted systems (including concurrency and managing multi-user agent interactions). Meanwhile, students' frequent exposure to GenAI is shifting expectations toward immediate, definitive answers, often at the expense of critical thinking [24]—raising risks of overreliance on LLM-generated content that may undermine pedagogical goals.

Limitations of This Review. We acknowledge the limitations of our literature review, including potential reproducibility issues with Google Scholar, the risk of relevant work being excluded using our filtering heuristic (i.e., citation graph pruning), and inconsistencies in versioning across published manuscripts. However, based on our comprehensive analysis of the corpus, we are confident that these factors do not compromise the validity of our findings or the inferences drawn. A detailed discussion of these limitations is provided in Supplementary Materials.

6.3 Future Work

Addressing these research gaps and challenges opens several promising directions for future work.

Active Environments. Multimodal research in physically active settings remains scarce despite strong potential for physical training and therapy. Active environments naturally produce rich signals, such as gesture, gaze, speech, and spatial coordination, that are highly relevant to embodied learning, rehabilitation, athletics, and hands-on skill training, where nuanced physical actions matter. Leveraging multimodality in these contexts could improve activity recognition, enable real-time feedback, and uncover behavioral patterns that unimodal methods often miss.

Longitudinal Analyses. Longitudinal research remains scarce in multimodal learning and training, with most studies predicting outcomes or identifying correlates at single time points. Few examine how learners or trainees evolve when interacting with multimodal data. Although longitudinal analyses exist with unimodal or digital trace data, multimodal applications are limited by scalability, data standardization, and the complexity of maintaining synchronized multimodal logs over extended periods [137]. Yet such work is essential: longitudinal MMLA can reveal developmental trajectories, shifts in engagement, and the dynamics of learning and skill acquisition. Addressing these challenges would enable more adaptive systems and a deeper understanding of progression in authentic, evolving contexts.

[2] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.

[3] Michael Agyemang Adarkwah. 2021. The power of assessment feedback in teaching and learning: a narrative review and synthesis of the literature. *SN Social Sciences* 1, 3 (March 2021), 75. https://doi.org/10.1007/s43545-021-00086-w

[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.* Springer International Publishing, Cham, 289–325. https://doi.org/10.1007/978-3-031-08076-0_12

[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] 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

[7] Matthew Armstrong, Cade Dopp, and Jesse Welsh. 2022. Design-based research. *N/A* N/A, N/A (2022), N/A.

[8] 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

[9] 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

[10] 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

[11] 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

[12] Paulo Blikstein. 2013. Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge.* Association for Computing Machinery, New York, NY, USA, 102–106.

[13] 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.

[14] Ulrich Boser and Abel McDaniels. 2018. Addressing the Gap between Education Research and Practice: The Need for State Education Capacity Centers. *Center for American Progress* N/A, N/A (2018), N/A.

[15] Chris A. Boulton, Emily Hughes, Carmel Kent, Joanne R. Smith, and Hywel T. P. Williams. 2019. Student engagement and wellbeing over time at a higher education institution. *PLOS ONE* 14, 11 (11 2019), 1–20. https://doi.org/10.1371/journal.pone.0225770

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

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[18] 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

[19] 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

[20] 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

[21] 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

[22] Wilson Chango, Juan A. Lara, Rebeca Cerezo, and Cristóbal Romero. 2022. A review on data fusion in multimodal learning analytics and educational data mining. *WIREs Data Mining and Knowledge Discovery* 12, 4 (2022), e1458. https://doi.org/10.1002/widm.1458 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1458

[23] 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.

[24] 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

[25] 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

Standardization. Standardization is a key challenge in multimodal learning research. Progress requires shared, open logging formats for complex and temporal data and greater adoption of existing norms [137]. These steps would minimize fragmentation, enhance reproducibility, and enable scalable, interoperable systems for research and deployment.

Interpretability. A major challenge in multimodal learning and training systems is limited interpretability, which undermines trust and impedes pedagogical integration. Emerging techniques—such as using LLMs to generate rationales or explanations—offer promising avenues to make system behavior more intelligible to educators and learners. Interpretability also supports pedagogical alignment by enabling educators to verify that feedback reinforces intended learning objectives. Although recent work improves stakeholder-facing explanations through chain-of-thought prompting [28, 130], the inner workings of foundation models and complex systems remain largely opaque.

Agentic and Multi-Agent Workflows. Interest is growing in agentic and multi-agent architectures that coordinate sensing, reasoning, and feedback. These pipelines enable diverse components—e.g., activity recognition, affect analysis, and instructional scaffolding—to operate semi-autonomously toward shared objectives. Emerging frameworks like Anthropic's Agent Skills [5] provide modular, reusable capabilities for composing adaptable, context-aware pedagogical agents. Likewise, Model Context Protocol (MCP) [4] structures contextual metadata alongside model behavior, improving traceability and interpretability across agent interactions. Together, these approaches address long-standing challenges of scalability, modularity, and pedagogical alignment in multimodal systems.

6.4 Implications

This review outlines key considerations for designing, implementing, and evaluating multimodal learning and training systems. As these systems advance—especially with LLM integration—we must assess not only technical performance but also educational value and alignment with effective teaching practices. Realizing the potential of multimodality requires context-aware integration strategies that are explicitly tied to instructional goals.

Multimodal integration has shifted in the post-LLM era, from handcrafted features and explicit cross-channel synchronization, to using foundation models to infer structure, relationships, and meaning from raw or weakly labeled inputs. These models enable explanation generation, instructional scaffolding, and streamlined feedback, accelerating development and lowering barriers to entry. However, their proliferation also creates new risks: model-generated errors can propagate across modalities, especially when early outputs are treated as ground truth.

Prior work in intelligent learning environments shows that early misclassifications can shape future predictions and lead to degenerate feedback loops [46]. Mechanisms for verification, revision, and human-in-the-loop oversight are therefore critical to prevent error compounding. Research also finds that LLMs can hinder learning when learners over-rely on generated content or accept it uncritically [66]. The fluency of LLM outputs may disincentivize effortful learning, encouraging passive consumption over active engagement. Similarly, student frustration and the expectation of direct answers can undermine metacognition and long-term retention [24].

A largely overlooked question is uptake: *what if students choose not to engage with the systems we design?* Technological sophistication is insufficient if learners do not attend to or act on feedback. LLM-generated feedback can yield substantial gains, but primarily for learners who actively engage with it [125]. Achieving meaningful impact requires collective buy-in from researchers, administrators, teachers, students, and parents; without it, even advanced systems may fail to support learning—or fail to be used at all.

The future of multimodal learning and training depends not only on improved models and architectures but also on integrating these technologies into learning experiences that augment, rather than replace, instruction. Progress

[26] Yi Han Victoria Chua, Justin Dauwels, and Seng Chee Tan. 2019. Technologies for automated analysis of co-located, real-life, physical learning spaces: Where are we now?. In *LAK19: 9th International Learning Analytics and Knowledge Conference*. ACM, Tempe AZ USA, 10. https://dl-acm-org.proxy.library.vanderbilt.edu/doi/10.1145/3303772.3303811

[27] 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

[28] 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 (2023, to appear)*. N/A, N/A, N/A.

[29] 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.

[30] 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.

[31] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

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

[33] Clayton Cohn, Nicole Hutchins, and Gautam Biswas. 2024. Chain-of-Thought Prompting with Stakeholders-in-the-Loop for Evaluating Formative Assessments in STEM+Computing. (May 2024). Submitted to Computers & Education. Currently under review..

[34] Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. *Proc. Conf. AAAI Artif. Intell.* 38, 21 (March 2024), 23182–23190.

[35] Clayton Cohn, Caitlin Snyder, Joyce Fonteles, Ashwin T S, Justin Montenegro, and Gautam Biswas. 2024. A Multimodal Approach to Support Teacher, Researcher, and AI Collaboration in STEM+C Learning Environments. (May 2024). Submitted to the British Journal of Educational Technology special section Hybrid Intelligence: Human-AI Co-evolution and Learning. Currently under review..

[36] Clayton Cohn, Caitlin Snyder, Justin Montenegro, and Gautam Biswas. 2024. Towards a human-in-the-loop LLM approach to collaborative discourse analysis. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*. Springer Nature Switzerland, Cham, 11–19.

