Leveraging LLMs for Dynamic Engagement Pattern Recognition in Collaborative Learning

Stella Xin Yin[1(✉)], Zhengyuan Liu[2(✉)], Dion Hoe-Lian Goh[1], and Nancy F. Chen[2]

[1] Nanyang Technological University, Singapore, Singapore
xin013@e.ntu.edu.sg
[2] Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

Abstract. Engagement in collaborative contexts is difficult and challenging to observe and analyze. Current multimodal learning analytics (MMLA) focus on individual learners, prioritize visual signals over conversational dynamics, and reduce rich behavioral data to numerical predictions. In this paper, we address these gaps by introducing a MMLA framework that leverages large language models to dynamically and hierarchically recognize engagement patterns. Our approach moves beyond static scoring by integrating verbal dialogue and non-verbal behavioral cues, and identifies and represents evolving engagement trajectories in an interpretable, descriptive form. We demonstrate its effectiveness on a dataset of 1,072 video recordings from 13 student groups in a computational thinking course. This work presents a novel method for gaining actionable, data-driven insights into collaborative learning processes, with significant implications for pedagogical practice and intervention design.

Keywords: Collaborative learning · Engagement · Multimodal learning analytics · Large language models

1 Introduction

Student engagement plays an important role in shaping learning experience, influencing both what and how learners achieve. Engagement level is closely related to participation depth, interaction richness, learning satisfaction, and educational outcomes that extend beyond test performance alone [7]. However, traditional engagement assessments, such as self-report questionnaires and teacher observations, are static, subjective, and retrospective, failing to capture the dynamic, moment-to-moment processes in learning [1,11]. In contrast, process-oriented analysis provides a more nuanced view, revealing specific behaviors and mechanisms that drive productive knowledge acquisition [4]. In collaborative contexts, engagement is not merely the sum of individual behaviors but

S. Xin Yin and Z. Liu—Equal contribution.

E. G. Blanchard et al. (Eds.): AIED 2026, LNAI 16582, pp. 499–508, 2027.
https://doi.org/10.1007/978-3-032-29755-6_37

also involves social interaction, shared goal pursuit, and the distributed construction of knowledge [12,14,17]. Thus, engagement is more complex to observe and analyze. Current research emphasizes non-verbal cues such as gaze, facial expression, and gesture [2,8], yet these provide only a partial account of the interaction. Verbal communication is central to collaborative learning, as dialogue shapes the trajectory of group activity. Moreover, existing models are constrained to static images or short video segments, limiting the ability to capture how engagement evolves over time [9]. Another limitation lies in the scoring-based scheme. Previous studies treat engagement as a unidimensional construct rated on Likert scales, which neglects the complexity of interaction and lacks interpretability [2]. As a result, teachers and curriculum designers receive aggregate scores with little actionable insight for targeted interventions.

To address these challenges, we propose a novel MMLA framework with a nested temporal structure, leveraging the capabilities of large language models (LLMs) to dynamically recognize engagement patterns in collaborative learning. Our approach integrates verbal and non-verbal modalities across group members, identifies evolving engagement trajectories at the task, weekly, and course levels, and provides interpretable patterns that can inform pedagogical practices. Through this approach, we aim to deepen the understanding of collaborative engagement and open new directions for data-driven educational research. We demonstrate our framework's effectiveness by analyzing 1,072 video recordings from 13 groups completing 102 pair-programming coding tasks. Through experiment and analysis, we address the research question: To what extent can the proposed MMLA framework accurately identify and synthesize multi-scalar engagement patterns from raw multimodal data?

2 Literature Review

Engagement in Learning Process. Student engagement, the investment of cognitive and emotional energy in learning tasks [18], shapes academic participation, success, and persistence [7]. High engagement serves as a protective factor against student boredom, disaffection, and dropout rates [21]. Learning theory conceptualizes engagement into three dimensions: *Behavioral*, *cognitive*, and *emotional* [7]. These dimensions refer to observable learning behaviors, depth of thinking, and affective responses, all of which significantly impact learning performance, satisfaction, competence, and future learning motivation [1,11].

MMLA in Collaborative Learning. MMLA has gained emerging interest [15] and been applied across diverse learning contexts, including computer science [16,23], mathematics education [13], and STEM project-based learning [20]. These studies analyzed scaffolding strategies [16], identified clusters among pair programmers [23], explored interplay between discourse behaviors and emotional indicators [13], and predicted team success in open-ended tasks [20], advancing understanding of learner interaction [24].

