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

CLAYTON COHN, Vanderbilt University, USA
EDUARDO DAVALOS, Vanderbilt University, USA
CALEB VATRAL, Vanderbilt 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.

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

This work was supported by National Science Foundation grants XXXXXXXX and XXXXXXXX..
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, caleb.m.vatral@vanderbilt.edu, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN, USA, 37235; 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

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), 55 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* [62]. Learning analytics focuses on gathering and evaluating data on learners' and trainees' behaviors—specifically, their approaches to learning and training tasks [95, 166]. For example, intelligent tutoring systems like Practical Algebra Tutor [79] focus on diagnosing student errors, open-ended environments like Betty's Brain [85] adaptively scaffold learning, and teacher-feedback tools (e.g., [73, 125]) 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* [109, 151]?. 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 [72, 109]. [72, 109].

Advances in sensor and data collection technologies are propelling learning analytics beyond traditional log-based analyses [109]. In physical learning spaces, traditional computer-based action logging is challenging and computer log data is insufficient to capture the entirety of learners' actions, affective states, and collaborative behaviors. Recognizing this limitation, researchers are integrating additional data collection devices in classroom environments to supplement log data; employing video to record physical interactions, microphones for conversations, biometric sensors for stress levels, and eye trackers for attention [151].

This enriched data collection enables a more complete understanding of students' affective, cognitive, psychomotor, and metacognitive states, contributing to the evolution of multimodal learning analytics (MMLA) [13, 14, 158]. MMLA has matured through a decade of extensive research, disseminated via journal special issues (e.g., [53, 97, 110]), conferences run by the Society for Learning Analytics Research [61], an edited volume [65], and several systematic literature reviews (e.g., [4, 23, 40, 51, 101, 130, 158]). This review concentrates on *applied research methods* in MMLA, building upon this substantial foundation.

1.2 Related Work

Recent work in MMLA research, surveys, and literature reviews have portrayed the MMLA landscape through multiple lenses that include: multimodal data fusion [23], conceptual model and taxonomy [51], statistical and qualitative assessment of literature [122, 131], virtual reality [119], technology and data engineering focused for automating MMLA [27], and impact and ethical considerations [4]. Our literature review augments these by focusing on applied methods being used to support data collection and analysis from multimodal learning and training environments. To our knowledge, no recent work has explicitly centered on multimodal learning and training methodologies, i.e., the approaches taken to collect, fuse, and analyze multimodal data and interpret that data using learning theories. Below,

we detail two foundational reviews for framing our paper. We extend and modify these existing taxonomies to reflect recent MMLA advances.

Di Mitri et al. [51] introduced the Multimodal Learning Analytics Model (MLeAM), a conceptual framework that outlines the relationship between behavior, data, machine learning, and feedback in MMLA. This work provided a taxonomy and introduced the concept of data observability, distinguishing between the input domain of quantifiable evidence and the hypothesis domain of inferred annotations like emotions and cognition. The *observability line* demarcates these domains, underpinning the AI-mediated transformation from input to hypotheses—pivotal in MMLA research. A second influential survey by Chango et al. [23] surveyed fusion methods in MMLA, categorizing studies by fusion type and application stage within the multimodal pipeline. They proposed a classification with three fusion types: *early* (feature-level integration), *late* (decision-level integration), and *hybrid* (a combination of both). This scheme clarifies fusion approaches and their relevance to educational data mining.

Integrating insights from both surveys, we propose a classification focused on *feature observability*, which can be used to identify the different processes by which individuals fuse different sources of information to create unified data representations. It further distinguishes between sensory data and human-inferred annotations. This adapted scheme, detailed in the following section, 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, and so on). We define a *modality* as a unique attribute derived from data from one or more data streams, where each modality conveys different information, even if derived from the same data collection medium [109]. We define *modality groups* as distinct sets of modalities that convey similar types of information, which we derive via inductive coding (see Section 1 for details). We define *multimodal* as a combination of either multiple modalities **or** multiple data streams (i.e., multiple data collection mediums). For instance, the same video data stream can be used for the affect and pose modalities (one-to-many), and the affect modality could be derived from separate audio and video data streams (many-to-one). Both examples are considered to be multimodal by our definition. Additionally, we use the terms "papers" and "works" interchangeably in this review, as we expand our definition of "paper" to include other publications outside of conference and journal submissions (e.g., books and book chapters).

Importantly, our literature review includes all papers gathered during our literature search that were not excluded for any of the reasons listed in our exclusion criteria (see Section B.2.2). This includes multimodal learning and training analysis that was done "in passing" and not as a core focus of the paper. Consider a paper whose sole focus is multimodal composing environments (i.e., creating works using multiple modalities rather than analyzing multimodal data) that perform multimodal learning analysis as a byproduct. Should this paper be included, despite not having multimodal analysis as one of its focuses? In this review, we argue yes, as we are interested in the different methods researchers have used to conduct multimodal analysis; we do not limit ourselves to papers where multimodal analysis is the primary focus. These distinctions are quite nuanced, which is why the definitions we use in this work are not meant to establish a "universal" definition of multimodality and multimodal analysis. They were chosen after much research, analysis, and discussion to characterize the scope of our review.

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 [38].

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;
- 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 of our literature review and our taxonomy for multimodal methods applied to 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 the framework depicted in Figure 1). This includes an analysis of each of the 5 modality groups (Section 4.2). Section 5 presents three categories of research (which we refer to as "archetypes") that, based on our analysis, best characterize the multimodal learning and training field as a whole. Section 6 presents our findings for the corpus in the aggregate, highlighting the field's current trends, state-of-the-art, results, challenges, and research gaps. It also addresses the limitations of our work and future research directions. Section 7 concludes our literature review 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.

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

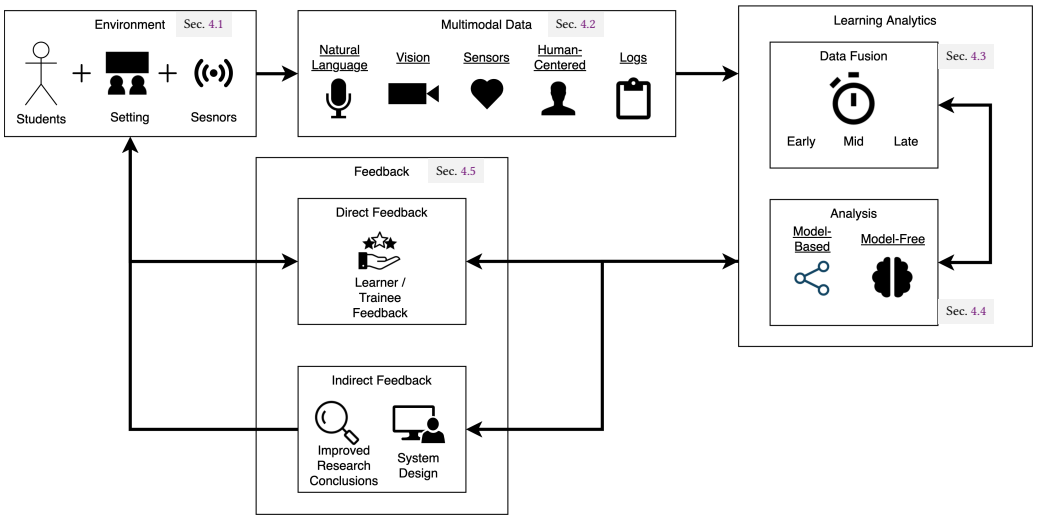


Fig. 1. Multimodal Learning and Training Environments Literature Review Framework

due to scalability issues [38]. 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 second component of our framework is *multimodal data*, which comprises the diverse data streams generated by the environmental sensors. We classify this data into five modality groups (Section 2.2.3): (1) natural language, (2) vision, (3) sensors, (4) human-centered, and (5) environment logs, each detailed in respective 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 the specific features extracted from our reading of the review corpus.

2.2.1 Environment Type. In examining MMLA literature, it is crucial to recognize the contexts where these methods are utilized. 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 [96, 105, 116, 155] and the physical-virtual space continuum [20, 39, 118].

Multimodal methods in learning environments aim to enhance educational outcomes by analyzing data to understand 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 [52],

athletics [96], and the military [70]. 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 [93, 159], 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.

