JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 12(1), 186–200. https://doi.org/10.18608/jla.2025.8595

Utilizing Multimodal Large Language Models for Video Analysis of Posture in Studying Collaborative Learning: A Case Study

Ridwan Whitehead[1], Andy Nguyen[2] and Sanna Järvelä[3]

Abstract
Incorporating non-verbal data streams is essential to understanding the dynamics of interaction within collaborative learning environments in which a variety of verbal and non-verbal modes of communication intersect. However, the complexity of non-verbal data — especially gathered in the wild from collaborative learning contexts — demands efficient and effective analysis. Methodological advancements are necessary to handle this complexity, enabling researchers to derive meaningful insights from these data streams. The advancement of Generative Artificial Intelligence (GenAI) has significantly broadened its accessibility, making it available to a diverse array of users and demonstrating its utility in aiding data analytics. However, the application of GenAI in multimodal learning analytics, particularly within the context of feature extraction for studying collaborative learning interactions, remains unexplored. This study aims to explore how multimodal large language models (MLLMs) can be utilized as part of the multimodal learning analytics (MMLA) process, focusing on the extraction of postural behaviour. The study focuses on an illustrative case study involving 52 pre-service teachers engaged in a physics-based collaborative learning task, demonstrating how MLLMs can be used for feature extraction. The integration of GenAI techniques in learning research promises a new horizon in understanding and enhancing collaborative learning interactions.

Notes for Practice

- Non-verbal data streams are crucial for understanding dynamics in collaborative learning environments, where verbal and non-verbal communication intersect. The advancement of generative AI (GenAI) with multimodal large language models (MLLMs) brings new opportunities to study collaborative learning interactions.
- The study presents a case from a co-located collaborative learning context that utilizes multimodal large language models (MLLMs) to analyze and interpret video data.
- The presented approach allows researchers to utilize MLLMs for feature extraction, facilitating the identification of patterns and insights that can inform the development of more effective educational interventions and adaptive learning environments.
- The study evaluates the reliability of MLLMs in annotating posture, identifying both reliable applications and lessons learned.

Keywords: Multimodal learning analytics, generative artificial intelligence, multimodal large language models (MLLMs), collaborative learning

Submitted: 07/08/2024 — **Accepted:** 13/02/2025 — **Published:** 19/03/2025

Corresponding author [1]Email: ridwan.whitehead@oulu.fi Address: Learning and Educational Technology Research Lab (LET Lab), Faculty of Education and Psychology, University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland. ORCID iD: https://orcid.org/0009-0002-2888-7304
[2]Email: andy.nguyen@oulu.fi Address: Learning and Educational Technology Research Lab (LET Lab), Faculty of Education and Psychology, University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland. ORCID iD: https://orcid.org/0000-0002-0759-9656
[3]Email: sanna.jarvela@oulu.fi Address: Learning and Educational Technology Research Lab (LET Lab), Faculty of Education and Psychology, University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland. ORCID iD: https://orcid.org/0000-0001-6223-3668

1. Introduction

In collaborative learning research, diverse analytical approaches have become essential for providing insights into the complex dynamics of learning environments where verbal and non-verbal modes of behaviours are intertwined. The extraction of non-verbal features in collaborative learning settings has been a topic of interest, with various methodologies employed, ranging from observations to advanced machine vision techniques (Radu et al., 2020; Kajamaa & Kumpulainen, 2020). However, an

inherent complexity arises when the integration of non-verbal behaviours, such as gestures, facial expressions, posture, and gaze, into the study of learning processes becomes of interest, since these modalities are often subject to unique methodological and analytical constraints (Schneider, 2024). Unlike verbal data, which can often be transcribed and analyzed systematically, non-verbal behaviours are inherently dynamic, context-dependent, and require specialized tools and frameworks for accurate capture and interpretation (Noëlet al., 2022). The sheer volume and complexity of data can quickly overwhelm traditional methods, making it difficult to process and derive meaningful insights (Ouhaichi et al., 2024). Additionally, the variety of non-verbal behaviours found in collaborative learning contexts require innovative approaches to successfully extract and interpret (Jeitziner et al., 2024). These challenges necessitate methodological advancements to handle the scale, heterogeneity, and contextual nuances of the data efficiently and effectively.

The fields of multimodal learning analytics (MMLA) and artificial intelligence (AI) in education have emerged to address these challenges, as non-verbal behaviours are typically emphasized as information-rich and valuable modalities to integrate in the MMLA process. Researchers have developed AI-driven systems capable of detecting and analyzing gestures, facial expressions, gaze patterns, and posture in real time, enabling a more nuanced understanding of group interactions (Çini et al., 2023; Radu et al., 2020; Zhou et al., 2024). These advancements facilitate the integration of diverse data streams, enhancing the granularity of non-verbal data capture and interpretation. By leveraging AI within MMLA, researchers are not only addressing the scale and complexity of non-verbal data but are also uncovering new insights into the intricate interplay between non-verbal cues and collaborative learning processes (Mu et al., 2020). These innovations represent a critical step forward in developing tools and frameworks that can capture the dynamic and context-dependent nature of non-verbal behaviours more effectively; however, challenges remain in ensuring accuracy, contextual sensitivity, and use of techniques for analyzing non-verbal interactions.

Recently, the advancement of generative AI (GenAI) with large language models (LLMs) and multimodal large language models (MLLMs) has further augmented the potential of AI in educational research. The first LLMs released to the public, such as OpenAI's GPT-3, demonstrated a remarkable ability to process and generate humanlike text. It served as an early example to how language models, as general models, trained on extensive amounts of data can perform well on tasks such as translation, question-answering, and more (Brown et al., 2020).

Building upon the capabilities of LLMs, MLLMs extend these analytical affordances to encompass multiple modalities beyond text, such as visual and auditory data. For example, models like OpenAI's GPT-4o have shown proficiency in understanding and generating visual content (Wu et al., 2023), which can be particularly relevant in educational contexts, where understanding the role of non-verbal behaviours can provide critical information regarding learning processes. Consequently, MLLMs can effectively interpret and integrate information from these various sources, enabling a comprehensive analysis of how different communication modes interact and contribute to the learning process.

The affordances of LLMs and MLLMs for the analysis of multiple modalities represent a significant opportunity for methodological advancement in MMLA, particularly in the extraction of non-verbal features from learners. While prior developments in MMLA research have progressed via classical machine learning and deep learning models to understand these interactions (e.g., Nguyen et al., 2022; Zhang et al., 2024), these approaches demand substantial technical expertise and frequently necessitate substantial interdisciplinary collaboration. By leveraging the generative and interpretative capabilities of LLMs and MLLMs, these challenges could be mitigated, including issues regarding heterogeneity and contextual nuances. This would enhance accessibility for educational researchers and enable deeper insights into the meanings derived from extracted non-verbal modalities.

Acknowledging the existing gap in understanding the potential utilization of MLLMs in research contexts, this paper explores through a case study how MLLMs can be utilized in studying collaborative learning. This approach extends the multimodal learning analytics system presented by Ochoa (2022) and is designed to guide researchers in the use of MLLMs as a tool to investigate video data found in contemporary educational settings. Specifically, postural behaviours are targeted as part of the feature extraction process, demonstrating how prompt engineering can be utilized to enhance data processing pipelines. We aim to answer the following research questions:

RQ1: How can MLLMs be reliably used as a scientific research tool in the MMLA process to extract posture behaviours?

RQ2: What are the considerations, including strengths and weaknesses, for using MLLMs for image annotation in a collaborative learning context?

Answering these research questions would allow for researchers to take advantage of the affordances of MLLMs for research and analysis, allowing for more holistic research to unlock deeper insights into the dynamics of learning.

