Multimodal Large Language Models as a Catalyst for Advancing Learning Analytics in Early Childhood Education

Yuanyuan YANG[a*], Tianchen SUN[b], Yuan SHEN[c] & Yangbin XIE[a]
[a]*School of Smart Education, Jiangsu Normal University, China*
[b]*Research Center for Scientific Data Hub, Zhejiang Lab, China*
[c]*College of Education, Zhejiang University of Technology, China*
*44.yangoo@gmail.com

Abstract: Early childhood education (ECE) involves rich, informal, and multimodal learning processes that are difficult to assess with traditional methods. This paper presents a novel framework that integrates Multimodal Large Language Models (MLLMs) into ECE learning analytics to capture children's spontaneous expressions—such as drawings, speech, gestures, and social interactions—in a scalable, child-centered manner. The system includes multimodal data collection, MLLM-based feature extraction, automated developmental analytics, and educator-in-the-loop feedback. A real-world case from a rural kindergarten illustrates the framework's ability to generate interpretable indicators and actionable insights. We discuss opportunities for individualized assessment and developmentally appropriate practices, as well as challenges related to interpretability, privacy, and equity. This work demonstrates the potential of MLLMs to support holistic, play-based learning analytics in early childhood settings.

Keywords: Multimodal Large Language Models (MLLMs), Early Childhood Education (ECE), Learning Analytics (LA), Children Development, Generative AI (GenAI)

1. Introduction

Early childhood is a crucial period for cognitive, emotional, social, and motor development, forming the foundation for lifelong learning and well-being (Burns & Bowman, 2001). Yet, capturing the complexity and individuality of children's learning remains challenging (Crescenzi-Lanna, 2020). Traditional assessments—like standardized tests, checklists, and observations—offer only fragmented views and often lack scalability, contextual depth, and real-time responsiveness (Shepard, 2000).

Learning analytics offers a promising response by providing data-driven insights into learning (Siemens & Long, 2011), but its use in early childhood education (ECE) is still limited. Although multimodal data—such as speech, drawings, and interactions—can be collected non-invasively, analyzing such complex, unstructured input remains difficult (Crescenzi-Lanna, 2020; Blikstein & Worsley, 2016).

Recent advances in Multimodal Large Language Models (MLLMs) present new opportunities for integrating learning analytics into ECE (Alayrac et al., 2022; Tsimpoukelli et al., 2021). These models can interpret diverse data forms, enabling scalable, child-centered insights into developmental trajectories.

In this paper, we introduce an MLLM-powered learning analytics framework grounded in play-based and constructivist theories (Vygotsky, 1978; Piaget, 1951). Our framework bridges raw multimodal data and meaningful developmental insights by analyzing children's drawings, verbal narratives, and movement patterns to infer indicators across cognitive, social, emotional, and motor domains.

2. Related Work

2.1 Learning Analytics and Multimodal Approaches in Early Childhood Education

Learning analytics has traditionally focused on structured digital traces—such as grades, logins, and content access—primarily in higher education or online settings (Nguyen et al., 2022). In early childhood education (ECE), however, learning is embodied, exploratory, and multimodal, making these approaches less effective. Most ECE applications involve educational games with predefined outcomes and structured tasks, offering limited insight into spontaneous and contextual learning (Agus et al., 2018).

To overcome the limitations of traditional approaches, researchers have explored multimodal methods that capture the richness of early learning environments (Ochoa & Worsley, 2016). By integrating video, audio, and visual artifacts, these methods offer deeper insights into children's cognitive and social behaviors (Blikstein & Worsley, 2016). Techniques such as facial recognition (Ramakrishnan et al., 2019), speech analysis (Oviatt et al., 2018), and eye-tracking (Giannakos et al., 2019) provide more comprehensive data, but often depend on complex, rule-based systems that are hard to scale. A more scalable and pedagogy-aligned approach is still needed for multimodal learning analytics in ECE.