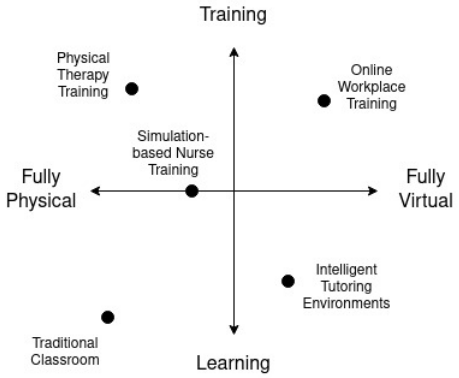


Fig. 2. Learning-Training Continuum

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 mediums* plays a pivotal role in gaining a comprehensive understanding of learners' progress, interactions, strategies, and struggles within these environments. We previously defined *data collection mediums* as the different types of raw data collected. 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.

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.

Medium	Definition
Video	Sequence of video frames from a camera source [28, 56, 118].
Audio	Audio signals captured by a microphone [115, 116, 143].
Screen Recording	Sequence of video frames of the contents displayed on a device screen [5, 75, 87].
Eye	Eye movement data and gaze points captured by tracking devices with sensors and cameras [22, 113, 144].
Logs	Environment logs containing data on participants' activities and interactions within the system [11, 121, 136].
Sensor	Specialized sensors used to gather participants' physiological data [70, 76, 88].
Interview	Structured or unstructured conversations between researchers and participants (e.g., interviews) [12, 96, 106].
Survey	Standardized sets of questions administered to participants [39, 44, 117].
Participant Produced Artifacts	Materials produced by research participants using various mediums. This can include things like physical objects created for an engineering task or written responses to formative assessment questions [9, 21, 107].
Researcher Produced Artifacts	Materials produced by the researchers that contribute to analysis and findings. This can include things like researchers' observational notes [70, 94, 139].
Motion	Raw motion data collected by different devices/technologies [52, 96, 155].
Text	Raw textual input [159].

Table 1. Definitions for data collection mediums.

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 [64, 65]. Similarly, reinforcement learning techniques are being developed for educational recommendations (e.g., Liu et al. [88]). Evaluating these methods is essential for understanding current 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.

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. [23], includes early, late, and hybrid fusion. *Early fusion* merges raw data from different sources at the initial processing stage, capturing

Modality	Description	Modality Group
Affect	The emotional or affective state exhibited by a participant [49, 121, 143].	NLP, Vision, Sensor
Pose	The physical position, geospacial location, or body posture of a participant [5, 137, 140].	Vision, Sensor
Gesture	The gestures and body language displayed by a participant [6, 116, 158].	Vision
Action	The observable actions or activities performed by a participant [63, 87, 120].	Vision, Sensor
Prosodic Speech	Prosodic speech information that includes volume, pauses, and intonation [105, 136, 138].	NLP
Transcribed Speech	Textual speech obtained through transcriptions, i.e., speech-to-text [12, 39, 86].	NLP
Qualitative Observations	Qualitative researcher observations on the participant's performance and behaviors, and the state of the environment [76, 93, 157].	Human-centered
Logs	Data related to the research context and participant's actions within the system [11, 66, 99].	Sensor
Gaze	Data on the direction and focus of the participant's eye gaze [55, 56, 162].	Vision, Sensor
Interview	Notes from interviews between researchers and participants [10, 54, 75].	Human-centered
Survey	Responses to surveys/questionnaires provided by the participant [113, 115, 117].	Human-centered
Pulse	The participant's pulse, indicating their heart rate [82, 83, 148].	Sensor
EDA	Participant's electrodermal activity, measuring skin conductance as an indicator of arousal [81, 92, 132].	Sensor
Temperature	Participant's body temperature [84, 113, 132].	Sensor
Blood Pressure	Participant's blood pressure [83, 113, 148].	Sensor
EEG	Participant's electroencephalography activity, recording the brain's electrical signals [66, 113, 132].	Sensor
Fatigue	The level of fatigue experienced by the participant during the activity [82, 83].	Vision, Sensor
EMG	Participant's electromyography activity, measuring muscles' electrical signals [50, 52].	Sensor
Participant Produced Artifacts	Artifacts produced by the participant [22, 100, 107].	Human-centered
Researcher Produced Artifacts	Artifacts produced by the researcher [28, 58, 106].	Human-centered
Spectrogram	Visual representation of audio frequencies in the form of a spectrogram [91].	NLP
Text	Raw text data generated by the participant in the study environment [159].	NLP
Pixel	Raw RGB pixel values representing visual information captured by cameras or sensors [120].	Vision

Table 2. Definition of modalities and references to papers where they were discussed.

Method	Definition
Classification	A supervised learning task in machine learning where the goal is to assign pre-defined labels or categories to input data based on feature analysis. It involves training a model on a labeled dataset to learn the mapping between input features and corresponding output labels, enabling the model to make predictions on new, unseen data [5, 121, 138].
Regression	A supervised learning task for predicting continuous numerical values. In contrast to classification, where the output is categorical, regression involves training a model to understand the relationships between input features and continuous target values, allowing for the prediction of numerical outcomes [49, 118, 136].
Clustering	An unsupervised learning technique that involves grouping similar data points based on their inherent patterns or similarities. Unlike classification, clustering does not rely on predefined labels, and the goal is to discover hidden structures within the data, grouping similar observations into clusters [6, 20, 28].
Qualitative	Qualitative analysis involves the examination and interpretation of non-numerical data, often in the form of textual or visual information. It aims to uncover patterns, themes, or meanings within the data that is context-dependent [75, 76, 93].
Statistical	Statistical methods encompass a broad range of techniques used to analyze data and draw meaningful conclusions. This includes descriptive statistics (such as mean, median, and standard deviation) and inferential statistics (e.g., via hypothesis testing, ANOVA, and correlation analysis), which help researchers make inferences about generalizability based on sample data [86, 90, 107].
Network analysis	The study of relationships and interactions among entities in a network. It includes graph-based approaches, where nodes represent entities (e.g., individuals, concepts) and edges represent relationships or connections between them [24, 39, 105].
Pattern Extraction	Refers to the process of identifying meaningful patterns or structures within data. This can include various techniques such as Markov analysis and sequence mining, where patterns in the temporal order of events are discovered, or other methods aimed at uncovering hidden structures that provide insights into the underlying dynamics of the data [103, 113, 144].

Table 3. Analysis methods and their definitions.

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. This has led us to introduce a new category, *mid fusion*, which involves moderately processed data

