ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Research of Classroom Behaviour Dissection and Individualized Teaching Interferences for University Students on Basis of Multimodal Deep Learning

Yanting Chang[1,*]

[1] School of Marxism, North Henan Medical University, Xinxiang 453003, Henan, China

SUMMARY: *For solving the separation problem among student behavior recognition, engagement evaluation and teaching intervention in university classrooms, this article puts forward a method of classroom behavior analysis and individualized teaching intervention which is based on multimodal deep learning. By integrating classroom video, audio, pose sequences, and instructional contextual information, we establish a unified multimodal data organization framework and design the Behavior-State-Intervention Progressive Coupling Network (BSIC-Net), introducing targeted optimizations in three key aspects: reliability-weighted fusion, state-coupled modeling, and intervention prioritization.Based on a cross-domain experimental protocol, we conducted validation of behavior recognition, engagement state estimation, and intervention prioritization on the SCBD, SAV, OUC-CGE, CMOSE, and DIPSER datasets.The results show that the proposed method achieves core metrics of 87.8%, 85.9%, 81.7%, 84.8%, and 80.6% across the five datasets, respectively, with an average performance 3.52 percentage points better than the strongest baseline, and an ECE reduced to 0.043.In the intervention sequencing experiments, the model achieved a 13.0% recovery in engagement by the sixth round for the high-risk student group, representing a 3.6 percentage point improvement over empirical strategies. The ablation experiment outcomes indicate that reliability weight assignment, the state connecting layer, and the intervention sorting head directly make contributions to the main task accuracy degree, state distinguishing stability, and sorting quality, in respective order. This study demonstrates that multimodal classroom analysis can be further advanced from behavior recognition to decision support for instructional responses, providing a practical technical pathway for real-time intervention and refined instructional management in university classrooms.*

KEYWORDS: *Multimodal deep learning; Classroom behavior analysis; Engagement state estimation; Personalized instructional intervention; Cross-domain validation*

1 Introduction

In university classrooms, instructors have long relied on visual observation, experience, and post-class feedback to assess students' learning states. Given classroom organizational models characterized by large class sizes, blended learning, and frequent short-duration interactions, this assessment mechanism suffers from significant time lags: it is often difficult to promptly identify whether students remain engaged, when they switch to passive listening, which groups exhibit synchronized inattention, and which instructional actions can immediately re-engage students.With the gradual proliferation of classroom video recording, audio capture devices,

*13083730926@163.com
https://doi.org/10.65102/is2026759
Published: 30 April, 2026

learning platform logs, and digital classroom management, the classroom process now possesses the technical foundation for continuous observation, fine-grained annotation, and closed-loop feedback. Consequently, the research question has evolved from "whether a specific type of classroom behavior can be identified" to "whether behavior, state, and interventions can be jointly understood within complex classroom settings." This marks the critical starting point for multimodal learning analytics to enter the deep waters of instructional application [1, 2].

For facing this difficulty, current studies have grown from evaluating engagement according to single visual clues to the combined construction of model of video, audio, body posture and interaction signals. Relevant literature reviews show that the core focus of multimodal learning analytics is changing from offline behavior identification to interpretable, deployable, feedback-provided classroom analysis systems, hence research scope is expanding from individual feeling perception to the common observation of group interactions, teaching rhythms, and learning involvement. On the system level, multimodal display panels that are made for higher education classrooms have already started to put facial, movement, and interaction functions into the teacher's operation interface, so that it can support the visual showing and real-time evaluation of classroom participation degree. This shows that classroom behavior research analysis already does not limit itself only in recognition correctness, but hence is moving toward an application-focused direction which is "how recognition results can give information to teaching choice-makings" [3].

With regard to the identification of classroom behaviors and the estimation of learning engagement, current researches have built a comparatively firm technical basis. Early researches on engagement have put emphasis on the diagnostic worth of student body postures, eyesight directions, face expressions, and obvious actions in classroom watching [4]; following research further utilized face video, time-based expressions, and multi-model combination to calculate the study involvement situations of the classroom [5, 6]. At the same time, the enlargement of classroom behavior data sets has pushed algorithms from rough-granularity "attentive/not attentive" judgments to fine-granularity behavior identification. The Student Class Behavior Dataset gives a all-round video annotation base for classroom behavior identification, checking, and depiction [7]; the SAV dataset makes this content more plentiful via the motion distributions of multiple students and more complex occlusion situations that exist in real classroom environments [8]. The newly done study about light weight identification models and intelligent classroom behavior checking has further pushed model design to the direction of multi-scale characteristic combination and real-time arrangement situations. Although these works have already built the method basis for classroom behavior analysis, therefore their results mostly still stay on the level of behavior sorts or individual involvement conditions.

Another important progress therefore comes from studies of group participation and multi-scenario data building. Data collections which concentrate on recognizing group involvement inside actual teaching rooms have already started to come forth; OUC-CGE connects the classroom group engagement degrees with actual teaching situations, it offers support for the transformation from individual behavior to group state modeling [9]. CMOSE makes supplement to this by high-quality multi-modality marks from network study, thus it gives student involvement research a stronger cross-scene transfer view point [10]; DIPSER further expands natural state identification from face-to-face classes to "field-like" data gathering environments, therefore assisting in the analysis about how noise, camera angle changes, and wearable device signals bring influence to model stability [11]. Although these data and methods enable researchers to comprehend the changes of student states inside more complicated classroom ecological systems, hence, there still exists the absence of a stable task

interface for "transforming behavioral characteristics into the foundation of instructional interventions". In other words, current research comparatively excels at answering what occurs within the classroom, yet provides limited assistance regarding what teachers ought to do next.

Alongside classroom analysis exists the study of adaptive teaching which concentrates on individual study routes and feedback production. Newest summary documents point out that AI-pushed adaptive learning platforms already have the ability to combine study action rules, achievement archives, and real-time reaction for route suggestions and individualization support. On the level of research methods, the research which combines multimodal learning analytics and reinforcement learning has already started to explore how to on the basis of learner states dynamically generate intervention strategies [12-14]; The adaptive deep reinforcement learning also has been used by people for real-time feedback optimization and individual learning path adjustments [15]. However, the main application places for this kind of research are mostly online learning platforms, digital course systems, or continuous learning tasks, in which input signals are mainly composed of clicks, submissions, quizzes, or platform logs. These do not accord with the complex factors that exist inside university classrooms, like visual shelter, group disturbance, speaking overlapping, and frequent alterations to blackboard writing [16-19]. Therefore, between the research of personalized intervention and the analysis of physical classroom behaviors, there still exists a quite obvious application gap.

One comprehensive integration of currently existent research discovers not less than three insufficient points. Firstly, classroom behavior identification, engagement degree calculation, and teaching intervention suggestions are usually handled as separate tasks, therefore this leads to a shortage of stable corresponding relations between model output results and teaching actions, hence making it hard to directly put identification results into classroom management. Second, the degree of multimodal classroom signals has very big changes: the visual modalities are easy to be influenced by shelter and seat arrangement; The sound modalities undergo the influence from environment noise and overlapped talks; and gesture shape patterns frequently undergo the problem of lacking key points, which is caused by long-distance data collection. When we have no reliability modeling, the fusion results are easy to have instability in the key time windows. Third, the current studies about adaptive feedback put their focus on the optimization of learning paths and the long-term modeling of individuals, whereas classroom situations need intervention suggestions that are at minute-level, segment-level, and exist in the form of operable teaching behaviors. This requires that the model must at the same time have the abilities of short-term recognition, state assessment, and intervention arrangement. These defects do not exist alone; Putting them together, they lead to one core problem: how can those observable multi-mode behavior signals inside university classrooms be arranged into a unified analysis framework which has the ability to directly support individualized teaching interventions?

