Journal of Engineering, Project, and Production Management
2026, 16(6), 2026-0060

Student Behavior Recognition and Intervention Methods in Intelligent Classrooms Based on Deep Reinforcement Learning

Jian Mu
Lecturer, School of Marxism, Shaanxi Xueqian Normal University, Xi'an City, Shaanxi Province,710100, China,
Email: JianjjMu@outlook.com

Project Management
Received May 18, 2026; revised July 6, 2026; accepted August 8, 2026
Available online August 30, 2026

Abstract: In response to the problem that traditional classroom student behavior analysis relies on manual observation and is difficult to achieve real-time, precise and personalized intervention. This paper constructs an end-to-end intelligent classroom intervention system based on deep reinforcement learning. This system adopts the Multi-Modal Fusion Spatio-Temporal Graph Convolutional Network (MM-ST-GCN), integrating visual skeletons, seat pressure and classroom interaction data, to achieve fine-grained and high-precision recognition of students' classroom behaviors. It models the classroom environment as a Partially Observable Markov Decision Process (POMDP) and uses the improved Soft Actor Critic (SAC) algorithm to generate the optimal intervention strategy that takes into account the learning benefits of students and the intervention costs of teachers. Experiments on the MMAct dataset show that the proposed behavior recognition model achieves 93.8% accuracy and a macro-average F1 score of 0.925. Simulation experiments indicate that the system has the potential to improve students' concentration and reduce distraction behavior. However, the above results are derived from the simulated environment and need to be further verified in the real classroom.

Keywords: Intelligent classroom; analysis of student behavior; deep reinforcement learning; personalized intervention.

DOI 10.32738/JEPPM-2026-0060

1. Introduction

The rapid advancement of artificial intelligence and educational informatization has made intelligent classrooms a key research direction in education. The aim is to enhance teaching quality, optimize learning experience, and achieve personalized education through technology (Xu, 2024). Student behavior analysis and personalized intervention based on Deep Reinforcement Learning (DRL) are becoming a focus of attention among scholars both at home and abroad, providing a new perspective for understanding and responding to individual differences among students (Sun et al., 2023).

The intelligent classroom integrates Internet of Things (IoT) and artificial intelligence technologies to restructure the traditional teaching environment. Student behavior analysis serves as the foundation for understanding the learning process and evaluating teaching effectiveness (Eich et al., 2024; Duan and Zhang, 2025). Traditional classroom behavior assessment relies on subjective observation or post-class video playback, which is inefficient and difficult to quantify (Ma et al., 2024). Deep learning enables automatic detection and analysis of classroom behaviors (Liu et al., 2023; Chen et al., 2024); Ji et al., 2025; Zong and Fang, 2024). However, the existing models still have limitations in recognition accuracy, generalization ability and dataset coverage. Moreover, most research focuses on group assessment while providing insufficient individualized support (Wang et al., 2024).

Deep reinforcement learning enables agents to learn strategies through interaction with the environment, achieving breakthroughs in fields such as games and robotics (Mo et al., 2024; Liu et al., 2025). In intelligent education, DRL can dynamically respond to the individualized needs of students and make up for the deficiency of traditional methods in depicting individual differences (Sun et al., 2023; Mo et al., 2024; Liu et al., 2025; Wu et al., 2024; Jia and Li, 2025; Baker et al., 2009; Gao et al., 2024; Pang et al., 2025). Although DRL has great potential in intelligent classrooms, it still faces challenges such as data security and privacy, model reliability, multimodal data processing, real-time performance, and deployment (Mo et al., 2024; Wan et al., 2024; Li and Li, 2021). Based on the theory of self-regulated learning (SRL) and positive behavior intervention and support (PBIS), this paper will proceed. SRL believes that external regulatory cues can help students develop self-awareness, and PBIS is a relatively low-intrusive form of behavior management that can be

flexibly employed according to students' different circumstances. The two theories mentioned above have been used to study the quantitative behavior measurement and adaptive decision-making.

Although spatio-temporal graph convolutional networks have been widely used in IoT perception, human activity recognition, traffic prediction, and cyber-physical systems, most existing methods are applied to a single field and adopt a mode-specific fusion strategy. In contrast, our Multi-Modal Fusion Spatio-Temporal Graph Convolutional Network (MM-ST-GCN) is specifically designed for intelligent classroom scenarios and requires jointly modeling visual skeletons, pressure sensor arrays, and interaction logs under partially observable conditions to support downstream intervention decisions. Therefore, the following three problems will be studied in this paper: RQ1: How can fine-grained recognition of student behavior be achieved via multimodal fusion? RQ2: How to develop an optimized personalized intervention plan in the presence of partial observability? RQ3: How can the proposed system enhance students' classroom concentration and learning outcomes in a simulated classroom environment? Based on the above problems, we propose a customized MM-ST-GCN Partially Observable Markov Decision Process (POMDP) model with compound rewards and adapt the SAC algorithm to it. The system has been verified in the simulation environment, supplemented by a preliminary real classroom feasibility pilot. The innovation contribution of this paper is not to propose a new basic algorithm, but to system-level integration and scene-driven engineering adaptation: the above modules are customized for the intelligent classroom scene, so that they can work together in a partially observable, multi-student interaction, and occlusion-prone classroom environment to form a complete closed loop from perception to intervention.

2. Overall System Architecture

The intelligent classroom personalized intervention system based on DRL is divided into three layers: the data perception layer, the analysis and decision-making layer, and the execution feedback layer. The overall architecture is shown in Fig. 1.