[37] Matt Cook, Zack Lischer-Katz, Nathan Hall, Juliet Hardesty, Jennifer Johnson, Robert McDonald, and Tara Carlisle. 2019. Challenges and strategies for educational virtual reality. *Information Technology and Libraries* 38, 4 (2019), 25–48.

[38] 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

[39] Lucrezia Crescenzi-Lanna. 2020. Multimodal Learning Analytics research with young children: A systematic review. *British Journal of Educational Technology* 51, 5 (2020), 1485–1504. https://doi.org/10.1111/bjet.12959 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.12959

[40] 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.

[41] 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.

[42] 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.

[43] 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

[44] E. Davalos, U. Timalsina, Y. Zhang, J. Wu, J. Fonteles, and G. Biswas. 2023. ChimeraPy: A Scientific Distributed Streaming Framework for Real-time Multimodal Data Retrieval and Processing. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, Los Alamitos, CA, USA, 201–206. https://doi.org/10.1109/BigData59044.2023.10386382

[45] Ángel del Blanco, Ángel Serrano, Manuel Freire, Iván Martínez-Ortiz, and Baltasar Fernández-Manjón. 2013. E-Learning standards and learning analytics. Can data collection be improved by using standard data models?. In *2013 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, Berlin, Germany, 1255–1261. https://doi.org/10.1109/EduCon.2013.6530268

[46] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv e-prints* N/A, N/A, Article arXiv:2305.14314 (May 2023), N/A pages. https://doi.org/10.48550/arXiv.2305.14314 arXiv:2305.14314 [cs.LG]

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* N/A, N/A (2018), N/A.

[48] 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

[49] 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

will require sustained cross-disciplinary collaboration, careful attention to learner behavior, and commitments to transparency and pedagogical grounding—serving educational values as much as technological advancement.

References

[1] 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.

[2] 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

[3] 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

[4] Anthropic. 2024. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol Accessed: 2025-12-14.

[5] Anthropic. 2025. Equipping Agents for the Real World with Agent Skills. Anthropic Engineering Blog. https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills Accessed: 2025-12-14.

[6] 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

[7] 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.

[8] 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

[9] 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.

[10] 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

[11] Roger Azevedo, Jason Harley, Gregory Trevors, Melissa Duffy, Reza Feyzi-Behnagh, François Bouchet, and Ronald Landis. 2013. Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In *International handbook of metacognition and learning technologies*. Springer, 427–449.

[12] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[13] 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.

[14] Berna Cengiz, Iliyasu Yahya Adam, Mehmet Ozdem, and Resul Das. 2025. A survey on data fusion approaches in IoT-based smart cities: Smart applications, taxonomies, challenges, and future research directions. *Information Fusion* 121 (2025), 103102.

[15] 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

[16] 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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[50] 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. https://doi.org/10.1111/jcal.12288

[51] Daniele Di Mitri, Jan Schneider, Kevin Trebing, Sasa Sopka, Marcus Specht, and Hendrik Drachsler. 2020. Real-Time Multimodal Feedback with the CPR Tutor. In *Artificial Intelligence in Education*, Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.). Vol. 12163. Springer International Publishing, Cham, 141–152. http://link.springer.com/10.1007/978-3-030-52237-7_12

[52] Hendrik Drachsler and Jan Schneider. 2018. JCAL Special Issue on Multimodal Learning Analytics. *Journal of Computer Assisted Learning* 34, 4 (2018), 335–337. https://doi.org/10.1111/jcal.12291 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12291

[53] 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

[54] 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

[55] 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

[56] 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.

[57] 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

[58] 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.

[59] Joyce Horn Fonteles, Celestine E Akpanoko, Pamela J. Wisniewski, and Gautam Biswas. 2024. Promoting Equitable Learning Outcomes for Underserved Students in Open-Ended Learning Environments. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (Delft, Netherlands) *(IDC '24)*. Association for Computing Machinery, New York, NY, USA, 307–321. https://doi.org/10.1145/3628516.3655753

[60] Society for Learning Analytics Research. N/A. Multimodal learning analytics across spaces (SOLAR crossmmla sig). https://www.solaresearch.org/community/sigs/crossmmla-sig/. [Accessed 07-02-2024].

[61] Society for Learning Analytics Research (SOLAR). N/A. What is Learning Analytics? https://www.solaresearch.org/about/what-is-learning-analytics/. [Accessed 07-02-2024].

[62] 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.

[63] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32, 5 (2020), 829–864.

[64] Michail Giannakos, Daniel Spikol, Daniele Di Mitri, Kshitij Sharma, Xavier Ochoa, and Rawad Hammad (Eds.). 2022. *The multimodal learning analytics handbook*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-08076-0

[65] 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

[66] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. 2021. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array* 10 (2021), 100057.

[67] 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

[68] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. https://doi.org/10.3102/003465430298487 arXiv:https://doi.org/10.3102/003465430298487

[69] 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.

[70] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[71] H Ulrich Hoppe. 2017. *Computational methods for the analysis of learning and knowledge building communities*. Society for Learning Analytics Research (SoLAR), Beaumont, Alberta, Canada, 23–33.

[72] Nicole Hutchins and Gautam Biswas. 2023. Using Teacher Dashboards to Customize Lesson Plans for a Problem-Based, Middle School STEM Curriculum. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (Arlington, TX, USA) *(LAK2023)*. Association for Computing Machinery, New York, NY, USA, 324–332. https://doi.org/10.1145/3576050.3576100

[73] Nicole Marie Hutchins et al. 2022. *Co-Designing Teaching Augmentation Tools to Support the Integration of Problem-Based Learning in K-12 Science*. Ph. D. Dissertation. Vanderbilt University.

[21] 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.

[22] 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

[23] 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.

[24] 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).

[25] 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).

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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

[31] 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).

[32] Lucrezia Crescenzi-Lanna. 2020. Multimodal Learning Analytics research with young children: A systematic review. *British Journal of Educational Technology* 51, 5 (2020), 1485–1504.

[33] 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.

[34] 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.

[35] 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

[36] 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

[37] 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

[38] 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.

[39] 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.

[40] 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

[41] 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.

[42] 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

[43] 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

[74] 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

[75] 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

[76] Suvarna Kadam and Vinay Vaidya. 2020. Review and analysis of zero, one and few shot learning approaches. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018. Volume 1*. Springer, N/A, N/A, 100–112.

[77] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004 (2004), 1–26.

[78] Kenneth R Koedinger, John R Anderson, William H Hadley, Mary A Mark, et al. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8, 1 (1997), 30–43.

[79] 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

[80] 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

[81] 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

[82] 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

[83] 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 (CoG)*. IEEE, Osaka, Japan, 17–24. https://doi.org/10.1109/CoG47356.2020.9231892

[84] 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.

[85] 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

[86] 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

[87] 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

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

[89] 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, 482–489.

[90] Yingbo Ma, Mehmet Celepkolu, and Kristy Elizabeth Boyer. 2022. Detecting Impasse During Collaborative Problem Solving with Multimodal Learning Analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, Online USA, 45–55. https://doi.org/10.1145/3506860.3506865

[91] 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

[92] 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

[93] 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

[94] Andino Maseleno, Noraisikin Sabani, Miftachul Huda, Roslee Bin Ahmad, Kamarul Azmi Jasmi, and Bushrah Basiron. 2018. Demystifying learning analytics in personalised learning. *International Journal of Engineering and Technology (UAE)* 7 (2018), 1124–1129.

[95] 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

[96] MDPI. 2021. New Trends on Multimodal Learning Analytics: Using Sensors to Understand and Improve Learning. https://www.mdpi.com/journal/sensors/special_issues/multimodal_learning_analytics_sensor. [Accessed 08-02-2024].

[97] 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

[44] Göran Folkestad. 2006. Formal and informal learning situations or practices vs formal and informal ways of learning. *British Journal of Music Education* 23, 2 (2006), 135–145.

[45] Joyce Horn Fonteles, Celestine E Akpanoko, Pamela J. Wisniewski, and Gautam Biswas. 2024. Promoting Equitable Learning Outcomes for Underserved Students in Open-Ended Learning Environments. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (Delft, Netherlands) *(IDC '24)*. Association for Computing Machinery, New York, NY, USA, 307–321. https://doi.org/10.1145/3628516.3655753

[46] 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.

[47] 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.