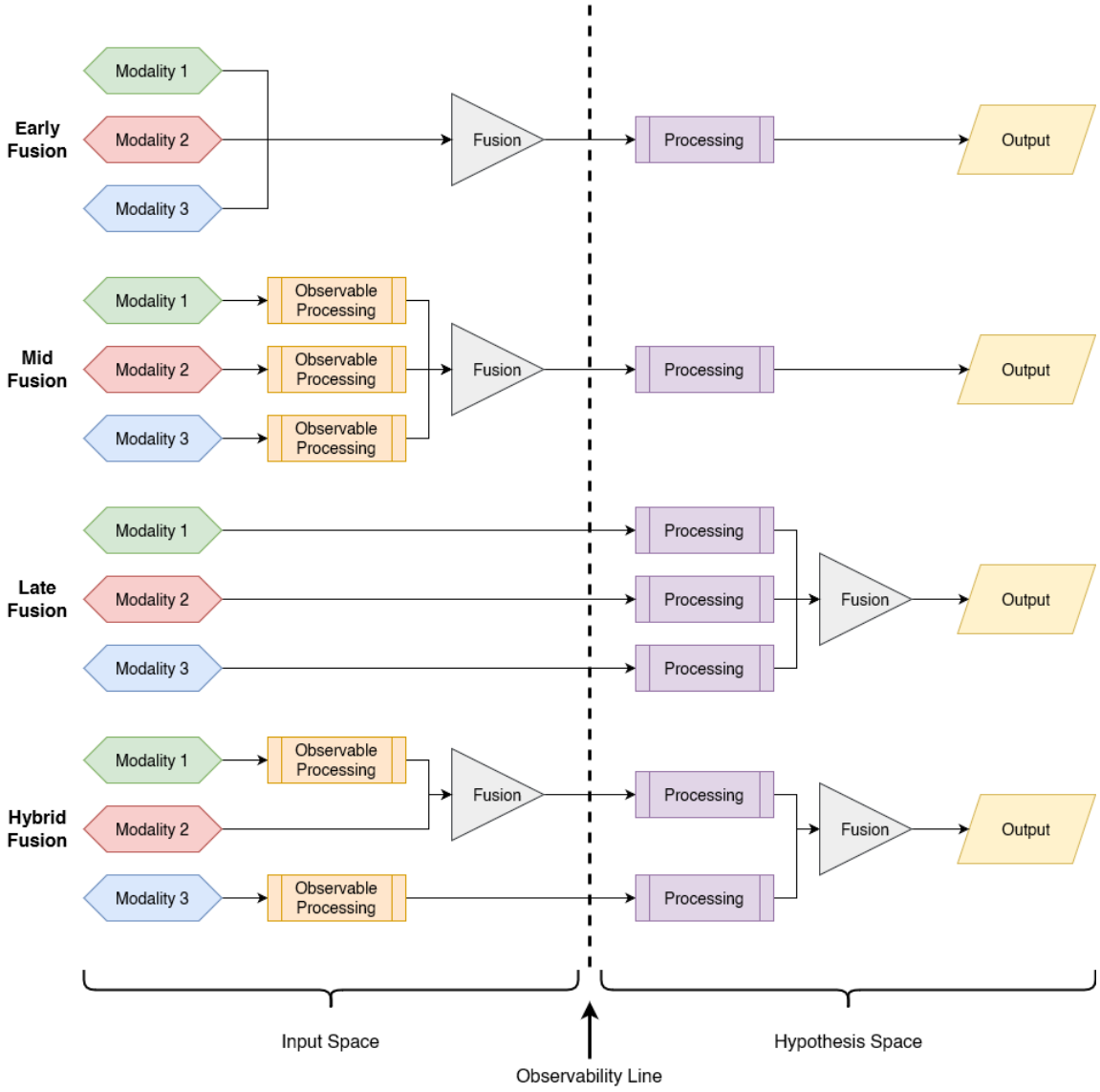


Fig. 3. Multimodal data fusion scheme according to when fusion is performed relative to the observability line.

integration, as conceptualized by Di Mitri et al. [51] using the observability line. To elaborate, Di Mitri et al. state "*The distinction between observable/unobservable is conceptual and can vary in practice.*" [51]. Here, *early fusion* combines unprocessed, 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 [23], 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.

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. [81, 142, 158].
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 [44, 55, 56].
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 [108, 118, 121].
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, 120].
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 [75, 76, 93].

Table 4. Categories of data fusion approaches.

2.2.6 Environment Setting. *Environment setting* refers to the overarching context in which learning and training activities take place. Analyzing the contextual settings in which these studies happen, we categorize these environments based on the nature of the setting, as shown in Table 5.

Setting	Definition
Virtual	Learning or training activities happen entirely within a virtual space, such as when participants engage with content and exercises in a computer-based system [5, 138, 143].
Physical	Activities unfold in a physical real-world environment. This encompasses traditional classroom settings and other tangible spaces where learning and training occur without reliance on digital technologies [116, 136, 158].
Blended	Learning or training activities blend elements of both the virtual and physical realms. Such environments are characterized by augmented reality, mixed reality, or a combination of digital and physical components [6, 49, 121].
Unspecified	Designated when a paper does not provide sufficient information about the learning or training environment [44, 88].

Table 5. Definitions for environment settings.

The reasons for considering environment setting in our analysis are twofold. First, we aim to unveil the contextual relevance of multimodal learning and training, discerning how these approaches manifest in computer-based (virtual) spaces, traditional classrooms (and other physical environments), and blended scenarios combining virtual and physical elements. Additionally, our categorization acknowledges instances where sufficient information about the environment setting is not provided, and this directs our attention to research gaps and unexplored areas within the current literature.

2.2.7 Domain of Study. While conducting our analysis, we recognized the importance of identifying the subject matter domain that study participants engage in during their learning and training experiences. This categorization helps us better contextualize the use of multimodal analytics to study learning performance and behaviors, as we can dynamically explore how these methods apply across a diverse array of domains. To circumscribe our literature review, we identified five subject areas that are defined in Table 6. Our categories are 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.

Subject	**Definition**
STEM+C	Participants engaged in learning or training related to Science, Technology, Engineering, Mathematics, and Computing disciplines. In our case, this also includes applied scientific disciplines such as healthcare and medicine [6, 136, 138].
Humanities	Participant activities focused on disciplines such as literature, debate, oral presentation, and other humanities-related fields [116, 121, 143].
Psychomotor Skills (PSY)	Learning or training activities emphasizing the development of motor skills and coordination [52, 66, 99].
Other	Subjects that do not fall into the above categories. This could include a wide range of topics beyond STEM, humanities, or psychomotor skills [100, 139].
Unspecified	Designated when the papers analyzed did not provide sufficient information about the subject matter to be conclusively categorized [9, 19, 88].

Table 6. Definitions of domains of study.

Importantly, papers reporting results from multiple studies are categorized with labels corresponding to the domain of each separate study [139, 155].

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 [19, 22, 83] or Multi-Person, which often emphasized collaborative or group dynamics [120, 124, 157]. It is noteworthy that some papers analyzed both the individual learner and groups of learners, reflecting the diversity in studies even within individual publications [9, 20].

2.2.9 Didactic Nature. Didactic nature 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. Our categorization scheme resulted in four categories that are summarized in Table 7.

2.2.10 Level of Instruction or Training. We sought to delineate the level of instruction or training for participants, providing valuable insights into the educational contexts targeted by the analyses in our corpus. The categories defined in Table 8 capture the diverse training and educational levels represented in our literature review.

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, reflecting the diversity inherent to the set of papers in our corpus.

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

Category	Description
Formal	Studies where the instruction is delivered in traditional classroom settings, online courses, or any structured learning environment with clear instructional objectives [20, 39, 76].
Informal	Informal environments involve learning or training that occurs in a more unstructured environment and may not include set learning goals and curricula [28, 55, 113]. For example, Worsley et al. [159] aimed to bridge multimodal interaction in the game Minecraft to support diverse learners in non-traditional learning environments.
Training	Studies that emphasize skill development, practical training, and professional development. This involves scenarios where participants are acquiring specific skills or competencies relevant to their profession or field [58, 96, 100].
Unspecified	This category implies that the papers reviewed did not provide sufficient information about the didactic nature of the study. It acknowledges instances where the training or educational context is not clearly defined or detailed in the literature [49].

Table 7. Categories of didactic nature.

Category	Description
K-12	Participants at the primary or secondary education levels, encompassing kindergarten through 12^th grade [56, 111, 155].
University	Participants engaged in university-level education, including undergraduate and graduate students [39, 66, 99].
Professional Development	Participants involved in learning or training experiences specifically designed for professional development and not tied to a K-12 or university curriculum [50, 52, 120].
Unspecified	Designated when papers did not provide sufficient information about the participants' level of instruction or training [20, 44, 93].

Table 8. Levels of instruction or training.

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 [78]. 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 [129], 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.

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.

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 [78], 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 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.

Algorithm 1 Citation Graph Pruning Algorithm

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

procedure DEGREE TRIMMING(G, n)
  $S, D \leftarrow \{\}, \{\}$
  for all $v \in V$ **do**
    if $\deg(v) <= n$ **then** $S = S \cup \{v\}$
  for all $v \in S$ **do**
    for all $e \in E$ **do**
      if $v \in e \wedge e \notin D$ **then** $D = D \cup \{e\}$
  return $(V \setminus S, E \setminus D)$
procedure SUBCONNECTED GRAPH TRIMMING(G)
  $[S_1, S_2, S_3, ..., S_n]$ = ConnectedComponent(G), where each $S_i = (V_i, E_i)$
  $j = \arg\max\{|V_1|, |V_2|, |V_3|, ..., |V_n|\}$
  return (V_j, E_j)
procedure ITERATIVE TRIMMING(G)
  while True **do**
    G' = 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'

3.3 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, 138, 143].

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%) [39, 55, 75] relative to training environments papers (16/73; 22%) [52, 54, 96]. This imbalance favoring learning environments 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* [141] 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%) [21, 58], with humanities (11/73; 15%) [108, 116] and psychomotor skills (5/73; 7%) [66, 99] 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%) [10, 139], compared to multi-person environments that are present in 31 (42%) papers [54, 111]. 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%) [87, 92, 144], followed by training (15/73; 21%) [54, 94, 107] and informal learning

(12/73; 16%) [28, 66, 159]. 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%) [22, 106], followed closely by K-12 environments (30/73; 41%) [83, 102]. Professional-level learning is less frequent (5/73; 7%) [52, 106]. 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%) [118, 157], followed by audio (37/73; 51%) [24, 154]. These two mediums indicate a preference for rich multimedia data that can capture the complexities of learning and training, as well as interactions within the environments. Logs (33/73; 45%) [66, 111] and participant-produced artifacts (30/73; 41%) [9, 81] are also popular, suggesting a strong inclination toward capturing learner behaviors and outputs directly from both the environments and the participants themselves.