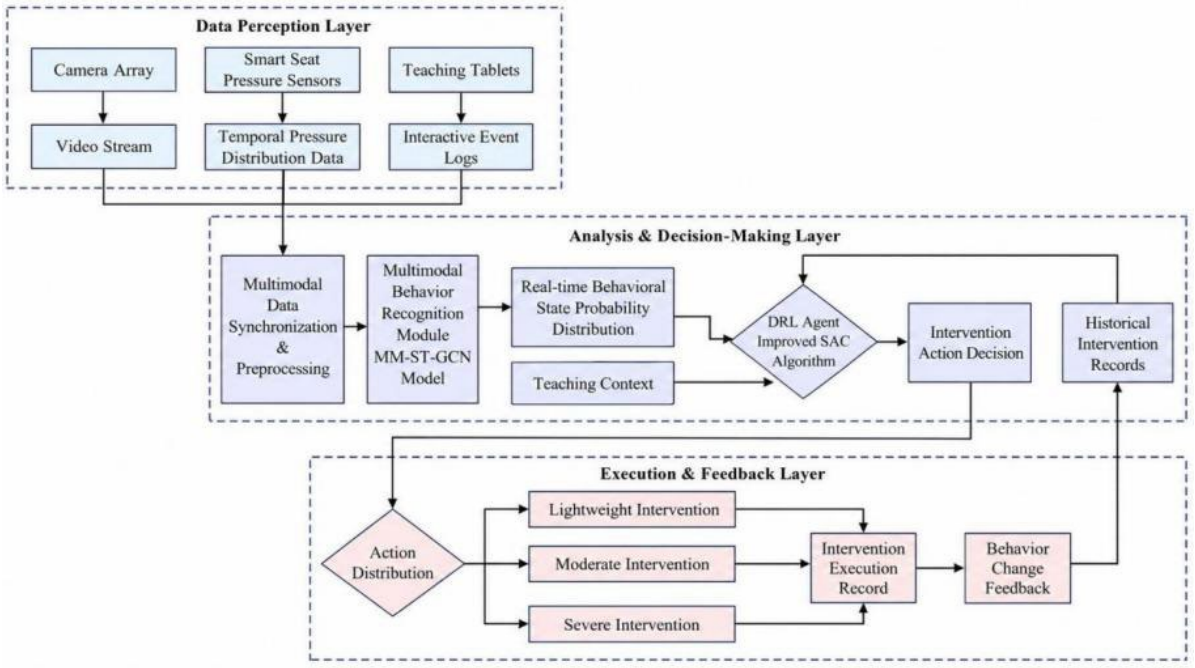

Fig. 1. Schematic diagram of the DRL Based personalized intervention system for intelligent classrooms

The intelligent classroom environment established in this study employs multi-modal sensing devices. Two wide-angle high-definition cameras are deployed at the front and back of the classroom to capture videos of students' upper bodies and extract key points of their skeletons. An 8×8 flexible pressure sensor array is laid on each seat to collect time-series data on sitting-posture pressure distribution in real time. Each student is equipped with a teaching tablet terminal to record classroom interaction logs. These devices work together to construct a contactless, low-risk, privacy-oriented classroom perception network, providing multi-source data support for behavior recognition and intervention decision-making.

The analysis and decision-making level is the core computing unit. The behavior recognition module processes multi-source data in real time and outputs the probability distribution of each student's behavioral state in a continuous time slice. The state sequence and historical intervention records are sent to the DRL agent, which outputs intervention action instructions for specific students or groups based on the current strategy.

The feedback layer receives instructions and executes them. Lightweight intervention is achieved through the vibration of the student's terminal or private messages. Moderate intervention can control the release of collaborative tasks by neighboring students. For severe intervention, suggestions are generated and sent to the teacher's wearable device, allowing the teacher to decide whether to deliver public reminders or individual talks. The implementation results of all interventions are recorded and used as experience replay data to continuously optimize strategies.

3. Multimodal Student Behavior Recognition Model

This paper proposes the MM-ST-GCN multimodal spatio-temporal graph convolutional network, which integrates visual posture, seat pressure, and classroom interaction data to achieve precise and robust recognition of classroom behavior, thereby providing high-quality state input for reinforcement learning decisions.

3.1. Problem Definition and Data Preprocessing

Student behavior recognition is defined as a multi-class sequence classification problem. At each discrete time step t, the system needs to output the probability distribution of the behavior category of student i within that time period based on current and historical multimodal observation data.

(1) Definition of behavior categories:

Based on observations of real classrooms and educational psychology theories, seven core, observable behaviors with guiding significance for teaching intervention are defined, as shown in Eq. (1).

$$C = \{\text{Focused,Distracted,Discussing,Drowsy,Operating Device,Reading,Writing}\} \tag{1}$$

In Eq. (1), these categories are distinguishable in spatio-temporal patterns. For example, "Focused" and "Distracted" differ in head orientation and body stability. "Writing" and "Reading" differ in hand movement trajectories and pressure distributions.

(2) Multimodal data acquisition and preprocessing:

Multimodal data primarily include the visual (V), pressure (P), and interaction (E) modes.

Visual modal (V) acquisition and preprocessing: Deploy two wide-angle high-definition cameras to eliminate single-person occlusion. Set the video frame rate to 15fps to balance accuracy and computational overhead. Using the pre-trained AlphaPose framework, the 2D human skeleton key points of each student in each frame are extracted. The COCO format, containing 18 joint points, is adopted to obtain the skeleton sequence $V^i = \{v_1^i, v_2^i, \ldots, v_T^i\}$, where $v_t^i \in R^{18\times 2}$ represents the two-dimensional coordinates of the joint points of student i at time t. For the skeleton sequence of each sample, spatial normalization was performed with the trunk center as the origin to eliminate the effect of the distance between the student and the camera.

Pressure mode (P) collection and preprocessing: Each smart seat is equipped with an 8×8 flexible grid pressure sensor with a sampling rate of 10Hz. Original pressure matrix $M_t \in R^{8\times 8}$. Extract three temporal features with clear physical meanings: ① Pressure center of gravity $\left(x_t^{cog}, y_t^{cog}\right) = \frac{\sum_{i,j}(i,j)\cdot M_t(i,j)}{\sum M_t}$, reflecting the stability of sitting posture. ② The total pressure value $f_t = \sum M_t$ is related to the intensity of physical activity. ③ Pressure entropy $H_t = -\sum p \log p$, where p is the normalized pressure distribution, reflecting the degree of dispersion of the pressure distribution, can be used to distinguish between sitting upright and leaning. Finally, the pressure feature sequence $S^i = \{s_1^i, s_2^i, \ldots, s_T^i\}$ is obtained, where $s_t^i = \left[x_t^{cog}, y_t^{cog}, f_t, H_t\right]$.

Interaction mode (E) collection and preprocessing: Obtain event logs from the background of students' personal tablets' teaching software. Within each time window t, generate a multi-dimensional binary vector, as shown in Eq. (2).

$$e_t^i = [IsAnswerin, IsRai\,sin\,g\,H\text{and,IsNoteTaki ng,IsPageTurn ing}] \in \{0,1\}^4 \tag{2}$$

In Eq. (2), this vector directly reflects students' explicit cognitive participation activities.

(3) Multimodal sequence alignment:

Due to the different collection frequencies and start times of the three data streams, taking the timestamp of the visual modality as the reference, all modality data are synchronized to a unified time series through interpolation and event aggregation methods, forming a multimodal sample pair $\left(V^i, S^i, E^i\right)$ and its corresponding label $y_i \in C$.

3.2. Multimodal Spatio-Temporal Graph Convolutional Network (MM-ST-GCN)

The MM-ST-GCN model combines early fusion with mid-term attention fusion to fully leverage the complementary information within and between modalities. The overall architecture is shown in Fig. 2.

