

ORIGINAL ARTICLE OPEN ACCESS

Leveraging Multimodal Large Language Models to Analyse Student Exploration Behaviours in Educational Game Environments

Yiqiu Zhou¹  | Xiner Liu²  | Andres Felipe Zambrano²  | Luc Paquette¹  | Amanda Barany²  |
Jaclyn Ocumpaugh²  | Ryan S. Baker²  | Jeff Ginger³

¹Curriculum and Instruction, University of Illinois Urbana-Champaign, Champaign, Illinois, USA | ²Learning Sciences and Technologies, University of Pennsylvania, Philadelphia, Pennsylvania, USA | ³Educational Psychology, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

Correspondence: Luc Paquette (lpaq@illinois.edu)

Received: 14 May 2025 | **Revised:** 29 April 2026 | **Accepted:** 12 May 2026

Keywords: educational games | generative AI | learning analytics | multimodal large language model | video analysis

ABSTRACT

Background: Video data provides rich opportunities to examine student behaviour in game-based learning environments, capturing not only observable actions but also subtle indicators of cognitive engagement. However, traditional video analysis is labor-intensive, and current applications of AI to this task are limited by models trained on general-purpose datasets, which often fail to capture the pedagogical meaning of student actions in authentic educational contexts.

Objectives: This study explores how multimodal large language models (MLLMs), specifically LLaVA-Video-7B-Qwen2, can support qualitative video analysis of student exploration behaviours in a Minecraft-based STEM learning environment.

Methods: We conducted an exploratory case study using screen recordings from a Minecraft-based STEM learning environment. We tested multiple prompt strategies for guiding MLLM-generated video descriptions and found that role-assignment prompting performed most effectively. We evaluated model outputs using a mixed-method framework that included quantitative scoring, GPT-based judgement, and researcher validation.

Results and Conclusions: Our findings show that MLLMs can reliably identify surface-level behaviours, such as navigation patterns and object interactions, but struggle to infer the intent or goals behind student actions, leading to a significant rate of over-interpretation (26.5% when explaining student strategies). The model's outputs are sensitive to prompt phrasing, underscoring the importance of prompt engineering. While current MLLMs show promise for streamlining parts of the video analysis workflow, their use in educational contexts requires structured oversight and careful interpretation to ensure reliability and relevance.

1 | Introduction

Analysing student behaviours through screen recordings offers a rich and ecologically valid source of data for examining how learners engage with digital learning environments (Wideman et al. 2007). Because these recordings capture high-fidelity details of student behaviour that are not always recorded in interaction logs, video screen capture has proven to

be a valuable data source for assessing implicit learning (Rowe et al. 2015, 2020) and identifying gameplay patterns (Wideman et al. 2007). For example, in game-based STEM learning environments such as Minecraft-based simulations, screen recordings can reveal subtle aspects of student exploration behaviours—including navigation strategies, environmental interactions, and tool usage—which are critical indicators of engagement, curiosity, and implicit learning processes.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Journal of Computer Assisted Learning* published by John Wiley & Sons Ltd.

Lay Summary

- What is currently known about this topic?
 - Video analysis is widely used in educational research but is time-consuming and resource-intensive.
 - Multimodal large language models (MLLMs) have shown potential in analysing video content but have mostly been evaluated on general tasks rather than educationally meaningful behaviours.
- What does this paper add?
 - It demonstrates how an MLLM (LLaVA-Video-7B-Qwen2) performs when analyzing gameplay screen captures from an educational context.
 - It introduces a structured approach to prompt design and evaluation for using MLLMs in educational video analysis.

Understanding these behaviours can directly inform instructional design, targeted interventions, and formative assessments (Calandra and Rich 2014; Cukurova et al. 2020; Lane et al. 2017, 20). Despite their potential, video analysis typically requires extensive manual transcription, annotation, and coding effort (Derry et al. 2010; Ramey et al. 2016). This labor-intensive process introduces both scalability and validity challenges. The volume of video often makes complete review impractical, compelling researchers to sample footage selectively. This selection process shapes which behaviours are studied and which remain unexamined (DeLiema et al. 2023). When only portions of a video are coded, it becomes difficult to determine whether observed patterns genuinely reflect students' typical behaviour or are the result of selection bias. This limitation undermines the credibility of inferences that can be drawn from the data (Blikstad-Balas 2017). As the scale of educational datasets grows, driven by advances in technology and increased digital interactions, there is an urgent need for more efficient and scalable approaches to analyse video data.

Recent advancements in Multimodal Large Language Models (MLLMs) (e.g., OpenAI GPT-4 with Vision) present a promising solution to this challenge. MLLMs combine the linguistic capabilities of traditional Large Language Models with visual encoders, enabling them to jointly process, reason about, and generate information from text, images, and video (Wu et al. 2023; Yang et al. 2023). These models have shown strong performance in tasks like object identification, video summarization, and vision-language alignment (Huang and Zhang 2024; Zhang, Yu, et al. 2024). As such, researchers could work with MLLMs to streamline video analysis, particularly in the initial phases of video annotation, transcription, and exploratory coding.

Despite these advances, the field still lacks a strong empirical foundation for understanding what these models can actually do in educational contexts. As noted by Bewersdorff et al. (2025), most existing work on MLLMs in education remains conceptual, outlining potential applications rather than empirically testing their performance in authentic learning settings. At the

same time, the broader MLLM literature offers limited guidance on how to evaluate or interpret model performance in educational video analysis, as model evaluations have been conducted on general-purpose benchmarks outside educational domains (Wu et al. 2023; Yin et al. 2024). These benchmarks typically involve visually salient actions, clear foreground-background separation, and relatively unambiguous ground truth labels, making them well suited for object or action recognition tasks in which surface-level visual patterns map clearly to predefined categories. In contrast, educational screen recordings, especially those from open-ended gameplay environments, are fundamentally different. The behaviours of interest are often subtle, context-dependent, and difficult to interpret without domain-specific knowledge. For instance, a student navigating a virtual biome could signify scientific inquiry or aimless wandering. Distinguishing between such possibilities requires more than visual recognition alone; it requires context-sensitive interpretation, where outputs must be not only accurate but also meaningful and instructionally relevant (Stoyanov and Kirschner 2023). This challenge is compounded by known limitations of MLLMs, including hallucination and difficulties with visual analysis and reasoning (Bewersdorff et al. 2025; Bai et al. 2024; Zhang, Yu, et al. 2024; Zhou et al. 2025). However, there remains little empirical guidance on their current capabilities in such settings or on how prompting strategies can be designed to support accurate and instructionally meaningful behavioural description. Together, these gaps point to the need for more systematic investigation of how video-based MLLMs perform in authentic learning environments and best practices for incorporating them into education research.

In response to these challenges, this paper presents a case study examining how a video-based MLLM (i.e., LLaVA-Video-7B-Qwen2) can support the analysis of screen recordings of students engaging with a Minecraft-based STEM learning environment WHIMC. We investigate two key research questions related to the use of LLaVA in describing and analysing student gameplay:

1. How do different prompting strategies influence the quality and educational relevance of MLLM-generated descriptions of student exploration behaviours? Specifically, we compare six prompting strategies to identify the best approach in transcribing student behaviours such as navigation, interaction with the environment, and use of in-game tools.
2. What are the potential and limitations of MLLMs in analysing student exploration behaviours? Through content analysis of MLLM outputs, we explore the insights these models can provide, the types of errors they generate (e.g., hallucinations or misinterpretations), and the practical implications for their use in video analysis in educational contexts.

This research contributes to the growing body of literature on AI-assisted qualitative research in three significant ways. First, we document a systematic approach to developing and testing prompts that guide MLLMs in documenting student behaviours, which can be adapted and extended by researchers in similar contexts. Second, we develop an evaluation framework for assessing the quality and utility of MLLM-generated

behavioural descriptions in qualitative research contexts. Third, we demonstrate concrete examples of what insights can be extracted from MLLM-generated documentation to support analysis of student exploration behaviours in educational games. In doing so, we contribute to ongoing discussions about how AI might be used to support trustworthy qualitative research.

2 | Literature Review

2.1 | Video Analysis in Educational Games

Video data has been a powerful tool for analysing student behaviours within game-based learning environments. Video recordings capture rich, multidimensional data that include both explicit actions (e.g., movements, interactions with game elements) and subtle behavioural cues (e.g., hesitation, strategic deliberation, and moment-to-moment decision-making) which traditional assessment methods or log data might miss (Rowe et al. 2015, 2020; Squire 2005). Prior case-based research has shown that video enables researchers to observe learning as it occurs in ecologically valid contexts (Squire 2005). Observing gameplay in situ enables researchers to capture how learners respond to game features dynamically, tracking how learners move from surface-level engagement to deeper forms of expertise, revealing changes in strategy use, content understanding, and metacognitive awareness (Wideman et al. 2007). These behavioural patterns are often accompanied by cognitive or emotional indicators such as confusion, frustration, or engagement—dynamics that are difficult to detect from logs alone but have been shown to impact learning outcomes in game-based environments (Magana et al. 2022).

Despite these strengths, video analysis itself remains a highly labour-intensive and time-consuming process. Researchers must iteratively segment footage, identify relevant behavioural patterns, develop or apply coding schemes, and repeatedly revisit the data to verify interpretations (Heath et al. 2010). These interpretive decisions are not trivial: determining which aspects of gameplay to attend to and how to classify or contextualise them requires careful theoretical and methodological deliberation (Derry et al. 2010; Ramey et al. 2016). The lack of standardised procedures for defining units of analysis further complicates the process and poses challenges to consistency and reliability across coders and research teams. As a result, even with the aid of software tools like computer-assisted qualitative data analysis software (CAQDAS), video analysis workflows often require substantial time investments, iterative revisions, and team coordination to ensure analytic rigour (Evers 2011; Pea and Hoffert 2014).

In summary, prior research demonstrates that video is particularly powerful for capturing the complexity of student behaviour in game-based learning environments, but the analytical processes remain labour-intensive, interpretively demanding, and difficult to scale. This inherent tension between the richness of video data and the scalability of its analysis represents the methodological challenge this study seeks to address, highlighting the need for tools that can support the early stages of video analysis such as transcription and exploratory coding, while

preserving the richness of the data. This motivates our investigation of whether video-based multimodal large language models can meaningfully augment these analytic processes.

