Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL36908

Research on Multimodal Behavior Representation Modeling Method in Kindergarten Free Play Context

Weili Yan

School of Education, Jilin International Studies University, Changchun, China
wellington589125@gmail.com

Abstract. Kindergarten free play is an important context in which children's social interaction, emotional expression, material exploration, and problem solving occur naturally. However, its behavioral process is open, continuous, and highly interactive. Traditional observation records are therefore limited in capturing the dynamic relations among movement, language, spatial position, and peer interaction. To address this issue, a multimodal behavior representation modeling method is constructed for kindergarten free play contexts. It combines public video dataset pretraining, small scale kindergarten video collection, area audio recording, and teacher event annotation. Pose estimation, voice activity analysis, behavior coding, and temporal fusion are used to identify children's behavioral states. The model takes a 10 second sliding window as the basic unit. It extracts visual, audio, and event annotation features, then uses a BiLSTM or Transformer encoder to analyze temporal relations in behavioral sequences. Preliminary results show that the fusion of video, audio, and teacher annotation achieves an Accuracy of 0.842 and a Macro F1 of 0.817, which is better than single modality input. Temporal results further show clear differences among collaborative construction, independent exploration, onlooker waiting, and conflict regulation in peer distance and voice activity ratio. This method provides traceable and interpretable process evidence for free play observation. It also offers computational support for teacher reflection and process based assessment.

Keywords: Kindergarten free play, Multimodal behavior representation, Learning analytics, Computer vision, Early childhood education

1. Introduction

Kindergarten free play is a natural context in which children's social interaction, problem solving, emotional expression, and autonomous exploration occur intensively. With the development of process-based assessment and digital observation, traditional teacher notes, rating scales, and video review are increasingly limited in continuously representing the dynamic relations among children's movement, language, spatial movement, and material manipulation [1]. Multimodal learning analytics provides a new technical pathway for observing free play by transforming video, audio, and teacher event annotations into continuous behavioral evidence [2]. Focusing on cooperation, exploration, waiting, conflict regulation, and material interaction in free play, a multimodal behavior representation modeling method can improve the granularity and interpretability of behavior

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL36908

recognition while offering computational support for teachers' understanding of children's play participation quality.

2. Literature review

2.1. Free play observation in early childhood education

Observation of free play has long focused on children's active choice, peer negotiation, rule generation, and situated learning in open environments. Existing studies usually rely on teacher observation, case records, and developmental scales to judge children's participation states. These approaches preserve the contextual meaning of educational interpretation, while they are also shaped by observation time, recording density, and teachers' professional experience [3]. When several children move, talk, and manipulate materials at the same time, a single observational perspective can hardly capture the full interaction chain. Digital observation does not weaken teachers' judgment, but turns short actions, interaction nodes, and material transitions into reviewable evidence, allowing free play assessment to move from outcome description toward process analysis [4].

2.2. Multimodal learning analytics for young children

Multimodal learning analytics emphasizes the integration of video, audio, body posture, spatial location, and human annotation into a unified analytical framework to describe changes in children's participation during learning and play. Young children's behavior is characterized by visible bodily action, fragmented language, and context-dependent emotional expression, so a single data source can usually explain only one aspect of behavior [5]. Video can present body posture, movement trajectories, and peer proximity, while audio can reflect speaking turns, negotiation intensity, and changes in group noise. Teacher annotations further supplement the educational meanings behind behavior [6]. The synchronized fusion of multisource information makes children's participation in free play less dependent on retrospective manual judgment and more traceable through behavioral sequences.

2.3. Behavior representation modeling and educational interpretation

The key task of behavior representation modeling is to transform raw video and audio data into understandable behavior units, temporal states, and interaction relations. Computer vision can extract action features such as standing, squatting, moving, turning, and approaching peers. Audio analysis can identify speaking turns, silence duration, and high-intensity sound segments, while human coding calibrates complex situations such as cooperation, conflict, help-seeking, and role play [7]. BiLSTM or Transformer encoders can further analyze the temporal relations among these features and judge how children's behavior shifts from independent exploration to peer negotiation or collaborative construction [8]. Model evaluation needs to consider both recognition accuracy and teacher interpretability, so that computational outputs can serve educational observation rather than replace professional judgment [9].