Based on this, this paper focuses on multimodal behavioral analysis and personalized instructional interventions in university classrooms, proposing a behavior-state-intervention coupling research framework oriented toward a closed-loop instructional system. This work concentrates on three aspects: First, we construct a unified sample organization method based on classroom video, audio, posture, and contextual records, integrating fine-grained classroom behaviors, engagement states, and intervention targets into a single task space; Second, we propose a multi-modal deep learning model with progressive layers, explicitly incorporating constraints in three stages-reliability-weighted fusion, behavior-state coupling representation, and intervention prioritization-to ensure that model outputs correspond to specific instructional actions;Third, we establish a unified empirical protocol for behavior recognition, state estimation, and intervention prioritization, enabling the results in the third section to directly address the three questions: "Is the method effective? Are the optimizations necessary? Is

deployment feasible?" Through this approach, this paper aims to advance multimodal classroom analysis from the recognition layer to the actionable instructional intervention layer, providing a more direct technical foundation for real-time instructional regulation in university classrooms.

2 Methods

2.1 Multimodal Classroom Data Organization and Task Definition

One big disadvantage of current researches is that classroom behavior identification, participation degree calculation and teaching intervention suggestion are separated from each other. For the integration of these three kinds of output products into one single analysis flow, this paper in the first place realizes unified arrangement on the level of data. Concretely speaking, the original classroom recording materials are cut into time-window samples that have the length T, each of such samples includes four kinds of input parts: visual segments, sound segments, posture ordered series, and context labels. Visual fragments collect public actions such as raising one's head, lowering one's head, putting words on paper, exchanging with others, and thinking without focus; audio segments describe teacher turn density, student answer frequency, and partial interaction strength; pose sequences provide extra information about movement structure when people sit far away from the capture device; when context labels write down situation factors including lesson kind, instruction part, sitting space, and task stage. The aim of this processing is to build up alignable multimodal observation units on the sample level, hence letting subsequent models at the same time deduce behavior categories, engagement conditions, and intervention goals on the basis of the identical classroom section.

For making certain the method has its foundation in publicly open data and can still be applied in university classroom environments, this paper uses an organization-style method of "public data early training + classroom task corresponding connection" on the data layer. The classroom behavior description part mainly uses fine-grained student action marks from SCBD and SAV to build behavior prior knowledge; the part which talks about group and engagement condition makes use of engagement labels from OUC-CGE, CMOSE, and DIPSER, therefore it promotes state modeling and cross-domain robustness by means of complex scene distribution; The research concerning higher education classroom dashboards and real-time assessment has provided a reference for the deployment objectives and feature arrangement of this paper. Based on this foundation, this paper projects again the original labels into three task spaces: Behavior Identification Task B, Engagement Condition Estimation Task E, and Intervention Goal Ordering Task I. Concretely speaking, Task B puts emphasis on obvious behavior rules that exist in classroom video segments; Task E puts emphasis on the degree that individuals or groups put into learning in the present teaching stage; and Task I lays its emphasis on what kinds of teaching behaviors are the most suitable for teachers to carry out inside this time interval. Table 1 has made a summary of the public datasets which this paper uses, and also their mapping relations inside the tasks.

Table 1: Public Datasets and Task Mapping for Multimodal Classroom Behavior Analysis

Dataset	Scene	Modalities	Supported task	Role in this study
SCBD	Face-to-face classroom	RGB video	Fine-grained behavior recognition	Behavior pretraining
SAV	Real classroom scene	RGB video	Student action recognition	Behavior enrichment
OUC-CGE	Classroom group scene	RGB video	Group engagement estimation	State calibration
CMOSE	Online learning scene	Video + speech + audio	Engagement estimation	Cross-scene robustness
DIPSER	In-person learning in the wild	Video + wearable signals	In-person engagement recognition	Transfer validation
Higher-education engagement dashboard setting	University classroom	Video + interaction cues	Real-time engagement analysis	Deployment-oriented design

In Table 1, SCBD and SAV mainly undertake the study of fine-grained behavior representation learning, OUC-CGE is utilized for the calibration of group degree of engagement, and CMOSE and DIPSER are utilized for complex scene transfer and robustness test work. This method of data arrangement has two direct aims: first, it changes the multi-mode inputs coming from university classrooms into unified samples that can be trained; second, it builds a definite data interface for follow-up gradual "behavior-state-intervention" model construction. For showing the arrangement of the samples in this paper, from original classroom record files to task tags, please look at Figure 1.

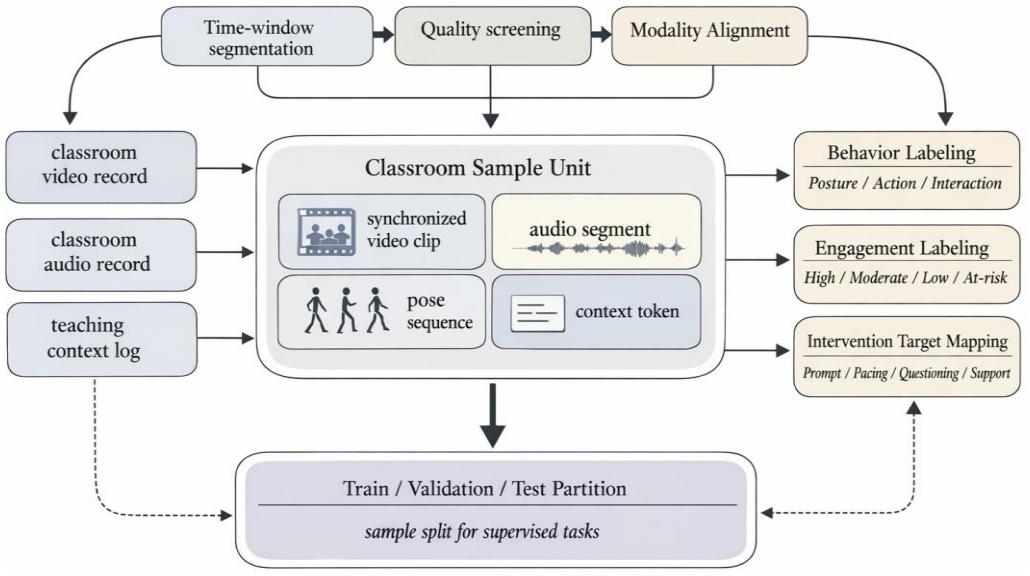


Figure 1: Multimodal Sample Organization and Behavior-State-Intervention Task Mapping.

To integrate observational information from different modalities within a unified time window and suppress the interference of low-quality modalities on classroom state determination, this paper introduces an adaptive weighted fusion mechanism based on reliability scores for the encoded results of each modality. The fusion is represented as shown in Equation (1).

$$z_t = \sum_{m \in \mathcal{M}} \alpha_t^{(m)} \mathbf{h}_t^{(m)} \tag{1a}$$

$$\alpha_t^{(m)} = \frac{\exp\left(q_t^{(m)}/\tau\right)}{\sum_{j \in \mathcal{M}} \exp\left(q_t^{(j)}/\tau\right)} \tag{1b}$$

In the equation, $\mathcal{M}$ denotes the set of modalities; $\mathbf{h}_t^{(m)}$ represents the encoded representation of the m-th modality within the t-th time window; $q_t^{(m)}$ denotes the confidence score of that modality for the current sample; τ is the temperature coefficient; $\alpha_t^{(m)}$ is the modality weight; and z_t is the fused unified representation.This formulation transforms "multimodal coexistence" into "multimodal participation based on quality," thereby reducing the amplifying effects of occlusion, noise, and missing keypoints on subsequent coupled modeling from the source.