2.2 | MLLMs And Prompt Engineering for Qualitative Analysis

Studies have demonstrated LLMs' potential in coding textual data (Kirsten et al. 2024; Zambrano et al. 2023), generating initial themes (Yan et al. 2024), and supporting codebook development (Barany et al. 2024). However, most of this research focuses on textual data, with non-textual data remaining largely unexplored. MLLMs extend beyond text-only processing to incorporate multiple modalities, including text, images, and video (Wu et al. 2023; Yang et al. 2023). Specifically, Video-capable Large Language Models (Vid-LLMs) have demonstrated capacities in content description, summarization, question-answering, and temporal localization (Tang et al. 2024). The automation of functionalities like video transcription and annotation could transform video analysis workflows, but challenges remain in using AI-generated outputs for research. Limitations such as hallucinations and incomplete contextual understanding by models (Bai et al. 2024; Liu, Xue, et al. 2024; Yang et al. 2023) necessitate careful prompt engineering.

Prompt engineering refers to the systematic design and refinement of natural-language inputs to shape model behaviour/responses (Chen, Zhang, et al. 2024; Marvin et al. 2023; Qi et al. 2023; Sahoo et al. 2024). Unlike simple prompt writing, this approach employs an iterative, structured workflow to test alternative prompts. Prompt engineering differs fundamentally from parameter- or model-driven optimization techniques (e.g., fine-tuning and architecture modification), as it operates at the input level without altering model weights or inference algorithms. However, it is important to acknowledge that LLMs are stochastic by nature and prompt refinement is more accurately understood as heuristic-guided exploration: a principled but iterative search through a large, partially opaque input space, rather than a deterministic optimization procedure (Sahoo et al. 2024; Qi et al. 2023). Prior research demonstrates that model performance can be enhanced by manipulating prompt structure (Whitehead et al. 2024), including context (Liu, Zhang, et al. 2024), assigning a specific role (Arlinghaus et al. 2024), and utilizing chain-of-thought reasoning to justify decisions (Wei et al. 2022). However, these insights mainly come from NLP tasks, such as text classification, summarization, or thematic coding (Chen, Zhang, et al. 2024; Liu, Zhang, et al. 2024; Sahoo et al. 2024). Prompting for MLLMs in video-based analysis introduces additional complexity due to the need to coordinate vision and language modalities. MLLMs must first interpret visual content before generating natural language responses (Wu et al. 2023; Zhang, Yu, et al. 2024). This cross-modal processing makes it more difficult to predict how the model will respond to a given prompt, as the final output depends on how the model interprets the visual scene (Qi et al. 2023). In educational settings, this challenge is further compounded by the need for outputs that connect observed behaviours to learning-relevant constructs. Effective prompts must therefore guide the model to attend to visually relevant elements and express them in a manner that aligns with educational research goals. Currently, there

is limited understanding of what constitutes a “useful” output in this context and how reliably MLLMs can produce educationally meaningful descriptions and interpretations. This lack of domain-specific guidelines for prompt design and output evaluation motivates the present study, which introduces a multi-phase engineering and evaluation framework to examine how a video-based MLLM can support educational video analysis.

2.3 | Evaluation of Prompt Engineering Approaches

Evaluating AI-generated output requires understanding how well it serves its purpose. Traditionally, natural language processing (NLP) tasks like text translation or summarization have been evaluated against reference outputs, using metrics like BLEU and ROUGE to measure how closely a model’s output matches human-written reference. However, these surface-level similarity measures are limited for open-ended responses, where qualities like coherence, flow, and informativeness matter more than exact word overlap (Akoury et al. 2020). Studies show that these metrics often correlate poorly with human judgements and struggle with meaning-preserving variations such as paraphrasing and synonyms (Becker et al. 2024; Gu et al. 2020). In response, researchers have explored alternative metrics, including lexical diversity measures that assess vocabulary richness (Jarvis 2013), distribution-based measures that compare semantic patterns against human writing (Wang et al. 2023), and preference-based metrics calculating alignment with human judgement (Gu et al. 2020).

Although these automated approaches offer scalability, they cannot fully replace human evaluation. Yet human assessment brings its own challenges, including inconsistent judgements that make it difficult to replicate results (Gehrmann et al. 2022). Thus, recent researchers have begun exploring hybrid evaluation strategies that combine human expertise with LLM as evaluation tools (Chen, Zhang, et al. 2024; Zheng et al. 2023). For example, Zheng et al. (2023) demonstrated that models like GPT-4 can achieve over 80% agreement with human preferences—comparable to inter-rater reliability among human evaluators—suggesting that LLM-based evaluation can approximate human judgement when designed carefully. Despite these promising results, concerns remain about LLM hallucinations and variability across domains, task types, or prompt styles.

In all likelihood, similar concerns are warranted when evaluating the output of MLLMs in our context, where the multimodal nature of the data and the educational context further complicate output assessment. Informed by prior evaluation efforts, our study adopts a multi-faceted evaluation strategy as different phases of prompt development require different forms of evidence. In the early stage of prompt refinement, automated indicators are useful because they allow us to compare prompt variants and identify responses that are sufficiently detailed and stable. In the strategy comparison stage, LLM-based judgement provides a scalable way to assess relative differences across prompting strategies. In the final stage, human coding is used to evaluate factual correctness, interpretive appropriateness, and educational usefulness in a domain-sensitive way. This blended

approach draws on the strengths and limitations identified in existing work, balancing scalability with the depth and rigour required for educational video research.

3 | Methods

This study adopts a qualitative, exploratory case-study design to examine how a video-based MLLM interprets and generates descriptions of educational gameplay screen captures. The Methods section details our research context, participants, data collection, video preprocessing, and model implementation.

3.1 | Learning System, Participants, and Other Data

Our analysis utilises screen recordings (Figure 2) from students participating in a Minecraft-based STEM learning environment (WHIMC; (Lane et al. 2017)) project, which leverages Minecraft Java Edition to create immersive astronomy learning environments. WHIMC consists of distinct virtual worlds, including ‘what-if’ earth scenarios, such as what earth would be like if it did not have a moon, as well as more radical environments like exoplanets and black holes. Students engage in scientific exploration and later work in teams to build habitats on mars. Screen recordings captured first-person gameplay experience, including their navigation through the virtual worlds, interactions with non-player characters (NPCs), and use of scientific measurement tools. Screen recordings did not capture participant faces or voices.

The study includes data from all 17 middle-school students (grades 6–8; age $M=11.63$ years, $SD=0.81$, range=10–13) who participated in a week-long summer camp program in the Western United States. This age range spans a critical period of cognitive and socio-emotional development and thus provides rich opportunities to observe how students interact with an open-ended learning environment like Minecraft. Of these participants, 11 identified as male, 4 as female, and 2 preferred not to respond. In terms of racial and ethnic background, 11 participants identified as white, 1 as African American, 1 as Native American, and 4 preferred not to respond. Throughout the camp participants completed several self-report surveys and engaged in learning activities on and off the computers. Data was collected with assent from the participants and consent from their guardians and approved by the Institutional Review Board for research.

3.2 | Multimodal Large Language Model and Implementation

To analyse screen recordings, we used LLaVA-Video-7B-Qwen2 (Zhang, Yu, et al. 2024; Zhang, Wu, et al. 2024), a video-based MLLM. Unlike general-purpose MLLMs such as GPT-4o or Gemini 1.5, which primarily process still images and treat short video clips as sequences of individual frames, LLaVA-Video is explicitly designed for temporal video understanding. This model builds on the Qwen2 architecture and incorporate additional modules for temporal understanding and frame-level

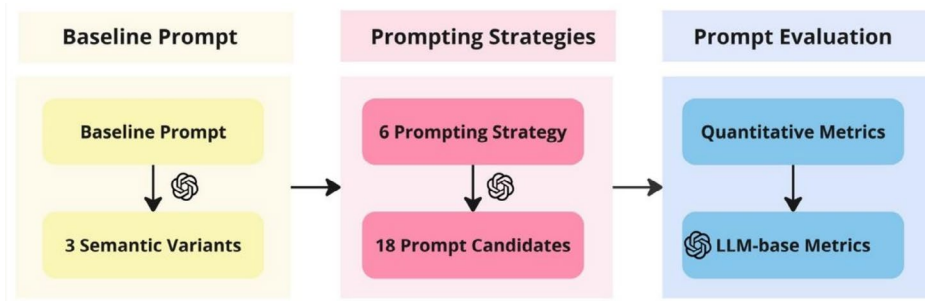


FIGURE 1 | Pipeline for prompt engineering (The GPT logo denotes steps where GPT was used, arrow indicates the sequential flow, and the pipeline was repeated separately for both tasks).

feature extraction, allowing it to process video as a continuous visual sequence rather than a collection of isolated snapshots. This architecture makes LLaVA especially suitable for tasks that require recognizing dynamic patterns of behaviour over time, such as movement, object interaction, or movement within digital environments. We selected this model based on several considerations. First, it is open-source and publicly reproducible, enabling transparent experimentation and future replication. Second, it was trained on large-scale, diverse set of video data as of October 2024, and it demonstrates strong performance across benchmark video tasks, including action recognition and temporal event localization (Fu et al. 2024; Zhang, Yu, et al. 2024; Zhang, Wu, et al. 2024). Third, and perhaps most importantly for our study, as a 7B parameter model, it is computationally lightweight enough to be tested across large batches of video clips without requiring specialised hardware or multi-GPU set-ups. While more advanced video-based models may continue to emerge, LLaVA has become a broadly adopted model in the open-source video MLLM community due to its accessibility and performance.