DOI: 10.54254/2755-2721/2026.GL36908

3. Experimental methods

3.1. Data collection and ethical control

Data collection follows a process that combines public dataset pretraining with small scale kindergarten based validation. Public data are mainly used to improve the basic recognition ability of the model for human posture, classroom activity, and audiovisual interaction. COCO Keypoints 2017 is used for pretraining human keypoint detection. This dataset contains a large number of images annotated with human keypoints. It supports the recognition of 17 human body keypoints, so it is suitable for extracting visual features such as standing, squatting, turning, and limb movement. The Ego4D dataset provides 3,670 hours of first person daily activity videos. It can support transfer learning for action continuity, object interaction, and social interaction segments. For the educational context, the OUC Classroom Group Engagement Dataset released in 2025 can be used as a reference. It focuses on group engagement recognition in real classroom settings and can provide visual support for multi person behavior analysis in classroom environments. The kindergarten validation data are collected from free play contexts in middle and senior classes of two kindergartens in one city in 2026. Data are collected for two consecutive weeks in each class, with 30 minutes recorded each day. The dataset includes about 40 hours of video, synchronized area audio, and teacher event record sheets.

3.2. Multimodal feature extraction and behavior coding

Feature extraction is conducted through the synchronization of three types of data, namely video, audio, and teacher event annotation. For video data, key frames are extracted at five frames per second. A pose estimation model is used to extract keypoints of the head, shoulders, elbows, wrists, hips, knees, and ankles. Body posture, movement speed, turning angle, peer distance, material contact area, and spatial staying time are further calculated. For audio data, a 10 second sliding window is used. Voice activity ratio, speaking turns, silence duration, volume peak, and duration of high intensity sound are extracted to distinguish negotiation, dispute, shared narration, and onlooker waiting [10]. Teacher event record sheets are used to supplement semantic information that is difficult for the computational model to judge directly. The coding categories include collaborative construction, independent exploration, help seeking, conflict regulation, role play, and onlooker waiting. Each 10 second window retains only one main label. When multiple behaviors occur at the same time, the behavior with the longest duration or the clearest educational meaning is selected as the label. Two coders first complete independent coding for 20% of the samples. After Cohen's kappa reaches 0.82, coding is extended to all samples.

3.3. Model construction and evaluation procedure

Model construction takes the 10 second time window as the basic unit of analysis. Video features, audio features, and teacher event annotations are concatenated into a unified input vector [11]. Let the video feature of the t th time window be x^v_t, the audio feature be x^a_t, and the teacher event annotation feature be x^e_t. The three types of features are concatenated and then input into the BiLSTM encoder. The model obtains a hidden state that contains forward and backward behavioral relations. It then uses softmax to output the child behavior category, as shown in Formula 1.

$$h_t = \mathrm{BiLSTM}([x^v_t; x^a_t; x^e_t]), \quad \hat{y}_t = \mathrm{softmax}(Wh_t + b) \tag{1}$$

DOI: 10.54254/2755-2721/2026.GL36908

Formula 1 shows that x_t^v represents the video feature of the t th time window, x_t^a represents the audio feature, and x_t^e represents the teacher event annotation feature. h_t represents the temporal hidden state output by BiLSTM. W and b represent the weight and bias of the classification layer. $\hat{y}_t$ represents the behavior category predicted by the model.

Another scheme adopts a Transformer encoder. It uses an attention mechanism to calculate the association strength among different behavior segments, so that the model can identify key behavioral signals in longer play sequences. For example, when high intensity vocalization, rapid peer approach, and material transition appear at the same time, their weight in judging conflict regulation or cooperative negotiation increases, as shown in Formula 2. The training set, validation set, and test set are divided in a ratio of 6 to 2 to 2. Stratification is based on class and date to avoid placing the same play episode into both training and testing. The evaluation indicators include Accuracy, Macro F1, AUC, and confusion matrix. Teachers then review the model outputs manually to examine whether the computational results match real educational observation.

$$\text{Attetion}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Formula 2 shows that Q represents the query matrix, K represents the key matrix, and V represents the value matrix. d_k represents the dimension of the key vector. Softmax is used to calculate the attention weight among different behavior time windows. The final output is a weighted feature representation for behavior state recognition.