[48] Nathan Gaw, Safoora Yousefi, and Mostafa Reisi Gahrooei. 2022. Multimodal data fusion for systems improvement: A review. *Handbook of Scholarly Publications from the Air Force Institute of Technology (AFIT), Volume 1, 2000-2020* (2022), 101–136.

[49] Michail Giannakos and Mutlu Cukurova. 2023. The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology* 54, 5 (2023), 1246–1267.

[50] 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

[51] 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.

[52] 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.

[53] Mazhar Hussain, Mattias O'Nils, Jan Lundgren, and Seyed Jalaleddin Mousavirad. 2024. A comprehensive review on deep learning-based data fusion. *IEEE Access* (2024).

[54] 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).

[55] 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

[56] 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).

[57] 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.

[58] 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

[59] 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

[60] Si Na Kew and Zaidatun Tasir. 2022. Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning* 27, 2 (2022), 405–427.

[61] Lin Sze Khoo, Mei Kuan Lim, Chun Yong Chong, and Roisin McNaney. 2024. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors* 24, 2 (2024), 348.

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

[63] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004 (2004), 1–26.

[64] Kirsty Kitto and Simon Knight. 2019. Practical ethics for building learning analytics. *British Journal of Educational Technology* 50, 6 (2019), 2855–2870.

[65] Kenneth R Koedinger, John R Anderson, William H Hadley, Mary A Mark, et al. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8, 1 (1997), 30–43.

[66] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint 2506.08872* (2025).

[67] 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

[68] 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

[98] 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.

[99] 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

[100] Su Mu, Meng Cui, and Xiaodi Huang. 2020. Multimodal Data Fusion in Learning Analytics: A Systematic Review. *Sensors* 20, 23 (2020), 6856. https://doi.org/10.3390/s20236856

[101] 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

[102] 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

[103] 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

[104] 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

[105] 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

[106] 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

[107] 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

[108] Xavier Ochoa, AWDG Charles Lang, and George Siemens. 2017. Multimodal learning analytics. *The handbook of learning analytics* 1 (2017), 129–141.

[109] Journal of Learning Analytics. 2015. Special section on multimodal learning analytics. https://learning-analytics.info/index.php/JLA/announcement/view/102. [Accessed 08-02-2024].

[110] 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

[111] OpenAI. 2023. GPT-4 Technical Report. *arXiv e-prints* N/A, N/A, Article arXiv:2303.08774 (March 2023), N/A pages. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs.CL]

[112] 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

[113] William R. Penuel, Jeremy Roschelle, and Nicole Shechtman. 2007. Designing Formative Assessment Software With Teachers: An Analysis of the Co-Design Process. *Research and Practice in Technology Enhanced Learning* 02, 01 (2007), 51–74. https://doi.org/10.1142/S1793206807000300 arXiv:https://doi.org/10.1142/S1793206807000300

[114] 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

[115] 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

[116] 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

[117] 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

[118] Stéphanie Philippe, Alexis D. Souchet, Petros Lameras, Panagiotis Petridis, Julien Caporal, Gildas Coldeboeuf, and Hadrien Duzan. 2020. Multimodal teaching, learning and training in virtual reality: a review and case study. *Virtual Reality & Intelligent Hardware* 2, 5 (Oct. 2020), 421–442. https://doi.org/10.1016/j.vrih.2020.07.008

[119] 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

[120] 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

[69] 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.

[70] 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.

[71] 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.

[72] 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).

[73] 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.

[74] 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

[75] 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

[76] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2024. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*. Springer, 126–142.

[77] 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.

[78] 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

[79] 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

[80] 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.

[81] 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

[82] 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.

[84] 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

[85] 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

[86] 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.

[87] Manuel Mondal, Mourad Khayati, Hông-Ân Sandlin, and Philippe Cudré-Mauroux. 2025. A survey of multimodal event detection based on data fusion. *The VLDB Journal* 34, 1 (2025), 9.

[88] 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

[89] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[90] Su Mu, Meng Cui, and Xiaodi Huang. 2020. Multimodal data fusion in learning analytics: A systematic review. *Sensors* 20, 23 (2020), 6856.

[91] 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.

[92] 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

[93] 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.

[94] 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*

[121] Umar Bin Qushem. 2020. *Trends of Multimodal Learning Analytics: A Systematic Literature Review*. Ph. D. Dissertation. UNIVERSITY OF EASTERN FINLAND. https://erepo.uef.fi/bitstream/handle/123456789/23508/urn_nbn_fi_uef-20201250.pdf?sequence=1

[122] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[123] 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, 333–339.

[124] María Jesús Rodríguez-Triana, Luis P Prieto, Alejandra Martínez-Monés, Juan I Asensio-Pérez, and Yannis Dimitriadis. 2018. The teacher in the loop: Customizing multimodal learning analytics for blended learning. In *Proceedings of the 8th international conference on learning analytics and knowledge*. Association for Computing Machinery, New York, NY, USA, 417–426.

[125] 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.

[126] 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.

[127] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press, N/A.

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

[129] Shashi Kant Shankar, Luis P. Prieto, María Jesús Rodríguez-Triana, and Adolfo Ruiz-Calleja. 2018. A Review of Multimodal Learning Analytics Architectures. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, Piscataway, NJ, USA, 212–214. https://doi.org/10.1109/ICALT.2018.00057

[130] 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. https://doi.org/10.1111/bjet.12993 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.12993.

[131] 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

[132] Caitlin Snyder, Nicole Hutchins, Clayton Cohn, Joyce Fonteles, and Gautam Biswas. 2023. Using Collaborative Interactivity Metrics to analyze students' Problem-Solving Behaviors during STEM+C Computational Modeling Tasks. (2023). Submitted to Learning and Individual Differences. Currently under review..

[133] 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* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 540–550. https://doi.org/10.1145/3636555.3636912

[134] Rustici Software. 2024. xAPI.com — xapi.com. https://xapi.com/?utm_source=google&utm_medium=natural_search. [Accessed 25-01-2024].

[135] 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.

[136] 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

[137] 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

[138] 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

[139] 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.

[140] Steven A. Stolz. 2015. Embodied Learning. *Educational Philosophy and Theory* 47, 5 (2015), 474–487. https://doi.org/10.1080/00131857.2013.879694

[141] 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), 1012–1027. https://doi.org/10.1109/TAFFC.2021.3127692

[142] 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

[143] 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

Access 6 (2018), 67783–67798. https://doi.org/10.1109/ACCESS.2018.2876801

[95] 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

[96] 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.

[97] 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

[98] 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

[99] 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

[100] 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

[101] 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

[102] L.P. Prieto, K. Sharma, Ł. Kidziński, 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

[103] 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

[104] 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.

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

[106] Dan Rosenzweig-Ziff. [n. d.]. New York City blocks use of the ChatGPT bot in its schools. *The Washington Post* ([n. d.]). https://www.washingtonpost.com/education/2023/01/05/nyc-schools-ban-chatgpt/ Accessed: 2025-12-21.

[107] Jan Schneider, Daniele Di Mitri, Bibeg Limbu, and Hendrik Drachsler. 2018. Multimodal learning hub: A tool for capturing customizable multimodal learning experiences. In *European conference on technology enhanced learning*. Springer, 45–58.

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

[109] 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.

[110] Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D Velásquez. 2024. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion* 102 (2024), 102040.

[111] Shashi Kant Shankar, Luis P. Prieto, María Jesús Rodríguez-Triana, and Adolfo Ruiz-Calleja. 2018. A Review of Multimodal Learning Analytics Architectures. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. https://doi.org/10.1109/ICALT.2018.00057

[112] 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.

[113] 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

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] 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

[144] 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.

[145] 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

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

[147] 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

[148] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models, 2023. *URL https://arxiv. org/abs/2307.09288* N/A, N/A (2023), N/A.

[149] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017), N/A.

[150] Caleb Vatral, Gautam Biswas, Clayton Cohn, Eduardo Davalos, and Naveeduddin Mohammed. 2022. Using the DiCoT framework for integrated multimodal analysis in mixed-reality training environments. *Frontiers in artificial intelligence* 5 (2022), 941825.

[151] Caleb Vatral, Naveeduddin Mohammed, Gautam Biswas, Nicholas Roberts, and Benjamin Goldberg. 2023. A Comparative Analysis Interface to Streamline After-Action Review in Experiential Learning Environments. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)*. US Army Combat Capabilities Development Command–Soldier Center, N/A, N/A, 101.