To prepare the screen recordings for model input, we segmented each video into 10-s clips and sampled each at 1 frame per second (fps). These frames, along with text prompts, are converted into tokens. This segmentation was designed to balance temporal coherence and token efficiency. Because the model's fixed context window limits the total number of tokens it can handle in one pass, there is an inherent trade-off between detail per frame (i.e., tokens dedicated to describing each frame) and overall video coverage (i.e., how many frames can be included). As pointed out in vision-language model evaluations (Fu et al. 2024), longer clips or denser sampling (e.g., 30s per clip or 5fps) can exceed the model's capacity, leading to either frame truncation or overly compressed text outputs. Conversely, shorter clips or sparse sampling (e.g., 1–2s or 0.25 fps) risk omitting context necessary for interpreting behaviours, especially in open-ended gameplay environments. Although there is no universally accepted standard for clip length and sampling rate, we selected a 10-s clip length with sampling at 1 fps based on iterative testing of alternative segmentation settings, including varying lengths (5, 10, and 20s) and frame rates (0.5, 1, 2, and 3 fps). In making this decision, we considered three criteria: whether the clip preserved a coherent unit of activity, whether the resulting frame sequence could be included in a single model input without exceeding its token limit, and whether the generated output retained sufficient detail without truncation or excessive compression. Shorter clips tended to split meaningfully connected ongoing actions across

multiple segments, whereas longer clips or denser sampling could exceed the model's input limits and produce more general, less specific summaries. The 10-s window at 1 fps consistently satisfied all three criteria and was therefore adopted as the standard configuration across all clips.

For reproducibility, we used the pre-trained model with temperature set to 0, which reduces randomness and makes the model's outputs more deterministic. Model inference was conducted on a FASTER cluster using a single NVIDIA A100 GPU with 40GB memory. We intentionally chose not to fine-tune the model for several reasons. First, our current research goal is not to optimise LLaVA for a single task, but rather to evaluate its out-of-the-box potential for this type of task and understand its baseline strengths and limitations in educational video analysis. Fine-tuning might artificially inflate performance and obscure the generalizability of the findings to broader use cases, especially for researchers and educators working with limited resources who rely on existing pre-trained models. Second, fine-tuning large multimodal models requires substantial annotated video datasets and computational resources, which could reduce reproducibility for other researchers attempting to replicate or build on this work.

4 | Analytical Framework

Traditional video analysis requires researchers to document granular observations and then synthesise them to understand learning processes (Derry et al. 2010). In this study, the analytical framework aims to examine how MLLMs might support this workflow. Consistent with our qualitative, exploratory case-study design, the framework focuses on an in-depth investigation of how a single MLLM behaves when interpreting gameplay video data from one authentic educational context. To evaluate the model's capacity to support different stages of video analysis, we tested its performance in two tasks: description and analysis. The description task prompts models to document low-level in-game behaviours, while the analytical task aims to distill high-level insights by identifying behavioural patterns. Because MLLM outputs can vary greatly depending on prompt formulation, we developed a multi-phase engineering pipeline (Figure 1) to probe how different prompt designs shape the model's responses. This pipeline follows research on prompt engineering strategies (Liu et al. 2023; Qi et al. 2023; Sahoo et al. 2024), which show the importance of testing both *semantic variants* (e.g., alternative phrasing of the same question) and *prompting strategies*. Through this structured exploration, we

aim to characterise the model's current strengths and limitations when applied to educational video analysis.

4.1 | Prompt Tuning

4.1.1 | Baseline Prompt Generation

We began by manually creating baseline prompts for description and analysis tasks respectively, labelled as Candidate 1 (C1) in Table 1. Since the wording and phrasing of prompts can affect how well models understand and respond to them (Liu et al. 2023; Qi et al. 2023), we created semantic variants of each baseline prompt to account for this variability. This issue, referred to as prompt sensitivity (Xu et al. 2023), refers to the tendency of prompt-based models—particularly MLLMs—to exhibit inconsistent or unpredictable behaviour in response to slight changes in phrasing, structure, or word choice in prompts. In practice, this means that two prompts expressing nearly identical semantic meanings can yield different outputs, complicating the efforts to assess whether a model's output quality reflects its actual capabilities or just sensitivity to minor wording differences.

To mitigate this confounding factor and better isolate the impact of prompt phrasing, we generated alternative versions of baseline prompts as semantic variants. Specifically, we followed a paraphrasing-based method (Jiang et al. 2020; Liu et al. 2023), instructing GPT-4o chatbot to generate two semantic variants (C2 and C3) of each baseline prompt as shown in Table 1. GPT-4o was selected because it was the most advanced widely accessible model at the time of this work and has been extensively used in educational research. After generating several rounds of prompts, we observed that additional variants tended to produce similar outputs. We therefore selected the three prompts (C1–C3) that showed the most variation in word choice and sentence structure,

ensuring a meaningful range of semantic diversity while keeping the evaluation scope manageable. This step enables us to compare and identify the most effective phrasing leading to detailed student behaviour description (detailed in Section 4.2).

4.1.2 | Prompting Strategy Implementation

Building on the semantic variants, we next implemented six prompting strategies to generate 36 candidate prompts (3 candidates $\times$ 6 strategies $\times$ 2 tasks) using GPT-4o chatbot. See [Supporting Information: Appendices A and B](#) for the complete prompts used. These strategies were detailed in Table 2, selected based on their demonstrated effectiveness in LLMs (Chen, Zhang, et al. 2024; Chen, Chen, et al. 2024; Sahoo et al. 2024) and potential applicability to MLLM. While few-shot prompting has proven effective in MLLM applications, we excluded it due to two considerations. First, there is a lack of 'ground truth' against which to design example prompts. Second, while few-shot prompting approaches have been shown to improve a model's alignment with the given task (Nguyen and Hayward 2024), prior research indicates that incorporating examples in prompts can lead to overgeneralization or overlearning from the provided examples (Liu, Zhang, et al. 2024). Since our primary goal was to assess LLaVA's inherent capabilities to analyse screen captures, we excluded examples to ensure that its responses were based on its pretrained knowledge rather than biasing toward specific examples introduced by the prompt.

4.2 | Prompt Evaluation

We evaluated LLaVA prompts in two phases. Phase 1 examined the impact of prompt phrasing (C1, C2, and C3). This reduced the risk that variations in phrasing would confound the comparison of prompting strategies. Phase 2 compared outputs from six prompting strategies. We randomly selected 50 video clips and processed each clip with six prompting strategies under both tasks (description and analysis), resulting in 600 MLLM-generated outputs (50 clips $\times$ 6 strategies $\times$ 2 tasks).

TABLE 1 | Semantic variants of baseline prompts across analytical tasks.

Task	Candidate prompts
Description	C1: Describe the student's physical movements and exploratory actions in detail
	C2: Provide a detailed account of the student's physical movements and exploratory actions
	C3: Provide a detailed description of the student's physical movements and exploratory actions
Analysis	C1: Explain the exploration strategies reflected in the student's movement and actions in detail
	C2: Provide a detailed explanation of the exploration strategies demonstrated by the student's movements and actions.
	C3: Explain in depth the exploration strategies reflected in the student's movement and actions.

TABLE 2 | Prompting strategy taxonomy and descriptions.

Prompt strategy	Description
Zero-shot	A baseline prompt without any examples or additional context
Role-assignment	A prompt that assigns a specific role or perspective to the model
Tone	A prompt that specifies the desired tone or style for the response
Contextual priming	A prompt that provides relevant context information
Action-aware priming	A prompt that provides relevant action information
Chain-of-thought	A prompt that facilitates step-by-step reasoning processes

4.2.1 | Quantitative Evaluation

In Phase 1, we conducted a quantitative evaluation using common NLP metrics to compare outputs from three semantic variants for each baseline prompt. This step aimed to identify prompt formulations that yielded consistent, rich, and informative outputs for student behaviour transcription. We focused on two categories of metrics: *output length* and *lexical diversity*.

Output length metrics included average, standard deviation, and range of output token count. These metrics provided insight into the length and consistency of model responses. While longer outputs are not inherently better, prompts that consistently generate responses of sufficient length are more likely to engage the model in producing complete and detailed behavioural descriptions. Conversely, excessively short outputs may indicate that a prompt is too vague or underspecified, failing to elicit the depth needed for video transcription. Similarly, high variability in output length across similar inputs may reflect instability in how the model interprets the prompt, which is problematic for achieving reliable, replicable outputs in downstream analysis. In our evaluation, we define consistency as the model's ability to produce responses that are comparable in length and behavioural focus across different video clips. This level of consistency is critical for ensuring that the generated outputs can be meaningfully compared and used in further analysis. By analysing these metrics, we aimed to filter out prompt variants that produced overly brief, inconsistent, or otherwise suboptimal responses.

In addition to length-based metrics, we also calculated *lexical diversity metrics*, including unique word count, Type-Token Ratio ($TTR = \text{Number of unique words} / \text{Total number of words}$) and distinct bigram ($\text{Distinct-bi} = \text{Count of unique bigrams} / \text{Total number of bigrams}$). These metrics quantified vocabulary richness and phrase-level variation in model outputs, allowing us to assess whether a prompt leads to output that is linguistically diverse and behaviourally specific. Prior research suggests that higher lexical diversity often correlates with improved quality, as it signals more complete, less redundant, and contextually rich content (Chung et al. 2023; Schmidová et al. 2024). In our context, increased diversity might suggest that the model is attending to a wider range of student behaviours rather than generic descriptions.

We computed each metric for all prompt variants and used a scoring approach to identify the most effective prompts. Specifically, we applied min-max scaling across six metrics: (1) average output length, (2) standard deviation of length, (3) output range (maximum minus minimum length), (4) number of unique words, (5) type-token ratio (TTR), and (6) distinct bigrams (Distinct-N). Metrics associated with output diversity and richness were positively weighted, while two length variability metrics (i.e., standard deviation and range) were inverted, as lower values indicate more consistent responses. Each metric was normalised using min-max scaling and then averaged to compute a final aggregated score per prompt variant.

The final prompt for each strategy was selected balancing depth, consistency, and vocabulary richness as proxies for behavioural detail and information quality. A manual review of a subset of model outputs confirmed that these metrics

effectively identified prompts that generated consistent and informative behavioural descriptions and analysis (see [Supporting Information: Appendix C](#) for representative examples). Based on this evaluation, we selected the best-performing variant for each prompt strategy. In total, 12 final prompt candidates (6 strategies $\times$ 2 tasks) were selected and then advanced to LLM-based evaluation.