2. Theoretical Foundation

In collaborative learning research, multimodal data streams enhance the capacity to analyze and interpret complex interactions. For instance, combining audio recordings with video footage and physiological data can provide insights into a student's engagement and emotional state during a learning activity (Di Mitri et al., 2018). This holistic approach allows educators and

ISSN 1929-7750 (online).

researchers to understand better how different modes of communication and interaction contribute to the learning process. By capturing multiple data streams, multimodal learning analytics can reveal patterns and correlations that might be missed when relying on a single data source (Blikstein & Worsley, 2016; Chango et al., 2021). Moreover, by integrating non-verbal behaviours in combination with verbal interactions, it is possible to gain more insight into the purpose of learner interactions, providing a more accurate picture of the individual and group levels of regulation of learning (Whitehead et al., 2024; Cukurova, Zhou, et al., 2020; Zhou et al., 2022).

The integration of multiple data streams also supports the development of more personalized and adaptive learning experiences. By analyzing diverse data types, educators can identify individual students' needs and preferences, tailoring instructional strategies to optimize learning outcomes. For example, if a student's facial expressions and body language indicate confusion or frustration, an adaptive learning system can provide explanations and/or additional resources to help them overcome these challenges. This dynamic, responsive approach to education leverages the non-verbal behaviours to create more effective and engaging learning environments (Cloude et al., 2022; Mangaroska et al., 2021; Watanabe et al., 2019).

Despite the advantages, Cukurova, Giannakos, and Martinez-Maldonado (2020) highlight three main concerns regarding the application of MMLA, including ethical, practical, and methodological challenges. Ethical concerns primarily revolve around privacy, potential biases in computational models, and the risk of fostering a surveillance culture. Practical challenges include the complexity and cost of gathering and synchronizing multimodal data, the need for sophisticated equipment and technical expertise, and ensuring seamless integration of learning experiences across various contexts. Methodologically, the field grapples with effectively collecting, pre-processing, and analyzing diverse data streams in ways that are meaningful and contextually relevant, while also addressing the issues of context awareness and generalizability. These challenges highlight the need to explore alternative ways of addressing the practical and methodological issues outlined, which are especially important in the context of collaborative learning.

Recent advances in MMLA methodologies provide opportunities to address these challenges by leveraging multimodal data in novel ways. As MMLA evolves, integrating diverse data sources, such as physiological, behavioural, and contextual data, requires not only technological innovation but also a stronger grounding in theoretical frameworks. By aligning methodologies with theoretical models, researchers can better ensure that the data modalities integrated are meaningfully interpreted, leading to deeper insights into the dynamics of learning (Giannakos & Cukurova, 2023). In this context, emerging tools such as MLLMs open new possibilities for expanding on capabilities to how non-verbal modalities are integrated, analyzed, and applied, further advancing the field of MMLA.

Addressing these issues, Ochoa's (2022) MMLA process systematically bridges learning theory and multimodal data to enhance the understanding and improvement of learning processes. It involves two reciprocal phases: mapping and execution. In the mapping phase, theoretical learning constructs are identified and linked to observable behaviours, which are further mapped to specific multimodal data traces such as gestures, gaze, or speech. This step addresses the "streetlight effect" by prioritizing theoretically grounded constructs over readily available data. The execution phase operationalizes this map by capturing data using diverse sensors and modalities, extracting relevant multimodal features, fusing these features to identify complex behaviours, and estimating constructs. Ultimately, this approach allows researchers to integrate learning theories with technological advancements, enabling a more nuanced and context-sensitive analysis of learning processes (Ochoa, 2022). Figure 1 shows the framework outlining the MMLA process.

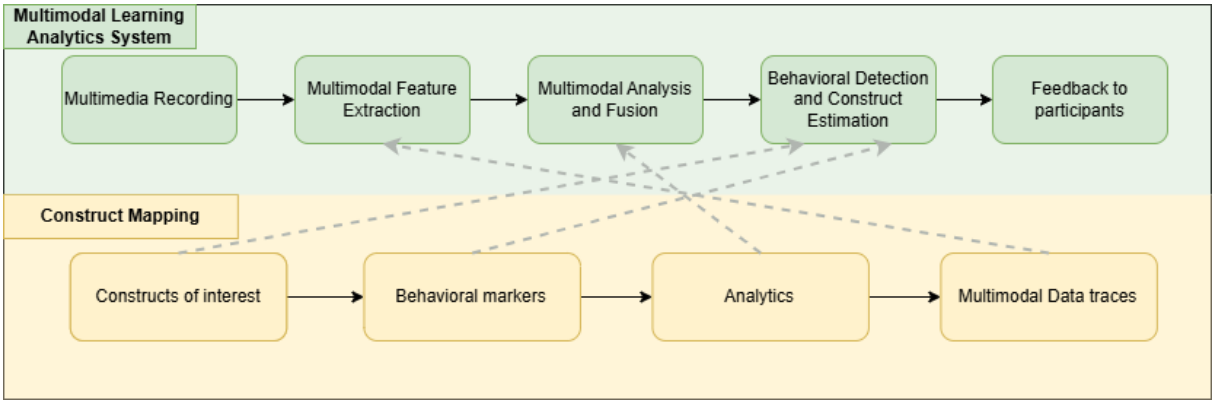

Figure 1. Multimodal learning analytics process (Ochoa, 2022).

However, these processes were not originally designed with the novel affordances of MLLMs in mind. Beyond the need for alignment with theory, existing MMLA methodologies do not yet fully leverage the cutting-edge capabilities that MLLMs offer, creating a growing need to explore what can be integrated within the MMLA process (Bewersdorff et al., 2025). The exemplification of such an approach (as we present in this paper) will enable researchers to effectively harness the enhanced

interpretative and scalable capacities of MLLMs, allowing for more sophisticated and dynamic analyses of collaborative learning interactions. To this end, understanding how MLLMs can be utilized is critical as researchers address the ethical, practical, and methodological challenges inherent in MMLA, enabling them to leverage non-verbal modalities present in collaborative learning contexts.

LLMs, including various generations with varying affordances, have demonstrated promising capabilities for collaborative learning interaction research. More specifically, the emergence of LLMs and, subsequently, MLLMs afford capabilities for researchers to integrate the various data modalities that learners produce into the MMLA process. While sometimes interchangeably used or mixed, it is important to note that LLMs and MLLMs are not the same, and that they differ in the types of data they can work with. While LLMs can typically only be inputted with text, MLLMs have additional capabilities such as understanding text, images, and audio.

Building upon this distinction, LLMs with sole capability of understanding text have demonstrated immense potential for enhancing the analysis of collaborative learning interactions. For instance, Suraworachet et al. (2024) compared the text-based affordances of GPT-4 to traditional natural language processing (NLP) approaches for predicting challenge moments in student discourse. Their findings suggest that GPT-4 can identify nuanced patterns in student interactions that traditional methods might lack, offering a more precise understanding of where students may require support.

Similarly, Xiao et al. (2023) explored how combining LLMs like GPT-3 with codebooks can support deductive coding in qualitative analysis. This combination allows for more efficient and accurate coding processes, reducing the manual effort required by researchers while maintaining high levels of reliability and validity. Brown et al. (2020) also highlighted the remarkable capacity of LLMs to process and generate humanlike text, offering nuanced insights into linguistic patterns and communication strategies within educational settings. Hutt et al. (2024) further demonstrated the strengths of GenAI by comparing classic NLP and GenAI methods for evaluating peer feedback, showing that GenAI can provide more detailed and contextually relevant feedback. LLMs can also be integrated into the learning analytics process. Yan et al. (2024) discussed the opportunities and challenges of using GenAI throughout the learning analytics cycle, emphasizing its potential to enhance various aspects of the process.

On the other hand, the multimodal capabilities of MLLMs afford researchers the ability to extract non-verbal behaviours more easily or accurately from video data (Zang et al., 2025). For example, in a study by Zeng et al. (2024), OpenAI's GPT-4o was utilized to synthesize gesture descriptions, context information, and user interaction history, employing advanced reasoning and contextual inference capabilities to map natural, free-form gestures to interactive functions without predefined datasets, thus showcasing MLLMs' ability to generalize from multimodal data inputs. Moreover, the classification of postures by utilizing MLLMs has also been a matter of interest. For example, Khan et al. (2024) leveraged language models to generate detailed pose descriptions for improving zero-shot classification tasks. Together, these studies demonstrate MLLM pipelines for diverse applications of leveraging non-verbal data for both general classification and domain-specific predictions.