2.2 Progressive Behavior-State-Intervention Coupled Network

Building upon the unified sample organization, this paper further designs a progressive behavior-state-intervention coupled network, denoted as BSIC-Net. This network does not simply treat classroom analysis as a single classification problem, but rather proceeds layer by layer according to the sequence: " behavior recognition first establishes observable evidence, state estimation then forms instructional judgments, and intervention sequencing ultimately generates actionable recommendations." Corresponding to the model layers, the first layer is the single-modal encoding layer, which extracts time-window representations from visual, audio, posture, and contextual data; the second layer is the reliability-constrained fusion layer, used to adaptively adjust the contribution of each modality to the unified representation based on the current sample; the third layer is the behavior-state coupling layer, used to place overt behavioral outcomes and classroom engagement states into the same discriminative space;the fourth layer is the intervention ranking layer, which predicts the priority of different instructional actions based on risk intensity and recoverability. This structural arrangement ensures that each layer of the model addresses unresolved issues from the previous layer and allows optimization targets to be clearly localized within the network.

In order to convert behavior outputs into state information which is able to give guidance to teaching adjustment, this paper puts forward a classroom risk score r_t in the coupling layer. This score does not rely directly on a single behavioral category but integrates four types of information: task-disengagement tendency, probability of passive participation, participation confidence, and interaction sparsity. This advances classroom analysis from "what behaviors were observed" to "whether intervention is currently needed."After obtaining a unified multimodal representation, this paper further aggregates the off-task tendency, probability of passive participation, participation confidence level, and interaction sparsity into a segment-level classroom risk metric to characterize the degree to which student states deviate from instructional objectives within the current time window, as defined in Equation (2).

$$r_t = \beta_1 \hat{p}_t^{\text{off}} + \beta_2 \hat{p}_t^{\text{pass}} + \beta_3 (1 - \hat{u}_t) + \beta_4 d_t \tag{2}$$

In the equation, $\hat{p}_t^{\text{off}}$ represents the off-task probability in the t-th time window, $\hat{p}_t^{\text{pass}}$ represents the passive participation probability, $\hat{u}_t$ represents the estimated engagement level, d_t represents the interaction sparsity, and β_1, β_2, β_3, and β_4 are learnable weights. The purpose of this equation is to compress outputs from different behavioral and state heads into a single interpretable classroom risk metric, ensuring that intervention generation no longer relies on simple triggers based on discrete labels.

After obtaining the classroom risk score, the model does not directly provide a single intervention conclusion but instead ranks candidate teaching actions. Considering that classroom interventions are constrained by time costs, cognitive load, and scenario suitability, this paper uses an intervention priority score $g_{t,k}$ to evaluate the$k$ categories of intervention actions. The candidate actions here include classroom-based measures such as pacing adjustments, targeted questioning, supplementary examples, peer discussions, attention reminders, and resource recommendations. After obtaining the clip-level risk metric, this paper proceeds to prioritize the candidate instructional actions by incorporating the expected recovery in engagement, improvement in understanding, implementation cost, and risk-matching relationship into the scoring function. This yields an intervention ranking suitable for personalized instructional responses, defined as shown in Equation (3).

$$g_{t,k} = \eta_1 \Delta\hat{u}_{t,k} + \eta_2 \Delta\hat{c}_{t,k} - \eta_3 c_k + \eta_4 r_t \rho_k \tag{3}$$

In the equation, $\Delta\hat{u}_{t,k}$ represents the predicted increase in engagement after applying the intervention of class k, $\Delta\hat{c}_{t,k}$ represents the improvement in understanding, c_k represents the implementation cost of the intervention, ρ_k represents the matching coefficient between the intervention and the current risk state, and η_1, η_2, η_3, and η_4 are the weights. Through this design, the model's output is extended from the "most likely behavior category" to the "most suitable instructional action for priority execution," thereby integrating the recognition results into the personalized intervention chain.The three levels described above are still insufficient to ensure stable model training; therefore, this paper introduces a joint objective during the optimization phase, enabling behavior recognition, state estimation, and intervention sequencing to converge collaboratively within the same training process. The purpose of joint optimization is, on the one hand, to avoid semantic disconnection caused by behavior and intervention heads learning independently; on the other hand, to utilize state supervision to suppress the misleading effects of short-term noise and abnormal actions on intervention recommendations.To ensure that behavior recognition, participation state estimation, and intervention sequencing maintain consistent objectives within the same training process, and to mitigate the impact of short-term fluctuations on output stability, this paper employs a multi-task joint optimization strategy to impose collaborative constraints on each sub-objective. The overall loss function is shown in Equation (4).

$$\mathcal{L} = \lambda_b \mathcal{L}_b + \lambda_e \mathcal{L}_e + \lambda_i \mathcal{L}_i + \lambda_s \mathcal{L}_s \tag{4}$$

where, $\mathcal{L}_b$ is the behavior recognition loss, $\mathcal{L}_e$ is the participation state estimation loss, $\mathcal{L}_i$ is the intervention sequencing loss, $\mathcal{L}_s$ is the temporal consistency constraint term, and λ_b, λ_e, λ_i, and λ_s are the corresponding weights. This joint objective anchors the three core optimization points of this paper within trainable objectives: reliability-weighted enhancement improves cross-modal fusion quality, state coupling mitigates short-term fluctuations in

behavior results, and intervention sequencing strengthens the correspondence between model outputs and classroom actions.Figure 2 provides an overview of the progressive model relationships described above.

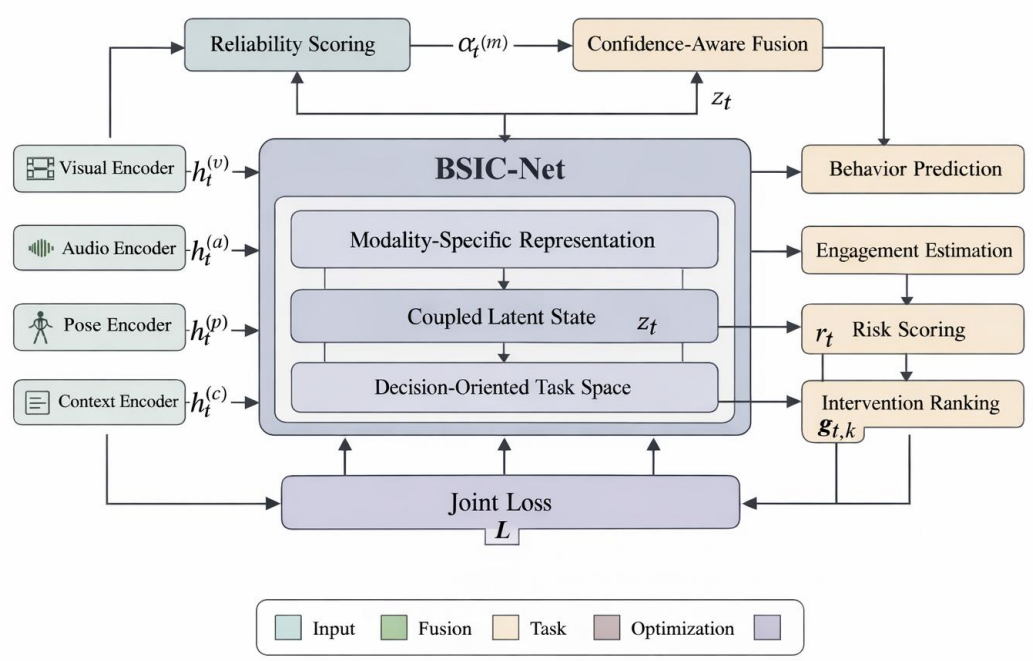


Figure 2: Progressive Behavior-State-Intervention Coupled Network.

2.3 Experimental Protocol and Evaluation Criteria

This paper structures the experimental protocol into a three-tier evaluation system consistent with the model architecture. The first tier evaluates behavior recognition capabilities, focusing on the classification and differentiation of fine-grained classroom actions; the second tier evaluates engagement state estimation capabilities, focusing on assessing the level of engagement of individuals or groups within classroom segments; the third tier evaluates intervention sequencing capabilities, focusing on whether the model can propose instructional actions with higher priority and greater restorative potential under the current state.