4.2.2 | LLM-Based Evaluation

Phase 2 employed GPT-4o through OpenAI's application programming interface to assess LLaVA outputs. We developed a rubric informed by current considerations about AI-generated content in educational contexts (Nguyen and Hayward 2024; Wang and Demszky 2023). The full GPT-4o prompt used for scoring is provided in [Supporting Information: Appendix D](#). This rubric includes four dimensions that we found useful: (1) comprehensiveness of behavioural documentation to ensure complete capture of student behaviours, (2) educational relevance to learning objectives, (3) sufficient level of detail, and (4) alignment with specific task goals (i.e., description and analysis). For each video clip, GPT-4o compared outputs from all six prompting strategies simultaneously, assigning relative scores (1–5) based on how well each output performed compared to others.

To ensure the reliability of GPT-4o's evaluation, we conducted stability and human agreement testing. First, we configured GPT-4o's temperature to 0. Given the inherent stochasticity of GPT models, which can yield varied outputs, we randomly selected 12 summaries (two per method) for each task and replicated the evaluation prompt three times to assess output consistency prior to the submission of all summaries to GPT. We confirmed that the scores assigned to the same summary across dimensions remained within an acceptable range of ± 1 across the three repetitions. Moreover, the rationales provided by GPT for each dimension, when represented as text embeddings, consistently achieved a cosine similarity exceeding 0.80 across all repetitions. Second, three researchers independently reviewed GPT-4o's ratings for 20% of the evaluation dataset. For each dimension, the researchers indicated whether they agreed or disagreed with GPT-4o's score. Final agreement was determined by majority vote, achieving 80% overall agreement across all cases ($n = 40$). We also noticed that agreement varied across dimensions, with educational relevance (90%) showing the highest reliability and behavioural documentation the lowest (60%).

4.3 | Content Analysis

Using the most promising prompts identified for each task, we conducted content analysis of LLaVA's outputs to evaluate their quality in terms of factual grounding, interpretive accuracy, and behavioural relevance. Our goal was to differentiate between grounded observations, plausible interpretations, and problematic content such as over-interpretation, irrelevance, or hallucination (Zhou et al. 2025). We analysed 90 selected video clips (45 per task), sampled independently from the broader dataset of LLaVA outputs. Clips were randomly selected from diverse students to ensure diversity across video content. To

maintain focus on active exploration behaviours, we excluded the first and last 5 min of each video, which might involve instruction (by the teacher) or wrap-up activities. This filtering step helped ensure that selected clips reflected meaningful gameplay and behavioural engagement. For each clip, all model-generated sentences were coded, resulting in a dataset of 926 sentences (492 for the description task and 434 for the analysis task). Sentences were segmented based on punctuation. In cases where the model output contained multi-line or

compound statements, we further divided them based on semantic completeness.

The sentence-level coding was conducted by three researchers, all co-authors of this article. All coders had prior experience with the WHIMC environment, a clear understanding of its learning goals, and the necessary domain knowledge to interpret student behaviours within this context. The coding scheme (Table 3) was developed inductively through repeated review of LLaVA-generated outputs, focusing on differentiating between grounded observations (factually accurate descriptions of visible actions), plausible interpretations (reasonable inferences beyond direct observation), and problematic content (such as hallucinations or irrelevant details). These distinctions are critical for evaluating the accuracy and usefulness of MLLM-generated descriptions of student behaviour. Prior to independent coding, the three coders completed a joint calibration session in which they collaboratively applied the coding scheme to a shared practice set of 30 sentences drawn from outside the main analysis corpus, resolving disagreements until consensus was reached.

Each sentence was then independently coded by all three researchers. The team followed an iterative process that involved multiple rounds of individual annotation, inter-rater reliability assessment, and collaborative discussion to resolve remaining discrepancies. After each round, the code definitions were refined to improve clarity and resolve discrepancies. After three rounds of refinement, the final coding scheme achieved pair-level Cohen's kappa ≥ 0.7 for all codes except plausible interpretation (Kappa ≥ 0.6) and over-interpretations (Kappa ≥ 0.35). We note that these two codes both represent edge cases—where ambiguity is a core part of the definition. Rather than collapsing these into broader categories, we accepted a lower agreement threshold to preserve important nuanced distinctions between reasonable inference and problematic over-interpretation. An example of how the finalised coding scheme was applied is shown in Figure 2.

TABLE 3 | Sentence-level coding scheme for analysing LLaVA output.

Code	Description
Reliable information	Accurate, specific, and contextually relevant information that contributes directly to understanding student behaviours
Plausible interpretations	Interpretations related to student behaviours that are reasonable
Over-interpretation	Interpretations that seem related to student behaviours but makes excessive assumptions or interpretations
Irrelevant information	Information unrelated to student behaviours, often focusing on aspects like game elements, health bars, or inventory
Hallucinations	Factually inaccurate or misinterpreted information

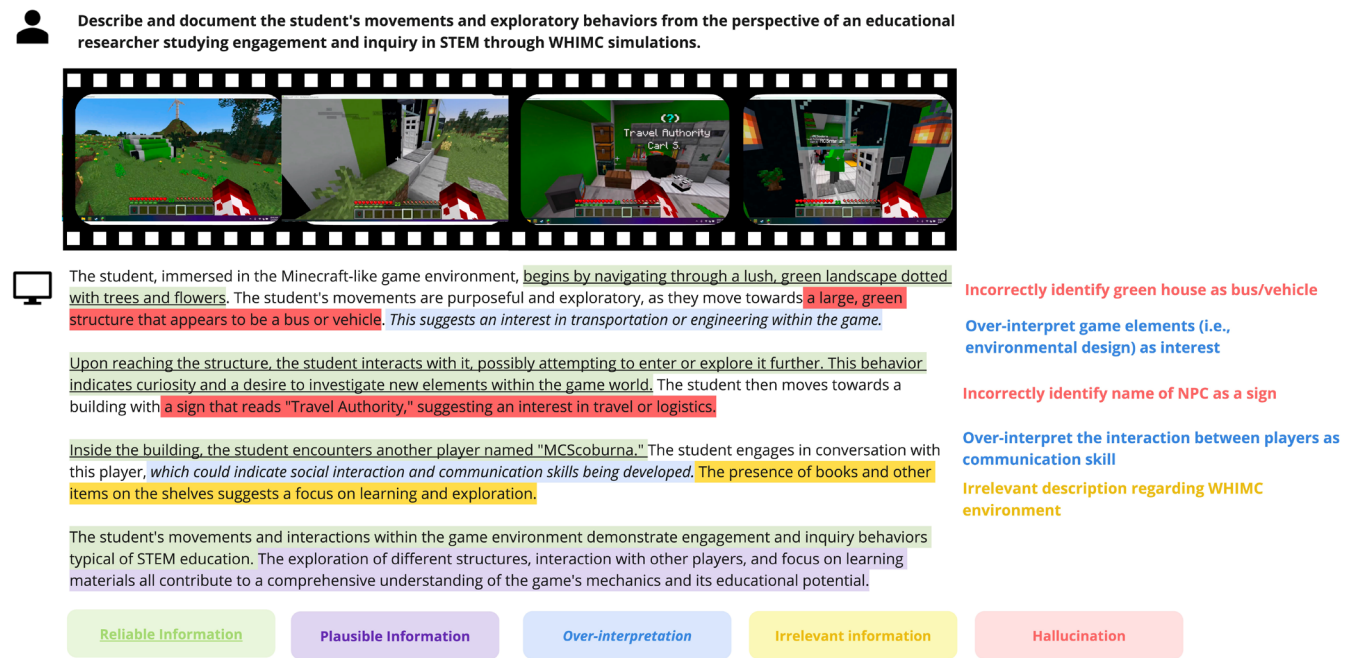


FIGURE 2 | Coded example of LLaVA output from description task.

TABLE 4 | Quantitative metrics across semantic variants for each prompting strategy (C1, C2, and C3 represent different prompt candidates. Bold entries indicate the selected optimal variant.)

Strategy	Variant	Avg length	Std length	Max	Min	Unique words	TTR	Distinct-N	Final score
Role-assignment	C1	204	59.2	314	80	1387	7.368	2.425	0.370
	C2	258	52.2	392	158	1784	7.223	2.382	0.667
	C3	157	59.6	374	108	901	8.717	3.314	0.333
Tone	C1	217	41.2	297	100	1273	8.530	2.714	0.849
	C2	80	25.8	175	49	741	5.375	2.162	0.333
	C3	151	64.9	335	71	1162	6.474	2.307	0.320
Context	C1	164	58.4	327	100	1129	7.261	2.578	0.635
	C2	169	55.9	277	87	1169	7.243	2.514	0.553
	C3	106	23.1	164	70	799	6.630	2.535	0.388

5 | Results

5.1 | Effects of Prompt Engineering on MLLM Output Quality

To determine the most promising prompting strategy (RQ1), we conducted a two-phase evaluation: (a) automated quantitative metrics comparing semantic prompt variants within each strategy, and (b) an LLM-based evaluation of the six prompting strategies. Table 4 presents selected examples of quantitative metrics used to evaluate semantic prompt variants across different prompting strategies in the description task. We applied min-max scaling and computed a final aggregated score for each variant to facilitate comparison across semantic variants (see Methods section for details). It is worth noting that the aggregated score should be interpreted as a screening heuristic rather than a definitive measure of output quality, given its assumption of equal metric weighting. We therefore complemented this quantitative screening with a manual review of a representative subset of outputs (Supporting Information: Appendix C) to verify that the highest-scoring prompts were indeed producing more informative and behaviourally grounded descriptions, rather than simply longer or more lexically diverse ones. Only a subset of the results is shown here for illustrative purposes. The complete table, including all six prompting strategies and both description and analysis tasks, is provided in Supporting Information: Appendices E and F. As shown in the table, prompt phrasing substantially influenced model output. For instance, within the tone strategy, average output length ranged from 80 words (C2) to 217 words (C1), while TTR values fluctuated between 5.4 and 8.5, highlighting how word choice impacts model behaviour. Based on these metrics, we selected optimal variants (indicated in bold) that consistently generated detailed and diverse outputs for each prompt strategy.