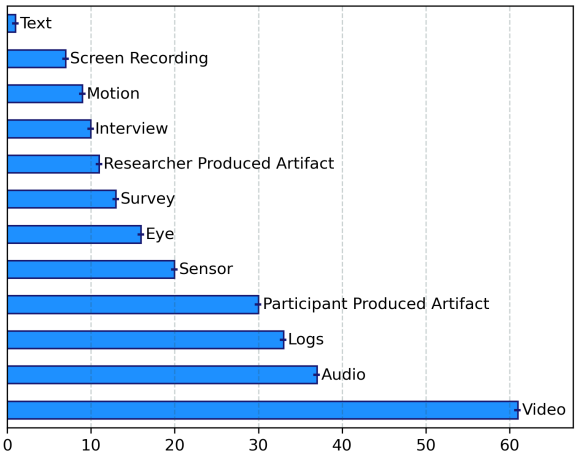


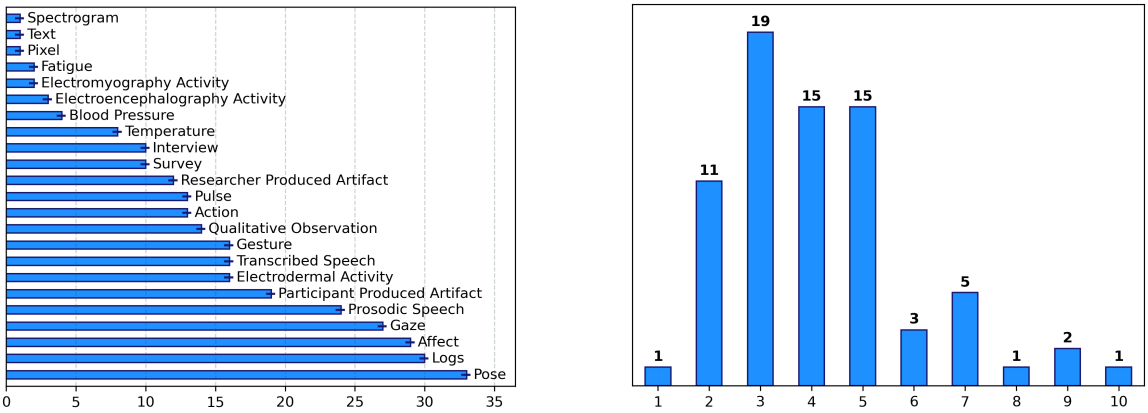
Fig. 4. Data collection mediums distribution.

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%) [88, 148] 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 [159] in the corpus), as learning and training environment research currently relies primarily on transcribed speech.

4.2 Multimodal Data

We present the breakdown of the different modalities used in our corpus in Figure 5.

"Pose" is the most prevalent modality, appearing in some form in roughly 45% of papers (33/73) [52, 94, 108]. 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



(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).

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 2-5 modalities due to that range offering 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 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 [91, 116, 137] and logistic regression models [44, 86, 115] were often used with natural language features. Other approaches like random forest [102], linear regression [136], and naive Bayes [137] 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

[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 [28] 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, which we still consider to be multimodal by our definition.

recurrent neural networks (e.g., LSTM models [71]), these were a relative rarity. Very few used transformer [150] 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 [149]. 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 [29, 31, 86, 87, 115]. These issues complicate the effective training of deep learning models [10, 30, 33, 115, 139]. Additional challenges include the complexity and time cost of cleaning, processing, and labeling data. Software packages like NLTK [89], openSMILE [57], and TAACO [41, 42] facilitate the programmatic extraction of audio- and text-based features, yet this can result in large, opaque feature sets [120]. Conversely, manual preprocessing and feature engineering can be time-intensive, potentially limiting the data researchers are willing to collect and analyze [80, 86]. 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 [54, 93, 108]. Many natural language studies focus on collaborative learning and training [90, 111, 136], 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 [80, 136, 159]. Learning environments often consist of multiple groups participating simultaneously, creating noisy conditions that hinder ASR accuracy, particularly in K-12 settings [80] and among non-English speakers [159].

Only one paper in the corpus used raw text as input [159], which is surprising given the prevalence of text-based transformer models [18, 48, 123]. 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 [44, 87, 143], these agents typically 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 [136, 137, 158]. In these collaborative settings, natural language features frequently were the most informative among all modalities [86].

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 [131]. Researchers often reported that including natural language features in the multimodal pipeline led to improved predictive performance [111]. 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%) within sensor-based learning and training research, covering a spectrum of modalities designed to capture and analyze various physiological and behavioral data. These papers were distributed across different modalities as follows: 11 papers focused on affective responses, 7 on body pose analysis, and 16 on electrodermal activity (EDA). Pulse rate was studied in 11 papers, activity in 5 papers, blood pressure (BP) in 4 papers, and temperature in 8 papers. Additionally, 3 papers dealt with electroencephalography (EEG), 2 with electromyography (EMG), another 2 with fatigue, and 8 papers focused on gaze tracking.

Of these 20 papers, 12 are learning-based, and 8 are training-based. This suggests sensors are more frequently employed for training-based research, which represents only 22% (16/73) of the full corpus but accounts for 40% (8/20) of papers employing sensors. Within MMLA, wearable sensors have been pivotal in monitoring learners' emotional and physiological states, predicting behavior and performance, providing real-time feedback, and enabling multimodal data integration [54, 70, 148]. The use of sensors in our corpus is diverse, extending from classroom environments to specialized training scenarios (e.g., CPR instruction [52]), where they serve as assessment tools and mechanisms for formulating educational interventions in real-time. However, integrating and interpreting data from sensors presents substantial challenges, particularly in terms of accurate and practical use with real-time applications [50, 132]. Ensuring sensor data maps to specific learning and training contexts, and managing the technical complexities of processing sensor data in multimodal data streams, also remain significant hurdles.

The state-of-the-art for sensor-driven multimodal learning and training analytics is characterized by advanced predictive modeling, real-time feedback systems, and multimodal data fusion. However, there is a pressing need for more granular data analysis to identify subtle patterns and correlations, particularly those not apparent through traditional analysis methods. Contextual and behavioral analytics are also often required to link physiological responses to specific learning activities in real-time. For example, signal processing methods aggregate sensory information into physical or learning characteristics such as relative learning gains [148], team dynamics [54], and shared physiological arousal

[103]. The field also requires robust, interactive visualizations that intuitively convey complex sensory data sets. Visual correlations with learning events would enable more profound pedagogical insights, revealing patterns in physiological responses and their triggers. Further, the use of Explainable AI (XAI) methods, particularly attribution-based methods, needs more research. These methods could clarify how different types of sensor data contribute to predictive models, enhancing educational and training research interpretability [126, 127].

Another gap in the literature is that research in sensor-based educational technologies often focuses on immediate or short-term effects; more longitudinal studies are needed to assess sustained impacts. Additionally, expanding sensor research to encompass diverse learning contexts and demographic groups will help us better understand the broader applications of these technologies. Sensor research also often occurs in controlled environments (e.g., using manikins), so scaling these technologies for widespread use and ensuring their generalizability across diverse educational settings remains challenging. One example is a study by Echeverria et al. [54] that applies accelerometer data to real-world nurse training simulations. However, this research is characterized by a narrow data scope concentrating solely on accelerometer metrics. To fully understand the learning process through movement and orientation during a training simulation, integrating additional sensory inputs such as gyroscope and magnetometer data may be beneficial. Such expansion would allow for a multidimensional analysis of physical interactions, potentially unveiling deeper insights into the kinesthetic aspects of learning. Investigation into the user experience and acceptance of wearable technologies in education are also needed, particularly concerning comfort, usability, perceived effectiveness, and, most importantly, privacy.

4.2.4 Human-Centered. Out of 73 papers in our corpus, 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 and perspectives. Participant-produced artifacts are the most common (19/73, 26%), followed by qualitative observation (14/73, 19%), researcher-produced artifacts (12/73, 16%), and both interview notes and survey responses (10/73 each, 14%). Participant artifacts often include diverse materials and standardized tests [81], with pre- and post-tests being the most prevalent for calculating learning gains [55, 124, 157]. The considerable use of qualitative observations highlights the importance of insights derived from direct human interpretation of participant behaviors. Common combinations include qualitative observations and participant artifacts [86, 87, 157, 158], participant artifacts and researcher artifacts [39, 124, 137, 140], and interview notes and qualitative observations [10, 75, 106, 158]. The only exception is a study that applied clustering, NLP, and linear modeling to a small dataset of researcher artifacts detailing student behaviors [28].