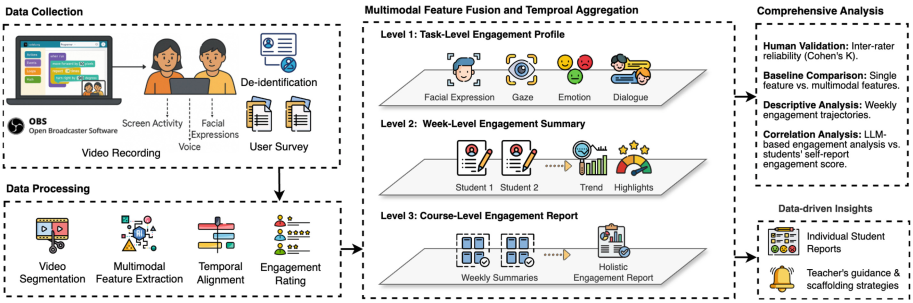

Fig. 1. MMLA framework for dynamic engagement pattern recognition.

Data Modalities in MMLA. Learning is inherently multimodal, with video data revealing non-verbal behaviors (e.g., postures, gestures) that indicate engagement and collaboration quality [14,22,25], while audio data captures spoken dialogue as a primary indicator of participation, including discourse patterns, talk moves, and off-task utterances that relate to collaboration performance [24].

3 Collaborative Learning Context

This study recruited 26 fifth-grade students (14 male, 12 female). Following our prior work [24], students attended a 4-week CT course on `Code.org` covering sequencing, debugging, loops, and events concepts, with 25 coding tasks per week. They alternated between "driver" and "navigator" roles in completing tasks through pair programming. We used OBS (obsproject.com) with built-in laptop cameras and microphones to record facial expressions, gaze, and verbal interactions, yielding 1,072 videos for analysis. At the end of the course, students completed a survey adapted from Fredricks et al. [7] to assess engagement levels (behavioral: α =.82; emotional: α = .90; cognitive: α = .79). Given potential inconsistencies in LLM predictions (due to demographic variation and noisy inputs), results should be interpreted with caution.

4 The Proposed MMLA Framework

The pipeline (Fig. 1) integrates multimodal data collection and processing to align behavioral, visual, and verbal signals. These signals are analyzed through a three-level architecture that progresses from task-level engagement profiles to week-level summaries and course-level reports. Table 1 includes the detailed steps of multi-modal and multi-dimensional data processing. Drawing on Fredricks et al.'s [7] theory, engagement is decomposed into three dimensions (i.e., cognitive, emotional, behavioral). We develop a 5-point evaluation rubric to quantify the frequency, intensity, and consistency of engagement indicators.

Multimodal Feature Fusion and Temporal Aggregation. We introduce a task-week-course structure that aligns with semester-based curriculum,