Figures 3 and 4 present the average LLM-based evaluation scores for six prompting strategies across four dimensions: *behavioural coverage*, *educational relevance*, *detail and specificity*, and *goal alignment*. The highest score in each dimension is bolded for clarity. Figure 3 summarises results for the description task, while Figure 4 results scores for the analysis

task. Across both tasks, the role-assignment strategy consistently achieved the highest average scores. It was especially effective in generating outputs that were educationally relevant and well-aligned with the task goals, suggesting that assigning the model a clear role enhances the quality and focus of its responses. In contrast, the zero-shot strategy consistently received the lowest scores across all dimensions and tasks, showing the importance of intentional prompt design when applying MLLMs to educational video analysis. Interestingly, the chain-of-thought-inspired strategy, while performing only moderately well on the description task, showed stronger results in the analysis task. This suggests that reasoning-oriented prompts may be more effective for tasks that require interpretation or explanatory insight. This superior performance, as judged by GPT-4o, is further illuminated by our qualitative content analysis (Section 5.2). The analysis suggests that role-assignment prompts may have helped the model focus on educationally relevant actions and reduced irrelevant output, though not eliminating the risk of over-interpretation.

5.2 | Strengths and Limitations of MLLM-Generated Behavioural Descriptions

To address RQ2, we conducted a sentence-level qualitative content analysis of LLaVA's outputs. Three researchers coded 926 sentences from the optimal prompt outputs using the five-category scheme (reliable, plausible, over-interpretation, irrelevant, and hallucination). Table 5 presents the coding results for description and analysis task. Overall, LLaVA shows promise in documenting behaviours, as most sentences were coded as either reliable or plausible (65.1% for description, 59.5% for analysis). Hallucinations occurred in 10.6% of cases in the description task and 9.0% in the analysis task. These errors often reflected domain confusion, in which the model misapplied Minecraft-specific gaming knowledge to the WHIMC environment. For example, some outputs interpreted student actions through entertainment-oriented Minecraft frames, such as resource gathering or survival-related play, even when the clip showed students engaging with educational

LLM-Based Evaluation Scores Across Prompting Strategies for Description Task

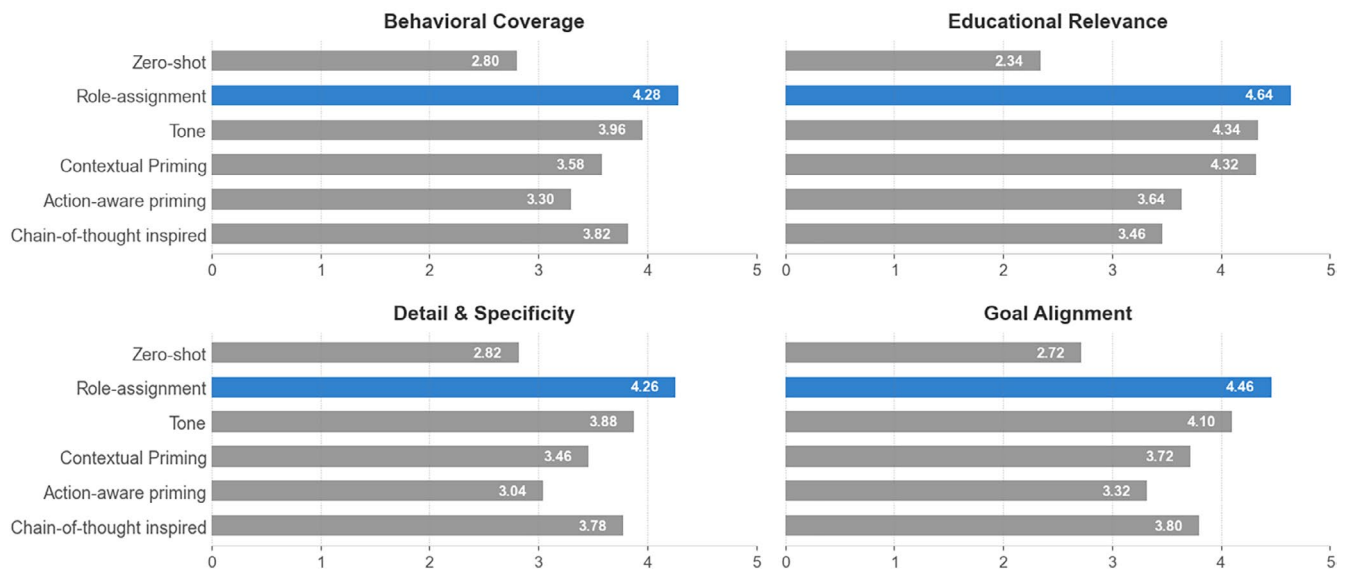


FIGURE 3 | LLM-based evaluation scores across prompting strategies for description task.

LLM-Based Evaluation Scores Across Prompting Strategies for Analysis Task

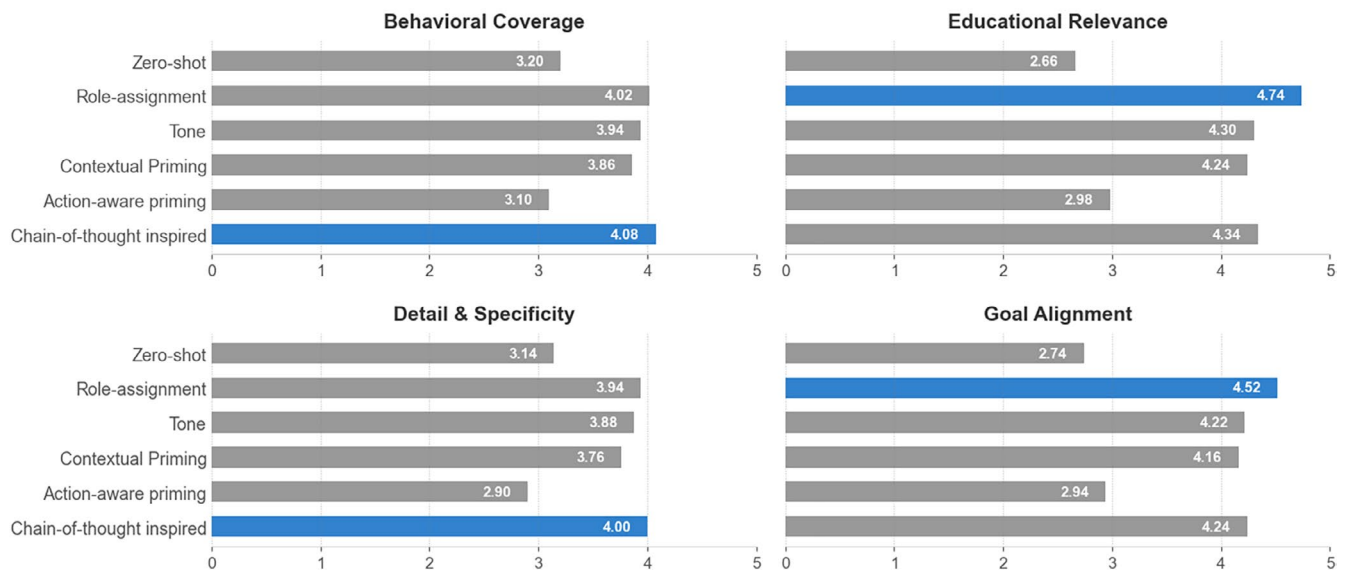


FIGURE 4 | LLM-based evaluation scores across prompting strategies for analysis task.

TABLE 5 | Count and proportion of each sentence-level code in LLaVA outputs.

Task	Reliable	Plausible	Over-interpretation	Irrelevance	Hallucination
Description	150(30.5%)	170(34.6%)	65(13.2%)	55(11.2%)	52(10.6%)
Analysis	111(25.6%)	147(33.9%)	115(26.5%)	22(5.1%)	39(9.0%)

tools or navigating the simulation. One output, for instance, stated that “the student is managing their inventory to collect resources for survival,” when the clip instead showed the student used a measurement tool designed for the specific WHIMC environment.

Differences between the two tasks emerge in the distribution of remaining codes. Analysis prompts led to higher proportion of over-interpretation (26.5% compared to 13.2% in description tasks), while reducing irrelevant content (5.1% vs. 11.2%). This suggests that while analysis prompt encourages the model to

focus more on target behaviours, it also increases the likelihood of unsupported or inflated conclusions about student's behaviours and motivations. For example, LLaVA sometimes interpreted simple actions like basic command usage as evidence of 'an interest in programming and coding, which are essential skills in STEM fields'. Similarly, student teleporting to a different location was interpreted as 'understanding of spatial awareness and optimization, which are critical in fields like engineering and architecture'. These over-interpretations can be more problematic compared to irrelevance. Irrelevant information, such as descriptions of game design, character appearance, or health bar, can be easily identified and filtered out since it clearly falls outside the scope of behaviour. However, over-interpretation is more difficult to detect as it often appears superficially plausible and requires domain knowledge about WHIMC to identify.

To better understand which student behaviours LLaVA could capture consistently, we conducted a thematic review of sentences coded as *reliable information* and identified recurring behavioural patterns. Table 6 presents the most common patterns with representative examples. Each pattern contributes a different kind of educational insight. For example, Social Interaction captures moments of peer engagement and provides evidence of how students interact with each other during gameplay. Exploration Behaviours document how students navigate the environment, what they attend to, and how they search for information in an open-ended environment, which allows researchers to distinguish between purposeful and unstructured exploration. Biome Description records which environmental zones students visit and offers contextual information about the spatial scope of their exploration, including whether they engage with a diverse range of biomes or remain within a limited area. Game Tool Usage captures instances in which students engage in deliberate data collection and measurement, both of which contribute meaningfully to forming and testing explanations in STEM inquiry.