2.2 Advances in Multimodal Large Language Models

Multimodal Large Language Models (MLLMs), such as Flamingo (Alayrac et al., 2022), GPT-4 with vision (Achiam et al, 2023), and MM-REACT (Yang et al., 2023), are designed to interpret and integrate multiple data types including text, images, audio, and video. These models have demonstrated impressive performance in tasks ranging from visual question answering to reasoning about diagrams and understanding speech in context (Tsimpoukelli et al., 2021). Their ability to generalize across modalities with minimal fine-tuning opens new possibilities for analyzing the spontaneous and creative outputs of young children.

Although applications of MLLMs in education are still emerging, there are early signs of their potential. For instance, Bewersdorff et al. (2024) proposed a MLLM-based framework to support multimodal content creation and personalized feedback in science education. Xing et al. (2024) highlights MLLMs' applications in language learning, STEM education, and medical training. Lee et al. (2024) developed an MLLM-based conversational agent to support art appreciation education through interactive multimodal dialogue. However, integrating MLLMs into ECE settings remains largely unexplored—particularly in ways that align with child-centered pedagogies and developmental theory.

3. MLLM-Based Learning Analytics Framework in ECE

We propose a child-centered framework using MLLMs to interpret young children's spontaneous multimodal outputs. It moves beyond rule-based systems, offering real-time, developmentally appropriate insights from data like drawings, narratives, and play.

3.1 Theoretical Foundation

Learning analytics framework for ECE should be grounded in developmental theories that reflect how young children learn and express themselves. Constructivist theory emphasizes learning through active, multimodal engagement (Piaget, 1951), while socio-cultural theory highlights the role of social interaction and individualized support within the Zone of Proximal Development (Vygotsky, 1978). Our framework aligns with these views by leveraging MLLMs to interpret children's spontaneous outputs—such as drawings and narratives—without disrupting natural play. It supports developmentally appropriate practice (Copple & Bredekamp, 2009), emphasizing transparency, educator interpretation, and ethical safeguards. Rather than scoring or labeling, it aims to surface meaningful patterns that inform

responsive teaching and holistic assessment. This framework operationalizes constructivist and sociocultural theories by mapping child-generated multimodal data to interpretable features that reflect developmental domains and support educator-guided scaffolding.

3.2 Framework Overview

This framework consists of four components that collectively support the automated and interpretable analysis of young children's multimodal learning data, integrating AI-driven processing with educator input as Figure 1 shown. **Multimodal Data Collection and Storage** gathers various forms of child-generated data—including activity videos, audio, drawings, transcripts, and environmental information—and stores them in a centralized database for integrated access and analysis. **MLLM-supported Feature Construction and Extraction** employs a Multimodal Large Language Model (MLLM) to generate interpretable feature schemas and structured feature matrices. These features are derived from multimodal inputs using MLLM-based prompts that extract values such as peer mentions and semantic alignment, verified through educator feedback. **Automated Analytics and Visualization** quantifies developmental indicators using clustering and similarity metrics, with results shown through interactive dashboards (e.g., radar and network diagrams) for longitudinal and group-level analysis. **MLLM-supported Feedback** generates constructive suggestions for children's development. Educators interact with the system to review AI-generated insights, contribute feedback, and apply the results in practice, completing the human-in-the-loop cycle.