Regarding analysis methods, a predominant strategy involves transforming human-centered modalities into quantifiable data for statistical analysis. López et al. performed statistical analysis using survey data [90]; Ochoa and Dominguez did the same with participant-produced artifacts [107]; and Bert et al. used both participant-produced artifacts and interview transcriptions [12]. This shows a preference for quantifiable insights from human-centered modalities. A substantial subset of papers (14/45) focuses on qualitative analysis, emphasizing rich, qualitative insights. Most papers adopt multiple analysis methods, with only 16/45 using one method exclusively. In contrast, 15/45 papers integrate two distinct methods, and 13/45 use three. Notably, Worsley and Blikstein [158] employ four analysis methods to identify correlations between multimodal data, experimental condition, design quality, and learning, using both human-annotated and automatically annotated data.

The human-centered approach offers insights into the participants' experiences, perceptions, and behaviors, often pinpointing subtle nuances that may be hard to derive from quantitative analyses. Qualitative observation can provide

rich contextual information, participant- and researcher-produced artifacts offer tangible artifacts, interview notes can capture in-depth discussions, and survey responses provide multiple participants' perspectives that collectively enrich the analyses.

While a human-centered approach to multimodal learning analytics yields valuable insights, it also poses several challenges related to subjectivity, scalability, resource intensiveness, and potential limitations in generalizability. Due to the inherent subjectivity of human-centered modalities, the analysis of this data may be susceptible to the influence of the researchers' perspectives, possibly introducing bias into the interpretation [98, 104]. Furthermore, these approaches often are resource-intensive, requiring trained researchers for data collection, coding, and analysis. Manual collection and human analysis can be time-consuming and may not scale well, especially in large-scale educational settings. Nevertheless, human-centered approaches may offer more transparent and interpretable insights than automated methods.

These insights underscore a significant gap in the seamless integration of qualitative and quantitative methods. Although human-centered modalities were frequently utilized for quantitative statistics and qualitative insights, more effective methodologies for combining qualitative interpretations with quantitative rigor are needed. Additionally, the lack of standardized coding practices for human-centered modalities (e.g., variability among raters) presents a substantial research gap, and the absence of common coding conventions across studies impedes the replicability and comparability of findings. Establishing standardized coding frameworks is essential to enhance the reliability and credibility of machine learning analyses, ensuring consistency in the interpretation and utilization of human-centered data. Finally, research is needed for automating human-coding processes.

4.2.5 Logs. Thirty papers (40%) from the corpus incorporated log data in their analysis (referred to as log-analysis papers). Logs have been used frequently in unimodal and multimodal learning analytics. Logs, typically from computer-based environments, provide context for linking complementary modalities to learning outcomes and behaviors. We observed that log data is often combined with multiple multimedia mediums, such as 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%). The diversity of data collection mediums highlights the many ways in which 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 for log-analysis features various approaches. Nearly all classification and regression papers used machine learning algorithms to predict students' achievement, engagement, or emotional state; including support vector machine, random forest, naive Bayes, and logistic regression [63, 92, 162]. These methods were popular for addressing smaller datasets and often implemented in ensembles. Only three papers used deep learning approaches, including CNNs [137] and LSTMs [99, 111]. Statistical methods were used to correlate learning variables (e.g., perceived student emotion) to outcome variables (e.g., learning gains and achievement).

Analyzing logs presents various hurdles, including time-cost, data scarcity, generalizability, and engineering expenses. Temporal aspects introduce difficulties in log analysis, such as aligning time frames, handling different sampling rates, and managing time-series data. These complexities make collecting multimodal data demanding, often resulting in smaller datasets that limit scope and scalability. Data scarcity is exacerbated by challenges in producing generalizable findings, as models and methods effective in one domain may not transfer to others. Additionally, high software development and engineering costs hinder integrating modern, innovative features like real-time collaboration tools in digital spaces.

These challenges create gaps in the log-analysis research. There is a noticeable deficiency in applying methods and findings from one educational setting to another within MMLA. This hesitancy is likely due to the diversity of educational environments and contexts. Embracing a standardized log format and consistent practices may overcome this barrier, leading to more unified research approaches and the broader applicability of insights and methodologies across different educational scenarios and domains.

Another aspect highlighting this challenge is the low adoption rate of established industry standards like xAPI [135], LTI [1], and Learning Management Systems (LMS) within educational technology research. This trend reflects a broader issue of aligning with best practices and norms established in the wider technology and education sectors. Embracing these standards could enhance interoperability, scalability of solutions, and more robust analysis of educational data. Addressing these gaps and embracing standardization could pave 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.

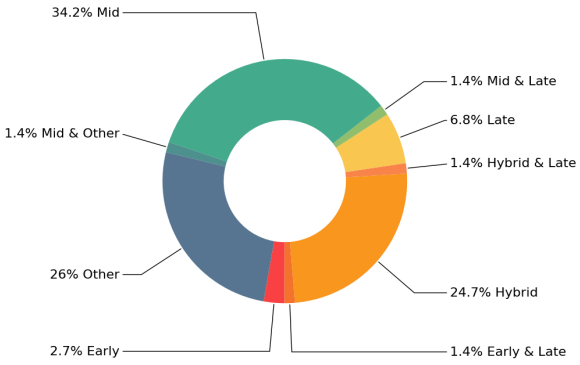


Fig. 6. Distribution of Fusion Types

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 fusion is observed only in 3 papers Late fusion, occurring after separate models are trained, is employed in 8 papers (11%). 20 papers (27%) adopt other types of fusion strategies, no fusion, or do not explicitly mention data fusion.

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 [21, 22, 142]. Except for 1 paper that used regression [118], 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.

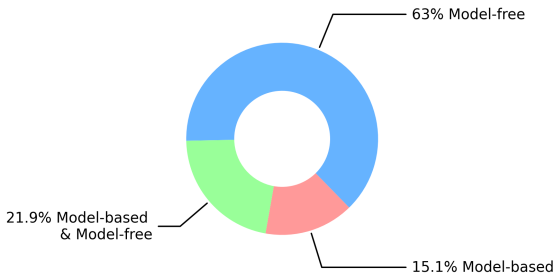


Fig. 7. Analysis approaches percentage distribution.

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, 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.

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 methods 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 [69] 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 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 presented in Section 4, we re-examined our corpus to classify the prevailing research objectives in the application of multimodal methods to learning and training environments. Following discussions, the authors achieved consensus on three primary research objectives that we term as *archetypes*: *Designing and Developing Methods*, *Analyzing Outcomes*, and *Exploring Behaviors*. These archetypes, detailed in subsequent subsections, often overlap within studies; for instance, research on method development may also yield insights into participant behaviors and outcomes. While these archetypes broadly define the field, they are not exhaustive, as certain 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 [55, 99, 132] and regression [56, 111, 118], and their efficacy is reported through performance metrics like F1-score [5, 11, 162]. Data collection often involves video, audio, and log data [86, 90, 139], targeting modalities such as affect, pose, prosodic speech, and logs [121, 139, 143], with data fusion techniques like mid or hybrid fusion being common [19, 49, 96]. The research predominantly employs model-based approaches [56, 99, 118].

Our corpus reveals a broad spectrum of tasks addressed by *Designing and Developing Methods* research, ranging from personalized feedback in CPR training [99] to engagement detection in educational games [121], and skill classification in sports [96]. 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 [50, 52, 108] and engagement detection [5, 19, 121] 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 [15]. This aspect will be further discussed in Section 6. Although some studies in our corpus do consider stakeholder impact [10, 108, 143], such instances are infrequent and not representative of the corpus as a whole.

5.2 Analyzing Outcomes

The Analyzing Outcomes archetype encompasses studies that aim to identify specific outcome metrics within learning and training environments. These metrics often include learning or training gains, engagement levels, error and accuracy rates, and other similar measures. The objective is to uncover findings that apply to broader populations beyond the sample studied. This focus on the implications of data for larger learner and trainee populations distinguishes the Analyzing Outcomes archetype from the Designing and Developing Methods archetype, which concentrates on refining specific analytical techniques. Outcome analysis typically employs supervised learning methods such as classification [20, 84, 99] and regression [56, 118, 143], with additional insights gained through the examination of model behaviors and the integration of statistical patterns and unsupervised methods [66, 86, 132].

Outcome analysis has been applied across various learning and training contexts, often focusing on constructs like attention and engagement [10, 56, 142], task performance and accuracy [11, 44, 96], learning outcomes [22, 50, 148], and collaborative outcomes [91, 137, 157]. Despite the diversity of environments, the commonality of outcome variables studied allows for the generation of insights with a degree of generalizability.