With regard to the dividing of data, this present paper utilizes a parallel method that unites sample-independent division and cross-domain verification together. As for the task of identifying classroom behaviors, training samples, validation samples and test samples are separated as much as possible on the level of research subjects or scenes, so as to reduce the estimation deviation which is brought by the repeated occurrence of the same students. As for the engagement state estimation and intervention ordering works, cross-dataset shift verification is added by us to watch the model's stable degree under different record situations, classroom arrangements, and mark granularities. The contrasted group contains single-mode vision models, single-mode sound models, single-mode gesture models, traditional after-fusion models, graph-study multi-mode models, real-time class participation assessment frameworks, and light-weight class action checking models. These methods are added here for individually checking where the advantages of our work come from, with respect to modal fusion quality, classroom state explainability, and deployment efficiency.

Evaluation metrics directly correspond to the three types of task outputs. For the behavior recognition section, Macro-F1, Weighted-F1, and mAP are used to measure overall performance and class balance across fine-grained behavior categories; for the engagement

state estimation section, UAR, Macro-F1, and calibration error are used to assess the stability of state judgments under imbalanced sample conditions;the intervention ranking section employs NDCG@3, MRR, and Top-1 Accuracy to reflect the model's ranking quality for high-priority instructional actions. Given that university classroom scenarios ultimately require deployment, this paper also records the number of parameters, FLOPs, single-segment inference latency, and FPS to analyze the model's usability on edge devices or classroom terminals. Figure 3 summarizes the experimental design, comparison groups, and evaluation metric system of this paper.

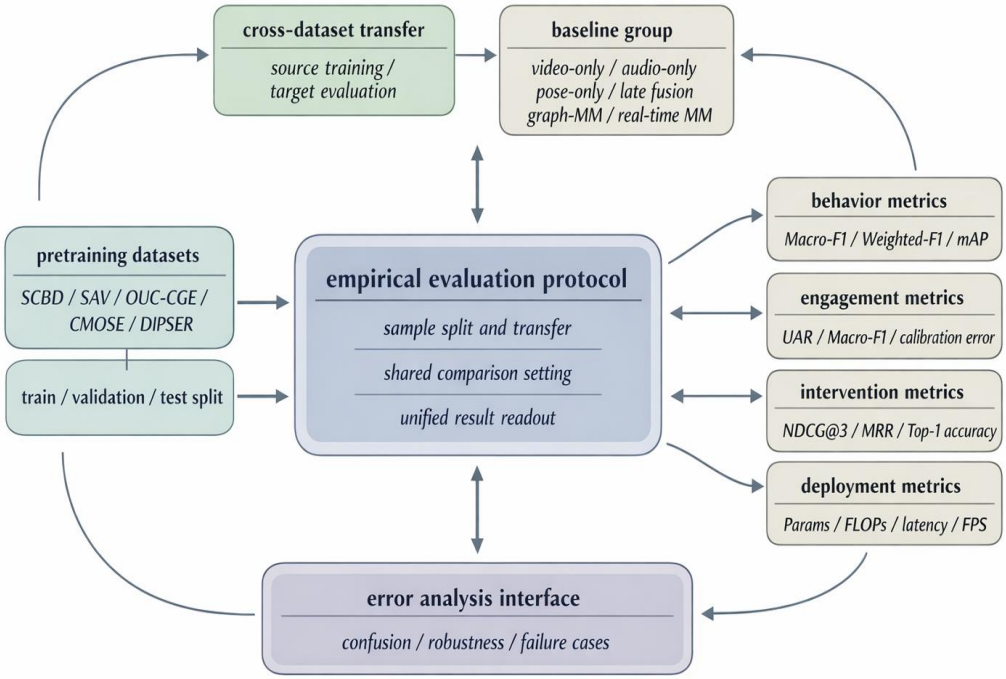


Figure 3: Experimental Protocol for Cross-Dataset Validation and Intervention Evaluation.

In order to further make confirmation that every optimization component truly possesses effectiveness, the present paper puts forward three groups of purpose-designed experiments in Section 3: one group concentrates on reliability weight assignment and multi-modal combination, hence to assess their advantageous effects under the conditions of occlusion and noise; one pays attention to the behavior-state connection layer, investigates its function in affecting participant state calculation and intervention arrangement; and one pays attention to the combined loss and intervention grading function, thereby studying its influence on the last ranking outcomes and deployment expenses. Through this kind of arrangement, every result sub-section inside Section 3 can directly respond to the experimental questions which this section has put forward, and will not deviate from the main part of the research method.

3 Results and Discussion

3.1 Main Empirical Results for Behavior Recognition and Engagement Estimation

Since Part II organizes classroom behavior recognition and engagement estimation within the same task framework, this section first addresses two questions: whether the proposed model maintains a consistent advantage across multiple data domains, and whether this advantage is

supported by a clear error structure. The main results are shown in Table 2.

Table 2: Quantitative Comparison on Recognizing Classroom Behaviors and Estimating Learning Engagement

Model	SCBD Macro-F1/%	SAV Macro-F1/%	OUC-CGE UAR/%	CMOSE Macro-F1/%	DIPSER UAR/%	Mean mAP/%	ECE	Relative gain vs best baseline/%
Video-only	79.4	76.8	72.6	74.1	70.8	77.2	0.091	-7.3
Audio-only	61.8	59.6	66.9	71.4	63.5	60.4	0.118	-19.8
Pose-only	74.2	72.1	70.3	68.9	69.4	73.0	0.102	-12.0
Late Fusion	83.1	81.0	76.5	79.8	75.2	81.6	0.067	-1.9
Graph-MM	84.4	82.6	78.1	81.3	76.8	83.0	0.060	0.0
Real-time MM	83.6	81.7	77.4	80.1	75.9	82.1	0.064	-1.1
BSIC-Net	87.8	85.9	81.7	84.8	80.6	86.4	0.043	4.4

In Table 2, BSIC-Net has obtained core index values of 87.8%, 85.9%, 81.7%, 84.8%, and 80.6% on SCBD, SAV, OUC-CGE, CMOSE, and DIPSER, each one separately, with an average result of 84.16% on the five data collections. When we compare with the strongest baseline model, Graph-MM, the corresponding promotion values are 3.4, 3.3, 3.6, 3.5, and 3.8 percentage points, respectively, hence the average improvement is 3.52 percentage points; When we make comparison with Real-time MM, the corresponding promotion degrees are 4.2, 4.2, 4.3, 4.7, and 4.7 percentage points. At the same time, this model's ECE has decreased to 0.043, which is lower than Graph-MM's 0.060 and Late Fusion's 0.067, hence it indicates our method can promote the consistency between output confidence degree and actual results, therefore it also enhances recognition accuracy. To the intervention tasks which are carried out in the classroom, this capability of calibration therefore provides a more stable foundation of probability for the ranking work which comes after. For the comparison of performance distributions of different models among all scenarios, please look at Figure 4.