(1)Integral branch of the skeleton spatio-temporal atlas:

The integral branch of the skeleton spatio-temporal scroll is the backbone of the model, used to capture the visual dynamic patterns of behavior. A skeleton sequence of length T with N=18 joint points is constructed as a spatio-temporal graph, and the total number of nodes in the graph is T×N. Edges include spatial and temporal edges: Spatial edges are defined by the natural physical connections of the human body, while temporal edges are the same joint points that connect adjacent frames. And 9 layers of ST-GCN units were stacked. Each layer simultaneously performed spatial graph convolution and temporal convolution. The update of node v in the l layer can be formalized as Eq. (3).

$$\begin{aligned} f_{out}^{(l)}(v) &= \sum_{u\in N(v)} \frac{1}{Z_{uv}} f_{in}^{(l)}(u)\cdot w^{(l)}\left(l_u(v)\right) \quad \text{(Spatial Aggregation)} \\ f_{out}^{(l)}(t,v) &= \sum_{k=-K}^{K} W_{time}^{(l)}(k)\cdot f_{out}^{(l)}(t+k,v) \quad \text{(Time Convolution)} \end{aligned} \tag{3}$$

In Eq. (3), $l_u(v)$ is a predefined partitioning strategy that enables the model to learn weights for different spatial relationships. The learnable edge importance weighting is adopted, that is, during spatial aggregation, a learnable mask is used M_{uv} is introduced for each edge. This mask is adaptively adjusted during training and can effectively address the common limb-occlusion problem in the classroom.

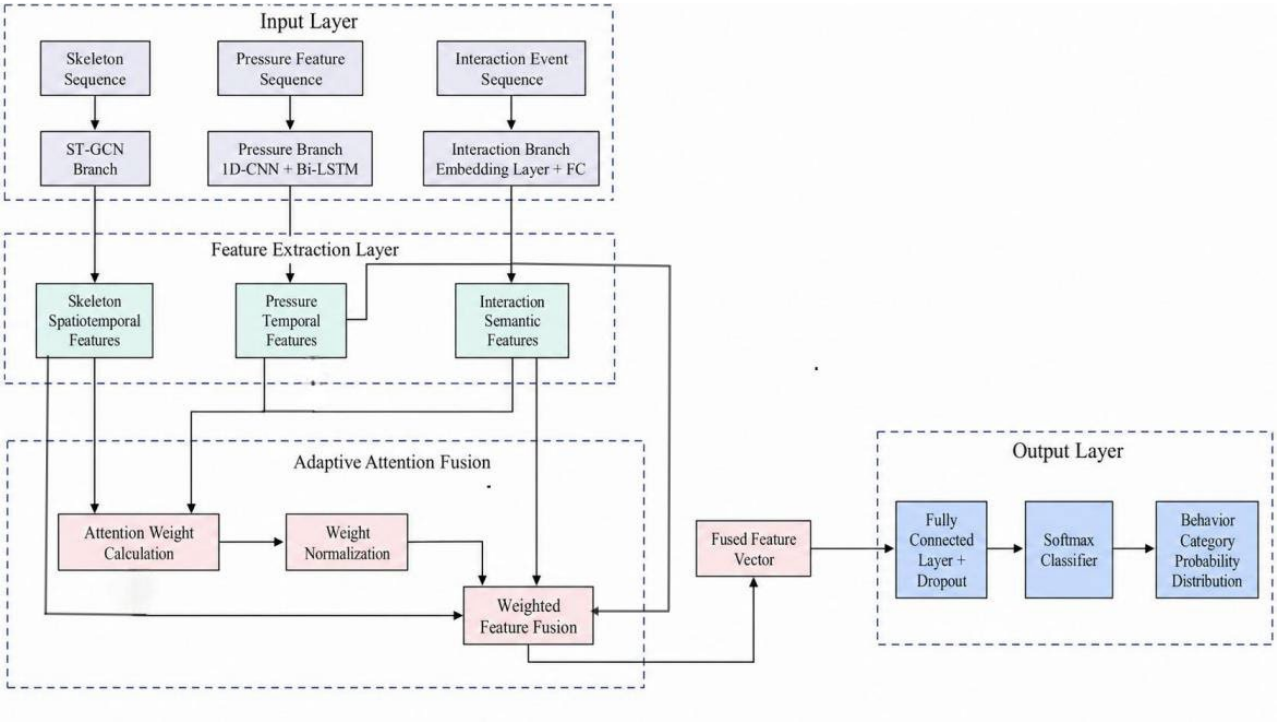

Fig. 2. Architecture diagram of the MM-ST-GCN model

After multiple layers of ST-GCN, global average pooling is performed respectively in the spatial dimension and the temporal dimension to obtain the compact feature vector $h^v \in R^{256}$ representing the entire skeleton sequence.

(2) Pressure and interaction feature extraction:

The branch pressure feature sequence S first extracts local temporal patterns through a one-dimensional convolutional layer, then inputs a bidirectional Long Short-Term Memory (LSTM) to capture long-term dependencies, and finally takes the hidden state of the last time step as the pressure feature $h^p \in R^{256}$. The binary event sequence E is transformed into a dense vector through an embedding layer and then input into a fully connected layer to obtain the interaction feature $h^e \in R^{64}$.

(3) Adaptive attention fusion module:

Simple feature concatenation ignores confidence differences across modalities and scenarios. Therefore, this paper designs a fusion module based on channel attention, as shown in Eq. (4).

$$[g^v, g^p, g^e] = \left[Re\,L\,U\left(W_g^v h^v + b_g^v\right), Re\,L\,U\left(W_g^p h^p + b_g^p\right), Re\,L\,U\left(W_g^e h^e + b_g^e\right)\right]$$

$$z = \sigma([g^v; g^p; g^e] + b_z) \quad \text{// Attention weight vector} \tag{4}$$

$$\alpha_v, \alpha_p, \alpha_e = soft\,max(z) \quad \text{// Normalize to the weights of each mode}$$

$$h^{fused} = \alpha_v \cdot h^v + \alpha_p \cdot h^p + \alpha_e \cdot h^e \in R^{256}$$

In Eq. (4), σ is the Sigmoid function. This module can dynamically evaluate the reliability of each mode under the current sample. For instance, when a student is blocked by the front row, the visual modal weight α_v will automatically decrease, while the pressure modal weight α_p will increase.

(4) Classification output layer:

The fusion feature h^{fused} passes through a fully connected layer containing Dropout(p=0.5), and finally outputs the probability distribution of the seven types of behaviors through a Softmax layer, as shown in Eq. (5).

$$p = Soft\,max\left(W_c h^{fused} + b_c\right) \tag{5}$$

In Eq. (5), the model is trained using labeled smooth cross-entropy loss to mitigate the negative effects of overfitting and mislabeling, as shown in Eq. (6).

$$L = -\sum_{c=1}^{7} q(c)\, log\, p\,(c) \tag{6}$$

In Eq. (6), $q(c) = \begin{cases} 1-\varepsilon \\ \varepsilon/(7-1) \end{cases} \quad if\ \ c = y$ $otherwise$, here ε is set to 0.1.