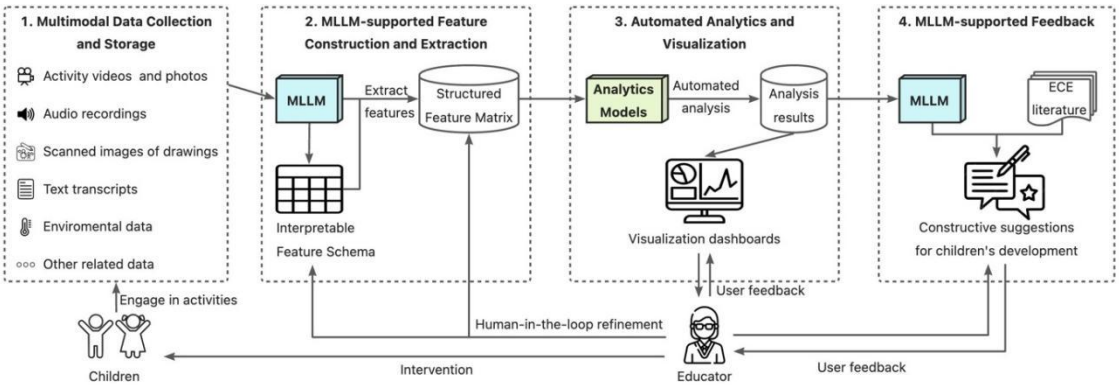


Figure 1. Overview of the MLLM-Based Learning Analytics Framework in ECE

3.3 Case Illustration

We collected data from a kindergarten class of 29 children, where children engaged in daily one-hour free play, then drew and described their experiences. Narratives were transcribed and corrected by their teacher. ChatGPT was used to generate a feature table (see example in Table 1) and analyze each child's drawing and narration (see Figure 2). In addition, we presented a social network and a radar chart displaying multiple developmental domains based on children's data over one semester, along with AI-generated suggestions based on results, as shown in Figure 3. All necessary consents were obtained from their parents and teachers, with data anonymized and stored securely under institutional ethical approval.

Table 1. Parts of *AI-generated Interpretable Feature Schema*

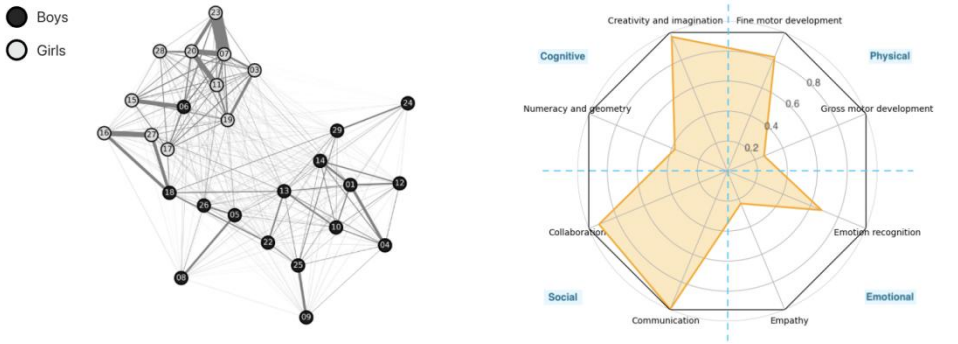
Feature Name	Data Source	Description	Example
Mentioned Peer Count	Narrative text	Number of peers mentioned in the narration	I played with A and B" → 2 peers
Number of Characters	Drawing	Total number of human figures in the drawing	1 (solo) vs 4 (group)
Text–Image Alignment Score	Narrative text & Drawing	Semantic consistency between narration and	Said "we built a castle" + drew a castle

drawing

Child's narrative: *Today we played a water maze. We used a small ball and rolled it to create a maze. Then we used pieces from a bag to guide it, and we treated that as our maze. At first, we started with something simple—a straight line. Sometimes there were obstacles. I felt that this water maze was starting to get a bit tricky. As we kept playing, I started to feel the ball wasn't very fun anymore, so we switched to a building block that could roll—a cylinder...*

Feature	Value	Reasoning
Mentioned Peer Count	1	Mentions "he/him" multiple times, suggesting at least one peer is involved
Number of Characters	2	Two human figures drawn
Text–Image Alignment Score	High	Narration describes a water maze, and the drawing clearly represents it
...	...	...

Figure 2. An Example of Child's Data from Real-world and AI-generated Result



Observation: Peer interactions are mostly gender-segregated, except for Students #06 and #18.