6 | Discussion

6.1 | Potential and Limitations of LLaVA for Analysing Student Behaviours

6.1.1 | Promising Capabilities in Low-Level Behaviour Documentation

LLaVA proved effective at documenting low-level student behaviours. In this study, we define low-level behaviours as observable and surface-level actions that typically require minimal inferencing to describe, such as biomes explored, building visits, or interactions with NPCs and other students (see examples in Table 6). These descriptions offer rich behavioural insights that may not be easily captured in log data. For example, while log data tracks a player's timestamped coordinates, it does not reveal *how* students approach the location. A student who frequently adjusts headings while running may be exploring surroundings, whereas one who moves straight toward a destination may already be familiar with it or following a specific goal. These subtle differences in movement style or visual attention provide insights into students'

TABLE 6 | Emerging patterns and their exemplary sentences from LLaVA output.

Identified patterns	Example sentences from LLaVA outputs
Social interaction	The student then moves toward the structure, climbing up the black blocks and reaching the top where they encounter another player. The student appears to be engaged in a simulation of exploration and inquiry, as they are moving toward two other characters.
Exploration behaviours	They move cautiously, their eyes scanning the terrain for any signs of adventure. Upon entering the area, Tanya interacts with the environment by examining the plants closely.
Biome description	The student then descends from the hill, moving toward the edge of the plain where the terrain meets the ocean. For instance, the player notices a large structure resembling a rocket or space shuttle.
Game tool usage (observation or measurements)	They seem to be observing and recording data, as evidenced by the text overlays that provide information about air flow and wind speed. The player places observations about the game environment, such as noting the growth of random plants.

underlying cognitive processes that are not discernible from logs but can be inferred through video-based analysis. By reliably generating such descriptions, MLLMs outputs could serve as a starting point for further inquiry. Researchers could use these outputs to identify and compare exploration strategies, track shifts in student engagement over time, or investigate how learners interact with specific features of the WHIMC environments. This highlights the potential of video-based MLLM analysis not only to complement log-based approaches but also to expand the scope of behavioural analytics in digital learning environments.

6.1.2 | Limitations in Analytical Tasks

Despite its promise in detail documentation, LLaVA struggles with high-level reasoning, particularly in pattern recognition and interpretation. We observed that LLaVA frequently defaults to broad descriptions like 'active engagement' or

‘curiosity-driven exploration’ across most analysed clips. One possible explanation is that the short clip length limits the model’s ability to detect meaningful differences in student behaviour. Another reason is LLaVA’s insufficient domain knowledge about WHIMC, resulting in overly broad interpretations that can apply indiscriminately to nearly every clip. More critically, we found a pattern of over-interpretation, especially when the model was prompted to analyse student behaviours. In several cases, LLaVA drew implausible conclusions that inflated the significance of minor behaviours without clear evidence. This problem of over-interpretation reflects a broader limitation of using a general-purpose MLLM to interpret situated educational activity. The model can often identify visible behaviour, but it cannot reliably interpret what that behaviour means in relation to the task, the learning goals, or the pedagogical context (Mathur et al. 2025). As a result, behaviours that were observable but ambiguous in meaning were at times presented as evidence of STEM reasoning, curiosity, or strategic engagement without sufficient support from the clip. This points to a fundamental challenge in using general-purpose AI for education analysis: the model does not simply record activity but interprets it through patterns learned from training data that lack the pedagogical norms, task structures, and domain knowledge required for appropriate judgements (Bai et al. 2024; Mathur et al. 2025). Without careful oversight, such over-interpretations can lead to inaccurate conclusions about students, thereby compromising the validity of LLaVA outputs.

6.1.3 | Hallucination as Behaviour Misclassification

Another issue is hallucination, where the model generates content inconsistent with visual evidence (Bai et al. 2024; Liu, Xue, et al. 2024; Liu, Zhang, et al. 2024). Notably, these hallucinations manifest primarily as misclassifications rather than complete fabrications. For instance, the model occasionally misclassifies student actions in WHIMC by mapping them to Minecraft gaming behaviours (e.g., picking up crops or resource gathering for survival). These hallucinations may occur due to LLaVA’s training data, which includes public, entertainment-focused Minecraft gameplay recordings. As a result, the model fails to distinguish between gameplay mechanics and learning activities. This misalignment represents a fundamental challenge when employing general-purpose GenAI models to educational research: they are not designed with contextual information required and therefore often lack the specialised knowledge to generate accurate, domain-specific outputs (Dempere et al. 2023; Nguyen et al. 2024).

6.1.4 | Occasional Omissions

LLaVA occasionally omits key information in the screen capture, with such omissions appearing in approximately 5% of the analysed clips. It fails to record when a student uses a thermometer to measure temperature or when a student nods toward another player. The former omission is relatively minor since it can be inferred from log data, but the latter is more significant, as video is the most reliable way to capture such interactions. This omission may stem from insufficient prompt specificity;

explicitly specifying relevant instruments may better direct the model’s attention to such details. Future research should therefore further investigate prompting strategies to improve the reliability and completeness of documentation.

6.2 | Implications for AI-Assisted Video Analysis

Our findings highlight several practical and pedagogical implications for integrating MLLMs into educational video analysis workflows. First, MLLMs can streamline initial phases of video analysis, particularly in video selection and preliminary annotation. Rather than manually reviewing large volumes of footage (Derry et al. 2010; Fazeli et al. 2023), researchers and educators can use model-generated outputs to identify clips relevant to their instructional or research objectives. For example, educators could efficiently locate clips featuring peer interaction or tool usage by searching through MLLM-generated descriptions rather than by reviewing hours of recordings. This automated initial screening can significantly reduce the manual workload and enable educators and researchers to dedicate more resources toward deeper, interpretive analysis. Similarly, MLLMs could facilitate video annotation and transcription, which involves adding descriptive, interpretive, or analytic metadata and records to the selected video segments (Derry et al. 2010). MLLM-generated descriptions may provide a useful first layer of behavioural documentation that researchers can review and build on in subsequent analysis. Using these records, researchers could apply LLMs to identify patterns, trends, or emerging themes (Xiao et al. 2023; Yan et al. 2024; Zambrano et al. 2023). These behavioural annotations could also inform stealth assessments (Shute et al. 2021) and adaptations (Hooshyar et al. 2018; Vanbecelaere et al. 2020), enabling educators to better understand and respond to student needs in near-real-time contexts, or differentiate instruction based on specific behavioural indicators. In this sense, MLLMs are best understood as tools that help redistribute labour within the research process. They reduce the burden of exhaustive initial review and allow researchers to focus more on deeper interpretive work.

However, the successful implementation of MLLMs depends on prompt design. Our comparative analysis reveals that prompt design functions at two levels: prompting strategy and phrasing. At the strategy level, the strong performance of the role-assignment strategy is consistent with prior work demonstrating that role framing can constrain model behaviour and promote more focused and coherent responses (Arlinghaus et al. 2024). At the phrasing level, our results highlighted the prompt sensitivity of MLLMs. This aligns with research on paraphrasing-based augmentation (Jiang et al. 2020; Liu et al. 2023) and broader work documenting the sensitivity of MLLMs to small linguistic changes (Xu et al. 2023). This sensitivity raises concern for replicability and consistency of MLLM-generated outputs, as small adjustments to phrasing (e.g., ‘describe’ vs. ‘explain’, ‘analyse’ vs. ‘examine’) may lead to unintended shifts in level of detail or analytical depth. These findings suggest that MLLMs should not be treated as transparent windows into student behaviour. The prompt shapes which aspects of student activity the model attends to and what kinds of inferences it draws. It therefore directly influences how behaviour is described and which meanings appear in the output. In our study, role-assignment prompts

did not simply improve performance in a general sense (Kong et al. 2024). They orient the model toward more educationally relevant behaviours, while not eliminating the risks of over-interpretation and domain confusion. MLLMs are therefore better understood as interpretive lenses than as neutral recorders: different prompts function like different lenses, foregrounding some aspects of activity while backgrounding or distorting others.

From this perspective, prompt engineering is not merely a technical step for improving output quality. It also has been argued to function as a form of methodological transparency (Zhang et al. 2025). Since MLLM outputs are sensitive to role framing, wording, and task formulation (Arlinghaus et al. 2024; Jiang et al. 2020; Liu et al. 2023; Xu et al. 2023), documenting prompt design decisions (e.g., which strategies were tested, how outputs shifted in response to different phrasings, and where the model consistently over- or under-interpreted) helps make the model's interpretive tendencies visible. This practice makes AI-assisted interpretation more systematic, traceable, and open to scrutiny by other researchers. Accordingly, educators and researchers should adopt systematic prompt development (and documentation) practices—such as iterative prompt refinement processes and the use of standardised prompt templates—to enhance output reliability and consistency. Future work should investigate how specific linguistic elements influence model outputs and whether standardised prompt templates (i.e., meta-prompts) that fix core structure while allowing limited contextual variation can offer a stable foundation.

Although this study uses screen recordings from a Minecraft-based STEM environment, these findings have broader implications for other digital learning environments, such as virtual laboratories, simulations, and VR training systems. The challenge identified in this study (i.e., overinterpretation or misinterpretation) is likely to persist across platforms. For instance, in a virtual chemistry lab, an MLLM might accurately describe a student 'mixing red and blue liquids' (low-level description) but fail to distinguish whether this action represents a systematic experiment or a random trial (high-level interpretation) without specific context. Therefore, the methodological pipeline introduced here provides a transferable framework for researchers seeking to leverage MLLMs for other educational domains. Overall, our findings suggest that while MLLMs can support video annotation and transcription, their effective use requires careful prompt engineering, human oversight, and professional training to ensure responsible implementation.

6.3 | Limitations and Future Work

As one of the earliest attempts to leverage MLLMs for educational video analysis, our work has several limitations. First, the decision to segment recordings into 10-s clips, while suitable for examining LLaVA's capacity to document low-level behaviours, may create artificial boundaries that obscure strategies developing across clips. This might result in the repetitive patterns we observed. Increasing clip duration might mitigate these gaps by allowing the model to observe longer behavioural sequences, yet at the cost of the possible risk of missing momentary, yet

meaningful events (e.g., brief NPC dialogues) due to coarse video processing granularity. In future studies, we plan to experiment with longer clips and coarser frame rates to see if this will allow us to capture more diverse strategies. Second, our evaluation framework has limitations that should be addressed in future research. The assumption that longer outputs usually contain more valuable information may not always hold and require empirical validation. Future research should explore other types of metrics to evaluate the open-ended text of MLLMs. Future work should also investigate how domain-specific contextual information can be incorporated more systematically into prompts. Many student actions in educational videos are observable, but their meaning remains ambiguous unless interpreted in relation to the learning goals and task structure. Providing richer contextual guidance may help reduce unsupported inferences and improve the educational relevance of model-generated interpretations.