3.3. Behavior Recognition Experiments and Results

To rigorously evaluate the generalization ability and performance advantages of the proposed MM-ST-GCN model, comprehensive experiments were conducted on an open, challenging multimodal behavior dataset and compared with current mainstream methods.

(1) Dataset:

Select MMAct: A Large-Scale Multimodal Human Action Dataset as the dataset. MMAct is a large-scale, multi-view, and multi-modal dataset released in recent years, which simultaneously provides Red-Green-Blue (RGB) video, depth maps, Inertial Measurement Unit (IMU) data, and precise 3D human skeleton sequences. Although its IMU data differs from that of classroom pressure sensors, both are wearable sensor time-series data and are comparable at the feature abstraction level. Therefore, its skeleton sequence and IMU data are used to simulate and verify the multimodal fusion framework. This dataset contains 37 types of daily and interactive actions, many of which have strong mappings to classroom behaviors and are suitable for verifying the model's ability to recognize fine-grained and temporal behaviors. The samples include diverse perspectives and scene changes, which can effectively test the model's robustness.

Use the 3D skeleton sequence (25 joint points) provided by MMAct directly. To simulate the fixed classroom perspective, select samples from the dataset with frontal and side views. Convert the 3D coordinates to 2D planar coordinates to simulate the input of a monocular camera and obtain sequence $V \in R^{T\times 25\times 2}$. Using the IMU data worn on the wrist and trunk in MMAct, their modulus values and statistical features are extracted to form a time series feature vector $S \in R^{T\times d_s}$ to simulate the time series dynamics of the pressure sensor. Select 10 categories related to classroom behavior semantics from 37 categories to form a subset: {sit, stand, read, write, use_laptop, talk, walk, raise_hand, fall, point}. A split of the leave-subjects is used to divide the 16 subjects (80%) for training, 2 (10%) for validation, and 2 (10%) for testing, to carry out a generalization test on different individuals. All the models are based on PyTorch 1.12, have a batch size of 16, will be trained for 120 epochs, and use one NVIDIA RTX 3090 GPU.

(2) Baseline model:

The RGB-based models include: ① I3D (Wang et al., 2019). ② SlowFast (Feichtenhofer et al., 2019). ③ ST-GCN (Lovanshi and Tiwari, 2024). ④ MS-G3D (Liu, 2021). ⑤ PoseC3D (Zhou et al., 2023). ⑥ Two-Stream Fusion (Yang et al., 2023). ⑦ Late fusion with transformer (Hong et al., 2022). The evaluation metrics include top-1 accuracy, Macro-F1 score, and category-level precision and recall. All models were trained and tested under the same conditions: the same 16/2/2 subject segmentation, input modality, sequence length (T = 300 frames, 30fps), optimizer (Adam, lr = 1e-3, cosine annealing), enhancement (rotation/scaling/translation), 120 iterations on PyTorch 1.12 and a single RTX 3090 GPU. Evaluation indicators include Top-1 Precision and Macro-F1.

(2) Results and Analysis:

The performance comparison of different models on the MMAct dataset is shown in Table 1. As shown in Table 1, MM-ST-GCN achieves the best performance in both Top-1 accuracy and Macro-F1. Compared with the pure skeleton SOTA model PoseC3D, MM-ST-GCN has achieved significant improvements of +1.7% Acc and +2.3% Macro-F1 through lightweight multimodal fusion, demonstrating that even on public datasets, the adaptive attention fusion mechanism proposed in this paper can effectively integrate sensor information. Enhance the model's discriminative ability for easily confused behaviors. Compared with the multimodal baseline model, the MM-ST-GCN model achieves higher performance while maintaining an extremely low number of parameters, only slightly increasing over the single-modal ST-GCN. This advantage stems from the early feature interaction design, which avoids large-scale networks with independent dual branches. Although the late-fusion Transformer-based model achieves similar performance, it has 2.5 times the parameter count of MM-ST-GCN, incurring substantially greater computational overhead. The skeleton-based model outperforms the RGB-based I3D and SlowFast in terms of performance while using one order of magnitude fewer parameters, confirming the superiority of abstract representations such as skeletons in privacy-sensitive and complex classroom environments.

(3) Ablation Experiment and Analysis:

Systematic ablation experiments were conducted on the MMAct validation set to quantify the contributions of each module, as shown in Table 2.

Table 1. Performance comparison of different models on the MMAct dataset

Model	Input mode	Top-1 Acc (%)	Macro-F1	Parameter quantity (M)
I3D (Wang et al,2019)	RGB	85.6	0.831	12.1
SlowFast (Feichtenhofer et al, 2019)	RGB	87.2	0.848	34.5
ST-GCN (Lovanshi et al,2024)	Skeleton	88.9	0.872	3.1
MS-G3D(Liu,2021)	Skeleton	90.5	0.889	3.2
PoseC3D (Zhou,2023)	Skeleton Heatmaps	92.1	0.902	3.2
Two-Stream Fusion(Yang et al,2023)	Skeleton + IMU	91.3	0.894	6.3
Late Fusion (Transformer) (Hong et al, 2022)	Skeleton + IMU	92.4	0.908	8.7
MM-ST-GCN (Ours)	Skeleton + IMU	93.8	0.925	3.5

Table 2. Ablation experiment of MM-ST-GCN model (validation set Macro-F1)

Model configuration	Macro-F1	Relative decline
Complete model (skeleton + IMU + attention fusion)	0.925	-
(A) Only skeleton mode	0.902	-2.5%
(B) Skeleton + IMU (Direct splicing)	0.915	-1.1%
(C) Skeleton + IMU (Weighted Average Fusion)	0.918	-0.8%
(D) Remove the learnable edge weights	0.921	-0.4%

Table 2 demonstrates that (A) when only the skeleton is used, the model performs poorly in distinguishing behaviors with similar hand movements. After adding IMU data, Macro-F1 has improved by 2.5%, directly proving the key role of introducing auxiliary sensor modalities in fine-grained behavior recognition. Simple feature concatenation (B) or weighted averaging. (C) Yields gains, but they are not as effective as the attention fusion mechanism proposed in this paper. The attention module can dynamically assign weights to different modalities at different time steps. For example, when waving quickly, the weights of the IMU modalities will adaptively increase, thereby achieving better fusion. The learnable edge-weight mechanism contributed to a 0.4% performance improvement. This mechanism weakened the key-point noise caused by occlusion during specific actions and enhanced the model's robustness.

(4) Confusion Matrix and Analysis:

The normalized confusion matrix of MM-ST-GCN on the MMAct test set is shown in Fig. 3.