AI Suggestion:

1. Use mixed-gender activities like building or role-play to promote collaboration.
2. Pair children across genders during free play to encourage inclusive interaction.

Observation: This domain shows one of the lowest scores in the cognitive area.

AI suggestion:

1. Introduce more play-based math activities, such as block sorting, counting games, and pattern-based puzzles.
2. Use everyday contexts (e.g., snacks, classroom objects) to embed basic math talk during routines.

Figure 3. An Example of Data Visualization and AI Suggestions

4. Opportunities, Challenges and Future Directions

The integration of MLLMs into early childhood learning analytics presents promising opportunities. By interpreting children's multimodal expressions—such as drawings, speech, and social interactions—these models offer a scalable, child-centered means to infer developmental trajectories. Such capabilities can enhance formative assessment, support individualized instruction, and reduce the burden on educators by generating actionable insights grounded in children's natural play.

However, several challenges must be addressed to ensure responsible and effective deployment. First, MLLMs need to adapt to the often ambiguous and diverse nature of young children's communication. We suggest co-designing prompt templates with educators, enhancing both interpretability and contextual relevance. Second, the lack of transparency in model outputs may hinder educator trust and application. Embedding educator feedback in the loop can improve clarity and alignment with pedagogical needs. Third, ethical concerns around privacy, consent, and cultural sensitivity remain critical. All data used in our framework is anonymized, locally stored, and encrypted to ensure confidentiality. To promote equity, future work will incorporate more culturally and linguistically diverse datasets, supporting inclusive and fair analytics across early learning contexts.

Acknowledgements

We acknowledge the financial support from the China Postdoctoral Science Foundation. We would like to express our sincere gratitude to all the teachers and children at Anji Kindergarten for their invaluable participation in this study.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv *preprint arXiv:2303.08774*.

Agus, R., & Samuri, S. M. (2018). Learning analytics contribution in education and child development: A review on learning analytics. *Asian journal of assessment in teaching and learning*, *8*, 36-47.

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., ... & 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*, 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.

Burns, M. S., Donovan, M. S., & Bowman, B. T. (Eds.). (2001). *Eager to learn: Educating our preschoolers*. National Academies Press.

Copple, C., & Bredekamp, S. (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8*. National Association for the Education of Young Children. 1313 L Street NW Suite 500, Washington, DC 22205-4101.

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

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. International Journal of Information Management, 48, 108-119.

Lee, U., Jeon, M., Lee, Y., Byun, G., Son, Y., Shin, J., ... & Kim, H. (2024). LLaVA-docent: Instruction tuning with multimodal large language model to support art appreciation education. *Computers and Education: Artificial Intelligence*, *7*, 100297.

Nguyen, Q., Rienties, B., & Whitelock, D. (2022). Informing learning design in online education using learning analytics of student engagement. *Open world learning: research, innovation and the challenges of high-quality education*, 189-207.

Piaget, J. (2013). *Play, dreams and imitation in childhood*. Routledge.

Ramakrishnan, A., Ottmar, E., LoCasale-Crouch, J., & Whitehill, J. (2019). Toward automated classroom observation: Predicting positive and negative climate. 14th IEEE Int. In *Conf. on Automatic Face and Gesture Recognition (FG 2019), Lille*.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, *29*(7), 4-14.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, *46*(5), 30.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S. M., Vinyals, O., & Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, *34*, 200-212.

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data . *Journal of Learning Analytics*, *3*(2), 213-219.

Oviatt, S., Grafsgaard, J., Chen, L., & Ochoa, X. (2018). Multimodal learning analytics: Assessing learners' mental state during the process of learning. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2* (pp. 331-374).

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

Xing, W., Zhu, T., Wang, J., & Liu, B. (2024). A Survey on MLLMs in Education: Application and Future Directions. *Future Internet*.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., ... & Wang, L. (2023). Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.