Table 1. Steps of Multi-modal and Multi-dimensional Data Processing

Step 1: Video Segmentation Video recordings are segmented at the task level, which serves as the minimal unit of analysis for engagement detection.
Step 2: Multimodal Feature Extraction Facial action, gaze pattern, emotion status, and verbal features are extracted from video data. These data sources provide evidence of the behavioral, emotional, and cognitive dimensions of engagement. **(1) Facial Action.** In the Facial Action Coding System (FACS), some Action Units (AUs) show significant correlations with engagement, indicating cognitive load, emotional valence, and attention focus [2,5,8]. Following previous work [2], we selected 12 AUs for the collaborative task. AU1 (inner brow raiser), AU2 (outer brow raiser), and AU5 (upper lid raiser) represent the "wide-eyed" expression related to moments of realization or confusion. AU4 (brow lowerer) and AU7 (lid tightener) indicate concentration and focus, serving as key components in expressions of cognitive effort. AU6 (cheek raiser), AU12 (lip corner puller), and AU14 (dimpler) are facial indicators of positive engagement. Conversely, AU15 (lip corner depressor) indicates sadness, and AU17 (chin raiser) indicates doubt and uncertainty. AU25 (lips part), AU26 (jaw drop), and AU27 (mouth stretch) represent mouth movements of active verbal communication during discussions. **(2) Gaze Direction Patterns.** Gaze direction is a reliable indicator of attention and learning engagement [19]. Sustained visual attention to learning interfaces indicates high engagement. Conversely, gaze patterns away from learning content indicate disengagement and correlate with reduced learning effectiveness. Moreover, joint attention coordination between collaborative partners represents high engagement quality, as synchronized attention during pair programming enhances collaboration effectiveness. **(3) Emotion.** The seven-emotion classification (neutral, happiness, sadness, surprise, anger, disgust, fear) from Paul Ekman's research [6] serves as a reliable engagement indicator [9]. This emotion status complements the discrete AU analysis, which can be measured by emotional intensity using 0-1 scale. The higher numerical values indicate more intense emotional states. **(4) Verbal Communication.** Conversations provide rich information about cognitive and collaborative engagement. There are some basic indicators including speaking time distribution and turn-taking behaviors. In the semantic space, linguistic markers such as reasoning language ("because", "therefore"), uncertainty expressions ("maybe", "I think"), and collaborative discourse ("let's", "what if") indicate depth of engagement and participation quality [24].
Step 3: Temporal Alignment All features are temporally aligned at the task level to ensure consistency across modalities and to support integrated multimodal analysis. Specifically, facial AUs and gaze directions are detected continuously as behavioral events assigned start and end timestamps. Dialogue segments are naturally bounded by utterance boundaries, marked by its temporal span. We align the heterogeneous streams using unified timestamps. This representation enables models to capture temporal dependencies across modalities.
Step 4 Engagement Rating **Cognitive engagement** reflects students psychological investment and self-regulation, indicated by concentration (e.g., AU4/7), sustained attention, and reasoning-focused dialogue. **Emotional engagement** captures affective involvement, with positive expressions and supportive communication indicating higher scores, and negative or neutral affect indicating lower scores. **Behavioral engagement** reflects active participation and effort, evidenced by expressive speech, coordinated gaze, and task-focused contributions, while disengagement and poor coordination indicating lower scores.

which mirrors students' learning progression and enables instructors to examine engagement at multiple temporal scales. **Level 1: Task-level Engagement Profile** generates individual engagement features of each pair programming task. For each session, facial AUs, gaze patterns, emotion status, and spoken utterances are listed chronologically. We then obtain each student's engagement score with representative evidence, according to the rating rubric. **Level 2: Week-level Engagement Summary** aggregates task-level profiles to weekly student summary, with engagement score synthesized from all task-level observations and evidence. This layer yields engagement trend insights for each student, identifies peak collaboration moments, and highlights specific areas where students could improve. **Level 3: Course-level Engagement Report** merges weekly summaries into the final engagement report. This layer characterizes both individual and group-level engagement and provides a final score.

5 Model Setup and Engagement Analysis

Model Setup. We use Qwen3-Omni-30B-A3B-Instruct for multimodal feature extraction with precise, timestamp-grounded event localization, due to privacy and ethical considerations. Our dataset contains sensitive audiovisual recordings of children, and closed-source, API-based models (e.g., Gemini, GPT-4o) would require transmitting raw biometric data to external servers, introducing unacceptable risks of data leakage and unauthorized processing. To address context window limitations that impact model performance on long videos, we split each recording into 45-second chunks, process them with sufficient context, then align timestamps and concatenate results to construct the complete task-level log. We adopt vision-enhanced speaker diarization [3]. For engagement analysis, we use Qwen3-30B-A3B-Thinking (temperature = 0.2) to reduce generation variance.

Engagement Analysis. We validated our framework through four dimensions. **(1) Human Validation.** Two raters annotated 10% (100 out of 1072 videos) recordings. Raters labeled facial expressions, gaze direction, and emotion categories at 5-second intervals, transcribed dialogue and annotated linguistic markers, and rated engagement on a Likert scale after each task. Inter-rater reliability was calculated using Cohen's kappa. In addition, two CT education experts were invited to evaluate the LLM-generated engagement report on accuracy, relevance, and actionability using a 5-point scale. **(2) Baseline Comparison.** We conducted ablation studies by testing each modality independently and comparing their engagement scores against the full multi-modal framework. **(3) Engagement Trajectory Analysis.** We demonstrate how multimodal behavioral indicators (facial engagement, gaze patterns, dialogue contributions) are reflected in an engagement report across time scales through a case study. **(4) Correlation Analysis.** We examined distributions of engagement scores across behavioral, emotional, and cognitive dimensions. Pearson correlation analysis assessed convergent validity between self-reported engagement scores and LLM-as-a-judge rated engagement scores.