However, this archetype's research is not without limitations. By concentrating on outcome variables, these studies may overlook the complexities of learning and training processes and their impact on the final outcomes. Consequently, educational technologies and interventions risk being tailored to high-performing learners, potentially neglecting the varied learning behaviors and individual differences among students [59]. Moreover, similar to the limitations noted in the *Designing and Developing Methods* archetype, *Analyzing Outcomes* studies often do not incorporate stakeholder perspectives into their analyses, which could lead to biased assumptions and conclusions. The importance of including stakeholder feedback in research is further discussed in Section 6.

5.3 Exploring Behaviors

The Exploring Behaviors archetype is characterized by studies that delve into human behavior and experiences within learning and training contexts, employing an exploratory approach to uncover the factors influencing these experiences. This research examines a variety of human signals that vary temporally, socially, and spatially, tailored to the specific objectives and contexts of the learning or training experience. Unlike the other archetypes, Exploring Behaviors research often incorporates qualitative observations [28, 75, 82, 83], and employs data exploration techniques such as correlation analysis [90, 105] and pattern recognition [6, 39, 103, 124]. There is also a higher concentration of human-centered modalities, such as researchers' and participants' artifacts, surveys, and interviews. This is largely due to the nature of behavior-focused research.

Data fusion in this context is typically qualitative [12, 76, 159], 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. [124] 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. [105] utilized correlation analysis alongside social network metrics and annotated behaviors to distinguish collaborative dynamics in a software engineering course. Closser et al. [28] 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.

Despite its strengths, Exploring Behaviors research faces challenges. The context-dependent nature of the findings may restrict their broader applicability. The qualitative methods used for behavior analysis also introduce the possibility of researcher bias, which can influence the interpretation of data and compromise the neutrality of the findings.

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 roughly 7:2, with the majority focusing on STEM instruction incorporating at least one virtual component (i.e., virtual or blended environment settings).
- *Participants.* Study participants tend to be either university or K-12 students learning or training in multi-person environments slightly more often than individual environments (3:2).
- *Data and Modalities.* Video, audio, environment logs, and participant-produced artifacts are the most common data collection mediums; and pose, logs, affect, gaze, and prosodic speech are the most popular modalities. The vast majority of the papers in our corpus incorporate 2-5 modalities for their analyses suggesting that it is more useful to focus on a few, meaningful modalities rather than several (potentially uninformative) modalities. A large number of papers performed at least one type of vision analysis (i.e., used features derived from video or gaze). A majority also incorporated human-centered modalities (e.g., participant- or researcher-produced

artifacts, surveys, and interviews), and nearly all of these papers incorporated more than one human-centered modality.

- *Analysis Methods and Approaches.* Classification, statistical analysis, and qualitative analysis were the most common. Classification was typically used for predicting an outcome (dependent) variable like learning or training gains, while statistical methods were often used to select features and correlate them with outcome variables. Qualitative analyses typically involve case study observations, qualitative coding, and thematic analyses. Model-based papers outnumbered model-free ones by about 3:1, suggesting that the community prefers to develop models to help explain learning and training processes.
- *Data Fusion.* Roughly, 75% of the papers performed early, mid, late, or hybrid fusion. Mid fusion was the most prevalent, followed by hybrid fusion. While not always the case, researchers often stated that fused modalities yielded results superior to unimodal or unfused features, so researchers should consider exploring data fusion while analyzing multimodal learning and training environments. Leveraging data fusion can provide a more holistic understanding of learner and trainee behaviors and outcomes, allowing researchers to develop more effective, personalized interventions and supports.
- *Publication Mediums.* The *British Journal of Educational Technology* (BJET) and *International Conference on Learning Analytics & Knowledge* (LAK) were by far the two most popular publication venues represented in our corpus.

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 [87, 137, 138]. 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 [154]. 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. [80]

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. [91] 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 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 [158], 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. [87] 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" [86].

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" [80]. 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. [106]
>
> The limited number of pair work EEs does not allow us to make any strong claims in terms of the framework's reliability. [100]
>
> ...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. [24]
>
> ...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. [80]

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 [112]. Resource and access limitations, along with privacy concerns, hinder the application of advanced AI methods in learning and training environments. Similarly, conversational agents are underrepresented, with few papers discussing agents and none employing interactive, multi-turn agents (one paper [143] 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 [34, 37] 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* [128] and *co-design* [114] 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 [112] and Gemini [145] 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 [35]. Smaller, open-source models can also be trained via parameter-efficient methods to ease the computational overhead endemic to large transformer models [47]. 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 [163] and diagram generation via LLM planning [164]. 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, 133, 134]. Given the recent proliferation of multimodal LLMs in core AI research, we expect to see LLM integration with multimodal learning and training environments increasingly in the coming months and years.

6.3.2 Data Scarcity Mitigation. Data scarcity is a significant issue plaguing the field, causing multimodal learning and training methods to lag behind state-of-the-art core AI approaches. Compiling massively open learning and training corpora could help mitigate this, but several challenges need resolution. First, collecting multimodal data for large-scale studies is more challenging than for unimodal studies. Sharma et al. found a significant negative correlation between the number of modalities analyzed and the sample size of MMLA studies, indicating difficulties in collecting, processing, and analyzing multimodal data [131]. Second, there are significant ethical limitations when collecting and publishing standardized datasets for education and training, particularly concerning privacy and surveillance [43]. These ethical concerns are amplified with educational datasets involving children and the publication of raw datasets containing identifiable data. One approach to overcome these limitations is to focus on designing generalizable methods requiring limited data, such as zero and few-shot learning approaches, which have seen significant development in core-AI domains [77].

6.3.3 Standardization. Blanco et al. [46] underscore the necessity for uniform coding standards and methodologies, highlighting a significant void in the literature concerning multimodal, temporal, and human-focused data. Prevailing e-learning norms such as xAPI [135] and LTI [1] have been incorporated within educational technology platforms like Canvas, Moodle, BlackBoard, and Udemy, yet these platforms predominantly utilize unimodal data and/or are restricted by proprietary licensure. Adapting these technical frameworks to accommodate and preserve multimodal information represents a colossal challenge, leading to minimal use in multimodal learning and training research.

Multimodal learning and training research merges AI, multimodal data, and authentic educational contexts, with researchers advocating for novel software fulfilling these criteria. This has led to a disparate landscape, with distinct teams gathering, examining, and devising new approaches for their specific settings [161]. Formulating uniform standards specific to multimodal learning and training research is crucial for enhancing the dependability and trustworthiness of machine learning techniques in the field, and for improving the analysis of human-centric data [45]. Adopting a unified log format and organizational structure for multimodal log data could mitigate the field's dependence on context-specific methodologies and bolster generalizability. It is also paramount that researchers and engineers proactively 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 [7]. Interaction analysis is common but challenging due to human analysts' cognitive limits and the fast-changing nature of embodied contexts [165]. 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 [60].

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 [160]. While some physical training works in our corpus leveraged multimodality [52, 96], 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 whose outputs lack explainability. This is problematic, as researchers and stakeholders (e.g., teachers, trainers, and study participants) often cannot understand the model's decision-making processes. This can hinder teachers' and trainers' ability to guide their students and foment distrust in AI algorithms and systems. Prior work has sought to create more explainable systems via data visualization tools that help make participants' learning and training processes more transparent [74, 152], including some of the papers in our corpus [94, 106]. More research is needed. We recognize the potential for LLMs to enhance explainability (*Chain-of-Thought* prompting [35, 156] elicits reasoning chains from the LLM alongside the model's assessment scoring decisions, for example). Recently, researchers gathered feedback from multiple stakeholders (teachers and students) about using LLMs in the classroom. Both parties indicated that they saw the potential for LLMs to improve learning outcomes, but that explainability is necessary to get teachers and students to "buy in" [34].

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 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 [16], but less frequently within multimodal studies. The challenges of scalability and standardization of multimodal logs have restricted longitudinal MMLA research [161], 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* [8], 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) [74] 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 the research methods currently applied to multimodal learning and training environments. In doing so, we developed a novel, programmatic approach to distilling our literature review corpus, *citation graph pruning*. We presented a taxonomy and framework reflecting the current advances in multimodal learning and training research, and identified and analyzed five modality groups of multimodal data (Natural Language, Vision, Sensors, Human-Centered, and Logs). We presented descriptive statistics, performed a qualitative thematic analysis, and discussed our findings for the current state-of-the-art, results achieved, challenges faced, and research gaps for modality groups and the corpus as a whole. We used this information to derive three archetypes that best characterize current multimodal learning and training research. In addition, we identified the need for a new type of data fusion, *mid fusion*, which is characterized by fusing derived, observable features. We concluded by offering several promising avenues for further research exploration and presented the limitations of our work. As the field of multimodal learning and training analytics continues to expand with the advent of generative AI, we believe that this literature review will serve as a springboard for new multimodal learning and training 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].