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

[153] 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

[154] 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

[155] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv e-prints* N/A, N/A, Article arXiv:2201.11903 (Jan. 2022), N/A pages. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs.CL]

[156] 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 New South Wales Australia, 365–369. https://doi.org/10.1145/3170358.3170420

[157] 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

[158] 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

[159] Kang Xia, Wenzhong Li, Shiwei Gan, and Sanglu Lu. 2024. TS2ACT: Few-Shot Human Activity Sensing with Cross-Modal Co-Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–22.

[160] Lixiang Yan, Linxuan Zhao, Dragan Gasevic, and Roberto Martinez-Maldonado. 2022. Scalability, Sustainability, and Ethicality of Multimodal Learning Analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 13–23. https://doi.org/10.1145/3506860.3506862

[161] 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

[162] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. 2023. Hierarchical Video-Moment Retrieval and Step-Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. N/A, N/A, 23056–23065.

[163] Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2023. DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning. *arXiv preprint arXiv:2310.12128* N/A, N/A (2023), N/A.

[164] Mengxi Zhou, Joyce Fonteles, Joshua Danish, Eduardo Davalos, Selena Steinberg, Gautam Biswas, and Noel Enyedy. 2024. Exploring artificial intelligence supported interaction analysis. In *Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024*. International Society of the Learning Sciences, NY, USA, 2327–2328.

[165] John Zilvinskis, James Willis III, and Victor M. H. Borden. 2017. An Overview of Learning Analytics. *New Directions for Higher Education* 2017, 179 (2017), 9–17. https://doi.org/10.1002/he.20239 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/he.20239

[120] 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

[121] 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.

[122] 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

[123] 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.

[124] Jing Ru Teoh, Jian Dong, Xiaowei Zuo, Khin Wee Lai, Khairunnisa Hasikin, and Xiang Wu. 2024. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *PeerJ Computer Science* 10 (2024), e2298.

[125] Danielle R Thomas, Conrad Borchers, Shambhavi Bhushan, Erin Gatz, Shivang Gupta, and Kenneth R Koedinger. 2025. LIm-generated feedback supports learning if learners choose to use it. In *European Conference on Technology Enhanced Learning*. Springer, 489–503.

[126] Umesh Timalsina, Eduardo Davalos, Nihar Purshottam Sanda, Yike Zhang, Joyce Horn Fonteles, TS Ashwin, and Gautam Biswas. [n. d.]. SyncFlow: A Scalable Platform for Multimodal Learning Analytics. ([n. d.]). https://www.researchgate.net/publication/397356107_SyncFlow_A_Scalable_Platform_for_Multimodal_Learning_Analytics

[127] 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

[128] 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

[129] Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Vol. 86. Harvard university press.

[130] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv e-prints* N/A, N/A, Article arXiv:2201.11903 (Jan. 2022), N/A pages. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs.CL]

[131] 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.

[132] 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.

[133] 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

[134] 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

[135] 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.

[136] Lixiang Yan, Linxuan Zhao, Vanessa Echeverria, Yueqiao Jin, Riordan Alfredo, Xinyu Li, Dragan Gašević, 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.

[137] Lixiang Yan, Linxuan Zhao, Dragan Gasevic, and Roberto Martinez-Maldonado. 2022. Scalability, Sustainability, and Ethicality of Multimodal Learning Analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 13–23. https://doi.org/10.1145/3506860.3506862

[138] 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

[139] 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.

[140] 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.

[141] 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.

[142] Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep multimodal data fusion. *ACM computing surveys* 56, 9 (2024), 1–36.

[143] Xianxun Zhu, Chaopeng Guo, Heyang Feng, Yao Huang, Yichen Feng, Xiangyang Wang, and Rui Wang. 2024. A review of key technologies for emotion analysis using multimodal information. *Cognitive Computation* 16, 4 (2024), 1504–1530.

A CORPUS TABLE

Table 5 enumerates the 73 papers in this literature review's corpus.

UUID	First Author	Title	Year	Publication
2456887548 [5]	Alyuz	An Unobtrusive And Multimodal Approach For Behavioral Engagement Detection Of Students	2017	MIE
818492192 [6]	Andrade	Understanding Student Learning Trajectories Using Multimodal Learning Analytics Within An Embodied-Interaction Learning Environment	2017	LAK
3637456466 [8]	Ashwin	Impact Of Inquiry Interventions On Students In E-Learning And Classroom Environments Using Affective Computing Framework	2020	UMUAI
3448122334 [9]	Aslan	Investigating The Impact Of A Real-Time, Multimodal Student Engagement Analytics Technology In Authentic Classrooms	2019	CHI
1886134458 [10]	Azcona	Personalizing Computer Science Education By Leveraging Multimodal Learning Analytics	2018	FIE
3146393211 [11]	Birt	Mobile Mixed Reality For Experiential Learning And Simulation In Medical And Health Sciences Education	2018	Information
1326191931 [19]	Chan	Multimodal Learning Analytics In A Laboratory Classroom	2019	MLPALA
2936220551 [20]	Chango	Multi-Source And Multimodal Data Fusion For Predicting Academic Performance In Blended Learning University Courses	2020	CEE
4277812050 [21]	Chango	Improving Prediction Of Students' Performance In Intelligent Tutoring Systems Using Attribute Selection And Ensembles Of Different Multimodal Data Sources	2021	JCHE
1426267857 [23]	Chen	Affect, Support, And Personal Factors: Multimodal Causal Models Of One-On-One Coaching	2021	JEDM
3809293172 [27]	Closser	Blending Learning Analytics And Embodied Design To Model Students' Comprehension Of Measurement Using Their Actions, Speech, And Gestures	2021	IJCCI
4019205162 [38]	Cornide-Reyes	Introducing Low-Cost Sensors Into The Classroom Settings: Improving The Assessment In Agile Practices With Multimodal Learning Analytics	2019	Sensors
1576545447 [43]	Cukurova	Artificial Intelligence And Multimodal Data In The Service Of Human Decision-Making: A Case Study In Debate Tutoring	2019	BJET
1609706685 [48]	Di Mitri	Learning Pulse: A Machine Learning Approach For Predicting Performance In Self-Regulated Learning Using Multimodal Data	2017	LAK

2070224207 [98]	Di Mitri	Detecting Medical Simulation Errors With Machine Learning And Multimodal Data	2019	CAIM
3009548670 [51]	Di Mitri	Real-Time Multimodal Feedback With The Cpr Tutor	2020	AIED
1763513559 [49]	Di Mitri	Keep Me In The Loop: Real-Time Feedback With Multimodal Data	2021	IJAIED
1296637108 [53]	Echeverria	Towards Collaboration Translucence: Giving Meaning To Multimodal Group Data	2019	CHI
1581261659 [55]	Emerson	Early Prediction Of Visitor Engagement In Science Museums With Multimodal Learning Analytics	2020	ICMI
1598166515 [54]	Emerson	Multimodal Learning Analytics For Game-Based Learning	2020	BJET
4035649049 [57]	Fernández-Nieto	Storytelling With Learner Data: Guiding Student Reflection On Multimodal Team Data	2021	TLT
483140962 [62]	Fwa	Investigating Multimodal Affect Sensing In An Affective Tutoring System Using Unobtrusive Sensors	2018	PPIG
4278392816 [65]	Giannakos	Multimodal Data As A Means To Understand The Learning Experience	2019	IJIM
8536806391 [69]	Henderson	Sensor-Based Data Fusion For Multimodal Affect Detection In Game-Based Learning Environments	2019	EDM
86191824 [74]	Jiang	Examining How Different Modes Mediate Adolescents' Interactions During Their Collaborative Multimodal Composing Processes	2019	ILE
3398902089 [75]	Järvelä	What Multimodal Data Can Tell Us About The Students' Regulation Of Their Learning Process?	2019	LAI
32184286 [79]	Kubsch	Once More With Feeling: Emotions In Multimodal Learning Analytics	2022	MMLA Handbook
2056607681 [80]	Larmuseau	Multimodal Learning Analytics To Investigate Cognitive Load During Online Problem Solving	2020	BJET
1877483551 [83]	Lee-Cultura	Motion-Based Educational Games: Using Multi-Modal Data To Predict Player'S Performance	2020	COG
3660066725 [81]	Lee-Cultura	Children'S Play And Problem Solving In Motion-Based Educational Games: Synergies Between Human Annotations And Multi-Modal Data	2021	IDC
3856280479 [82]	Lee-Cultura	Children'S Play And Problem-Solving In Motion-Based Learning Technologies Using A Multi-Modal Mixed Methods Approach	2021	IJCCI
804659204 [87]	Liu	Towards Smart Educational Recommendations With Reinforcement Learning In Classroom	2018	TALE