However, the success of GenAI in research applications, specifically the use of MLLMs, has been mixed, depending on the specific context and parameters used. Recent studies have investigated the use of MLLMs for text classification tasks, yielding varied results. Reiss (2023) examined ChatGPT's performance and found significant variation based on different prompts and parameters. On this note, Pangakis et al. (2023) argued that LLMs must be validated against human annotations to ensure their effectiveness, noting the variability in prompt quality, data complexity, and task difficulty. These findings indicate that while LLMs have potential, their performance can be inconsistent and context dependent. Yet, Yan et al. (2024) conducted a comprehensive study on MLLMs, demonstrating high test–retest reliability and specifically demonstrating GPT-4o's performance in minimizing hallucinations across various categories. By focusing on a specific use case, the study also addresses practical challenges that may not emerge in abstract or generalized analyses. A benchmark study found significant performance gaps between humans and MLLMs in understanding abstract and complex images, highlighting shortcomings of MLLMs (Liu et al., 2024). Consequently, a study interviewing qualitative researchers on the use of LLMs has highlighted the need for tools and guides for ethical use of LLMs for research, which further exemplifies the gap in the field (Schroeder et al., 2024).

Furthermore, how MLLMs can be used as part of the MMLA research process has not been explored, and there is a need to further explore how MLLMs can be utilized for advanced analysis of learner data, specifically within the learning analytics cycle (Yan et al., 2024). The additional capabilities offered by these tools can complement MMLA, potentially enhancing their effectiveness, or addressing their limitations. By integrating GenAI into the MMLA process, researchers and educators can achieve a more comprehensive understanding of learning processes, leading to more effective interventions and improved learning outcomes.

3. Methodology

Case studies are particularly valuable for exploring complex, context-specific phenomena where in-depth understanding and practical insights are required (Yin, 2018). In the context of this research, a case study approach is essential for illustrating

how MLLMs can be integrated into the MMLA process for annotating posture behaviours, offering a focused and detailed examination of their potential and limitations. This approach allows for the exploration of methodological advancements, such as prompt engineering, within a real-world collaborative learning environment, providing actionable insights for broader applications (Stake, 1995).

3.1. Participants and Procedure

The participants included 52 pre-service teachers who participated in small groups of three to four members, totalling 14 groups, who were engaged in a collaborative learning task. They worked in an experimental setting, where they completed a 60-minute collaborative task focused on physics education. The experiment was conducted as part of their higher education course. The participants were assumed to be heterogenous in terms of prior knowledge related to the task. The task was structured around a collaborative learning script in the GoLab learning environment (Go-Lab, n.d.). The script aimed to teach the concept of energy conservation through diverse technological tools, mainly utilizing simulations as part of the teaching and learning process. The participants were assigned to groups randomly. They were seated in a crescent formation, with a large TV screen in front of them. They had access to a mouse and keyboard to navigate the learning platform displayed on the TV screen. The collaborative script, which the participants were instructed to follow linearly, included a brief introductory video about the concept of energy conservation, a skateboard simulation complemented by guiding questions and answers, and a hands-on experiment, where participants dropped balls from varying heights, observing and answering questions related to energy conservation. In other words, the task was designed to include both the learning platform and hands-on elements.

3.1.1. Ethical Considerations

Ethical procedures were integral to this study, particularly given the use of sensitive video data with MLLMs. Ethical approval was obtained from the university ethics committee, and the study adhered to the guidelines of the Finnish national board on research integrity. Informed consent was obtained from all participants, ensuring that they understood the study's purpose, the use of their data, and their rights, including the option to withdraw. While a complete anonymization of video data is not feasible due to the visibility of faces and other identifying features, steps were taken to minimize risks, ensuring secure storage and restricted access to the data. To further protect privacy, a commercial alternative was utilized (Microsoft Azure) with settings configured to prevent the model from using the data for further training and enhance data security, giving us the ability to not store the uploaded images any longer than necessary on the servers. Recognizing the inherent biases in MLLMs and their potential ethical implications, particularly in handling sensitive data (Schlagwein & Willcocks, 2023), this study emphasizes the importance of ethical reflection and responsible AI use in research.

3.1.2. Video Recording and Pre-Processing

The data in this case study included a total of 14 videos. The total time of the videos for all the groups, where the task was included, was 11:30:41 (hh:mm:ss). The videos were shot using Konftel 3840 × 2160 resolution conference cameras at 2560 × 1440 resolution. Since each entire group was recorded on a single camera, we needed to find a way to extract each person individually from every frame for further analysis. To achieve this, a python script was utilized, which used YOLOv8 (Jocher et al., 2023) to track objects in a video, process every 30th frame, extract and save cropped images of detected objects with persistent unique IDs, and display the annotated frames in real time. Although the algorithm works well, in a dynamic environment, where the participants were free to move around, and where obstructions from the camera view were common, the results were not perfect; thus, the extracted images were manually reviewed for inconsistencies in the images. Some issues that surfaced were related to participants moving around: the object detection algorithm would mix up the persistent unique IDs, causing the IDs for two participants to switch, even mixing up objects like the chair or the legs of the table. Also, due to the nature of the setting, other participants from other groups, or researchers asking questions would come into frame, and would be assigned a unique ID. In cases where two participants were visible in one frame, the quality of image was not clear, and/or the image did not clearly show the participant, those frames were removed from the dataset. After necessary manual corrections, the entire dataset was reviewed using XnViewMP, an image viewer, with the capability to load and view many, many images. After this process, we ended up with the frames of each person cropped, with unique persistent IDs, for the whole dataset with a timestamp (presented as the frame number as part of the image file), which amounted to around 150k total frames at one second granularity. Hence, in the context of this case study, this process shows how non-verbal data can be processed to be ready for analysis by MLLMs.