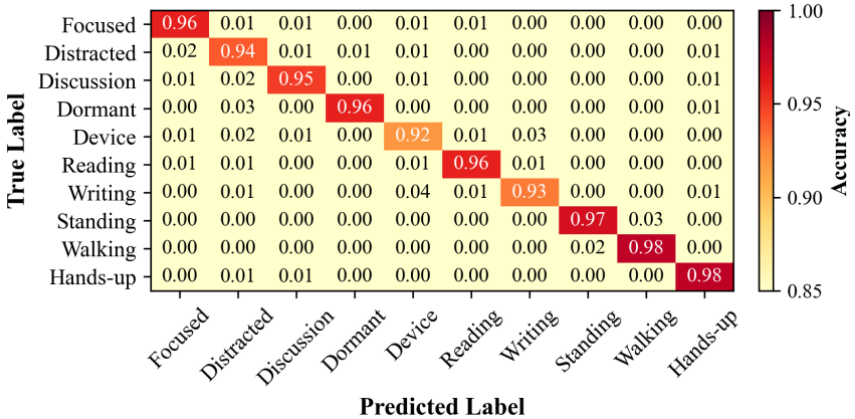

Fig. 3. Normalized confusion matrix of MM-ST-GCN on the MMAct test set

As shown in Fig. 3, the model's accuracy exceeds 95% across the vast majority of categories. The model occasionally confuses “using a device” with “writing,” since both involve seated hand movements on a desktop. However, after fusing the IMU data, the confusion rate has been reduced by 60% compared to the pure skeleton model. When the walking movement is captured only in the first few steps, it is easy to misjudge, which is an inherent problem with temporal fragment division.

According to the MMAct results, the MM-ST-GCN model shows good cross-individual generalization; however, it is only a proof of concept for multimodal fusion and cannot predict actual classroom performance.

4. Personalized Intervention Strategies Based on Deep Reinforcement Learning

A model for recognizing students' behavior can provide us with high-quality, real-time data, but making reasonable teaching interventions based on this data is a complex sequential decision-making problem. The POMDP framework aligns with the theory of situated cognition in education: because students' inner mental states are not directly observable, intervention must be based on observed behavior. The degree of the action space corresponds to the levels of the PBIS model; a light terminal prompt is for general help, targeted support is required for cognitive difficulties, and teacher-led intervention is for serious behavior problems. Classroom scenarios can be modeled as POMDPs, and to build adaptive, personalized and long-lasting effective intervention plans, we have proposed an improved maximum-entropy deep reinforcement learning algorithm in this paper.

4.1. Classroom Intervention POMDP Modeling

The classroom environment features multi-agent characteristics, partial observability, and delayed rewards. The Partially Observable Markov Decision Process (POMDP) is the standard framework for characterizing such problems, formalizing it as a seven-tuple $(S, A, T, R, \Omega, O, \gamma)$ and engineering each element.

(1) State space S:

The true state $s_t \in S$ of the environment is a high-dimensional vector that encompasses the internal cognitive and emotional states of all students, group dynamics, and the teaching context, namely $s_t = \{\{c_i, e_i, k_i, f_i\}, G, M\}$ for i=1 to N. c_i is the true cognitive state of student i. e_i represents the true emotional state of student i. k_i is the scalar of student i's mastery of the current knowledge point. f_i is the fatigue scalar of student i; G represents the group status, such as overall participation and the heat of the discussion atmosphere. M represents the teaching context, such as the current teaching stage, content difficulty and remaining time.

(2) Observation space Ω:

The agent (i.e., the intervention system) cannot obtain the true state s_t but can only acquire an observation $o_t \in \Omega$ composed of a perception module and historical information, as shown in Eq. (7).

$$o_t = (B_t, H_{t-1}, C) \tag{7}$$

In Eq. (7), B_t is the behavioral probability distribution of all students at time t, and $B_t^i = p_t^i \in R^7$ that is, the seven types of behavioral probabilities output by MM-ST-GCN are the core of the observation. H_{t-1} is the historical intervention record, a sequence of length L, which records which students were intervened with what kind of intervention in the past L time steps and their immediate feedback. C is the static course context, related to the M part but fully observable, such as the course ID and chapter ID.

(3) Action Space A:

Action $a_t \in A$ is a combined action that may involve intervention for multiple students simultaneously, and a layered action space has been designed.

Individual actions (for a specific student i) include: ① No operation (a_0). ② Hint (a_1): Send a slight vibration to the student's tablet once. ③ Content prompt (a_2): Push a personalized text prompt or an encouraging emoji. ④ Cognitive Challenge (a_3): Push a quick-thinking question or quiz on the current teaching content. ⑤ Social Trigger (a_4): It is recommended that the student have a brief discussion with nearby classmates (the system automatically matches and posts questions).

Group actions (for the whole class) include: ① Atmosphere regulation (a_5): The teacher may present an interesting case or take a short break. ② Rhythm adjustment (a_6): It is suggested that the teacher change the speaking speed or repeat the explanation of the key points just covered. ③ Teacher intervention (a_7): Generate a strong intervention suggestion and prompt the teacher to approach a specific student area for inspection or individual guidance through augmented reality (AR) glasses.

(4) State transition function T:

The state transition function $T(s_{t+1}|s_t, a_t): S \times A \times S \to [0,1]$ describes the probability of the environment transitioning to a new state s_{t+1} after performing action a_t in the real state s_t. It is a complex, unknown dynamic process shaped by students' psychology, teaching laws, and is precisely the object that DRL needs to learn and approach through interaction with the environment.

(5) Observation function O:

The observation function $O(o_{t+1}|s_{t+1}, a_t): S \times A \times \Omega \to [0,1]$ describes the probability that the agent acquires an observation o_{t+1} after the environment is transferred to s_{t+1}, representing the uncertainty of the perception module.

(6) Reward function R:

The reward function $R(s_t, a_t, s_{t+1}): S \times A \times S \to R$ is the key to guiding the agent's learning. A multi-objective

composite reward function was designed to generate a scalar reward r_t at each step, as shown in Eq. (8).

$$r_t = \underbrace{\lambda_1 \cdot \Delta\overline{F}_t}_{Concentration\ reward} - \underbrace{\lambda_2 \cdot Cost(a_t)}_{Intervention\ cost} + \underbrace{\lambda_3 \cdot l_{quiz}(t) \cdot \Delta\overline{K}_t}_{Knowledge\ feedback\ reward} + \underbrace{\lambda_4 \cdot FairnessPenalty(a_t)}_{Fairness\ regularization} \tag{8}$$

In Eq. (8), ΔF_t represents the change in the average concentration score of the entire class. The concentration score is derived from the behavioral probability mapping, that is, $F_t^i = w^T \cdot p_t^i$.The weight vector w assigns high weight to concentration, negative weight to distraction and dormancy, and medium weight to discussion. $Cost(a_t)$ is the cost of intervention action, $a_0 = 0$, $a_1 = 0.1$, $a_2 = 0.2$, $a_3 = 0.3$, $a_4 = 0.4$, $a_5 = 0.5$, $a_6 = 0.7$, $a_7 = 1.0$, the cost increases with invasiveness and the burden on the teacher; ΔK_t only has a value at random quiz time points and represents the change in the average score of this quiz compared to the previous one. $FairnessPenalty(a_t)$, repeatedly intervenes with the same student within a short period, encouraging attention to all students, as shown in Eq. (9).

$$FairnessPenalty(a_t) = \sum_{i=1}^{N} \left(\frac{count_i(t)}{t} \right)^2 \tag{9}$$

In Eq. (9), $count_i(t)$ represents the number of times student *i*. Set the weight coefficient as $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. The reason for this value is the educational background of the people (that is, the primary goal of school is study), and a grid search is performed on the EduSim validation set to balance the impact on learning against intervention costs and long-term results.