[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] 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.* N/A, N/A, N/A.

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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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.

[14] 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.

[15] 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.

[16] 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

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

[18] 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.

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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.

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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..

[35] Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2023. A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. (2023). Accepted to The 14th Symposium on Educational Advances in Artificial Intelligence. In press..

[36] Clayton Cohn, Caitlin Snyder, and Gautam Biswas. 2024. A Human-in-the-Loop LLM Approach to Collaborative Discourse Analysis. (February 2024). Submitted to the 25th International Conference on Artificial Intelligence in Education. Currently under review..

[37] 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 Brtish Journal of Educational Technology special section Hybrid Intelligence: Human-AI Co-evolution and Learning. Currently under review..

[38] 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.

[39] 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

[40] 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

[41] 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.

[42] 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.

[43] 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.

[44] 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

[45] 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

[46] 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

[47] 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]

[48] 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.

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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.

[58] 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

[59] Joyce Fonteles, Celestine Akpanoko, Pamela Wisniewski, and Gautam Biswas. 2024. Promoting Equitable Learning Outcomes for Underserved Students in Open-Ended Learning Environments. (2024). https://doi.org/10.1145/3628516.3655753 Accepted to the 23rd Annual ACM Interaction

Design and Children Conference. In press..

[60] Joyce Horn Fonteles, Eduardo Davalos Anaya, Ashwin T. S., Yike Zhang, Mengxi Zhou, Efrat Ayalon, Alicia C. Lane, Selena Steinberg, Joshua Danish, Noel Enyedy, and Gautam Biswas. N.D.. A First Step to Using Machine Learning Methods to Enhance Interaction Analysis in Embodied Learning Environments. (N.D.). Under Review.

[61] 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].

[62] 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].

[63] 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.

[64] 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.

[65] 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

[66] 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

[67] 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.

[68] 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

[69] 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

[70] 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.

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

[72] 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.

[73] 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

[74] 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.

[75] 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

[76] 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

[77] 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.

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

[79] 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.

[80] 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

[81] 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

[82] 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

[83] 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

[84] 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

[85] 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.

[86] 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

[87] 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

[88] 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

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

[90] 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.

[91] 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

[92] 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

[93] 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

[94] 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

[95] 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.

[96] 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

[97] 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].

[98] 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

[99] 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.

[100] 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

[101] 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

[102] 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

[103] 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

[104] 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

[105] 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

[106] 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

[107] 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

[108] 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

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

[110] 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].

[111] 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

[112] 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]

[113] 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

[114] 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

[115] 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

[116] 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

[117] 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

[118] 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

[119] 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

[120] 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

[121] 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

[122] 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

[123] 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.

[124] 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.

[125] María Jesüs 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.

[126] 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.

[127] 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.

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

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

[130] 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

[131] 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.

[132] 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

[133] Caitlin Snyder, Nicole Hutchins, Clayton Cohn, Joyce Fonteles, and Gautam Biswas. 2023. Analyzing Students Collaborative Problem-Solving Behaviors in Synergistic STEM+C Learning. (2023). Accepted to the 14th International Learning Analytics and Knowledge Conference..

[134] 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..

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

[136] 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.

[137] 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

[138] 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

[139] 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

[140] 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.

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

[142] 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

[143] 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

[144] 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

[145] 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.

[146] 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

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

[148] 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

[149] 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.

[150] 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.

[151] 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.

[152] 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.

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

[154] 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

[155] 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

[156] 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]

[157] 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

[158] 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

[159] 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

[160] 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.

[161] 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

[162] 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

[163] 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.

[164] 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.

[165] Mengxi Zhou, Joyce Horn Fonteles, Joshua Danish, Eduardo Davalos, Selena Steinberg, Gautam Biswas, and Noel Enyedy. N.D.. Exploring Artificial Intelligence Supported Interaction Analysis. (N.D.). Under Review.

[166] 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

A CORPUS TABLE

Table 9 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 [9]	Ashwin	Impact Of Inquiry Interventions On Students In E-Learning And Classroom Environments Using Affective Computing Framework	2020	UMUAI
3448122334 [10]	Aslan	Investigating The Impact Of A Real-Time, Multimodal Student Engagement Analytics Technology In Authentic Classrooms	2019	CHI
1886134458 [11]	Azcona	Personalizing Computer Science Education By Leveraging Multimodal Learning Analytics	2018	FIE
3146393211 [12]	Birt	Mobile Mixed Reality For Experiential Learning And Simulation In Medical And Health Sciences Education	2018	Information
1326191931 [20]	Chan	Multimodal Learning Analytics In A Laboratory Classroom	2019	MLPALA
2936220551 [21]	Chango	Multi-Source And Multimodal Data Fusion For Predicting Academic Performance In Blended Learning University Courses	2020	CEE
4277812050 [22]	Chango	Improving Prediction Of Students' Performance In Intelligent Tutoring Systems Using Attribute Selection And Ensembles Of Different Multimodal Data Sources	2021	JCHE
1426267857 [24]	Chen	Affect, Support, And Personal Factors: Multimodal Causal Models Of One-On-One Coaching	2021	JEDM
3809293172 [28]	Closser	Blending Learning Analytics And Embodied Design To Model Students' Comprehension Of Measurement Using Their Actions, Speech, And Gestures	2021	IJCCI
4019205162 [39]	Cornide-Reyes	Introducing Low-Cost Sensors Into The Classroom Settings: Improving The Assessment In Agile Practices With Multimodal Learning Analytics	2019	Sensors
1576545447 [44]	Cukurova	Artificial Intelligence And Multimodal Data In The Service Of Human Decision-Making: A Case Study In Debate Tutoring	2019	BJET
1609706685 [49]	Di Mitri	Learning Pulse: A Machine Learning Approach For Predicting Performance In Self-Regulated Learning Using Multimodal Data	2017	LAK

2070224207 [99]	Di Mitri	Detecting Medical Simulation Errors With Machine Learning And Multimodal Data	2019	CAIM
3009548670 [52]	Di Mitri	Real-Time Multimodal Feedback With The Cpr Tutor	2020	AIED
1763513559 [50]	Di Mitri	Keep Me In The Loop: Real-Time Feedback With Multimodal Data	2021	IJAIED
1296637108 [54]	Echeverria	Towards Collaboration Translucence: Giving Meaning To Multimodal Group Data	2019	CHI
1581261659 [56]	Emerson	Early Prediction Of Visitor Engagement In Science Museums With Multimodal Learning Analytics	2020	ICMI
1598166515 [55]	Emerson	Multimodal Learning Analytics For Game-Based Learning	2020	BJET
4035649049 [58]	Fernández-Nieto	Storytelling With Learner Data: Guiding Student Reflection On Multimodal Team Data	2021	TLT
483140962 [63]	Fwa	Investigating Multimodal Affect Sensing In An Affective Tutoring System Using Unobtrusive Sensors	2018	PPIG
4278392816 [66]	Giannakos	Multimodal Data As A Means To Understand The Learning Experience	2019	IJIM
853680639 [70]	Henderson	Sensor-Based Data Fusion For Multimodal Affect Detection In Game-Based Learning Environments	2019	EDM
86191824 [75]	Jiang	Examining How Different Modes Mediate Adolescents' Interactions During Their Collaborative Multimodal Composing Processes	2019	ILE
3398902089 [76]	Järvelä	What Multimodal Data Can Tell Us About The Students' Regulation Of Their Learning Process?	2019	LAI
32184286 [80]	Kubsch	Once More With Feeling: Emotions In Multimodal Learning Analytics	2022	MMLA Handbook
205660768 [81]	Larmuseau	Multimodal Learning Analytics To Investigate Cognitive Load During Online Problem Solving	2020	BJET
1877483551 [84]	Lee-Cultura	Motion-Based Educational Games: Using Multi-Modal Data To Predict Player'S Performance	2020	COG
3660066725 [82]	Lee-Cultura	Children'S Play And Problem Solving In Motion-Based Educational Games: Synergies Between Human Annotations And Multi-Modal Data	2021	IDC
3856280479 [83]	Lee-Cultura	Children'S Play And Problem-Solving In Motion-Based Learning Technologies Using A Multi-Modal Mixed Methods Approach	2021	IJCCI
804659204 [88]	Liu	Towards Smart Educational Recommendations With Reinforcement Learning In Classroom	2018	TALE