4. Results

4.1. Multimodal feature performance in free play behavior recognition

The preliminary experiment processed 40 hours of kindergarten free play videos. After segmentation with a 10 second sliding window, 14,400 sample windows were obtained. After removing segments with severe occlusion, missing audio, and no child activity, 13,260 valid sample windows were retained. The experiment compared the performance of four input types, including video only, audio only, video and audio, and the fusion of video, audio, and teacher annotation. As shown in Table 1, the video only model reached an Accuracy of 0.761 and a Macro F1 of 0.724. Its main advantage was the recognition of movement, posture change, and material operation. However, multi person occlusion affected the judgment of peer interaction. The audio only model reached an Accuracy of 0.683 and a Macro F1 of 0.641. It was sensitive to speaking turns and dispute related sounds, but it was difficult to locate specific children. The Accuracy of the video and audio model increased to 0.809, which indicates that the combination of movement and sound better distinguishes cooperation, dispute, and onlooker behavior. After teacher event annotation was added, Accuracy reached 0.842, Macro F1 reached 0.817, and AUC reached 0.884.

Table 1. Behavior recognition results under different input modalities

Model input type	Accuracy	Macro F1	AUC	Main advantage	Main error source
Video only	0.761	0.724	0.803	Stable recognition of movement, posture, and material operation	Occlusion and overlapping movement among children

DOI: 10.54254/2755-2721/2026.GL36908

Table 1. (continued)

Audio only	0.683	0.641	0.742	Sensitive recognition of speaking turns and dispute sounds	Difficulty in locating specific children
Video and audio	0.809	0.781	0.856	Simultaneous capture of movement and communication changes	Group noise affects negotiation recognition
Video, audio, and teacher annotation	0.842	0.817	0.884	Clearer semantic boundaries of behavior	Dependence on the quality of human annotation

4.2. Temporal patterns and interaction representation results

The temporal results show that free play behaviors formed a relatively stable transition structure rather than appearing as separate events. Among 13,260 valid time windows, independent exploration accounted for 31.6%, followed by collaborative construction at 24.8%, onlooker waiting at 15.7%, role play at 12.9%, help seeking at 7.6%, and conflict regulation at 7.4%. As shown in Figure 1, the horizontal axis represents average peer distance, the vertical axis represents voice activity ratio, and bubble size indicates the number of valid windows for each behavior. Collaborative construction showed a short peer distance of about 0.55 meters and a voice activity ratio of 56%, suggesting close interaction and continuous communication. Conflict regulation showed a peer distance of about 0.75 meters and the highest voice activity ratio at 70%, indicating stronger vocal expression and higher interaction density. In contrast, independent exploration had a longer peer distance of about 1.6 meters and a voice activity ratio of 20%, while onlooker waiting showed about 1.85 meters and 10%. These results indicate that peer distance and voice activity can effectively distinguish high-interaction and low-interaction play states.