6 Experimental Results

6.1 Human Validation and Baseline Comparison

Human Validation. Inter-rater reliability reached 0.86 (human-human) and 0.78 (human-LLM). Sub-scale κ values ranged from 0.71 (facial/gaze) to 0.93 (interaction patterns). Expert reviews further confirmed the LLM's utility, with high mean scores for accuracy (3.9), relevance (4.2), and actionability (4.1), validating the system for automated engagement analysis.

Baseline Comparison. We compare single-modality approaches by feeding individual modalities as input to the models for task-level engagement analysis. Experimental results showed that single modalities achieved lower agreement with human ratings (facial expression: .49, gaze pattern: .54, emotion: .47; conversation: .57). In contrast, multimodal approaches demonstrated higher agreement as .78, indicating that single modalities capture only a limited portion of authentic engagement status and cannot provide detailed descriptions.

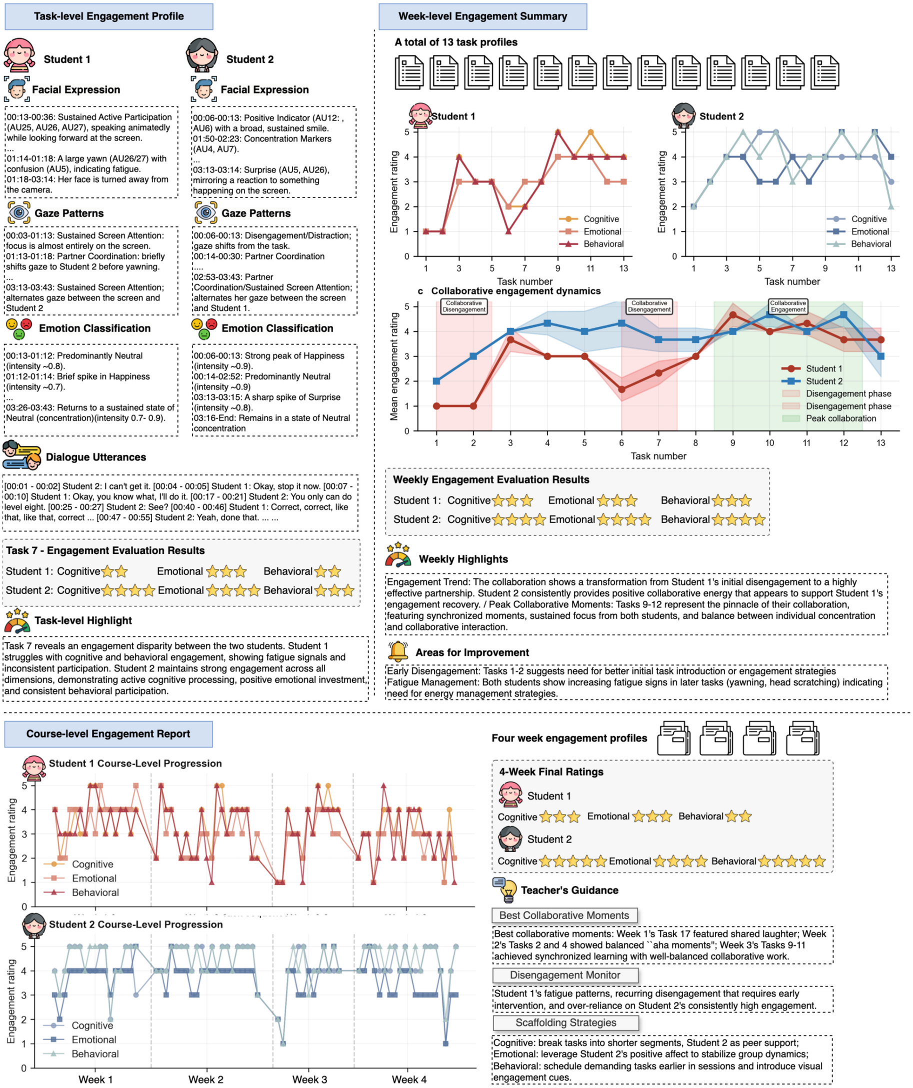

Fig. 2. Task-, week-, and course-level engagement report for Group 8.