Another important direction concerns the design of tools for human–AI collaboration. These tools should help researchers evaluate how well claims that AI generates align with the actual video. Future systems could link each claim to specific frames or visual cues and indicate when an output carries greater uncertainty or relies on interpretation. Such designs could support more transparent AI-assisted analysis and invite closer researcher scrutiny, rather than encouraging passive acceptance of model-generated descriptions.

7 | Conclusion

This study presents an early exploration of how MLLMs may support qualitative video analysis in a Minecraft-based STEM environment. Regarding RQ1, our systematic comparison of semantic variants and prompting strategies demonstrates that prompt design substantially shapes MLLM output quality. Specifically, the role-assignment strategy consistently yielded the most educationally relevant and goal-aligned responses, supporting prior work on role framing as an effective mechanism to align model attention with educational constructs. The observed prompt sensitivity, where minor phrasing variations altered output quality, underscores the necessity of systematic prompt engineering to ensure consistent outputs. Regarding RQ2, our content analysis reveals that LLaVA can generate reliable descriptions of low-level student behaviours such as navigation patterns, biome exploration, and interaction with game elements. These behaviours, which are often difficult to extract from log data alone, offer insights into students' engagement in open-ended learning environments. However, the model struggled with higher-level interpretations, often defaulting to vague and generic claims or producing over-interpretations that risk mischaracterizing student behaviours. These challenges highlight the importance of human oversight and domain-specific prompting when using MLLMs in educational contexts. Overall, this study shows that video-based MLLMs can contribute meaningfully to educational video analysis. They support scalable documentation of observable student behaviour that would be difficult to capture through manual coding alone. At the same time, they struggle with higher-level interpretation, especially when a behaviour is ambiguous and interpretation of its meaning depends on task and pedagogical context. These findings

suggest that the value of MLLMs in qualitative educational research lies in how they reshape the research process. Rather than functioning as independent analysts, MLLMs operate more effectively as constrained analytic collaborators. They can extend the scale of observation, yet they also highlight the need for contextual judgement, methodological transparency, and critical scrutiny. In this sense, their contribution does not simply rest on increased efficiency in qualitative inquiry, but in moving the field toward more explicit, inspectable, and rigorous forms of AI-assisted analysis.

Author Contributions

Amanda Barany: writing – review and editing, project administration, supervision. **Andres Felipe Zambrano:** methodology, validation, writing – review and editing, formal analysis. **Jaclyn Ocumpaugh:** writing – review and editing, project administration, supervision. **Ryan S. Baker:** writing – review and editing, project administration, supervision. **Xiner Liu:** conceptualization, formal analysis, methodology, validation, writing – review and editing. **Yiqiu Zhou:** conceptualization, writing – original draft, methodology, formal analysis, writing – review and editing, validation, visualization. **Luc Paquette:** conceptualization, writing – review and editing, methodology. **Jeff Ginger:** writing – review and editing, project administration, supervision.

Acknowledgements

This work was supported by the National Science Foundation (NSF; DRL-2301172). Any conclusions expressed in this material do not necessarily reflect the views of the NSF. The authors acknowledge the assistance of ChatGPT, a large language model, in the editing and proof-reading process of this manuscript.

Funding

This research was funded by the National Science Foundation (NSF; DRL-2301172). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

Ethics Statement

This study was reviewed and approved by the Institutional Review Board (IRB) at the University of Illinois Urbana-Champaign. All participants provided informed consent prior to participation in the study, in accordance with the ethical standards of the IRB.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Peer Review

For transparency, the peer review documents associated with this article are available at <https://doi.org/10.1002/jcal.70269>.

References

Akoury, N., S. Wang, J. Whiting, S. Hood, N. Peng, and M. Iyyer. 2020. “STORIUM: A Dataset and Evaluation Platform for

Machine-In-The-Loop Story Generation.” <https://doi.org/10.48550/arXiv.2010.01717>.

Arlinghaus, C. S., C. Wulff, and G. W. Maier. 2024. “Inductive Coding With ChatGPT—An Evaluation of Different GPT Models Clustering Qualitative Data Into Categories.” <https://doi.org/10.31219/osf.io/gpnye>.

Bai, Z., P. Wang, T. Xiao, et al. 2024. “Hallucination of Multimodal Large Language Models: A Survey.” arXiv:2404.18930. <http://arxiv.org/abs/2404.18930>.

Barany, A., N. Nasir, C. Porter, et al. 2024. “ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development.” In *Artificial Intelligence in Education*, edited by A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, 134–149. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_10.

Becker, J., J. P. Wahle, B. Gipp, and T. Ruas. 2024. “Text Generation: A Systematic Literature Review of Tasks, Evaluation, and Challenges.” <https://doi.org/10.48550/arXiv.2405.15604>.

Bewersdorff, A., C. Hartmann, M. Hornberger, et al. 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: 102601.

Blikstad-Balas, M. 2017. “Key Challenges of Using Video When Investigating Social Practices in Education: Contextualization, Magnification, and Representation.” *International Journal of Research & Method in Education* 40, no. 5: 511–523. <https://doi.org/10.1080/1743727X.2016.1181162>.

Calandra, B., and P. J. Rich. 2014. *Digital Video for Teacher Education*. Routledge. <https://doi.org/10.4324/9781315871714>.

Chen, B., Z. Zhang, N. Langrené, and S. Zhu. 2024. “Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review.” arXiv:2310.14735; Version 5. <http://arxiv.org/abs/2310.14735>.

Chen, D., R. Chen, S. Zhang, et al. 2024. “MLLM-As-a-Judge: Assessing Multimodal LLM-As-a-Judge With Vision-Language Benchmark.” arXiv:2402.04788. <http://arxiv.org/abs/2402.04788>.

Chung, J., E. Kamar, and S. Amershi. 2023. “Increasing Diversity While Maintaining Accuracy: Text Data Generation With Large Language Models and Human Interventions.” In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, edited by A. Rogers, J. Boyd-Graber, and N. Okazaki, vol. 1, 575–593. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.34>.

Cukurova, M., Q. Zhou, D. Spikol, and L. Landolfi. 2020. “Modelling Collaborative Problem-Solving Competence With Transparent Learning Analytics: Is Video Data Enough?” *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*: 270–275. <https://doi.org/10.1145/3375462.3375484>.

DeLiema, D., A. Hufnagle, V. N. V. Rao, J. Baker, J. Valerie, and J. Kim. 2023. “Methodological Innovations at the Intersection of Video-Based Educational Research Traditions: Reflections on Relevance, Data Selection, and Phenomena of Interest.” *International Journal of Research & Method in Education* 46, no. 1: 19–36. <https://doi.org/10.1080/1743727X.2021.2011196>.

Dempere, J., K. Modugu, A. Hesham, and L. K. Ramasamy. 2023. “The Impact of ChatGPT on Higher Education.” *Frontiers in Education* 8: 1206936. <https://doi.org/10.3389/feduc.2023.1206936>.

Derry, S. J., R. D. Pea, B. Barron, et al. 2010. “Conducting Video Research in the Learning Sciences: Guidance on Selection, Analysis, Technology, and Ethics.” *Journal of the Learning Sciences* 19, no. 1: 3–53. <https://doi.org/10.1080/10508400903452884>.

Evers, J. C. 2011. “From the Past Into the Future. How Technological Developments Change Our Ways of Data Collection, Transcription