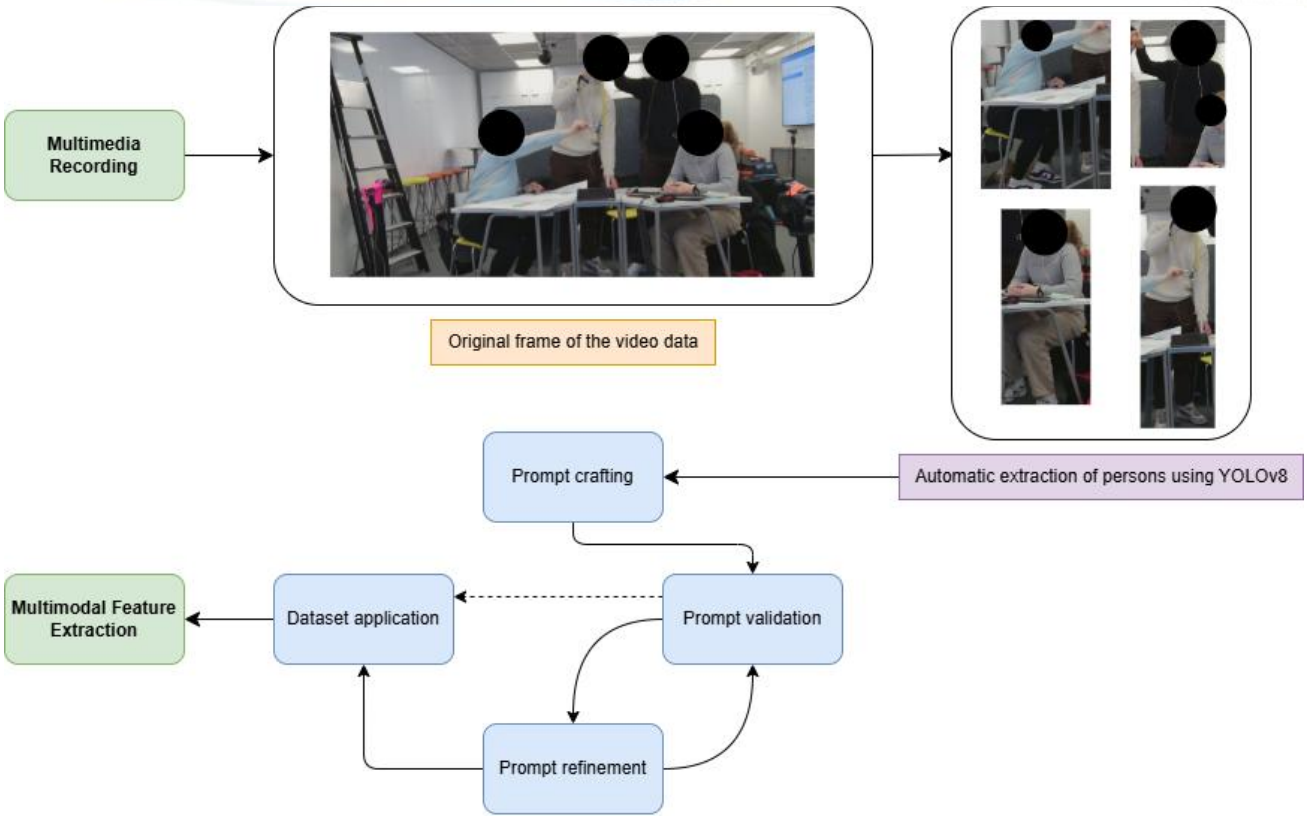

Figure 2. Data pipeline of the case study.

3.2. Data Analysis

The procedure for our analysis is presented in this section. Construct mapping is detailed in relation to how the categories aimed to be extracted were determined. Then, the justification behind the prompt engineering approach, along with procedure for it, is detailed.

3.1.3. Construct Mapping

This case study focuses on socially shared regulation of learning (SSRL) within collaborative learning contexts. SSRL involves the coordination of cognitive, behavioural, and emotional processes among group members to achieve shared learning goals (Järvelä et al., 2023). In this study, SSRL was connected with nine observable posture-related behaviours based on their relevance to regulatory processes: 1) sitting; 2) standing; 3) engaged with computer peripherals; 4) hands touching each other; 5) holding a task-related object; 6) hands near the face area; 7) hands resting on laps; 8) leaning on the table; and 9) arms resting on the table. These behaviours were selected because they potentially reflect, either on their own or in combination with other categories/modalities, important dimensions of SSRL and collaborative learning, such as task engagement, socio-emotional states, and task roles (Guo et al., 2019; Kajamaa & Kumpulainen, 2020; Taylor, 2016). While this study does not aim to conduct a complete and exhaustive process of construct mapping, posture behaviours are considered here as a demonstration of their relevance to collaborative learning and SSRL processes. Additionally, the case study does not analyze these behaviours but rather illustrates how they can be extracted using MLLMs.

To determine the categories, the constant comparison method was used (Onwuegbuzie et al., 2009) to work from the ground. The categories were determined by subject matter experts watching one of the 14 videos and noting down possibly significant postural behaviours. Initially, there were 30 categories, which was refined and reduced to nine through a process of grouping similar behaviours and conducting pilot tests. During these tests, smaller batches of data were analyzed with various prompts to evaluate which of these categories the MLLM could accurately interpret, ensuring the final set was both manageable and reliable before proceeding with the complete analysis. The authors were satisfied with the categories when they believed an adequate representation of the phenomenon (posture behaviours representing SSRL) was reached, and an initial impression was made of the interpretive capabilities of the MLLM. Validity — achieved through the saturation of the methodology — was an iterative process; when new observations no longer contributed meaningful changes to the category set, a stable and comprehensive representation of posture behaviours related to SSRL was indicated. Table 1 shows the categories used, along with their implications.

Table 1. Posture Characterizing Categories

Full name	Abbreviation	Description	Implications
Sitting	SIT	The person is sitting.	Whether the person is sitting or standing, depending on the phase of the task, could indicate if and how the learners are engaged with the task.
Standing	STD	The person is standing.	
Engaged with computer peripherals	ECP	The person is actively engaged with a computer peripheral such as mouse or keyboard.	Engagement with peripherals would indicate information regarding roles. Moreover, it would show the students engaging in answering a task or exploring task instructions.
Hands touching each other	HTC	The person's hands are touching each other.	Clasped hands can be a self-soothing gesture or indicate nervousness, impacting the collaborative climate and possibly signalling a need for regulation of learning (Navarro & Karlins, 2008).
Holding task-related object	HTR	The person is holding an object related to the task, such as pen, marker, paper, etc.	Handling objects related to the task can signal active participation and engagement, influencing perceptions of commitment to group tasks (Sinha et al., 2015).
Hands near face area	HFN	The person has one or both hands near the face/neck or ear area.	Often associated with contemplation or stress, this behaviour can affect group perceptions of an individual's emotional state and readiness to engage (Ekman & Friesen, 1969).
Hands resting on laps	HRL	The person is resting one or two hands on their lap.	This might suggest relaxation or disengagement, depending on the context. It can serve as a non-verbal cue in deciphering an individual's readiness to participate in learning tasks (Burgoon et al., 2011).
Leaning on table	LTB	The person is leaning on the table, visibly giving their weight.	Leaning forward can indicate interest and engagement. (Knapp et al., 2014).
Arms resting on table	ART	The person is resting their arms on the table.	This posture can suggest openness to interaction and readiness to participate, affecting collaborative dynamics (Mehrabian, 1972).

3.1.4. Prompt Engineering

Prompt engineering is an emerging technique for optimizing the performance of language models on specific tasks by designing and refining input instructions or prompts (Amatriain, 2024; Gu et al., 2023). It involves structuring input text to guide LLM outputs and can include methods such as role-prompting, one-shot, few-shot, chain-of-thought, and tree-of-thoughts prompting (Chen et al., 2023). The technique enables task adaptation of the model without updating its parameters and facilitates easier application of pre-trained models in real-world scenarios (Gu et al., 2023).

Prompt engineering as part of the data pipeline of the present case study (see Figure 2) included the following four steps: 1) prompt crafting, 2) prompt validation, 3) prompt refinement, and 4) dataset application. Prompt crafting involves drafting text-based instructions for MLLMs, guided by construct mapping steps: identifying constructs of interest, their behavioural markers, and related analytics. Before applying prompts to the main dataset, they undergo validation through reliability trials, where prompts are tested, refined, and retested until they meet desired reliability criteria. Once validated, these prompts enable MLLMs to extract the features for MMLA process.

For the present case study, Microsoft Azure's OpenAI service was used to access OpenAI's GPT-4o model. This model was used since it was reported to have the highest multimodal capabilities at the time of this study (Cui et al., 2024). We accessed the model through a python software development kit (SDK). This approach connects to the API of OpenAI's services. Unlike using their chatbot interface, this approach allows for high degrees of customization and automation. There are many parameters that can be customized, however, relevant to our case, we adjusted the "detail" and "temperature"

parameters. The detail parameter, which has low, high, and auto settings, lets you control the model's image processing and text generation. By default, the auto setting chooses between low or high based on the input size. In low mode, the model receives a 512 × 512 image and generates a 65-token description for faster responses. High mode begins with the low-res image, then creates detailed 512 × 512 crops, using up to 129 tokens for richer details. We chose the "high" detail setting to obtain the most accurate interpretation possible from the model. The temperature parameter ranges from 0 to 2, where zero is more "deterministic" and two is more "creative." Since our purpose for the vision analysis is related to the classification of images, we used a relatively more deterministic setting of "0.1." For this case study, zero shot prompting was utilized.

While sending the frames for the analysis, there are two fields must be filled: system message and text prompt. For the system message, we kept the default of "You are a helpful assistant." The text prompt is what accompanies the image and allows you to customize the information you want to extract from the image. Aside from the multimodal capabilities of MLLM, the prompt is a significant component in that this is where you insert the coding scheme. We constructed our prompt to be structured to ensure systematic and comprehensive image analysis, following a logical progression of instructions that clearly delineate the expected tasks and outcomes.

To achieve this, we adopted a codebook-based approach, crafting the prompt to incorporate expert-drafted coding schemes aligned with deductive coding principles. This follows evidence from recent studies demonstrating that utilizing predetermined codebooks with LLMs, as opposed to training task-specific models, can enhance the efficiency and reliability of qualitative coding tasks (Xiao et al., 2023). The complete prompt can be found in the supplementary materials of this study.