3783339081 [87]	Liu	A Novel Method For The In-Depth Multimodal Analysis Of Student Learning Trajectories In Intelligent Tutoring Systems	2018	JLA
3796180663 [86]	Liu	Learning Linkages: Integrating Data Streams Of Multiple Modalities And Timescales	2018	JCAL
518268671 [90]	López	Using Multimodal Learning Analytics To Explore Collaboration In A Sustainability Co-Located Tabletop Game	2021	ECGBL
566043228 [19]	Ma	Automatic Student Engagement In Online Learning Environment Based On Neural Turing Machine	2021	IJIET
3754172825 [91]	Ma	Detecting Impasse During Collaborative Problem Solving With Multimodal Learning Analytics	2022	LAK
147203129 [92]	Mangaroska	Multimodal Learning Analytics To Inform Learning Design: Lessons Learned From Computing Education	2020	JLA
1847468084 [93]	Martin	Computationally Augmented Ethnography: Emotion Tracking And Learning In Museum Games	2019	ICQE
2879332689 [94]	Martinez-Maldonado	From Data To Insights: A Layered Storytelling Approach For Multimodal Learning Analytics	2020	CHI
2155422499 [100]	Morell	A Multimodal Analysis Of Pair Work Engagement Episodes: Implications For Emi Lecturer Training	2022	JEAP
2273914836 [102]	Nasir	Many Are The Ways To Learn Identifying Multi-Modal Behavioral Profiles Of Collaborative Learning In Constructivist Activities	2022	IJCSCL
1469065963 [103]	Nguyen	Examining Socially Shared Regulation And Shared Physiological Arousal Events With Multimodal Learning Analytics	2022	BJET
2345021698 [105]	Noël	Exploring Collaborative Writing Of User Stories With Multimodal Learning Analytics: A Case Study On A Software Engineering Course	2018	Access
2609260641 [106]	Noël	Visualizing Collaboration In Teamwork: A Multimodal Learning Analytics Platform For Non-Verbal Communication	2022	DAMLE
2497456347 [108]	Ochoa	The Rap System: Automatic Feedback Of Oral Presentation Skills Using Multimodal Analysis And Low-Cost Sensors	2018	LAK
2634033325 [107]	Ochoa	Controlled Evaluation Of A Multimodal System To Improve Oral Presentation Skills In A Real Learning Setting	2020	BJET

3051560548 [111]	Olsen	Temporal Analysis Of Multimodal Data To Predict Collaborative Learning Outcomes	2020	BJET
123412197 [113]	Papamitsiou	Utilizing Multimodal Data Through Fsqca To Explain Engagement In Adaptive Learning	2020	TLT
85990093 [116]	Petukhova	Multimodal Markers Of Persuasive Speech : Designing A Virtual Debate Coach	2017	INTERSPEECH
957160695 [115]	Petukhova	Virtual Debate Coach Design: Assessing Multimodal Argumentation Performance	2017	ICMI
1374035721 [117]	Pham	Attentivelearner2: A Multimodal Approach For Improving Mooc Learning On Mobile Devices	2017	AIED
2836996318 [118]	Pham	Predicting Learners' Emotions In Mobile Mooc Learning Via A Multimodal Intelligent Tutor	2018	ITS
3135645357 [120]	Prieto	Multimodal Teaching Analytics: Automated Extraction Of Orchestration Graphs From Wearable Sensor Data	2018	JCAL
3408664396 [121]	Psaltis	Multimodal Student Engagement Recognition In Prosocial Games	2017	T-CIAIG
3308658121 [124]	Reilly	Exploring Collaboration Using Motion Sensors And Multi-Modal Learning Analytics	2018	EDM
3625722965 [96]	Sanusi	Table Tennis Tutor: Forehand Strokes Classification Based On Multimodal Data And Neural Networks	2021	Sensors
2000036002 [132]	Sharma	Predicting Learners' Effortful Behaviour In Adaptive Assessment Using Multimodal Data	2020	LAK
1118315889 [136]	Spikol	Using Multimodal Learning Analytics To Identify Aspects Of Collaboration In Project-Based Learning	2017	CSCL
3339002981 [138]	Spikol	Estimation Of Success In Collaborative Learning Based On Multimodal Learning Analytics Features	2017	ICALT
1637690235 [137]	Spikol	Supervised Machine Learning In Multimodal Learning Analytics For Estimating Success In Project-Based Learning	2018	JCAL
3796643912 [139]	Standen	An Evaluation Of An Adaptive Learning System Based On Multimodal Affect Recognition For Learners With Intellectual Disabilities	2020	BJET
2181637610 [140]	Starr	Toward Using Multi-Modal Learning Analytics To Support And Measure Collaboration In Co-Located Dyads	2018	ICLS
1315379489 [142]	Sümer	Multimodal Engagement Analysis From Facial Videos In The Classroom	2021	TAC

3093310941 [143]	Tanaka	Embodied Conversational Agents For Multimodal Automated Social Skills Training In People With Autism Spectrum Disorders	2017	PLOS
1345598079 [144]	Tancredi	Intermodality In Multimodal Learning Analytics For Cognitive Theory Development: A Case From Embodied Design For Mathematics Learning	2022	MMLA Handbook
433919853 [148]	Tisza	Understanding Fun In Learning To Code: A Multi-Modal Data Approach	2022	IDC
1770989706 [154]	Vrzakova	Focused Or Stuck Together: Multimodal Patterns Reveal Triads' Performance In Collaborative Problem Solving	2020	LAK
2055153191 [155]	Vujovic	Round Or Rectangular Tables For Collaborative Problem Solving? A Multimodal Learning Analytics Study	2020	BJET
3095923626 [158]	Worsley	A Multimodal Analysis Of Making	2017	IJAIED
3309250332 [157]	Worsley	(Dis)Engagement Matters: Identifying Efficacious Learning Practices With Multimodal Learning Analytics	2018	LAK
666050348 [159]	Worsley	Multicraft: A Multimodal Interface For Supporting And Studying Learning In Minecraft	2021	HCII
1019093033 [162]	Yang	Prime: Block-Wise Missingness Handling For Multi-Modalities In Intelligent Tutoring Systems	2019	MMM

Table 9. 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 [26] and gscholar [153]. Ultimately, we employed SerpAPI [129], 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 10.[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 10. 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 11 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 11.

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 11. 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 11, 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 11, 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 [147], 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 [68] 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 11, 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 11, 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 12. 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 11, 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 11, Step ID 5.1), 74 nodes in the second iteration (Table 11, Step ID 5.2), 19 nodes in the third iteration (Table 11, Step ID 5.3), and 2 nodes in the fourth and final iteration (Table 11, 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).

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

Table 12. Disconnected DAG components by number of nodes in the component (size), frequency of occurrence (#), and total number of papers (papers).

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

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 11, 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" [67]; and keeping one titled, "supervised machine learning in multimodal learning analytics for estimating success in project-based learning" [137]. 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 [78], 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 11, 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 11, Step ID 8), several works were excluded due to their inaccessibility (Table 11, 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 11, 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 13. 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.

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 13. 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.

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.

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 11, 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 11, 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 14:

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 14. 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 11, 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 11, 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 15.[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 15. 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 [146], 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 16. 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 11, 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 16. 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 11, 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 15 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 11, 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 [32]). 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 [25] 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 11, 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 17 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 [17], 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 17. 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 2^{nd} 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 [129]. 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".

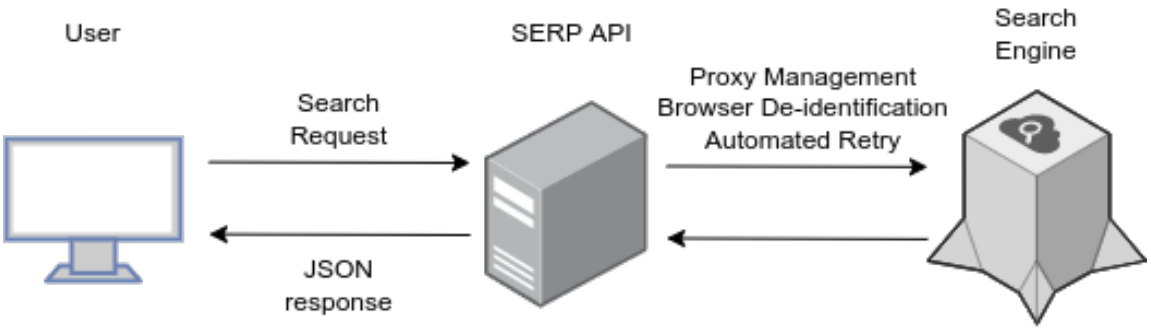


Fig. 8. Searching Google Scholar via SerpAPI.

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.

Received 00 Month 2023