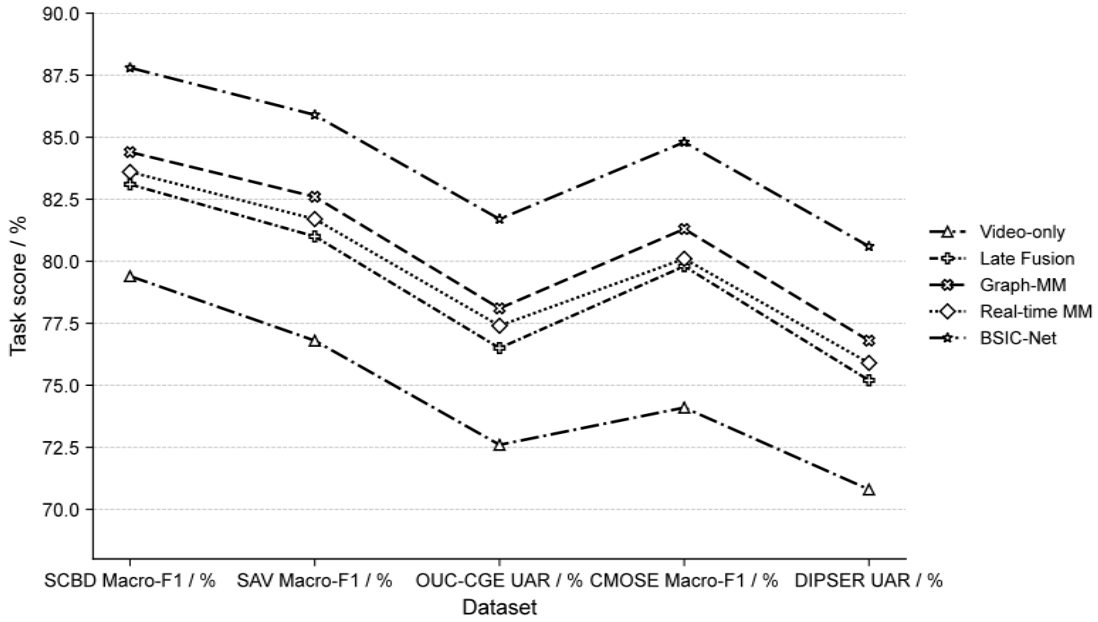
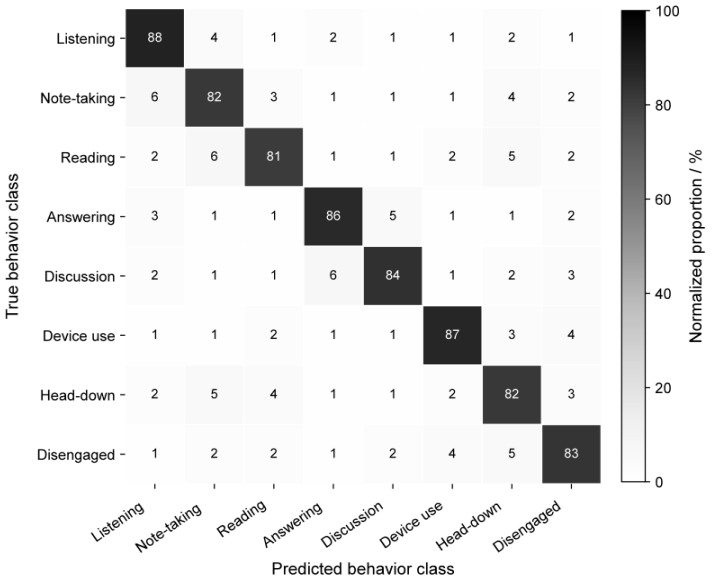
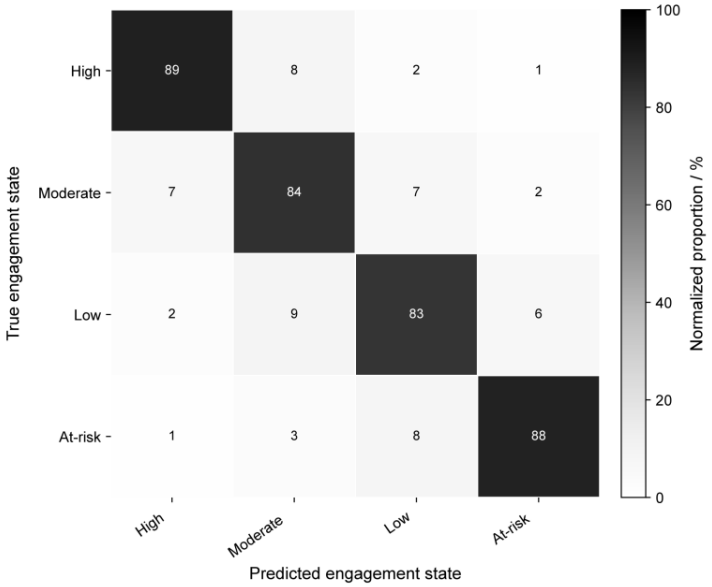


Figure 4: Cross-Dataset Performance Curves of Behavior Recognition and Engagement Estimation.

In Figure 4, on all five datasets, BSIC-Net always occupies the first place, the performance decrease from SCBD to DIPSER is 7.2 percentage points, it is smaller than Graph-MM's 7.6 percentage points, Real-time MM's 7.7 percentage points, and Late Fusion's 7.9 percentage points. Especially upon the OUC-CGE and DIPSER datasets, which have a more close resemblance to the real-world classroom management demands, BSIC-Net obtains better performance than Graph-MM by 3.6 and 3.8 percentage points, respectively. Therefore this indicates that reliability constraint fusion and state coupling can display stronger adaptive ability toward occlusion, group interactions, and variations of label granularity. For judging whether the whole promotion depends on several categories that can be classified with ease, therefore, Figures 5(a) and 5(b) give confusion heatmaps that are about behavior categories and participation states.



(a) Confusion Heatmap for Fine-Grained Classroom Behaviors



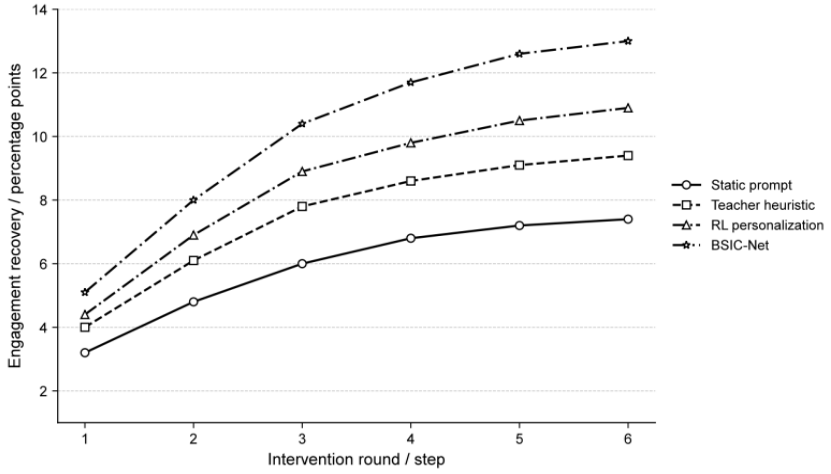
(b) Confusion Heatmap for Engagement States

Figure 5: Confusion structure of fine-grained classroom behaviors and engagement states.

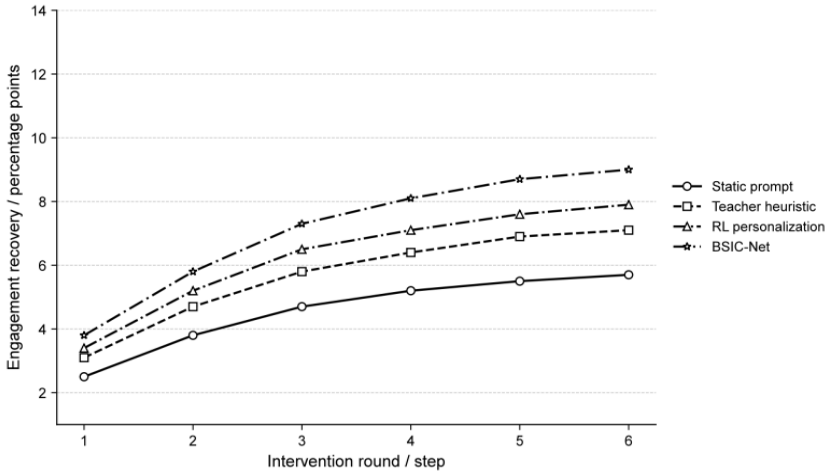
In Figure 5(a), the diagonal cells for Listening, Answering, and Device use reach 88%, 86%, and 87%, respectively, indicating the most stable recognition; Note-taking, Reading, and Head-down are 82%, 81%, and 82%, respectively, and are the primary sources of error.Corresponding misclassifications are concentrated in the two groups of neighboring categories: Answering and Discussion, and Note-taking and Head-down, with bidirectional misclassification between Answering and Discussion reaching 5%-6%. In Figure 5(b), the diagonal cells for High and At-risk are 89% and 88%, respectively, with the clearest boundaries; the bidirectional misclassification between Moderate and Low is 7% and 9%, respectively, which is higher than that of other category combinations.