3783339081 [86]	Liu	A Novel Method For The In-Depth Multimodal Analysis Of Student Learning Trajectories In Intelligent Tutoring Systems	2018	JLA
3796180663 [85]	Liu	Learning Linkages: Integrating Data Streams Of Multiple Modalities And Timescales	2018	JCAL
518268671 [89]	López	Using Multimodal Learning Analytics To Explore Collaboration In A Sustainability Co-Located Tabletop Game	2021	ECGBL
566043228 [18]	Ma	Automatic Student Engagement In Online Learning Environment Based On Neural Turing Machine	2021	IJIET
3754172825 [90]	Ma	Detecting Impasse During Collaborative Problem Solving With Multimodal Learning Analytics	2022	LAK
147203129 [91]	Mangaroska	Multimodal Learning Analytics To Inform Learning Design: Lessons Learned From Computing Education	2020	JLA
1847468084 [92]	Martin	Computationally Augmented Ethnography: Emotion Tracking And Learning In Museum Games	2019	ICQE
2879332689 [93]	Martinez-Maldonado	From Data To Insights: A Layered Storytelling Approach For Multimodal Learning Analytics	2020	CHI
2155422499 [99]	Morell	A Multimodal Analysis Of Pair Work Engagement Episodes: Implications For Emi Lecturer Training	2022	JEAP
2273914836 [101]	Nasir	Many Are The Ways To Learn Identifying Multi-Modal Behavioral Profiles Of Collaborative Learning In Constructivist Activities	2022	IJCSCL
1469065963 [102]	Nguyen	Examining Socially-Shared Regulation And Shared Physiological Arousal Events With Multimodal Learning Analytics	2022	BJET
2345021698 [104]	Noël	Exploring Collaborative Writing Of User Stories With Multimodal Learning Analytics: A Case Study On A Software Engineering Course	2018	Access
2609260641 [105]	Noël	Visualizing Collaboration In Teamwork: A Multimodal Learning Analytics Platform For Non-Verbal Communication	2022	DAMLE
2497456347 [107]	Ochoa	The Rap System: Automatic Feedback Of Oral Presentation Skills Using Multimodal Analysis And Low-Cost Sensors	2018	LAK
2634033325 [106]	Ochoa	Controlled Evaluation Of A Multimodal System To Improve Oral Presentation Skills In A Real Learning Setting	2020	BJET

3051560548 [110]	Olsen	Temporal Analysis Of Multimodal Data To Predict Collaborative Learning Outcomes	2020	BJET
123412197 [112]	Papamitsiou	Utilizing Multimodal Data Through Fsqca To Explain Engagement In Adaptive Learning	2020	TLT
85990093 [115]	Petukhova	Multimodal Markers Of Persuasive Speech : Designing A Virtual Debate Coach	2017	INTERSPEECH
957160695 [114]	Petukhova	Virtual Debate Coach Design: Assessing Multimodal Argumentation Performance	2017	ICMI
1374035721 [116]	Pham	Attentivelearner2: A Multimodal Approach For Improving Mooc Learning On Mobile Devices	2017	AIED
2836996318 [117]	Pham	Predicting Learners' Emotions In Mobile Mooc Learning Via A Multimodal Intelligent Tutor	2018	ITS
3135645357 [119]	Prieto	Multimodal Teaching Analytics: Automated Extraction Of Orchestration Graphs From Wearable Sensor Data	2018	JCAL
3408664396 [120]	Psaltis	Multimodal Student Engagement Recognition In Prosocial Games	2017	T-CIAIG
3308658121 [123]	Reilly	Exploring Collaboration Using Motion Sensors And Multi-Modal Learning Analytics	2018	EDM
3625722965 [95]	Sanusi	Table Tennis Tutor: Forehand Strokes Classification Based On Multimodal Data And Neural Networks	2021	Sensors
2000036002 [131]	Sharma	Predicting Learners' Effortful Behaviour In Adaptive Assessment Using Multimodal Data	2020	LAK
1118315889 [135]	Spikol	Using Multimodal Learning Analytics To Identify Aspects Of Collaboration In Project-Based Learning	2017	CSCL
3339002981 [137]	Spikol	Estimation Of Success In Collaborative Learning Based On Multimodal Learning Analytics Features	2017	ICALT
1637690235 [136]	Spikol	Supervised Machine Learning In Multimodal Learning Analytics For Estimating Success In Project-Based Learning	2018	JCAL
3796643912 [138]	Standen	An Evaluation Of An Adaptive Learning System Based On Multimodal Affect Recognition For Learners With Intellectual Disabilities	2020	BJET
2181637610 [139]	Starr	Toward Using Multi-Modal Learning Analytics To Support And Measure Collaboration In Co-Located Dyads	2018	ICLS
1315379489 [141]	Sümer	Multimodal Engagement Analysis From Facial Videos In The Classroom	2021	TAC

3093310941 [142]	Tanaka	Embodied Conversational Agents For Multimodal Automated Social Skills Training In People With Autism Spectrum Disorders	2017	PLOS	
1345598079 [143]	Tancredi	Intermodality In Multimodal Learning Analytics For Cognitive Theory Development: A Case From Embodied Design For Mathematics Learning	2022	MMLA Handbook	
433919853 [147]	Tisza	Understanding Fun In Learning To Code: A Multi-Modal Data Approach	2022	IDC	
1770989706 [153]	Vrzakova	Focused Or Stuck Together: Multimodal Patterns Reveal Triads' Performance In Collaborative Problem Solving	2020	LAK	
2055153191 [154]	Vujovic	Round Or Rectangular Tables For Collaborative Problem Solving? A Multimodal Learning Analytics Study	2020	BJET	
3095923626 [157]	Worsley	A Multimodal Analysis Of Making	2017	IJAIED	
3309250332 [156]	Worsley	(Dis)Engagement Matters: Identifying Efficacious Learning Practices With Multimodal Learning Analytics	2018	LAK	
666050348 [158]	Worsley	Multicraft: A Multimodal Interface For Supporting And Studying Learning In Minecraft	2021	HCII	
1019093033 [161]	Yang	Prime: Block-Wise Missingness Handling For Multi-Modalities In Intelligent Tutoring Systems	2019	MMM	

Table 5. Each of the 73 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 view and distill the initial search results to the 73 papers in our final corpus.

B.1 Literature Search

Our literature search consisted of 42 search strings defined, discussed, and agreed upon by the authors as being representative of the body of works this literature review would be conducted on. Instead of performing our queries manually, we opted to perform our queries programmatically via an API-based Google Scholar web scraping tool. There are several available tools for scraping Google Scholar, such as scholarly [25] and gscholar [152]. Ultimately, we employed SerpAPI [128], a third-party Google Scholar web scraping API, for its most essential feature: organic web results. Other API tools' results are not organic, i.e., a query made via the API and one manually queried in a browser-based environment will produce two different sets of results.

Queries were posed via API request to Google Scholar for papers published between 1/1/2017 and 10/22/2022 (the date of our literature search). 2017 was collectively agreed upon as being the best cutoff date for inclusion in our search due to the rapid technological advancements in the field over the past 5 years. Several papers prior to 2017 are discussed in Section 1, as they are seminal works; however, they are not considered for inclusion in our corpus.

For the literature search, this review's authors decided on 14 distinct search phrases, and each phrase was searched 3 times with a different spelling of the word *multimodal* — multimodal, multi-modal, and multi modal — prepended to it. The 14 search phrases are enumerated in Table 6.[2]

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

Table 6. Search strings used for the literature search.

For each of the 42 search strings, the top 5 pages (100 publications) deemed most relevant by Google Scholar were collected. The top-5 cutoff was financially imposed because of our subsequent citation graph construction (see Appendix B.2.1). To build the citation graph, each individual paper's citation information is queried, but each query is capped at 20 citations per API call by SerpAPI. This means that a paper with 100 citations requires 5 additional API calls to gather all of its citation information. The number of API calls needed to construct the citation graph would be intractable (and unaffordable) if the initial search was left unbounded; therefore, the top-5 cutoff was put in place.