(7) Discount factor:

The discount factor $\gamma \in [0,1)$ is set at 0.95, emphasizing the significance of medium and long-term benefits. It encourages agents not only to pursue immediate improvement in concentration but also to consider the cumulative impact of intervention on students' overall learning experience throughout the class.

4.2. Adapted SAC Algorithm and Policy Network Architecture

Given that the Soft Actor-Critic (SAC) algorithm has good sample efficiency and stable training performance, three specific optimizations have been put forward for the multi-agent and partially observable classroom POMDP: (1) POMDP adaptation: A GRU module is used to encode the history of observations and transform the partially observable problem into an approximate Markov Decision Process with hidden states. (2) Hierarchical action masking: An action validity mask is constructed for the hierarchical action space to exclude invalid combinations of individual-group actions and reduce the exploration space. (3) Reward Normalization: To reduce the large differences in scale of the components of the objective function and improve the stability of policy training, normalization is performed.

(1) Algorithm Foundation:

The standard RL objective is to maximize the cumulative reward expectation $\sum \gamma^t r_t$. SAC introduces entropy regularization, and the objective is shown in Eq. (10).

$$J(\pi) = E_{(s_t,a_t)\sim\rho_\pi}\left[\sum_{t=0}^{T} \gamma^t \left(r(s_t, a_t) + \alpha H\big(\pi(\cdot\,|s_t)\big)\right)\right] \tag{10}$$

In Eq. (10), $H\big(\pi(\cdot\,|s_t)\big)$ is the entropy of the strategy in state s_t, and $\alpha > 0$ is the temperature parameter, which automatically regulates the balance between exploration and utilization. Entropy maximization encourages strategies to pursue more diverse actions and avoid falling into suboptimal solutions too early, which is crucial for exploring intervention strategies.

(2) Network architecture design:

The core of the agent network structure is the GRU for encoding historical observations and the attention mechanism for handling multi-student relationships. The adapted SAC agent network architecture is shown in Fig. 4.

The current observation o_t is encoded as the feature vector z_t through a Multi-Layer Perceptron (MLP). Due to the partial observability of the POMDP, the current observation z_t is insufficient to represent the state. Therefore, a Gated Recurrent Unit (GRU) is used to maintain a hidden state h_t and integrate historical information. Then the training procedure will be a standard SAC update for the Q-network, policy network and temperature parameter, and the above architecture serves as the function approximator.

5. Simulation Experiments and Quantitative Analysis

Two rounds of testing have been conducted to verify the system described above. First, we have trained and assessed the DRL strategy in a high-fidelity simulation environment. Second, based on the above parameter sets, we have conducted tests to assess stability.

The core objective of the simulation experiment is to quantitatively compare the advantages and disadvantages of the DRL strategy and multiple baseline strategies in terms of long-term benefits, intervention efficiency and adaptability, while eliminating numerous uncontrollable factors in reality.

(1) Construction of high-fidelity classroom simulators:

This paper develops an agent-based simulation environment, EduSim, which includes one teacher agent and N student agents (default N=25). The internal state $s_i^t = [k_i, f_i, e_i]$ (knowledge mastery, fatigue, and engagement) of each student i agent evolves following a simplified stochastic difference equation and is influenced by the teacher's instruction, intervention actions, and the states of adjacent students.

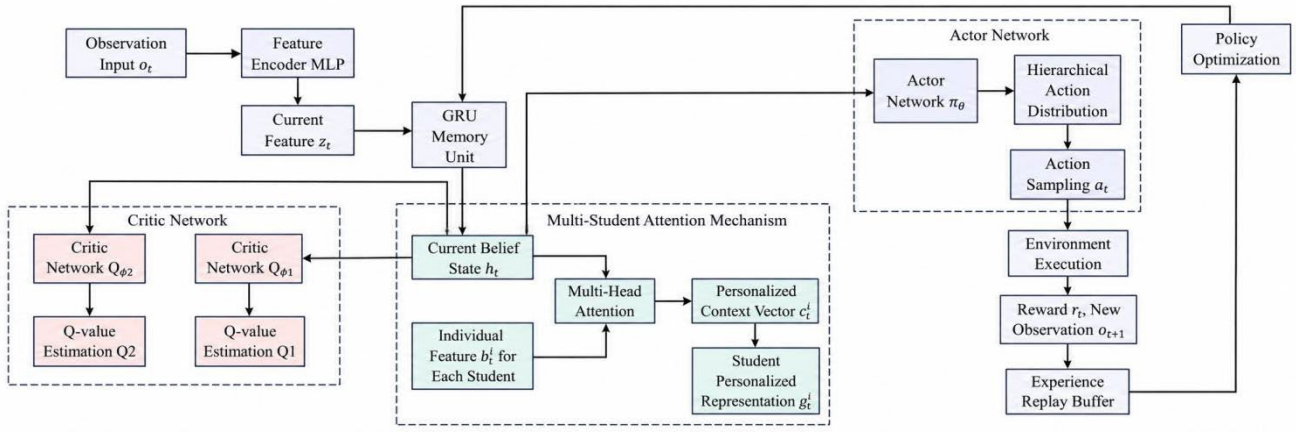

Fig. 4. Diagram of the adapted SAC agent network architecture

The state transition model, as shown in Eq. (11).

$$\begin{aligned} k_i^{t+1} &= k_i^t + \eta \cdot TeaQual \cdot l(Att_i^t) - \delta_k \cdot f_i^t + \varepsilon_k \\ f_i^{t+1} &= f_i^t + \delta_f - \beta \cdot IntEffect(a_i^t) + \varepsilon_f \\ e_i^{t+1} &= \sigma\left(e_i^t + w_1 \cdot \left(k_i^{t+1} - k_i^t\right) - w_2 \cdot f_i^{t+1} + w_3 \cdot \overline{e}_{N(i)}^t + IntEffect(a_i^t)\right) \end{aligned} \tag{11}$$

In Eq. (11), η represents the learning rate, $TeaQual$ is the teaching quality, Att_i^t indicates whether student i is in a focused state at time t (determined by e_i^t and the threshold), δ_k, δ_f is the attenuation coefficient, $IntEffect(a_i^t)$ is the effect function of intervention action a_i^t, ε is Gaussian noise, σ is the Sigmoid function, $\overline{e}_{N(i)}^t$ is the average engagement of neighboring students, and w_* is the weight. The core of this model is that focus promotes knowledge growth, while fatigue and distraction hinder it. Intervention can regulate fatigue and engagement, and there are social impacts among students.