The prompt begins with an initial directive that establishes the overarching goal of analyzing the image for categorizing visible actions and postures according to predefined categories (Analyze the image provided and categorize…). This sets a clear objective and prepares the model for the subsequent detailed guidelines. Next, the prompt provides guidance on observation techniques, emphasizing the importance of focusing on the individuals' body language to identify and list only the relevant categories observable in the image (…the image should be scanned for the individuals' body language…). This step is crucial in limiting the analysis to relevant information. The prompt further instructs on the presentation of results, mandating that responses should only include clear observations without elaboration, ensuring clarity in reporting. A comprehensive list of categories is then provided, each accompanied by an abbreviation to standardize the terminology and streamline the identification process (…Sitting (SIT)," "One or two arms resting on table (ART)," and "Hands touching each other/clasped OR hand(s) touching their arm (HTC)…). Then, step-by-step instructions are presented based on the categories listed (Begin by determining if they are sitting or standing…). The prompt concludes with a specification of the expected output format, providing an example to reinforce clarity and consistency in reporting (…Output as the following example: SIT, HTR, LTB.).

The images are then embedded and combined with the prompt and parameter settings and then run. The API returns the output of the model in JSON (JavaScript Object Notation) format. Each of the JSON files are named based on the image file names. These are then parsed and converted into .csv (comma-separated values) format, which makes it suitable for further analysis.

4. Results and Findings

The results of this study provide insights into the reliability and potential of MLLMs for annotating posture behaviours in collaborative learning contexts. This section highlights the outcomes of inter-rater and test–retest reliability tests, identifying both the strengths and limitations of using MLLMs for feature extraction. Additionally, we discuss the reasons for the discrepancies between the human and MLLM, and give practical suggestions to overcome these discrepancies.

4.1. Reliability Testing of the MLLM

Scientific research requires us to ensure that the reliability and validity of coded categories are met. Two main criticisms of MLLMs exist in relation to their outputs. Firstly, referred to as "AI hallucinations," MLLMs can provide factually incorrect information with absolute determination (Ji et al., 2023). This is true for both text generation and image generation. In the context of MLLMs, one of the fundamental issues is how the model misinterprets or fabricates details from unclear or ambiguous images due to limitations in resolution and incomplete training data (OpenAI, 2023). This can lead to hallucinations where the AI generates plausible but incorrect or nonsensical descriptions or interpretations of visual inputs. To evaluate the reliability and validity of the MLLM analysis, we assessed inter-rater reliability using a random sample of 1,312 frames, which represents approximately 2.5% of the dataset. This evaluation compared annotations produced by the MLLM with those generated by a human rater, as well as annotations between two human raters. Although the sample size of the subset is lower than the commonly recommended range of 10–20% for inter-rater reliability assessments (Lombard et al., 2002), this selection was shaped by practical considerations, including the large volume of data available and the pilot nature of this study. Despite this limitation, the reliability test provided valuable insights into the performance of the MLLM analysis. Using the sklearn.metrics.cohen_kappa_score python package, Cohen's Kappa was calculated. Agreement scores were interpreted based on the thresholds suggested by Landis and Koch (1977), where values 0.21–0.40 indicate fair agreement, 0.41–0.60 moderate

agreement, 0.61–0.80 substantial agreement, and >0.81 almost perfect agreement. While no additional statistical tests for significance were performed, Cohen's Kappa inherently accounts for random agreement, making it a reliable indicator of agreement quality.

Moreover, to validate the sufficiency of using 2.5% of the dataset for inter-rater reliability testing, we conducted a simulation-based power analysis on the categories. Simulating 1,000 iterations with a Cohen's Kappa of 0.8 (indicating substantial agreement) and a significance level of 0.05, we observed near perfect power (<0.99) across all categories. This result demonstrates that the sample size was sufficient to detect meaningful inter-rater agreement. As demonstrated in Table 2, the inter-rater reliability results are quite mixed, ranging from 0.959 to 0.325 (N=9, M=0.625, SD=0.243). The possible reasons for these results are discussed in the next subsection, where we outline common problems related to the use of MLLMs for feature extraction.

Table 2. Inter-Rater Reliability of Posture Characterizing Categories of MLLM–Human and Human–Human

Categories	MLLM–Human	Human–Human	Mismatch count of MLLM–Human
SIT	0.959	0.989	12
STD	0.956	0.986	13
ECP	0.629	0.937	100
ART	0.347	0.834	401
HTR	0.624	0.871	249
HFN	0.810	0.871	46
HTC	0.577	0.780	170
LTB	0.325	0.888	36
HRL	0.400	0.837	81
COMBINED CATEGORIES*	0.305	0.756	851

**Note: Combined categories refer to the combination of the categories selected for each image. For example, a single image may have more than one category present. This category considers whether the combination of these categories for each image agrees.*

Secondly, another issue is related to how MLLMs do not provide consistent output. They are often referred to as black boxes and often have low explainability as to why a certain output was reached. This issue results in not knowing what the models will produce, resulting in possible variations in outputs given to the same prompts. For this reason, test–retest reliability was conducted to ensure that the output by the MLLMs can be reproduced, and that it is not per chance that it gives a particular output. As it can be seen in Table 3, the first pass and second pass of the MLLM output have near perfect agreement on all categories, suggesting that MLLMs indeed can provide consistent output for image annotation tasks, ranging from 0.990 to 0.870 (N=9, M=0.944, SD=0.041).

Table 3. Test–Retest Reliability of Posture Characterizing Categories by MLLM

Categories	MLLM first pass–second pass
SIT	0.990
STD	0.990
ECP	0.959
ART	0.891
HTR	0.964
HFN	0.958
HTC	0.956
LTB	0.927
HRL	0.870
COMBINED CATEGORIES*	0.897

**Note. Combined categories refer to the combination of the categories selected for each image. For example, a single image may have more than one category present. This category considers whether the combination of these categories for each image agrees.*

4.2. Lessons Learned for Using MLLMs for Image Annotation

How we can consistently and efficiently utilize MLLMs for image annotation has been and will continue to be a question as the models used are being developed day by day. In this section, we demonstrate examples of where human annotators and the MLLM disagree. We show examples from different types of reasons, explain possible issues, and provide suggestions. Based on a systematic analysis of randomly sampled mismatches (N = 90, 10 per category — see more in appendix B) we quantified the distribution of these errors across three main sources, namely, limb intrusion, insufficient context, and insufficient guidance.

Limb intrusion: As seen in Figure 3, limb intrusion happens when the part of the body of another person intrudes within the frame of the person of interest, making it seem like it belongs to the person of focus. This causes the MLLM to wrongly annotate the image. In our analysis, we found that this error most occurred with ECP, HTC, SIT, and STD. To overcome this issue, care must be taken for the camera angle to clearly show the limbs, and their connecting person, which must be considered while pre-processing the data.

Figure 3. Example of limb intrusion (left) and insufficient context (right).

Insufficient context: This issue arises when an image does not provide enough contextual information for the MLLM to make an accurate annotation. This happens because the model processes each image independently, without any memory of prior images or context. For this reason, every image should have enough context to provide an accurate judgment of the contents. Our analysis showed that most categories suffered from this issue, including ART, ECP, HRL, HTC, HTR, LTB, and most frequently with SIT and STD. For example, as seen in Figure 3, overcropping caused the MLLM to misinterpret the person who was sitting to be standing. One way to overcome this issue would involve ensuring that contextual clues necessary for the interpretation of the image are readily available. In other words, ensure that the images you work with include the full body or sufficient visual context to clearly identify the posture, such as visible legs, arms, and torso, as well as surrounding elements that help distinguish between different postures.

Insufficient guidance: This refers to when not enough guidance has been given to the MLLM to accurately extract the categories from the image in line with the human raters and general coding scheme. This is usually the case for specific categories, as they may have not been defined accurately, or there is too much variation of a particular category in the dataset. In our analysis, this is reflected in the less concrete categories, such as ART, HFN, HRL, HTC, HTR, and LTB. These categories present themselves in various ways in the learning context and require extensive definition. To overcome this, special care needs to be given to the definitions, ensuring that they encompass various presentations.