Our initial search yielded a total of 4,200 papers (14 unique search terms * 3 spellings of multimodal * 100 publications per search string). The distillation procedure we used for corpus reduction is enumerated in Table 7 and discussed in

[2]The term "xai" was included in the search due to the authors' interest in exploring explainable AI methods applied to learning and training environments. Unfortunately, the field is still nascent, and no usable query results were returned with this search string.

the following subappendices. Throughout this appendix, each step of our corpus reduction procedure is identified via its Step ID in Table 7.

Step ID	Procedure	Removed	Remaining
0	Literature search	0	4200
1	Remove duplicates	2079	2121
2	Remove non-English	1	2120
3	Remove degree-0 nodes	488	1632
4	Remove disconnected components	101	1531
5	Iteratively remove degree-1 nodes		
5.1	Iteration 1	373	1158
5.2	Iteration 2	74	1084
5.3	Iteration 3	19	1065
5.4	Iteration 4	2	1063
6	Remove titles with keywords	204	859
7	Title reads	471	388
8	Abstract reads		
8.1	Remove inaccessible abstracts	10	378
8.2	First abstract round	211	167
8.3	Second abstract round	40	127
9	Full paper reads		
9.1	First full paper round	52	75
9.2	Feature discretization and extraction	2	73
9.3	Second full paper round	0	73
9.4	Second feature extraction round	0	73

Table 7. Our corpus reduction procedure. Step ID 0 is the literature search. Steps 1 and 2 used programmatic filtering via Python packages. Steps 3-5 were performed quantitatively via CGP. Step 6 uses human-in-the-loop regex filtering. Steps 7-9 were performed qualitatively via our quality control procedures. At each step of the corpus reduction procedure, the number of papers pruned and number of papers remaining are listed.

Our initial corpus contained 2,079 duplicates, which were removed by hashing paper titles (Table 7, Step ID 1). If a paper had multiple versions (or other duplicates), we used the official source (e.g., journal or conference) of publication. We removed 1 non-English paper (Table 7, Step ID 2) due to pragmatism (English is the only language shared between all of this review's authors). Non-English papers were identified using spaCy FastLang [146], where any paper whose title was identified as having less than a 100% chance of being English was selected for manual review and potential exclusion. In total, our initial search yielded 2,120 unique English papers published within our search window.

B.2 Study Selection

To reduce our corpus to a reviewable body of works, we employed both quantitative ans qualitative methods. After the initial search, we distilled the corpus quantitatively via CGP, which we discuss in Appendix B.2.1. Subsequent distillation was performed via qualitative means and is discussed in Appendix B.2.2.

B.2.1 Citation Graph Pruning (Quantitative Corpus Reduction). For visualization, analysis, and distillation purposes, we used NetworkX [67] to create and display a *citation graph* of the initial 2,120 works considered for inclusion in this review. The citation graph is a directed acyclic graph (DAG), where each node is a paper uniquely identifiable by its UUID (universally unique identifier) on Google Scholar, and each directed edge from A to B indicates paper A cites paper B. For the purposes of this paper, we consider the degree of each node (paper) p to be the sum of both incoming and outgoing edges, i.e., papers citing p and papers cited by p, respectively. We again used SerpAPI for collecting the list of works that cited each paper. The citation search did not need to be conducted in both directions, as any paper citing another paper in our corpus would already have been identified by the "cited by" list of the paper being cited. Citations by papers not included in our initial search (i.e., in the DAG) were ignored. Initially, our DAG contained a 3-node cycle. This was due to different editions of the same book chapter and papers by the same author citing each other during preprint. Once the cycle was identified, the cycle's edges were removed from the edge set. No nodes were removed as a result of correcting the cycle.

Once the DAG was constructed, we removed all 0-degree nodes (Table 7, Step ID 3) (i.e., nodes with no edges coming in or going out). We felt it reasonable that if a paper did not cite (or was not cited by) any other papers in the field (as determined by our literature search), then the paper was either not relevant to the field or did not yield methods or findings referenced by subsequent works. Importantly, our approach strikes a balance between incoming and outgoing citations, as earlier works are unable to *cite* many works in the corpus, and later works are unable to *be cited by* many works in the corpus. For example, works from early 2017 may not have any outgoing edges simply due to being some of the earliest works in the corpus, which would have prevented them from citing papers that had not yet been published. However, these same papers had a greater opportunity to be cited by subsequent papers, which is why we felt it important to consider both incoming and outgoing edges: we expect earlier papers to have more incoming edges and later papers to have more outgoing edges. Altogether, pruning 0-degree nodes from the DAG reduced our corpus by 488, dropping our corpus count to 1,632 works.

After removing 0-degree nodes, we examined the DAG's connectivity (Table 7, Step ID 4) to identify disconnected components not relevant to our literature search. This had to be done to account for overlapping terminology across domains. For example, a cursory look at our initial search results included several "multimodal training" papers related to deep learning (DL), where artificial neural networks (ANNs) are trained using data across multiple modalities but are not applied to multimodal learning or training environments. Our hypothesis, based on our search strings, was that the works relevant to this review would comprise the largest component of the DAG, leaving other smaller, disconnected components to be discarded as irrelevant because they lacked any edge to or from the DAG's primary component.

Evaluating the DAG's connectivity, we found one large component consisting of 1,531 nodes (papers) and 44 smaller, disconnected components of various sizes totaling 101 papers. The sizes of the disconnected components, their frequencies of occurrence in the DAG, and the total number of papers for each component size are listed in Table 8. All 101 papers were removed from the corpus by pruning the DAG's disconnected components, which left 1,531 papers represented by a single, connected graph.

Once we had our single component graph, we removed 1-degree nodes to further prune it. This created new 1-degree nodes, which were also removed. This process of removing 1-degree nodes was repeated four times (Table 7, Step ID 5) until the graph was stable (i.e., removing 1-degree nodes did not create any new 1-degree nodes). By iteratively removing 1-degree nodes, we felt we could effectively identify and remove works outside the scope of our literature review without losing works directly related to multimodal learning and training environments. This is because the field of multimodal learning and training environments spans several sub-fields across computer science, education, and cyberphysical systems, and the authors agreed it was unlikely papers with so few edges would be relevant to our review if they had not cited (or been cited by) more than a few other works in our corpus. We removed 373 nodes in the first iteration (Table 7, Step ID 5.1), 74 nodes in the second iteration (Table 7, Step ID 5.2), 19 nodes in the third iteration (Table 7, Step ID 5.3), and 2 nodes in the fourth and final iteration (Table 7, Step ID 5.4). Altogether, we removed 468 papers over four iterations, which reduced our corpus from 1,531 papers to 1,063. The CGP pseudocode is presented in Section 3.2.1 (Algorithm 1).

It was at this point we concluded our quantitative pruning procedure and began qualitatively reducing the corpus, which we discuss in the next subappendix.

Size	#	Papers
2	35	70
3	6	18
4	2	8
5	1	5

Table 8. Disconnected DAG components by number of nodes in the component (size), frequency of occurrence (#), and total number of papers (papers). For instance, the first row indicates that there were 35 disconnected components of size 2 in the graph, totaling to 70 papers.

B.2.2 Quality Control (Qualitative Corpus Reduction). Manually examining the remaining 1,063 titles informed us that a large part of our corpus was still outside the scope of our review. First, we noticed there were still many papers related to training multimodal neural networks. We also noticed many works applying multimodal methods to the medical field, usually in terms of medical imaging. To remove papers pertaining to multimodal neural network training and multimodal medical applications, we programmatically identified 217 titles via regex keyword search (Table 7, Step ID 6) that contained at least one of the six following words: neural, deep, machine, medical, medicine, and healthcare. We then evaluated the selected titles by hand. Of the 217, 13 were kept in the corpus due to their potential relevance to our review. Papers employing deep learning methods in MMLA and applying multimodal methods to medical learning or training environments were within the scope of our review, for example. Some specific examples include removing one paper titled, "deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues" [66]; and keeping one titled, "supervised machine learning in multimodal learning analytics for estimating success in project-based learning" [136]. The remaining 204 papers were removed from the corpus, reducing it to 859 potentially relevant works.