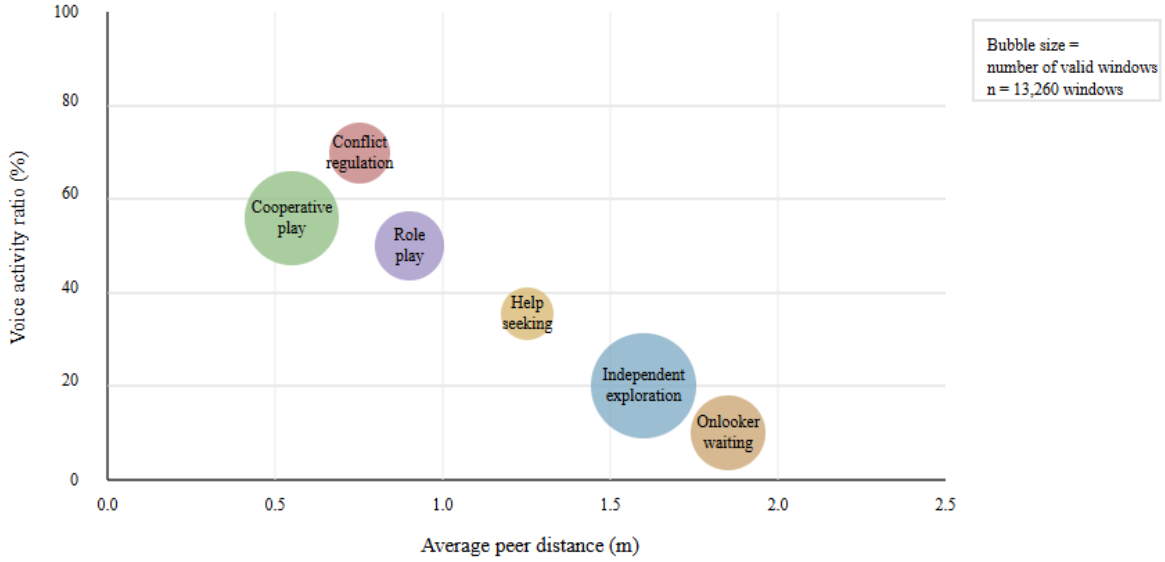

Figure 1. Distribution of peer distance and voice activity in free play behaviors

DOI: 10.54254/2755-2721/2026.GL36908

5. Discussion

The experimental results show that multimodal behavior representation can describe behavioral changes in kindergarten free play in greater detail. Video features are suitable for recognizing movement, posture, and material operation. Audio features supplement speaking turns, negotiation intensity, and dispute signals. Teacher event annotation helps calibrate educational meanings that the model cannot judge directly. After the fusion of these three types of data, the model recognizes collaborative construction, independent exploration, onlooker waiting, and conflict regulation more stably. It is also necessary to note that computational results still need to be interpreted by teachers within the actual context. This is especially important for silence, onlooking, role transition, and short conflicts.

6. Conclusion

A multimodal behavior representation model for kindergarten free play can transform video, audio, and teacher event annotation into continuous behavioral evidence. Through pose estimation, voice activity detection, behavior coding, and temporal modeling, children's behaviors such as collaborative construction, independent exploration, onlooker waiting, role play, help seeking, and conflict regulation can be recognized with relative stability. Preliminary results show that multimodal fusion performs better than video only or audio only input. This indicates that movement, sound, and teacher semantic annotation play complementary roles. The method is suitable for play observation, teacher reflection, and process based assessment in early childhood education. In practical implementation, low interference data collection, data security, and education first interpretation should remain the core principles.

References

[1] Caskurlu, S., Ocak, C., & Dai, C.-P. (2025). The scope of multimodal learning analytics in K-8: A systematic review. *Journal of Learning Analytics, 12*(2), 224–236.
[2] Ouhaichi, H., Spikol, D., & Vogel, B. (2023). Research trends in multimodal learning analytics: A systematic mapping study. *Computers and Education: Artificial Intelligence, 4*, 100136.
[3] Mohammadi, M., et al. (2025). Artificial intelligence in multimodal learning analytics: A systematic literature review. *Computers and Education: Artificial Intelligence, 8*, 100426.
[4] Esterhazy, R., et al. (2025). Advancing multimodal collaboration analytics: A scoping review. *Journal of Learning Analytics, 12*(2), 105–124.
[5] Whitehead, R., Nguyen, A., & Järvelä, S. (2025). Utilizing multimodal large language models for video analysis of posture in studying collaborative learning: A case study. *Journal of Learning Analytics, 12*(1), 186–200.
[6] Lu, W., et al. (2025). A video dataset for classroom group engagement recognition. *Scientific Data, 12*(1), 644.
[7] Liu, Q., Jiang, X., & Jiang, R. (2025). Classroom behavior recognition using computer vision: A systematic review. *Sensors, 25*(2), 373.
[8] Zhang, X., et al. (2024). Smart classrooms: How sensors and AI are shaping educational paradigms. *Sensors, 24*(17), 5487.
[9] Chen, J. J. (2024). A scoping study on AI affordances in early childhood education: Mapping the global landscape, identifying research gaps, and charting future research directions. *Journal of Artificial Intelligence Research, 81*, 701–740.
[10] Pollarolo, E., et al. (2024). Play with coding toys in early childhood education and care: Teachers' pedagogical strategies, views and impact on children's development. A systematic literature review. *Entertainment Computing, 50*, 100637.
[11] Vartiainen, J., Sormunen, K., & Kangas, J. (2024). Relationality of play and playfulness in early childhood sustainability education. *Learning and Instruction, 93*, 101963.