6.2 Engagement Trajectory and Correlation Analysis

Engagement Trajectory Analysis. We use Group 8 students as a case to illustrate the structure and analytical insights provided by engagement report. As shown in Fig. 2, the report is organized into three parts. **Level 1: Task-level Engagement Profile.** In our example, the profiles of two students are presented side by side, with features listed from top to bottom. Three non-verbal features are displayed as sequential timestamps with brief summaries of student behavior. Dialogue analysis complements behavioral features by capturing interaction dynamics. Based on these descriptive results, the LLM assigns engagement scores across cognitive, emotional, and behavioral dimensions and also provides justifications grounded in the observed indicators. The task-level report concludes with a highlight, which is generated by synthesizing patterns from all feature categories to identify the most salient moments of engagement or disengagement during the task. **Level 2: Weekly Engagement Summary.** The weekly summary aggregates task-level profiles to reveal trends and distinct patterns. All 13 tasks were compiled into engagement trends, highlighting disengagement (red) and peak engagement (green) periods. This visualization allows instructors to rapidly assess individual trajectories and group collaboration quality. Weekly highlights identify key aspects of student engagement evolution (e.g., Student 1's recovery from early disengagement, supported by Student 2's consistent positivity) and peak collaboration episodes where both students demonstrated synchronized high engagement. Finally, the "Areas for Improvement" helps instructors adjust intervention, especially when students are disengaged or fatigue. **Level 3: Course-level Engagement Report.** The course-level report consolidates engagement data into line graphs with final ratings and actionable insights to present engagement evolution. This course report helps instructors develop a comprehensive understanding of students' learning journey at both individual and group level. Importantly, the report is not limited to a single course but can aggregate data from multiple courses to generate semester-based engagement reports. This allows instructors to identify cross-curricular patterns, thereby informing curriculum planning and long-term student support strategies.

Correlation Analysis Results. LLM-rated engagement significantly correlated with self-reported emotional ($r = 0.66$), cognitive ($r = 0.64$), and behavioral ($r = 0.59$) dimensions ($p < .01$). However, the LLM identified more low-end outliers, suggesting high sensitivity to disengagement than questionnaires.

7 Discussion and Implication

Collaborative Engagement Patterns. In the case study, while early engagement seems asynchronous between partners, one partner's sustained emotional engagement serves as a regulatory anchor that gradually increases the other's cognitive and behavioral participation over time. In our example, Student 2's consistent positivity acted as a stabilizing factor for the group. Student 1, on the other hand, although less engaged, showed gradual increases during tasks. This pattern provides empirical evidence for socially regulated learning theory [10,12], which posits that group members collectively monitor and adjust emotional states to maintain productive conditions.

Ambiguity in Visual Cue Interpretation. Prior research critiques that automated facial analysis systems have higher error rates for children [9]. For example, a furrowed brow may indicate frustration, confusion, or physical discomfort, but also indicate deep cognitive processing. To address this challenge, our framework triangulates multiple indicators to reduce misclassification. First, we integrate multiple modalities. A furrowed brow with sustained screen attention and task-focused dialog is interpreted more reliably as concentration. Second, our LLM-based analysis generates descriptive engagement profiles rather than binary classifications, allowing educators to apply contextual judgment. Third, hierarchical aggregation from task to week to course level smooths momentary errors, allowing engagement patterns to emerge from accumulated evidence.

Implications. By generating theory-grounded narratives, our LLM-based approach bridges the MMLA "interpretability gap", mapping computational patterns to established engagement constructs. This provides a replicable template for research across diverse contexts. For practitioners, our reports enable real-time, evidence-based interventions and actionable scaffolding. The report does not intend to increase teachers' burden to read them one by one; instead, it aims to help educators efficiently identify key patterns and access relevant details only when needed. Finally, the framework also supports professional development by providing concrete exemplars of productive collaboration and helping teachers develop observational skills to recognize students' disengagement signals.

8 Conclusion

We introduce a novel MMLA framework that leverages LLMs to dynamically recognize and interpret engagement patterns in collaborative contexts. By integrating verbal dialogue with non-verbal behavioral cues, our approach moves beyond traditional static scoring methods to capture the evolving, multidimensional nature of student engagement. Nonetheless, we acknowledge several limitations. The small sample size reduces statistical power and limits generalizability. In addition, automated facial analysis may introduce bias. Future work should validate the framework with larger and diverse samples, develop child-specific models, and establish fairness-aware MMLA protocols.

Acknowledgments. This research is supported by National Research Foundation Singapore under AISG Programme (Award No: AISG3-RPGV-2025-018).