Next, we selected papers for exclusion based on consensus. Pursuant to Kitchenham [77], we initially excluded works based on reading papers' titles, then abstracts, and eventually full manuscripts. The first five authors of this review acted as reviewers (henceforth referred to as "the Reviewers") for the quality control procedure. For the title reads (Table 7, Step ID 7), four of the Reviewers read all 859 titles. For each title, each Reviewer independently determined whether the title was likely to fall inside the scope of the review. The results were tallied, and papers were then selected for inclusion/exclusion based on majority voting, i.e., papers with at least three votes "for" were automatically included, and papers with at least three votes "against" were automatically excluded. For the papers with a 2-2 tie, a fifth reviewer was used as a tie breaker. The Reviewers selected 347 papers for inclusion and 372 papers for exclusion. 140 papers were tied, and a fifth reviewer selected 41 of those for inclusion. In total, 388 papers were selected for inclusion after the title reads — 347 by majority vote, and 41 by tie-breaker.

Before conducting the abstract reads (Table 7, Step ID 8), several works were excluded due to their inaccessibility (Table 7, Step ID 8.1). While gathering the abstracts, we noticed not all papers were publicly available. Several were defined by invalid URLs or behind paywalls. Whenever a paper's abstract (or introduction, in the case of a book or book chapter) was unavailable via its SerpAPI URL, a Google search was conducted in order to obtain the abstract manually through websites such as ResearchGate and other academic repositories. When this failed, we relied on the [Anonymous] University Library's proxy to access papers behind paywalls. If we were unable to freely access a paper's abstract online through Google search or via Vanderbilt's proxy, the paper was excluded from the corpus. Altogether, 10 papers were removed due to inaccessibility, leaving 378 papers for the abstract reads.

The "abstracts" quality control procedure consisted of two rounds. Similar to the procedure for the title reads, each of the remaining 378 abstracts was first assigned to two Reviewers, and a majority voting scheme was employed (Table 7, Step ID 8.2). Papers were then selected for inclusion or exclusion based on a predefined set of exclusion criteria. The exclusion criteria for the abstracts is listed in Table 9. Exclusion criteria are cumulative, so each criterion applies to subsequent steps in our corpus reduction procedure. An exclusion criterion for the abstracts will similarly apply to full paper reads later on, for example.

Because this literature review focuses on multimodal methods applied to learning and training environments, any paper not dealing with a learning or training environment was not considered for this review. As mentioned in Section 1, VR environments were also not considered for inclusion in our corpus due to issues with scaling this technology in classroom settings and the lack of semantic meaning with respect to the environment for video analysis. If a paper does not analyze multimodal data, it is similarly out-of-scope for this review. Papers must also include systematic methods for analyzing the multimodal data, and those methods must be original, applied research. Papers that are literature reviews, pedagogical tools, theoretical foundations, doctoral consortiums, etc., may be used for reference in our Introduction and Background, but they are not considered for inclusion in the actual review corpus unless they additionally provide original, applied research via multimodal methods and analysis.

1. Paper does not deal with learning or training environments
2. Paper's environment is VR-only
3. Paper does not analyze multimodal data
4. Paper does not apply multimodal analysis methods
5. Paper is not original applied research

Table 9. Exclusion criteria for the abstract reads. Each of the 378 abstracts was assigned to two different Reviewers. Each reviewer was instructed to exclude works based on this set of criteria.

Of the 378 abstracts, Reviewers agreed to keep 96 papers (i.e., both Reviewers selected the work for inclusion) and discard 211 (i.e., both Reviewers selected the work for exclusion). 71 were selected for further review (i.e., one reviewer selected the work for inclusion and one reviewer selected the work for exclusion). To address the 71 abstracts that did not receive unanimous agreement among Reviewers, a second round of abstract reads was performed (Table 7, Step ID 8.3). This round consisted of each of the 71 abstracts without unanimous agreement receiving three additional reads: one read from each of the three Reviewers who did not read the abstract in the initial abstract round. Each of the 71 papers was subsequently included or excluded based on majority voting (i.e., papers were kept if and only if at least two out of the three second abstract round Reviewers elected to keep the abstract in the corpus). Of the 71 second abstract round papers, 31 were selected for inclusion, and 40 were removed from the corpus. With 96 papers selected for inclusion from the first round of abstract reads, and 31 papers selected from the second round, 127 papers in total were kept in the corpus for the next round of quality control: full paper reads.

The "full paper" quality control procedure also involved two rounds of review. To conduct full paper reads (Table 7, Step ID 9), the 127 papers kept from the abstract round were split into 5 approximately equal partitions and randomly assigned to the 5 Reviewers. Conducting full paper reads took several weeks, during which two additional exclusion criteria were defined. They are enumerated in Table 10:

1. Paper's results are not informative with respect to learning or training
2. Paper's analysis methods are not able to be determined from the manuscript

Table 10. Exclusion criteria for the full paper reads. Each of the 127 papers was assigned to two different Reviewers. Each reviewer was instructed to exclude works based on this set of criteria (in addition to the previously established exclusion criteria).

Certain papers deal with learning or training environments but are outside the scope of this review because they are not informative with respect to learning or training. Consider a paper presenting a novel neural network architecture that uses a classroom dataset as a performance benchmark. While the classroom constitutes a learning environment, the paper itself is not conducting research to inform learning or training, but rather is using a dataset collected from a learning environment to evaluate its "core AI" approach. We elected not to include these types of works in our review, as we aim to focus on multimodal methods that are explicitly used to inform learning or training. Additionally, a few papers we encountered did not have analysis methods that were well-defined enough for feature extraction (i.e., we were unsure of their exact methods for analyzing the data). This often included short workshop papers whose method details were unable to be determined without referencing an external work[3]. Because these types of papers would be very difficult to reproduce on their own, we elected to exclude them from our review.

During the first round of full paper reads (Table 7, Step ID 9.1), Reviewers marked each paper as "immediate exclude," "immediate accept," "borderline exclude," or "borderline accept." Papers marked as "immediate exclude" were discussed by all 5 Reviewers and excluded only if all agreed. These were papers with easily identifiable reasons for exclusion based on our criteria (for instance, a proposed theoretical framework with no analysis or a doctoral consortium presenting ideas for future research). No papers were ever excluded from our corpus during full paper reads without unanimous agreement from all five Reviewers. Papers marked as "immediate accept" were kept in the corpus for the second full paper read round. Papers marked as "borderline exclude" or "borderline accept" were assigned to a separate reader for further review and were subsequently discussed. Similar to papers marked for immediate exclusion, borderline papers were excluded prior to the second full paper read round only if all Reviewers agreed. Altogether, 52 papers were excluded during the first round of full paper reads, which left 75 works remaining in the corpus.

B.3 Feature Extraction

During the first full paper read round, several features were extracted from each paper (Table 7, Step ID 9.2). Features included identifying information (e.g., title, first author, publication year), and information related to the paper's methods (e.g., data collection mediums, modalities, and analysis methods). The extracted features and their descriptions are found in Table 11.[4]

[3]This does not include all workshop papers. Only those papers whose analysis methods could not be determined from the manuscript

[4]For the "Year" category, we used the date the manuscript was first publicly available (if listed, otherwise we used the publication date) in order to most accurately represent when the methods were performed. In some instances, the first date of online availability preceded the official publication date by over a year. Additionally, only data that was ultimately used in the paper's analysis was considered for the "Data Collection Mediums" category (i.e., if it was collected but never analyzed, we did not include it).

Feature	Description
UUID	Universally unique identifier on Google Scholar
Title	Publication title
First Author	Publication's first author
Year	Year publication was first publicly available
Environment Type	Type of environment analyzed in the publication
Data Collection Mediums	Types of data collected from the environment
Modalities	List of the different modalities used during analysis
Analysis Methods	List of the analysis methods used in the publication
Fusion Type	List of the types of data fusion used in the publication
Publication Source	Publication journal, conference, workshop, etc.

Table 11. Initial features extracted from each paper.

After the first read, the Reviewers discussed their extracted features. To ensure alignment and understanding between the Reviewers with respect to the features, feature categories were discretized via inductive coding [145], where four Reviewers each extracted initial feature sets from 25% of the corpus's papers. For example, the initially extracted *data collection mediums* feature included instances of video camera, web camera, and Kinect camera, all of which were mapped to the "VIDEO" data collection medium. Once the Reviewers agreed on the discrete sets of features, papers were reread by their original Reviewers, and their features were extracted into the discrete sets. The initial feature-space is described below in Table 12. We call these features *circumscribing features* to delineate them versus the identifying features (e.g., UUID, paper title, author, etc.) that were extracted for identification purposes but not used during analysis. In total, two sets of circumscribing features were extracted from the corpus to gather the information needed to conduct our analysis (Table 7, Step IDs 9.2 and 9.4).