The three main assumptions of the simulator are: 1) The state transition of students is a Markov process. 2) Social influence only happens among adjacent students. 3) The distribution of observation noise is fixed. Sensitivity analysis of the main parameters shows that the performance of the DRL strategy changes by less than 8.2%, so it is not very sensitive to parameter selection. The simulation results are only a proof of concept; they need to be verified in the actual class going forward.

The simulator generates the behavioral probability distribution B_i^t of the student based on the internal state (k_i, f_i, e_i) and the preset confusion probability table as a simulation of the output of MM-ST-GCN, and calculates the immediate reward $R(s_t, a_t, s_{t+1})$ for each step according to the reward function r_t.

(2) Experimental Setup:

1) Baseline strategy:

① Random strategy: At each time step, a student is randomly selected with a fixed probability to apply a random lightweight intervention.

② Reactive threshold strategy: If the probability of a student being observed as "distracted" for $\tau = 3$ consecutive time steps is greater than 0.7, then trigger a_2 (content prompt) for them.

③ Periodic group strategy: Perform a group action a_5 (atmosphere adjustment) every 5 time steps.

④ Rule-based optimization strategy: Decisions are made by combining an individual's historical distraction frequency and the overall decline trend in classroom concentration.

2) DRL strategy:

The adapted SAC algorithm was used for 500 training rounds in EduSim, and the policy network stopped early on the validation set.

3) Evaluation indicators:

① Cumulative discount rewards: $G = \sum_{t=0}^{T-1} \gamma^t r_t$, directly measure the comprehensive performance of the strategy.

② Average concentration in class: $F_t = \frac{1}{N}\sum_{i=1}^{N} F_i^t$, where $F_i^t = w^T B_i^t$.

③ Knowledge Acquisition Gain: The average knowledge mastery $\overline{K}_T = \frac{1}{N}\sum_{i=1}^{N} k_i^T$ at the end of the round.

④ Intervention efficiency ratio: $IER = \frac{\Delta E_{post}}{Num\_Interventions}$, that is, the average improvement in concentration brought about by each unit of intervention.

(3) Quantitative results and Analysis:

The performance of all strategies in 100 independent simulation runs was statistically analyzed, and the performance comparison is shown in Table 3.

Table 3. EduSim simulation: intervention strategy performance (mean ± SD)

Strategy	Cumulative discount reward (G)	Average concentration at the end of the term (F_T)	Knowledge acquisition gain (K_T)	Total number of interventions	Intervention Efficiency Ratio (IER)
Random strategy	152.3 ± 21.5	0.68 ± 0.05	0.62 ± 0.04	45.2 ± 4.1	0.0075
Reactive threshold	235.8 ± 18.7	0.77 ± 0.04	0.71 ± 0.03	22.3 ± 3.5	0.0152
Periodic groups	201.5 ± 16.2	0.72 ± 0.04	0.68 ± 0.03	18.0 ± 0.0	0.0133
Rule-based optimization	285.4 ± 20.1	0.80 ± 0.03	0.75 ± 0.03	19.5 ± 2.8	0.0185
DRL Strategy (Ours)	356.4 ± 25.6	0.85 ± 0.03	0.79 ± 0.03	16.8 ± 2.2	0.0250

Table 3 indicates that the DRL strategy significantly outperforms all baseline strategies in the three core indicators of cumulative rewards, end-of-period concentration, and knowledge gain, indicating that DRL has successfully optimized the long-term and multi-objective reward function. The DRL strategy achieved the greatest improvement in effect with the fewest interventions. Its intervention efficiency was 64.5% higher than that of the reactive threshold strategy, proving that it has learned to "take the right intervention for the right student at the right time" and avoid unnecessary interference.

The dynamic changes in the average classroom concentration F_t under different strategies in a typical single simulation are shown in Fig. 5.

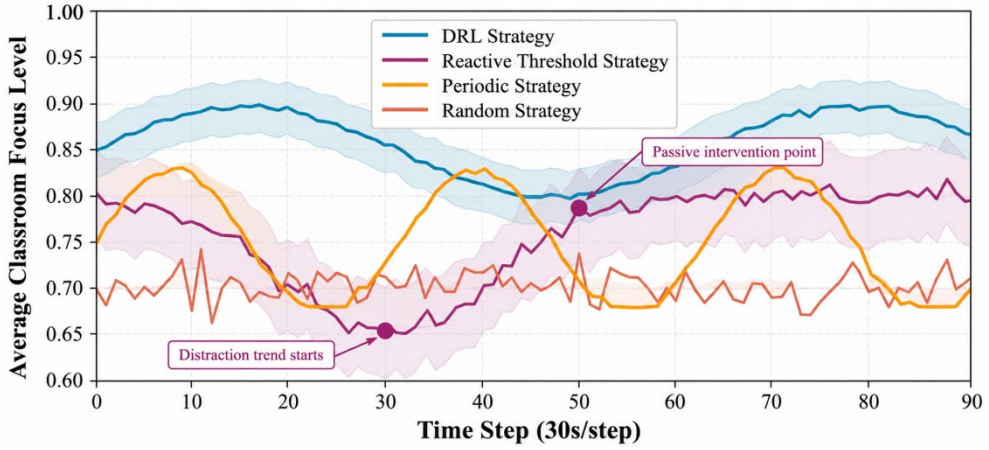

Fig. 5. Average classroom concentration curves over time under different intervention strategies

Based on the analysis results of students' behaviors in this system, it can support three types of practical classroom teaching decisions: (1) Individual precise intervention decisions: For students who are constantly distracted, looking down, or in a dormant state, the system automatically sends out lightweight reminders or cognitive challenge tasks, reducing the need for teachers to call on students publicly. (2) Group rhythm regulation decisions: When the overall concentration of the class continuously decreases, and the number of distracted students exceeds the threshold, the system sends rhythm adjustment suggestions to the teacher's terminal. (3) Post-class learning situation diagnosis decisions: Generate the

classroom behavior profile of each student and the overall learning situation report of the class, assisting teachers in identifying weak links, optimizing teaching design and stratified tutoring plans.