5. Discussion

This case study builds on the field of MMLA by exploring the utility of MLLMs in analyzing collaborative learning processes, with a focus on posture behaviours. The primary aims were to evaluate the reliability of MLLMs as scientific research tools in annotation of postural behaviours and identify considerations, including strengths and weaknesses, of their application in image annotation. Additionally, this work extends Ochoa's (2022) MMLA framework by integrating prompt engineering and addresses methodological, practical, and theoretical challenges in the field.

5.1. Annotating Posture Behaviours with MLLMs

The findings illustrated that MLLMs, with their own considerations and limitations, can serve as powerful tools for annotating posture behaviours, enabling scalable and efficient feature extraction. By embedding a codebook-based approach within

prompt engineering (Xiao et al., 2023), the study successfully extracted certain observable posture behaviours, such as sitting, leaning on a table, or hands resting on the lap, which can inform theoretical constructs measuring collaborative learning processes such as SSRL. The case study demonstrates how MLLMs can be employed to interpret video data with a high degree of granularity, moving beyond traditional, resource-intensive methods. By leveraging the interpretative capabilities of MLLMs, the study underscores the feasibility of integrating posture behaviours into collaborative learning research, in contrast to manual annotation or pre-trained, domain-specific models (Nguyen et al., 2022; Zhang et al., 2024). The case study expands existing examples where MLLMs have been used to annotate non-verbal behaviours in behavioural research (Khan et al., 2024; Zeng et al., 2024). Moreover, the present study also demonstrates a feasible alternative to joint position-based interpretation of posture behaviours, where often the learning setting is constrained to allow for the calculation of relevant postures (Radu et al., 2020). Overall, this approach highlights an early effort to expand the methodological toolkit for analyzing non-verbal behaviours in collaborative learning contexts by demonstrating a codebook guided prompt engineering process for non-verbal behaviour extraction. These signal how the MMLA process can be enhanced by MLLMs for a scalable and context-aware analysis of face-to-face collaborative learning environments. This has significant implications for how to achieve accessible and valuable insights, enhancing understanding of and support for collaborative learning.

Although we position MLLMs as promising tools to enhance the MMLA process, their use should not be taken for granted, nor should they be applied uncritically without addressing potential limitations. The inter-rater reliability between human annotators and the MLLM yielded mixed results, demonstrating substantial agreement for categories such as “sitting” and “standing” but lower agreement for complex categories like “arms resting on table” and “leaning on table.” These discrepancies underscore the challenges MLLMs face in interpreting ambiguous or nuanced postures, aligning with broader concerns regarding their difficulty in processing abstract and complex visual information (Liu et al., 2024; Zang et al., 2025). Nonetheless, beyond accuracy alone, we examined the consistency of MLLM-generated outputs, both correct and incorrect, as consistency is a crucial aspect of scientific research. Test–retest reliability analysis revealed near-perfect agreement across all categories, even those with lower inter-rater agreement. While our test–retest assessment is not exhaustive, our findings align with the extensive study by Yan et al. (2024), which demonstrated that GPT-4o outperformed other models in test–retest reliability, exhibiting an overall hallucination rate of 17.4%, and only 8% in the “actions” category, the primary focus of our study. This suggests that, within the context of our case study, MLLMs can produce systematic and reproducible results, with hallucinations posing minimal concern. These findings highlight the potential of MLLMs to serve as reliable tools for scalable posture annotation in collaborative learning research, provided their limitations are carefully considered and integrated into methodological frameworks.

Overall, this study extends Ochoa’s (2022) MMLA framework by integrating prompt engineering as a key methodological step in feature extraction. Typically, for the extraction of posture behaviours, classical machine learning and deep learning approaches, which require task-specific models trained on predefined behaviours and constrained to contexts similar to their training data, would be used (Cukurova, Zhou, et al., 2020; Radu et al., 2020; Watanabe et al., 2019). In contrast, MLLMs, when guided effectively, enable the direct extraction of complex, context-dependent posture behaviours that would otherwise require specialized sensors or external tracking tools. For example, a category such as “hands task related,” indicating that a student is physically engaged with a task-related object, would be difficult to capture directly from video using traditional methods, but an MLLM can infer such behaviours based on its interpretive capabilities. While traditional approaches excel in precision for well defined, narrowly scoped tasks, their effectiveness diminishes when applied to novel or highly contextualized behaviours. MLLMs, on the other hand, offer greater flexibility and generalizability by leveraging contextual understanding and natural language guidance, though their reliability remains dependent on effective prompt design and interpretability constraints. This signifies the contribution of MLLMs in expanding the range of analyzable features in MMLA, making it possible to work with more meaningful and high-level behavioural categories without additional algorithms or sensors, thereby addressing key challenges within the MMLA field (Cukurova, Zhou, et al., 2020).

5.2. Considerations for Using MLLMs for Annotating Postures

MLLMs offer a scalable alternative for feature extraction in MMLA (Khan et al., 2024; Zang et al., 2025; Zeng et al., 2024); however, their practical use requires careful methodological considerations to ensure validity and reliability. The observed discrepancies between MLLM-generated annotations and human raters underscore the importance of data preprocessing, particularly in mitigating issues such as limb intrusion and insufficient context. Since MLLMs process images independently, errors arise when key visual cues signifying context, such as the full-body posture or distinguishing elements, are missing, emphasizing the necessity of well framed, contextually rich data. Furthermore, the model’s performance is highly dependent on the specificity and clarity of the prompt engineering process, particularly for non-verbal behaviours that can manifest in multiple ways. This highlights a trade-off in using MLLMs: while they provide consistent and systematic outputs (test–retest reliability), their accuracy is contingent on external variables, necessitating a structured approach to prompt engineering and coding scheme design, while also considering data quality during the pre-processing step. From a methodological standpoint, this raises critical questions about the balance between automation and human oversight, suggesting that MLLMs, in their

current stage, are most effective as annotation tools, with the condition that human raters refine ambiguous or complex cases. Essentially, we argue that MLLMs do not necessarily replace the whole process but rather have the potential to complement existing methods by addressing key limitations. For instance, while YOLOv8's object detection algorithm can crop individuals, MLLMs can interpret these cropped images through prompt engineering. More broadly, these findings contribute to ongoing discussions on the role of MLLMs in behavioural research, illustrating how their integration can enhance scalability while reinforcing the need for rigorous validation protocols to mitigate biases and misclassifications (Pangakis et al., 2023; Reiss, 2023; Zheng et al., 2024). By addressing these challenges, MLLMs can serve as a powerful augmentation to existing annotation methodologies, offering a structured yet flexible approach to non-verbal analysis in face-to-face collaborative learning contexts.

While this study demonstrates the feasibility of using MLLMs for feature extraction in MMLA, several areas require further exploration. First, improving the accuracy and reliability of MLLM-based annotations remains a key challenge. Future work should investigate further how to refine codebooks, and how to effectively communicate them with the MLLMs to achieve more desirable results. Additionally, integrating MLLMs with traditional machine learning models or deep learning approaches may provide more robust annotation pipelines (Khan et al., 2024). Another important direction is the expansion of multimodal feature extraction beyond posture behaviours. Future studies could explore the ability of MLLMs to recognize gestures and proxemics, which are critical for understanding collaborative learning interactions. Furthermore, generalizability remains an open question, where evaluating how MLLMs perform across different educational settings, cultural contexts, and student populations is essential for broader applicability. This would include various conditions such as different camera angles, lighting conditions, or including diverse learner groups. For example, cultural and demographic differences in body language and learning behaviours may impact posture interpretations, requiring model adaptations or informed codebooks to mitigate potential biases. Finally, while this study focused on feature extraction rather than theoretical interpretation, future research should examine what insights can be extracted from automated posture annotations by MLLMs, and how they can be linked to collaborative learning processes. Combining MLLM extracted features with other data sources, such as speech or gaze behaviour, could provide deeper insights into these learning processes. Addressing these directions will help refine the role of MLLMs in MMLA, ensuring their effectiveness as scalable, interpretable tools for educational research.