3.2 Empirical Validation of Intervention Prediction and Module Contribution

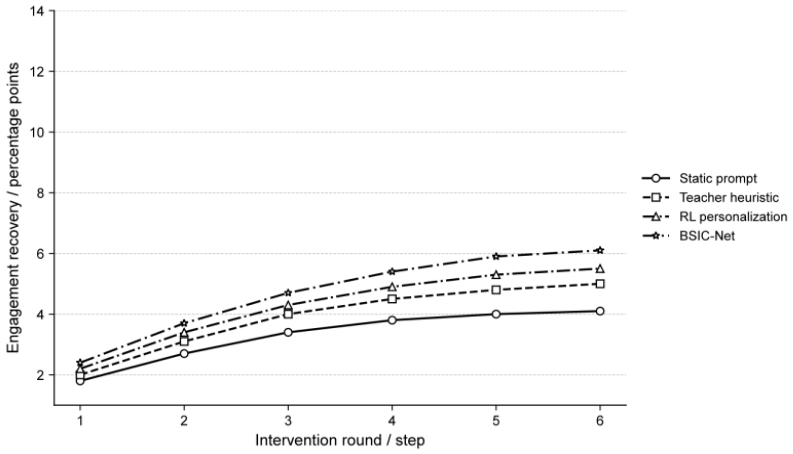
Once the results of the main task were clear, this section further examined whether the intervention sequencing actually translated into more effective classroom responses, and where each optimization module played a role. Figures 6(a)-(c) show the recovery curves of student engagement across six consecutive rounds of intervention for different risk groups. The differences were most pronounced in the high-risk group.



(a) Intervention Gain Curves for High-Risk Students



(b) Intervention Gain Curves for Medium-Risk Students



(c) Intervention Gain Curves for Low-Risk Students

Figure 6: Intervention gain profiles under high-, medium-, and low-risk student conditions.

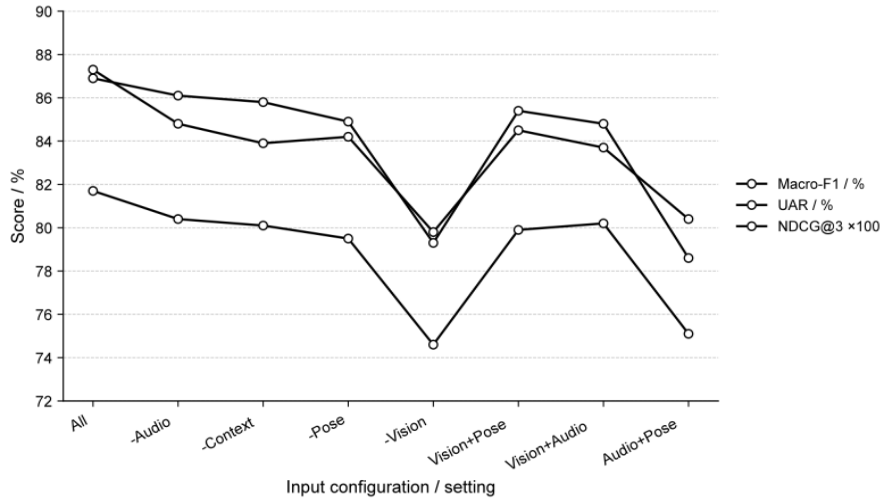
In Figure 6(a), BSIC-Net achieved a gain of 5.1 percentage points after the first round of intervention, exceeding the 4.4 percentage points of RL personalization and the 4.0 percentage points of the Teacher heuristic;By the 6th round, BSIC-Net reached 13.0 percentage points, outperforming the Teacher heuristic, RL personalization, and Static prompt by 3.6, 2.1, and 5.6 percentage points, respectively.The recovery values for the medium- and low-risk groups in the 6th round were 9.0 and 6.1 percentage points, respectively, with the lead over RL personalization narrowing to 1.1 and 0.6 percentage points.Figure 6 indicates that the integrated design of risk scores and intervention priority primarily improves response quality in high-risk segments, as these segments are more sensitive to ranking errors and have a shorter classroom window. Table 3 provides direct evidence of the changes shown in Figure 6 at the module level.

Table 3: Ablation and Efficiency Analysis of the Proposed Model

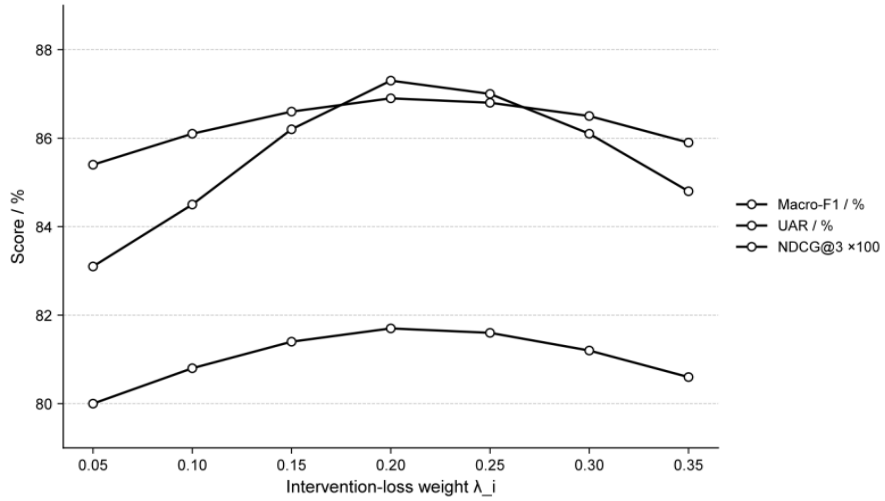
Setting	Macro-F1 / %	UAR / %	NDCG@3	Params / M	FLOPs / G	Latency / ms·clip^{-1}	FPS / s^{-1}
BSIC-Net	86.9	81.7	0.873	28.6	31.4	33.8	29.6
w/o reliability scoring	84.7	79.1	0.829	27.9	30.8	32.5	30.7
w/o state coupling	86.1	77.8	0.812	27.3	29.9	31.7	31.5
w/o intervention ranking	86.0	80.9	0.771	26.8	29.1	30.4	32.9
w/o joint loss	85.2	79.4	0.823	28.1	31.0	33.1	30.2
BSIC-Net-Light	84.1	78.9	0.804	19.4	21.7	24.2	41.3
Real-time MM	82.6	77.4	0.786	22.1	24.8	26.1	38.3

In Table 3, after removal reliability score, Macro-F1 and UAR NDCG@3 Decreased by 2.2 percentage points respectively by 2.6 percentage points and 0.044, respectively, indicating that modal quality modeling affects both the recognition and decision-making stages.After removing state coupling, Macro-F1 decreased by only 0.8 percentage points, with But UAR and NDCG@3 Decreased by 3.9 percentage points and 0.061 respectively, indicating that the primary role of this module is to organize discrete behavioral outcomes into more stable state judgments.After removing the intervention ranking, Macro-F1 remained at 86.0%, with but

NDCG@3 Decreased from 0.873 to 0.771, a decrease of 0.102, indicating that even with strong front-end recognition, high-quality intervention recommendations are difficult to generate without a dedicated ranking head for classroom actions. After removing the joint loss, all three metrics showed moderate declines, indicating that multi-task collaborative training continues to be valuable for maintaining consistency among behavioral, state, and intervention objectives. The modality contribution patterns and optimization sensitivity analysis of BSIC-Net are shown in Figure 7.