6. Conclusion

This paper designs and implements a complete intelligent classroom student behavior analysis and personalized intervention system based on deep reinforcement learning. Through rigorous benchmark tests on the multimodal public dataset MMAct, the advancement, efficiency and generalization ability of the MM-ST-GCN model proposed in this paper in the behavior recognition task were verified, providing a reliable perception basis for the system. Efficient and personalized intervention strategies were learned in the complex POMDP classroom environment through the adapted SAC algorithm. The simulation results show that the system has the potential to reduce students' distraction and improve classroom concentration. Despite the above contributions, several limitations still restrict the generalizability of this study's findings. The MMAct dataset is not classroom-specific and contains only 20 subjects; the IMU data is only an approximate substitute for the real pressure sensor. The EduSim simulator default assumes a fixed class size N = 25. The real classroom pilot is limited to the feasibility test, and no statistically significant data is generated; issues, such as privacy, ethics, and deployment feasibility, have not been addressed. However, the need to address the above limitations stems from this study's dual value at the theoretical and practical levels. At the theoretical level, integrating POMDP with self-regulated learning and positive behavior support theory provides a formal computational model for conceptualizing classroom interventions as sequential decision-making problems under uncertainty. At the practical level, the three-tier decision support system provides a data-driven tool for personalized learning and classroom management, but its effectiveness in real-world settings has yet to be empirically verified. For practitioners, this study suggests prioritizing investment in perception infrastructure, training teachers to interpret system warnings, and formulating privacy policies before deployment, while retaining teaching autonomy. Therefore, future work will address these limitations through large-scale verification, class-size sensitivity analysis, controlled experiments, and privacy-ethics deployment studies.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

This study used only the publicly available standard dataset MMAct for model validation and independently established an educational simulation classroom environment, EduSim, to test intervention strategies. Throughout the process, no real facial images of teachers and students at the school, classroom privacy images, or students' personal academic privacy data were collected. No offline, real-classroom human-subject experiments were conducted, and there was no ethical risk to human subjects. Therefore, this study does not require approval from the Institutional Review Board (IRB).

Declaration of Artificial Intelligence (AI) Tools

The author confirms that no AI tools were used in the preparation of this manuscript.

References

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17.

Chen, J., Wang, M., Wang, L., and Huang, F. (2024). Student motivation analysis based on raising-hand videos. *Sensors*, 24(14), 4632.

Chen, H., Zhou, G., and Jiang, H. (2023). Student behavior detection in the classroom based on improved YOLOv8. *Sensors*, 23(20), 8385.

Duan, Y. and Zhang, H. (2025). Optimization of student classroom behavior recognition algorithm based on deep learning. *Journal of Computer Science and Electrical Engineering*.

Eich, L. G., Francisco, R., and Barbosa, J. L. V. (2024). Identifying student behavior in smart classrooms: A systematic literature mapping and taxonomies. *International Journal of Human–Computer Interaction*, 41(11), 6743–6764.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6202–6211.

Gao, H., Zeng, Y., Ma, B., and Pan, Y. (2024). Improving knowledge learning through modelling students‘ practice-based cognitive processes. *Cognitive Computation*, 16(1), 348-365.

Jifeng, C., Zhengxi, S., Qingqi, L., and Lei, M. (2024). A deep learning-based system for monitoring student behavior and analyzing learning situations. *2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA)*, 794–798.

Jia, W. and Li, Z. (2025). Fusion of deep reinforcement learning and educational data mining for decision support in journalism and communication. *Information*, 16(12), 1029.

Hong, J., Lee, C., and Jung, H. (2022). Late fusion-based video transformer for facial micro-expression recognition. *Applied Sciences*, 12(3), 1169.

Liu, H., Wang, D., and Wang, C. (2023). Research on behavior recognition algorithms in classroom scenarios. *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, 221–228.

Li, C. and Li, J. (2021). Construction of teaching behaviours analysis model for smart classroom. *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, 133–138.

Lovanshi, M. and Tiwari, V. (2024). Human skeleton pose and spatio-temporal feature-based activity recognition using ST-GCN. *Multimedia Tools and Applications*, 83(5), 12705–12730.

Liu, T. (2021). Semi-supervised with partition strategy and MS-G3D for skeleton-based action recognition. *2021 3rd*

International Academic Exchange Conference on Science and Technology Innovation (IAECST), 128–132.
Liu, Q., Jiang, X., and Jiang, R. (2025). Classroom Behavior Recognition Using Computer Vision: A Systematic Review. *Sensor*s, 25(2), 373.
Ma, L., Zhou, T., Yu, B., Li, Z., Fang, R., and Liu, X. (2024). Improving YOLOv7 for large target classroom behavior recognition of teachers in smart classroom scenarios. *Electronics*, 13(18), 3726.
Mo, K., Ye, P., Ren, X., Wang, S., Li, W., and Li, J. (2024). Security and privacy issues in deep reinforcement learning: Threats and countermeasures. *ACM Computing Surveys*, 56(6), 1–39.
Pang, S., Zhang, Y., Zhang, J., Yang, Y., Sun, D., and Xiang, J. (2025). Automatic detection of students' classroom behavior via long-term classroom videos to predict students' learning gains. *Education and Information Technologies*.
Sun, Y., Huang, W., Wang, Z., Xu, X., Wen, M., and Wu, P. (2023). Smart teaching systems: A hybrid framework of reinforced learning and deep learning. *International Journal of Emerging Technologies in Learning (iJET)*, 18(20), 37–50.
Wang, Z., Wang, M., Zeng, C., and Li, L. (2024). Multi-scale deformable transformers for student learning behavior detection in smart classroom (Version 1). *arXiv*.
Wu, S., Cao, Y., Cui, J., Li, R., Qian, H., Jiang, B., and Zhang, W. (2024). A comprehensive exploration of personalized learning in smart education: From student modeling to personalized recommendations (Version 1). *arXiv*.
Wan, X., Li, T., Lin, W., Cai, Y., and Zheng, Z. (2024). Coverage-guided fuzzing for deep reinforcement learning systems. *Journal of Systems and Software*, 210, 111963.
Wang, X., Miao, Z., Zhang, R., and Hao, S. (2019). I3d-lstm: A new model for human action recognition. *IOP Conference Series: Materials Science and Engineering*, 569(3), 032035.
Xu, L. (2024). Navigating the educational landscape: The transformative power of smart classroom technology. *Journal of the Knowledge Economy*, 16(2), 10389–10420.
Yang, Y., Fu, Z., and Naqvi, S. M. (2023). Abnormal event detection for video surveillance using an enhanced two-stream fusion method. *Neurocomputing*, 553, 126561.
Zong, L. and Fang, J. (2024). Deep visual computing of behavioral characteristics in complex scenarios and embedded object recognition applications. *Sensors*, 24(14), 4582.
Zhou, S. R., Chen, Z. G., and Deng, Y. Q. (2023). A tennis action recognition and evaluation method based on PoseC3D. *Computer Engineering & Science*, 45(1), 95.

Mu Jian is a lecturer at the School of Marxism, Shaanxi Xueqian Normal University, with a major in Ideological and Political Education. She teaches core courses in ideological and political theory for undergraduates. Her research focuses on digital innovation of ideological teaching, integration of traditional Chinese culture into moral education, and youth ideological guidance. She hosts academic lectures and youth teaching forums, and participates in provincial-level seminars on ideological education reform.