6. Final Remarks

This study demonstrates the potential of MLLMs for analyzing non-verbal behaviours in collaborative learning by automating posture annotation from video data. Our findings highlight both the promise and limitations of using MLLMs in the MMLA process, particularly in terms of practicality, reliability, and interpretability. While MLLMs streamline large-scale feature extraction, their effectiveness depends on data quality, careful prompt engineering, and at this stage requires validation against human-coded data. Future research should refine these methods to enhance accuracy and explore broader applications of MLLMs in educational settings.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The data collection of the present study was carried out with the support of the LeaF Research Infrastructure, University of Oulu, Finland. This work was supported by the Academy of Finland Profi7 352788 and Academy of Finland projects 324381 and 350249.

References

Amatriain, X. (2024). *Prompt design and engineering: Introduction and advanced methods.* arXiv. https://doi.org/10.48550/arXiv.2401.14423

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences, 118*, Article 102601. https://doi.org/10.1016/j.lindif.2024.102601

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics, 3*(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winters, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Burgoon, J. K., Guerrero, L. K., & Manusov, V. (2011). Nonverbal signals. In M. L. Knapp & J. A. Daly (Eds.), *The SAGE handbook of interpersonal communication* (4th ed., pp. 239–280). SAGE Publications.

Chango, W., Cerezo, R., Sanchez-Santillan, M., Azevedo, R., & Romero, C. (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), 614–634. https://doi.org/10.1007/s12528-021-09298-8

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). *Unleashing the potential of prompt engineering in large language models: A comprehensive review.* arXiv. https://doi.org/10.48550/arXiv.2310.14735

Çini, A., Järvelä, S., Dindar, M., & Malmberg, J. (2023). How multiple levels of metacognitive awareness operate in collaborative problem solving. *Metacognition and Learning, 18*(3), 891–922. https://doi.org/10.1007/s11409-023-09358-7

Cloude, E. B., Azevedo, R., Winne, P. H., Biswas, G., & Jang, E. E. (2022). System design for using multimodal trace data in modeling self-regulated learning. *Frontiers in Education, 7*, Article 928632. https://doi.org/10.3389/feduc.2022.928632

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., & Zheng, C. (2024). A survey on multimodal large language models for autonomous driving. In K. Derpanis, H. Kuehne, S. Maji, V. Morariu, R. Souvenir, T. Hassner, & L. Verdoliva (Eds.), *Proceedings of the 2024 IEEE Winter Conference on Applications of Computer Vision Workshops: WACVW 2024* (pp. 958–979). IEEE. https://doi.org/10.1109/WACVW60836.2024.00106

Cukurova, M., Giannakos, M., & Martinez-Maldonado, R. (2020b). The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology, 51*(5), 1441–1449. https://doi.org/10.1111/bjet.13015

Cukurova, M., Zhou, Q., Spikol, D., & Landolfi, L. (2020a). Modelling collaborative problem-solving competence with transparent learning analytics: Is video data enough? In C. Rensing, H. Drachsler, V. Kovanović, N. Pinkwart, M. Scheffel, & K. Verbert (Eds.), *LAK '20: Proceedings of the 10th International Conference on Learning Analytics & Knowledge* (pp. 270–275). ACM Press. https://doi.org/10.1145/3375462.3375484

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning, 34*(4), 338–349. https://doi.org/10.1111/jcal.12288

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*(1), 49–98. https://doi.org/10.1515/semi.1969.1.1.49

Giannakos, M., & Cukurova, M. (2023). The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology, 54*(5), 1246–1267. https://doi.org/10.1111/bjet.13320

Go-Lab. (n.d.). *Golabz: The portal for inquiry learning spaces.* https://www.golabz.eu/

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., & Torr, P. (2023). *A systematic survey of prompt engineering on vision-language foundation models.* arXiv. https://doi.org/10.48550/arXiv.2307.12980

Guo, Z., Yu, K., Pearlman, R., Navab, N., & Barmaki, R. (2019). *Collaboration analysis using deep learning.* arXiv. https://doi.org/10.48550/arXiv.1904.08066

Hutt, S., DePiro, A., Wang, J., Rhodes, S., Baker, R. S., Hieb, G., Sethuraman, S., Ocumpaugh, J., & Mills, C. (2024). Feedback on feedback: Comparing classic natural language processing and generative AI to evaluate peer feedback. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th International Conference on Learning Analytics & Knowledge* (pp. 55–65). ACM Press. https://doi.org/10.1145/3636555.3636850

Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. *British Journal of Educational Technology, 54*(5), 1057–1076. https://doi.org/10.1111/bjet.13325

Jeitziner, L. T., Paneth, L., Rack, O., & Zahn, C. (2024). Beyond words: Investigating non-verbal indicators of collaborative engagement in a virtual synchronous CSCL environment. *Frontiers in Psychology, 15*, Article 1347073. https://doi.org/10.3389/fpsyg.2024.1347073

Jocher, G., Qui, J., & Chaurasia, A. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. Ultralytics. https://github.com/ultralytics/ultralytics

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kajamaa, A., & Kumpulainen, K. (2020). Students' multimodal knowledge practices in a makerspace learning environment. *International Journal of Computer-Supported Collaborative Learning, 15*(4), 411–444. https://doi.org/10.1007/s11412-020-09337-z

Khan, M. S. U., Naeem, M. F., Tombari, F., Van Gool, L., Stricker, D., & Afzal, M. Z. (2024). *Human pose descriptions and subject-focused attention for improved zero-shot transfer in human-centric classification tasks.* arXiv. https://doi.org/10.48550/arXiv.2403.06904

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2014). *Nonverbal communication in human interaction* (8th ed.). Wadsworth, Cengage Learning.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Liu, Z., Fang, F., Feng, X., Du, X., Zhang, C., Wang, Z., Bai, Y., Zhao, Q., Fan, L., Gan, C., Lin, H., Li, J., Ni, Y., Wu, H., Narsupalli, Y., Zheng, Z., Li, C., Hu, X., Xu R., ... & Ni, S. (2024). *II-Bench: An image implication understanding benchmark for multimodal large language models*. arXiv. https://doi.org/10.48550/arXiv.2406.05862

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Mangaroska, K., Martinez-Maldonado, R., Vesin, B., & Gašević, D. (2021). Challenges and opportunities of multimodal data in human learning: The computer science students' perspective. *Journal of Computer Assisted Learning, 37*(4), 1030–1047. https://doi.org/10.1111/jcal.12542

Mehrabian, A. (1972). Some subtleties of communication. *Language, Speech, and Hearing Services in Schools, 3*(4), 62–67. https://doi.org/10.1044/0161-1461.0304.62

Mu, S., Cui, M., & Huang, X. (2020). Multimodal data fusion in learning analytics: A systematic review. *Sensors, 20*(23), Article 6856. https://doi.org/10.3390/s20236856

Navarro, J., & Karlins, M. (2008). *What every body is saying: An ex-FBI agent's guide to speed-reading people.* Collins.

Nguyen, A., Järvelä, S., Wang, Y., & Rosé, C. P. (2022). Exploring socially shared regulation with an AI deep learning approach using multimodal data. In Chinn, C., Tan, E., Chan, C., & Kali, Y. (Eds.), *Proceedings of the 16th International Conference of the Learning Sciences — ICLS 2022* (pp. 527–534). International Society of the Learning Sciences. https://repository.isls.org//handle/1/8836

Noël, R., Miranda, D., Cechinel, C., Riquelme, F., Primo, T. T., & Munoz, R. (2022). Visualizing collaboration in teamwork: A multimodal learning analytics platform for non-verbal communication. *Applied Sciences, 12*(15), Article 7499. https://doi.org/10.3390/app12157499

Ochoa, X. (2022). Multimodal learning analytics: Rationale, process, examples, and direction. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *Handbook of learning analytics* (2nd ed., pp. 54–65). SoLAR.