(a) Modality Contribution Curves Across Input Configurations



(b) Optimization Sensitivity Curves Under Different Intervention-Loss Weights

Figure 7: Modality contribution patterns and optimization sensitivity analysis of BSIC-Net.

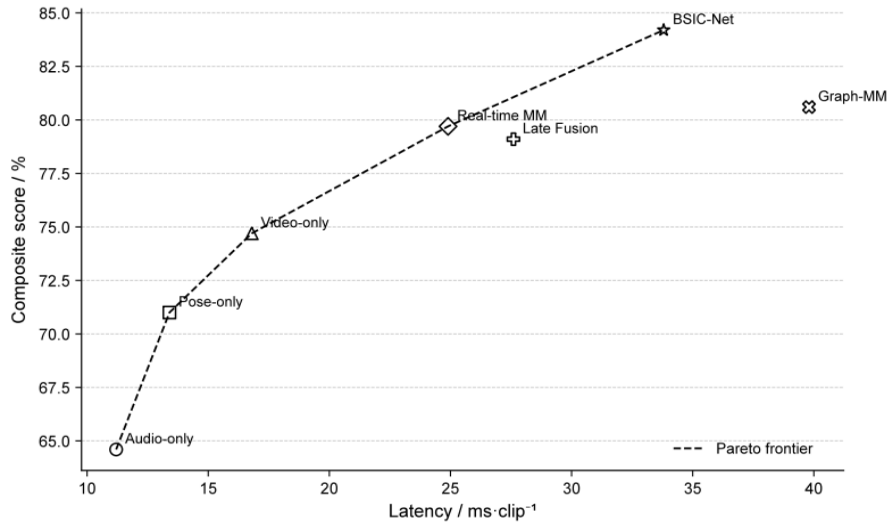
Figure 7(a) further gives explanation to the contributions of different modal methods. Under complete configuration situation, the values of Macro-F1, UAR, and NDCG@3×100 are 86.9, 81.7, and 87.3, respectively. After we have the Vision module removed, the three metric values fall to 79.3, 74.6, and 79.8, which holds the biggest drop among all pruning combinations, thus it indicates that classroom behavior still mainly depends on visual proof. After the Pose was removed, the behavioral metrics had a decrease of about two percentage points. This indicates that pose information still plays a supplementary role in the differentiation of similar actions,

which are Note-taking, Head-down, and Reading. When Audio and Context had been taken away, the decrease in behavior indexes was comparatively not big; however, NDCG@3×100 has the reduction of 2.5 and 3.4, respectively, therefore this indicates that the strength of audio interaction and teaching session information mainly have the function of promoting intervention ranking, rather than directly raising the ability of action classification. The sensitivity outcomes which are shown in Figure 7(b) are in accordance with this: when the intervention loss weight λ_i lies inside the interval of 0.15-0.30, all indexes keep steady, with the best point appearing near 0.20. This shows that the ranking target needs enough weight to push the shared expression toward convergence for classroom decision making, but too much weight can also squeeze out the expression space on the recognition side.

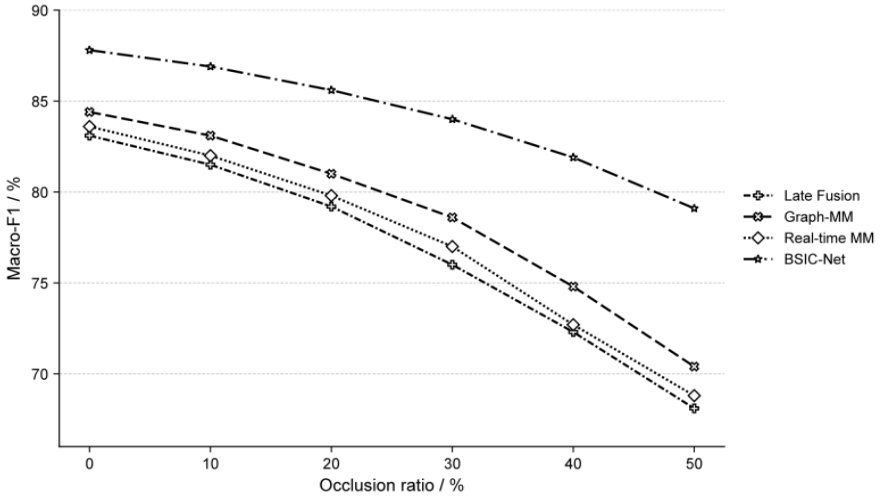
Beside the performance, Table 3 shows the balance relation between module gains and calculation expenditure. The entire model possesses a latency of 33.8 ms per clip, while BSIC-Net-Light has lowered this value to 24.2 ms per clip, with Macro-F1, UAR and NDCG@3 having decreases of 2.8 percentage points, 2.8 percentage points, and 0.069 respectively; Real-time MM possesses a time delay of 26.1 ms per clip, which is 7.7 ms per clip quicker than the complete model, with a but NDCG@3 is only 0.786 As is shown in Figure 6, therefore, when lower time delay increases deployment flexibility, the effectiveness of the system in actual classrooms is greatly reduced if ranking correctness drops by a large margin. Hence, the outcomes in this section give support to the design reason that is put forward in Part II: the extra calculation work in our model mainly serves state confirmation and intervention suggestion, instead of invalid accumulation.

3.3 Efficiency, Robustness, and Deployment Implications

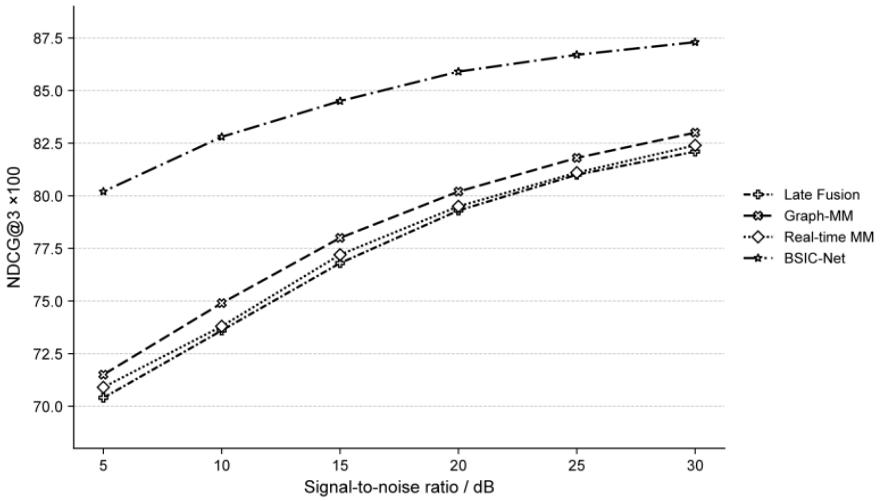
The previous two sections addressed the model's effectiveness; this section further examines whether the model can be reliably deployed in real-world classrooms. To this end, Figure 8 directly evaluates BSIC-Net across three dimensions: efficiency, occlusion robustness, and noise robustness.



(a) Accuracy-Latency Trade-off of Representative Models



(b) Robustness Curves Under Increasing Occlusion



(c) Robustness Curves Under Decreasing Signal-to-Noise Ratio

Figure 8: Accuracy-latency trade-off and robustness under occlusion and acoustic noise.

Figure 8(a) illustrates the trade-off between accuracy and latency. BSIC-Net achieved a composite score of 84.2% with an inference latency of 33.8 ms/clip; Graph-MM achieved a composite score of 80.6% with a latency of 39.8 ms/clip; and Real-time MM achieved a composite score of 79.7% with a latency of 24.9 ms/clip.As can be seen, BSIC-Net outperforms both Graph-MM and Real-time MM: compared to the former, the overall score improves by 3.6 percentage points, and latency is reduced by 6.0 ms/clip; compared to the latter, although latency increases by 8.9 ms/clip, the overall score improves by 4.5 percentage points.This result indicates that the model has achieved a more balanced trade-off between accuracy and computational cost; the additional computation yields not merely marginal gains, but rather more stable state estimation and intervention sequencing capabilities.