References

1. Bergdahl, N., Nouri, J., Fors, U.: Engagement, disengagement and performance when learning with technologies in upper secondary school. Comput. Educ. **149**, 103783 (2020)
2. Chen, Y., Zhou, J., Gao, Q., Gao, J., Zhang, W.: MDNN: predicting student engagement via gaze direction and facial expression in collaborative learning. Comput. Model. Eng. Sci. (CMES) **136**(1) (2023)
3. Cheng, L., et al.: Integrating audio, visual, and semantic information for enhanced multimodal speaker diarization. arXiv preprint arXiv:2408.12102 (2024)

4. Di Mitri, D., Schneider, J., Specht, M., Drachsler, H.: From signals to knowledge: a conceptual model for multimodal learning analytics. J. Comput. Assist. Learn. **34**(4), 338–349 (2018)
5. D'mello, S.K., Graesser, A.: Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Model. User-Adapt. Interact. **20**(2), 147–187 (2010)
6. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the human face: Guidelines for research and an integration of findings, vol. 11, Elsevier (2013)
7. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. Rev. Educ. Res. **74**(1), 59–109 (2004)
8. Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.: Automatically recognizing facial expression: predicting engagement and frustration. In: Educational Data Mining (2013)
9. Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)
10. Hadwin, A., Järvelä, S., Miller, M.: Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In: Handbook of Self-Regulation of Learning and Performance, pp. 83–106. Routledge (2017)
11. Hazzam, J., Wilkins, S.: Student engagement in online learning: a systematic review. Educ. Inf. Technol. **28**(5), 5425–5462 (2023)
12. Järvelä, S., et al.: Socially shared regulation of learning in CSCL: understanding and prompting individual-and group-level shared regulatory activities. Int. J. Comput.-Support. Collab. Learn. **11**(3), 263–280 (2016)
13. Moon, J., Yeo, S., Banihashem, S.K., Noroozi, O.: Using multimodal learning analytics as a formative assessment tool: exploring collaborative dynamics in mathematics teacher education. J. Comput. Assist. Learn. **40**(6), 2753–2771 (2024)
14. Nasir, J., Norman, U., Bruno, B., Dillenbourg, P.: Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. Int. J. Comput.-Support. Collab. Learn. **17**(2), 231–275 (2022)
15. Ouhaichi, H., Spikol, D., Vogel, B.: Research trends in multimodal learning analytics: a systematic mapping study. Comput. Educ. Artif. Intell. **4**, 100136 (2023)
16. Ouyang, F., Dai, X., Chen, S.: Applying multimodal learning analytics to examine the immediate and delayed effects of instructor scaffoldings on small groups' collaborative programming. Int. J. STEM Educ. **9**(1), 45 (2022)
17. Scardamalia, M., Bereiter, C.: Knowledge Building. Cambridge (2006)
18. Schunk, D.H., Mullen, C.A.: Learning theories: An educational perspective, Pearson .6th edn (2012)
19. Sharma, K., et al.: Eye-tracking and artificial intelligence to enhance motivation and learning. Smart Learn. Environ. **7**(1), 1–19 (2020)
20. Spikol, D., Ruffaldi, E., Landolfi, L., Cukurova, M.: Estimation of success in collaborative learning based on multimodal learning analytics features. In: 17th International Conference on Advanced Learning Technologies, pp. 269–273. IEEE (2017)
21. Wang, M.T., Fredricks, J.A.: The trajectories of student emotional engagement and school burnout with academic and psychological development: Findings from finnish adolescents. Learn. Instr. **36**, 57–65 (2014)
22. Whitehead, R., Nguyen, A., Järvelä, S.: Utilizing multimodal large language models for video analysis of posture in studying collaborative learning: a case study. J. Learn. Anal. **12**(1), 186–200 (2025)
23. Xu, W., Wu, Y., Ouyang, F.: Multimodal learning analytics of collaborative patterns during pair programming in higher education. Int. J. Educ. Technol. High. Educ. **20**(1), 8 (2023)

24. Yin, S.X., Liu, Z., Goh, D.H.L., Quek, C.L., Chen, N.F.: Scaling up collaborative dialogue analysis: an ai-driven approach to understanding dialogue patterns in computational thinking education. In: the 15th LAK Conference, pp. 47–57 (2025)
25. Zhou, Q., Suraworachet, W., Cukurova, M.: Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. Educ. Inf. Technol. **29**(1), 1071–1098 (2024)