Onwuegbuzie, A. J., Dickinson, W. B., Leech, N. L., & Zoran, A. G. (2009). A qualitative framework for collecting and analyzing data in focus group research. *International Journal of Qualitative Methods, 8*(3), 1–21 https://doi.org/10.1177/160940690900800301

OpenAI. (2023). *GPT-4 technical report.* arXiv. https://doi.org/10.48550/arXiv.2303.08774

Ouhaichi, H., Bahtijar, V., & Spikol, D. (2024). Exploring design considerations for multimodal learning analytics systems: An interview study. *Frontiers in Education, 9*, Article 1356537. https://doi.org/10.3389/feduc.2024.1356537

Pangakis, N., Wolken, S., & Fasching, N. (2023). *Automated annotation with generative AI requires validation.* arXiv. https://doi.org/10.48550/arXiv.2306.00176

Radu, I., Tu, E., & Schneider, B. (2020). Relationships between body postures and collaborative learning states in an augmented reality study. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education: 21st international conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, proceedings, part II* (pp. 257–262). Springer Cham. https://doi.org/10.1007/978-3-030-52240-7_47

Reiss, M. V. (2023). *Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark.* arXiv. https://doi.org/10.48550/arXiv.2304.11085

Schlagwein, D., & Willcocks, L. (2023). 'ChatGPT et al.': The ethics of using (generative) artificial intelligence in research and science. *Journal of Information Technology, 38*(3), 232–238. https://doi.org/10.1177/02683962231200411

Schneider, B. (2024). Three challenges in implementing multimodal learning analytics in real-world learning environments. *Learning: Research and Practice, 10*(1), 103–112. https://doi.org/10.1080/23735082.2023.2270611

Schroeder, H., Le Quéré, M. A., Randazzo, C., Mimno, D., & Schoenebeck, S. (2024). *Large language models in qualitative research: Can we do the data justice?* arXiv. https://doi.org/10.48550/arXiv.2410.07362

Sinha, S., Rogat, T. K., Adams-Wiggins, K. R., & Hmelo-Silver, C. E. (2015). Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer-Supported Collaborative Learning, 10*(3), 273–307. https://doi.org/10.1007/s11412-015-9218-y

Stake, R. E. (1995). *The art of case study research.* SAGE Publications.

Suraworachet, W., Seon, J., & Cukurova, M. (2024). Predicting challenge moments from students' discourse: A comparison of GPT-4 to two traditional natural language processing approaches. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th International Conference on Learning Analytics & Knowledge* (pp. 473–485). ACM Press. https://doi.org/10.1145/3636555.3636905

Taylor, R. (2016). The multimodal texture of engagement: Prosodic language, gaze and posture in engaged, creative classroom interaction. *Thinking Skills and Creativity, 20*, 83–96. https://doi.org/10.1016/j.tsc.2016.04.001

Watanabe, E., Ozeki, T., & Kohama, T. (2019). Modeling of non-verbal behaviors of students in cooperative learning by using OpenPose. In H. Nakanishi, H. Egi, I.-A. Chounta, H. Takada, S. Ichimura, & U. Hoppe (Eds.), *Collaboration technologies and social computing: 25th international conference, CRIWG+CollabTech 2019, Kyoto, Japan, September 4–6, 2019, proceedings* (pp. 191–201). Springer Cham. https://doi.org/10.1007/978-3-030-28011-6_13

Whitehead, R., Nguyen, A., & Järvelä, S. (2024). Exploring the role of gaze behaviour in socially shared regulation of collaborative learning in a group task. *Journal of Computer Assisted Learning, 40*(5), 2226–2247. https://doi.org/10.1111/jcal.13022

Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal large language models: A survey. In A. Cuzzocrea & R. Agrawal (Eds.), *2023 IEEE International Conference on Big Data (BigData)* (pp. 2247–2256). IEEE. https://doi.org/10.1109/BigData59044.2023.10386743

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In F. Chen, M. Billinghurst, M. Zhou, & S. Berkovsky (Eds.), *IUI '23 Companion: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 75–78). ACM Press. https://doi.org/10.1145/3581754.3584136

Yan, L., Martinez-Maldonado, R., & Gašević, D. (2024). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the14th International Conference on Learning Analytics & Knowledge* (pp. 101–111). ACM Press. https://doi.org/10.1145/3636555.3636856

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

Zang, Y., Li, W., Han, J., Zhou, K., & Loy, C. C. (2025). Contextual object detection with multimodal large language models. *International Journal of Computer Vision, 133*(4), 825–843. https://doi.org/10.1007/s11263-024-02214-4

Zeng, X., Wang, X., Zhang, T., Yu, C., Zhao, S., & Chen, Y. (2024). GestureGPT: Toward zero-shot free-form hand gesture understanding with large language model agents. *Proceedings of the ACM on Human–Computer Interaction, 8*(ISS), Article 545. https://doi.org/10.1145/3698145

Zhang, F., Possaghi, I., Sharma, K., & Papavlasopoulou, S. (2024). High-performing groups during children's collaborative coding activities: What can multimodal data tell us? In S. Pera, T. Bekker, T. Huibers, J. Good, C. Sylla, & S. Papavlasopoulou (Eds.), *IDC '24: Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (pp. 533–559). ACM Press. https://doi.org/10.1145/3628516.3655805

Zheng, Q., Lu, X., Jin, Q., Jain, J., Meadan-Kaplansky, H., Shi, H., Xion, J., & Huang, Y. (2024). Towards responsible use of large multi-modal AI to analyze human social behaviors. In R. Farzan, C. López, D. C. Llach, D. Quercia, M. Mustafa, S. Niu, & M. Wong-Villacrés (Eds.), *CSCW Companion '24: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 663–665). ACM Press. https://doi.org/10.1145/3678884.3687137

Zhou, Q., Suraworachet, W., & Cukurova, M. (2024). Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and Information Technologies, 29*(1), 1071–1098. https://doi.org/10.1007/s10639-023-12315-1

Zhou, Q., Suraworachet, W., Celiktutan, O., & Cukurova, M. (2022). What does shared understanding in students' face-to-face collaborative learning gaze behaviours "look like"? In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, proceedings, part I* (pp. 588–593). Springer Cham. https://doi.org/10.1007/978-3-031-11644-5_53

Appendix A: Complete Final Prompt Used for Posture Analysis

Analyze the image provided and categorize the visible actions and postures according to the following list. The image should be scanned for the individuals' body language. You are to identify and list only the categories that apply to each individual depicted in the image. Ensure no explanations or elaborations are included — only list the categories that are clearly observed in the image. Categories to Identify: Sitting (SIT), Standing (STD), engaged with computer peripherals (ECP), One or two arms resting on table (ART), Hands holding task-related object (HTR), One or two hands on face/neck/ear area (HFN), Hands touching each other/clasped OR hand(s) touching their arm (HTC), Leaning on table with one or both hands (LTB), one or both hands resting on laps (HRL). Begin by determining if they are sitting or standing. Once the posture is established, focus on the placement and activity of the hands and arms. Check where the hands are placed in comparison to the keyboard and mouse, if any. The categories are only applicable if their hands are directly engaged with them. Pay attention to the resting position of the arms, whether on a table or elsewhere. Look for any objects that might be held in the hands and are relevant to the task at hand. These would include, stationary, paper, balls, etc. Evaluate the facial area to determine if hands are present there. Identify if hands are in contact with each other, suggesting clasping or other forms of touch. Check if there are any hands resting on their lap. Check if they are giving their weight on the table with their hands. The result should be the short version of their names listed. Example: SIT, HTR, LTB.

Appendix B: Quantification Results for Common Errors in Posture Analysis

Row Labels	Count of Category
Insufficient Context	**36**
ART	5
ECP	1
HRL	7
HTC	1
HTR	5
LTB	2
SIT	7
STD	8
Insufficient Guidance	**31**
ART	5
HFN	8
HRL	2
HTC	7
HTR	2
LTB	7
Limb Intrusion	**7**
ECP	1
HTC	1
SIT	3
STD	2
Other	**16**
ECP	8
HFN	2
HRL	1
HTC	1
HTR	3
LTB	1
Grand Total	**90**