Feature	Feature Set
Environment Type	learning, training
Data Collection Mediums	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 value
Analysis Methods	Classification, regression, clustering, qualitative, statistical methods, network analysis, pattern extraction
Fusion Type	Early, mid, late, hybrid, other

Table 12. The first round of circumscribing features and their corresponding feature sets. For *Environment Type*, items in the feature set are mutually exclusive (i.e., an environment can either be a learning *or* training environment for the purposes of this paper, but it cannot be both). All other circumscribing features can consist of multiple items in the feature set (e.g., each paper in our corpus will contain multiple data collection mediums or modalities). For feature set acronyms, see Section 2.1.

During feature discretization and extraction (Table 7, Step ID 9.2), additional papers were newly identified for possible exclusion pursuant to our aforementioned criteria. After discussing each paper selected for possible exclusion, 2 papers were removed from the corpus due to all five Reviewers agreeing that each paper violated at least one exclusion criterion. After the two removals, 73 papers remained in the corpus, all of whose features were extracted into discrete sets pursuant to Table 11 by the first full paper read round reviewer. At this point, a second and final quality control round was performed for full paper reads (Table 7, Step ID 9.3), where each of the 73 papers remaining in the corpus was assigned to a reviewer who had not yet read that particular paper. For this round, Reviewers were instructed to perform two tasks: identify any papers remaining in the corpus that violated any of the exclusion criteria (to discuss later for possible exclusion), and perform a round of feature extraction (to determine inter-rater reliability, or IRR, with respect to the initial feature extraction via Cohen's k [31]). For this round, no additional papers were identified for exclusion, resulting in a final corpus of 73 works. Each paper's discrete feature sets were ultimately determined via consensus coding [24] by the two Reviewers who read that particular paper (i.e., for each paper, both Reviewers needed to agree on the presence or absence of each item in each feature's feature set). For reference, Cohen's k before consensus for the first round of feature extraction was $k = 0.873$.

Once our corpus was finalized, we performed one additional round of feature extraction (Table 7, Step ID 9.4) to allow for greater insight into the corpus via a more in depth analysis. These features are: Environment Settings, Domains of Study, Participant Interaction Structures, Didactic Natures, Levels of Instruction, Analysis Approaches, and Analysis Results (the findings reported from each paper). All of these features are explained in Section 2.1 and presented again here in Table 13 for readability alongside their discrete values. The one exception is Analysis Results, which was not discretized due to the wide degree of variability across each paper's findings. Instead, we noted each paper's findings, and used them in our thematic analysis [16], which we describe in Section 3.4.

Circumscribing 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

Table 13. The second set of circumscribing features, all of which are multi-label, and their corresponding feature sets.

Similar to our initial round of feature extraction, we began with inductive coding, where four Reviewers first extracted the new circumscribing features for the same papers he or she performed inductive coding on during the previous round of feature extraction. We then discussed each paper's extracted features and formulated discrete sets for the new circumscribing features (with the exception of Analysis Results). Next, we conducted two rounds of full paper reads to extract the second set of circumscribing features. During the first round, Reviewers revisited the same papers they read during inductive coding and extracted the new circumscribing features pursuant to the agreed-upon feature sets devised during inductive coding. During the second round, Reviewers reread (and extracted the additional features from) the same set of papers they were the 2nd reviewer for during the initial round of feature extraction. At this point, for each paper, the two Reviewers who extracted that paper's additional features performed consensus coding to define

that paper's final set of features. For reference, Cohen's $k = 0.71$ for the second round of feature extraction prior to consensus coding.

Each item in each of the circumscribing feature sets is described in Sections 2.2.1 (Environment Type), 2.2.2 (Data Collection Mediums), 2.2.3 (Modalities), 2.2.4 (Analysis Methods), 2.2.5 (Data Fusion), 2.2.6 (Environment Setting), 2.2.7 (Domain of Study), 2.2.8 (Participant Interaction Structure), 2.2.9 (Didactic Nature), 2.2.10 (Level of Instruction or Training), and 2.2.11 (Analysis Approach).

C LITERATURE REVIEW LIMITATIONS

The limitations of this work involve the use of Google Scholar to conduct the literature search, the use of a citation graph for programmatic corpus reduction, and a lack of screening for peer reviewed papers. All are discussed below.

C.1 Google Scholar.

While Google Scholar is widely used by researchers across both academia and industry, it poses a challenge for reproducibility. Like Google Search, Google Scholar is a proprietary search algorithm that is assumed to vary its results based on context. Factors such as the individual user conducting the search, the user's geolocation, the date the search is conducted, and the user's search history may all affect how Google Scholar collates search results. Google may also perform A/B testing in live environments to determine which version of its algorithm users deem more effective. The algorithm is also (presumably) continually evolving, and users are unable to know exactly which version of the algorithm was used to conduct a particular search. As such, there is little expectation that our initial corpus will be able to be reconstructed *in its exact form* without at least some degree of variability.

However, the authors are confident the degree of variability from different Google Scholar searches does not prohibit the *overall* reproducibility of the initial corpus. While SerpAPI's web scraping method is proprietary, its creators address several of our concerns in their documentation [128]. The API's search does not use information from any individual user's Google account when conducting the web scrape, as no Google account is attached to the SerpAPI account, API key, or API calls themselves. Instead, calls are made via proxy and random headers, as illustrated in Figure 8. When trying to reproduce the API's results via manual search, SerpAPI recommends using the URL in the API's JSON results in "incognito mode".

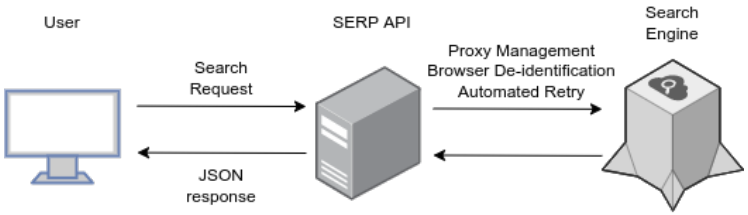


Fig. 8. Searching Google Scholar via SerpAPI.

Additionally, we reached out to SerpAPI directly and asked, "Does SerpAPI attach personal or identifying information when making request?", to which SerpAPI responded, "No, we don't add any personal information." SerpAPI also stated, "...others can reproduce your results by using Google Scholar website, if they use the same search criteria...", but we believe this to be an overstatement given Google's lack of transparency. While we cannot guarantee perfect reproducibility due to the aforementioned issues, we can state with a reasonable degree of confidence that our own individual search biases did not influence the initial search results (outside of the choosing of the search terms) due to how SerpAPI handles API calls to Google Scholar. For reference, this review's literature search was conducted by an author of this paper in Nashville, TN, USA.

C.2 Citation Graph Pruning.

As discussed in Section 3.2.1, we initially distilled our corpus quantitatively via citation graph pruning. In doing so, it is possible we excluded relevant works from our corpus based on them only having cited or been cited by a minimal number of other works in our corpus. However, this paper is a literature review of the prominent methods researchers are applying to multimodal learning and training environments. As such, the authors agreed that if a work did not utilize a large degree of previous research (i.e., cite several other works in the corpus) or serve as a base from which a large degree of other research has built upon (i.e., be cited by several other works in the corpus), then that work was, by definition, outside the scope of our review. Considering our corpus was still largely comprised (over 50%) of works later deemed to be outside the scope of this review after CGP, the authors are confident that few papers (if any) directly pertaining to multimodal learning and training environments were discarded as a result of CGP.

C.3 Peer Review.

Due to the prevalence of papers being published to open, non-peer-reviewed platforms like arXiv in recent years (particularly in computer science), we did not screen for non-peer-reviewed works during study selection (i.e., we did not adopt a paper's not being peer-reviewed as an exclusion criterion). To the best of our knowledge, all papers in our corpus underwent formal peer-review, with one possible exception. There is one paper in the corpus that was submitted to a workshop that none of this review's authors are familiar with. We are, therefore, unsure of whether or not the paper underwent formal peer review. However, the workshop includes submission, notification, and camera ready dates, so we are confident that the workshop was at least refereed.