- and Analysis." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12, no. 1: 1. <https://doi.org/10.17169/fqs-12.1.1636>.
- Fazeli, S., J. Sabetti, and M. Ferrari. 2023. "Performing Qualitative Content Analysis of Video Data in Social Sciences and Medicine: The Visual-Verbal Video Analysis Method." *International Journal of Qualitative Methods* 22: 16094069231185452. <https://doi.org/10.1177/16094069231185452>.
- Fu, C., Y. Dai, Y. Luo, et al. 2024. "Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis." arXiv:2405.21075. <http://arxiv.org/abs/2405.21075>.
- Gehrmann, S., E. Clark, and T. Sellam. 2022. "Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text." <https://doi.org/10.48550/arXiv.2202.06935>.
- Gu, J., Q. Wu, and Z. Yu. 2020. "Perception Score, A Learned Metric for Open-Ended Text Generation Evaluation." <https://doi.org/10.48550/arXiv.2008.03082>.
- Heath, C., J. Hindmarsh, and P. Luff. 2010. *Video in Qualitative Research*. Sage Publications.
- Hooshyar, D., M. Yousefi, M. Wang, and H. Lim. 2018. "A Data-Driven Procedural-Content-Generation Approach for Educational Games." *Journal of Computer Assisted Learning* 34, no. 6: 731–739. <https://doi.org/10.1111/jcal.12280>.
- Huang, J., and J. Zhang. 2024. "A Survey on Evaluation of Multimodal Large Language Models." <https://doi.org/10.48550/arXiv.2408.15769>.
- Jarvis, S. 2013. "Capturing the Diversity in Lexical Diversity." *Language Learning* 63, no. s1: 87–106. <https://doi.org/10.1111/j.1467-9922.2012.00739.x>.
- Jiang, Z., F. F. Xu, J. Araki, and G. Neubig. 2020. "How Can we Know What Language Models Know?" *Transactions of the Association for Computational Linguistics* 8: 423–438. https://doi.org/10.1162/tac1_a_00324.
- Kirsten, E., A. Buckmann, A. Mhaidli, and S. Becker. 2024. "Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding." <https://doi.org/10.48550/arXiv.2403.06607>.
- Kong, A., S. Zhao, H. Chen, et al. 2024. "Better Zero-Shot Reasoning With Role-Play Prompting." <https://doi.org/10.48550/arXiv.2308.07702>.
- Lane, H. C., S. Yi, B. Guerrero, and N. Comins. 2017. "Minecraft as a Sandbox for STEM Interest Development: Preliminary Results."
- Liu, H., W. Xue, Y. Chen, et al. 2024. "A Survey on Hallucination in Large Vision-Language Models." arXiv:2402.00253. <http://arxiv.org/abs/2402.00253>.
- Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Computing Surveys* 55, no. 9: 1–35. <https://doi.org/10.1145/3560815>.
- Liu, X., J. Zhang, A. Barany, M. Pankiewicz, and R. S. Baker. 2024. "Assessing the Potential and Limits of Large Language Models in Qualitative Coding." In *Advances in Quantitative Ethnography*, edited by Y. J. Kim and Z. Swiecki, vol. 2278, 89–103. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76335-9_7.
- Magana, A. J., J. Hwang, S. Feng, S. Rebello, T. Zu, and D. Kao. 2022. "Emotional and Cognitive Effects of Learning With Computer Simulations and Computer Videogames." *Journal of Computer Assisted Learning* 38, no. 3: 875–891. <https://doi.org/10.1111/jcal.12654>.
- Marvin, G., N. Hellen, D. Jjingo, and J. Nakatumba-Nabende. 2023. "Prompt Engineering in Large Language Models." In *International Conference on Data Intelligence and Cognitive Informatics*, 387–402. Springer Nature Singapore. <https://doi.org/10.22318/icds2025.704957>.
- Mathur, L., M. Qian, P. P. Liang, and L.-P. Morency. 2025. "Social Genome: Grounded Social Reasoning Abilities of Multimodal Models." <https://doi.org/10.48550/arXiv.2502.15109>.
- Nguyen, H., and J. Hayward. 2024. "Applying Generative Artificial Intelligence to Critiquing Science Assessments." *Journal of Science Education and Technology* 34: 199–214. <https://doi.org/10.1007/s10956-024-10177-x>.
- Nguyen, H., V. Nguyen, S. Ludovise, and R. Santagata. 2024. "Misrepresentation or Inclusion: Promises of Generative Artificial Intelligence in Climate Change Education." *Learning, Media and Technology* 50, no. 3: 393–409. <https://doi.org/10.1080/17439884.2024.2435834>.
- Pea, R. D., and E. Hoffert. 2014. "Video Workflow in the Learning Sciences: Prospect of Emerging Technologies for Argmenting Work Practices." In *Video Workflow in the Learning Sciences*, 427–460. Lawrence Erlbaum Associates.
- Qi, S., Z. Cao, J. Rao, L. Wang, J. Xiao, and X. Wang. 2023. "What Is the Limitation of Multimodal LLMs? A Deeper Look Into Multimodal LLMs Through Prompt Probing." *Information Processing & Management* 60, no. 6: 103510. <https://doi.org/10.1016/j.ipm.2023.103510>.
- Ramey, K. E., D. N. Champion, E. B. Dyer, et al. 2016. "Qualitative Analysis of Video Data: Standards and Heuristics."
- Rowe, E., J. Asbell-Clarke, and R. S. Baker. 2015. "Serious Games Analytics to Measure Implicit Science Learning." In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, edited by C. S. Loh, Y. Sheng, and D. Ifenthaler, 343–360. Springer International Publishing. https://doi.org/10.1007/978-3-319-05834-4_15.
- Rowe, E., J. Asbell-Clarke, E. Bardar, et al. 2020. "Advancing Research in Game-Based Learning Assessment: Tools and Methods for Measuring Implicit Learning." In *Advances in Educational Technologies and Instructional Design*, edited by E. Kennedy and Y. Qian, 99–123. IGI Global. <https://doi.org/10.4018/978-1-7998-1173-2.ch006>.
- Sahoo, P., A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. 2024. "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications." arXiv:2402.07927; Version 1. <http://arxiv.org/abs/2402.07927>.
- Schmidtová, P., S. Mahamood, S. Balloccu, et al. 2024. "Automatic Metrics in Natural Language Generation: A Survey of Current Evaluation Practices." <https://doi.org/10.48550/arXiv.2408.09169>.
- Shute, V., S. Rahimi, G. Smith, et al. 2021. "Maximizing Learning Without Sacrificing the Fun: Stealth Assessment, Adaptivity and Learning Supports in Educational Games." *Journal of Computer Assisted Learning* 37, no. 1: 127–141. <https://doi.org/10.1111/jcal.12473>.
- Squire, K. 2005. "Changing the Game: What Happens When Video Games Enter the Classroom?" *Innovate: Journal of Online Education* 1, no. 6: 5.
- Stoyanov, S., and P. A. Kirschner. 2023. "Text Analytics for Uncovering Untapped Ideas at the Intersection of Learning Design and Learning Analytics: Critical Interpretative Synthesis." *Journal of Computer Assisted Learning* 39, no. 3: 899–920. <https://doi.org/10.1111/jcal.12775>.
- Tang, Y., J. Bi, S. Xu, et al. 2024. "Video Understanding With Large Language Models: A Survey." arXiv:2312.17432. <http://arxiv.org/abs/2312.17432>.
- Vanbecelaere, S., K. van den Berghe, F. Cornillie, D. Sasanguie, B. Reynvoet, and F. Depaepe. 2020. "The Effectiveness of Adaptive Versus Non-Adaptive Learning With Digital Educational Games." *Journal of Computer Assisted Learning* 36, no. 4: 502–513. <https://doi.org/10.1111/jcal.12416>.
- Wang, R. E., and D. Demszky. 2023. "Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance for Scoring and Providing Actionable Insights on Classroom Instruction." arXiv:2306.03090. <http://arxiv.org/abs/2306.03090>.
- Wang, S., Y. Song, A. Drozdov, A. Garimella, V. Manjunatha, and M. Iyyer. 2023. "KNN-LM Does Not Improve Open-Ended Text Generation." <https://doi.org/10.48550/arXiv.2305.14625>.

Wei, J., X. Wang, D. Schuurmans, et al. 2022. "Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models."

Whitehead, R., A. Nguyen, and S. Järvelä. 2024. "Exploring the Role of Gaze Behaviour in Socially Shared Regulation of Collaborative Learning in a Group Task." *Journal of Computer Assisted Learning* 40, no. 5: 2226–2247. <https://doi.org/10.1111/jcal.13022>.

Wideman, H. H., R. D. Owston, C. Brown, A. Kushniruk, F. Ho, and K. C. Pitts. 2007. "Unpacking the Potential of Educational Gaming: A New Tool for Gaming Research." *Simulation & Gaming* 38, no. 1: 10–30. <https://doi.org/10.1177/1046878106297650>.

Wu, J., W. Gan, Z. Chen, S. Wan, and P. S. Yu. 2023. "Multimodal Large Language Models: A Survey." *IEEE International Conference on Big Data (BigData)* 2023: 2247–2256. <https://doi.org/10.1109/BigData59044.2023.10386743>.

Xiao, Z., X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer. 2023. "Supporting Qualitative Analysis With Large Language Models: Combining Codebook With GPT-3 for Deductive Coding." In *28th International Conference on Intelligent User Interfaces*, 75–78. Association for Computing Machinery. <https://doi.org/10.1145/3581754.3584136>.

Xu, Z., Y. Shen, and L. Huang. 2023. "MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning." <https://doi.org/10.48550/arXiv.2212.10773>.

Yan, L., V. Echeverria, G. M. Fernandez-Nieto, et al. 2024. "Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations." <https://doi.org/10.1145/3613905.3650732>. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 1–7.

Yang, Z., L. Li, K. Lin, et al. 2023. "The Dawn of LMMs: Preliminary Explorations With GPT-4V(ision)." arXiv:2309.17421. <http://arxiv.org/abs/2309.17421>.

Yin, S., C. Fu, S. Zhao, et al. 2024. "A Survey on Multimodal Large Language Models." arXiv:2306.13549. <http://arxiv.org/abs/2306.13549>.

Zambrano, A. F., X. Liu, A. Barany, R. S. Baker, J. Kim, and N. Nasir. 2023. "From nCoder to ChatGPT: From Automated Coding to Refining Human Coding." In *Advances in Quantitative Ethnography*, edited by G. Arastoopour Irgens and S. Knight, 470–485. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47014-1_32.

Zhang, D., Y. Yu, J. Dong, et al. 2024. "MM-LLMs: Recent Advances in MultiModal Large Language Models." arXiv:2401.13601. <http://arxiv.org/abs/2401.13601>.

Zhang, H., C. Wu, J. Xie, Y. Lyu, J. Cai, and J. M. Carroll. 2025. "Harnessing the Power of AI in Qualitative Research: Exploring, Using and Redesigning ChatGPT." *Computers in Human Behavior: Artificial Humans* 4: 100144. <https://doi.org/10.1016/j.chbah.2025.100144>.

Zhang, Y., J. Wu, W. Li, et al. 2024. "Video Instruction Tuning With Synthetic Data." arXiv:2410.02713. <http://arxiv.org/abs/2410.02713>.

Zheng, L., W.-L. Chiang, Y. Sheng, et al. 2023. "Judging LLM-As-a-Judge With MT-Bench and Chatbot Arena." <https://doi.org/10.48550/arXiv.2306.05685>.

Zhou, Y., J. Kang, and H. Nguyen. 2025. "Can we Trust Large Language Models for Video Analysis: An Exploration of Hallucination in Multimodal LLMs." In *Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025*, 110–118. International Society of the Learning Sciences (ISLS). <https://10.22318/icls2025.704957>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix A** Complete list of LLaVA prompts used in the description task. **Appendix B**: Complete list of LLaVA prompts used in the analysis task. **Appendix C**: Sample output from Llava on a single video clip (output generated using the semantic

prompt that yielded the best performance in the evaluation phase is shown in bold). **Appendix D**: GPT-4o prompt used for LLaVA response evaluation. **Appendix E**: Quantitative metrics and aggregated scores for all prompt variants in the description task (metric corresponding to the highest-performing prompt variant is shown in bold). **Appendix F**. Quantitative metrics and aggregated scores for all prompt variants in the analysis task (metric corresponding to the highest-performing prompt variant is shown in bold).