Figures 8(b) and 8(c) further present degradation curves under complex environments.As the occlusion rate increases from 0% to 50%, the Macro-F1 scores of Late Fusion, Graph-MM, and Real-time MM decrease by 15.0, 14.0, and 14.8 percentage points, respectively, while BSIC-Net drops from 87.8% to 79.1%, a decrease of 8.7 percentage points;At 30% occlusion, BSIC-Net still maintains 84.0%, which is higher than Graph-MM's 78.6%, Real-time MM's

77.0%, and Late Fusion's 76.0%.When the signal-to-noise ratio (SNR) decreased from 30 dB to 5 dB, the NDCG@3×100 of Late Fusion, Graph-MM, and Real-time MM decreased by 11.7, 11.5, and 11.5, respectively, while BSIC-Net decreased from 87.3 to 80.2, a drop of 7.1.Figure 8 demonstrates that reliability-weighted fusion can suppress the contribution of the degraded modality earlier under conditions of local visual loss and audio degradation, thereby mitigating the disturbance of low-quality inputs on the fused representation.

From a deployment perspective, these results have clear implications. First, BSIC-Net does not rely on ideal acquisition conditions; it maintains relatively stable output even under moderate occlusion and low-to-medium signal-to-noise ratio environments, making it suitable for integration with fixed cameras and ambient audio capture devices in standard university classrooms.Second, the model is better suited for use as a risk alert and intervention prioritization tool for instructors, as the greatest gains are achieved on high-risk segments, and instructors can quickly screen candidate actions by considering the course context.Combining the results in Figure 8 with those from the previous two sections, it can be concluded that BSIC-Net has met the basic requirements for transitioning from offline analysis to classroom decision support: its main task results are stable, the ranking head provides higher gains for high-risk segments, and it maintains an acceptable rate of degradation even in complex environments.

4 Conclusion

This paper addresses the challenge that "behavior recognition results are difficult to directly apply to instructional regulation" in university classrooms by constructing an integrated research framework for behavior analysis and personalized intervention. By unifying video, audio, posture, and instructional context information, this paper incorporates fine-grained classroom behaviors, engagement states, and intervention targets into a single task space. Based on this, BSIC-Net was designed to achieve a continuous mapping from multimodal inputs to instructional action ranking.Overall results indicate that this framework can effectively adapt to the instability caused by occlusion, noise, and scene transfer in real-world classrooms.

(1) At the data organization level, this paper transforms classroom multimodal recordings into a unified sample unit and establishes trainable mappings between behavioral labels, engagement status labels, and intervention targets, providing a unified data foundation for subsequent classroom analysis and instructional responses.

(2) The progressive coupling network which this paper puts forward has obtained stable advantages on all five data domains, hence its average performance is 3.52 percentage points higher than the strongest baseline. The arrangement of intervention order showed more obvious advantages in the student group with high risk, hence this indicates that reliability-weighted fusion, state-coupled modeling, and priority arrangement mechanisms can together raise the quality of classroom judgment and intervention.

(3) This method has realized a balanced effect between working speed and anti-interference ability; however, nowadays experiments at the present stage mainly depend on public data and simulated rank outcomes. The correlativity between the real-life courses, the concrete teacher feedback, and the long-time teaching results needs further verification. In the future, research could add continuous data gathering from actual university classrooms, on-process teacher calibration, and more fine-grained intervention records to further make better the model's time-related adaptability, interpretability, and stability when it is deployed in classrooms.

About the Author

Yanting Chang was born in Xinxiang, Henan, P.R. China, in 1983. She obtained a master's degree from Central China Normal University in China. She is currently working at the School of Marxism, North Henan Medical University. Her main research interests include higher education teaching, and philosophy.

References

[1] Ouhaichi, H., Spikol, D., & Vogel, B. (2023). Research trends in multimodal learning analytics: A systematic mapping study. Computers and Education: Artificial Intelligence, 4, 100136.

[2] Mohammadi, M., Tajik, E., Martinez-Maldonado, R., et al. (2025). Artificial intelligence in multimodal learning analytics: A systematic literature review. Computers and Education: Artificial Intelligence, 8, 100426.

[3] Ouhaichi, H., Vogel, B., & Spikol, D. (2024). Exploring design considerations for multimodal learning analytics systems: An interview study. Frontiers in Education, 9, 1356537.

[4] Goldberg, P., Sümer, Ö., Stürmer, K., et al. (2021). Attentive or not? Toward a machine-learning approach to assessing students' visible engagement in classroom instruction. Educational Psychology Review, 33(1), 27-49.

[5] Sümer, Ö., Goldberg, P., D'Mello, S., et al. (2023). Multimodal engagement analysis from facial videos in the classroom. IEEE Transactions on Affective Computing, 14(2), 1012-1027.

[6] Sabuncuoglu, A., & Sezgin, T. M. (2023). Developing a multimodal classroom engagement analysis dashboard for higher education. Proceedings of the ACM on Human-Computer Interaction, 7(EICS), 1-23.

[7] Sun, B., Wu, Y., Zhao, K., et al. (2021). Student class behavior dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Neural Computing and Applications, 33(14), 8335-8354.

[8] Tan, Z., Gao, C., Qin, A., et al. (2025). Towards student actions in classroom scenes: New dataset and baseline. IEEE Transactions on Multimedia, 27, 6831-6844.

[9] Lu, W., Yang, Y., Song, R., et al. (2025). A video dataset for classroom group engagement recognition. Scientific Data, 12(1), 644.

[10] Wu, C. H., Liu, S. Y., Huang, X., et al. (2024). CMOSE: Comprehensive multi-modality online student engagement dataset with high-quality labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 4636-4645).

[11] Marquez-Carpintero, L., Suescun-Ferrandiz, S., Lorenzo Álvarez, C., et al. (2025). DIPSER: A dataset for in-person student engagement recognition in the wild. arXiv,

arXiv:2502.20209.

[12] Li, M., Zhuang, X., Bai, L., et al. (2024). Multimodal graph learning based on 3D Haar semi-tight framelet for student engagement prediction. Information Fusion, 105, 102224.

[13] Kayande, D., & Kukreja, S. (2025). Design of an integrated multi-modal machine learning framework for real-time student engagement evaluation and learning outcome optimizations. MethodsX, 15, 103588.

[14] Sharif, M., & Uckelmann, D. (2024). Multi-modal learning assessment in personalized education using a deep reinforcement learning-based approach. IEEE Access, 12, 54049-54065.

[15] Ruan, S., & Lu, K. (2025). Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization. Computers and Education: Artificial Intelligence, 9, 100463.

[16] Tan, L. Y., Hu, S., Yeo, D. J., et al. (2025). Artificial intelligence-enabled adaptive learning platforms: A review. Computers and Education: Artificial Intelligence, 9, 100429.

[17] Kerimbayev, N., Adamova, K., Shadiev, R., et al. (2025). Intelligent educational technologies in individual learning: A systematic literature review. Smart Learning Environments, 12(1), 1.

[18] Wang, C., Mohamed, A. S. A., Yang, X., et al. (2025). Enhancing classroom behavior recognition with lightweight multi-scale feature fusion. Computers, Materials & Continua, 85(1), 855-874.

[19] Wang, J., Sun, Y., & Tian, S. (2025). Deep learning for student behavior detection in smart classroom environments. Information, 16(11), 949.

[20] Valdes-Ramirez, D., Beltran-Sanchez, J. A., Conant-Pablos, S. E., et al. (2026). A deep learning approach to estimating interaction levels in face-to-face lessons. Computers and Education: Artificial Intelligence, 10, 100528.