Journal of Educational Technology and Innovation
https://jeti.thewsu.org/index.php/cieti

Article

Real-Time Classroom Behavior Detection and Visualization System Based on an Improved YOLOv11

Jiajun Li [1], Nannan Wang [1,2], Junhao Zhang [1], Xiaozhou Yao [3] and Wei Wei [1,*]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China
[2] Department of Information Engineering, Shandong Management University, Jinan 250100, China
[3] Institut de la Communication, Université Lumière Lyon 2, 69007 Lyon, France
* Correspondence: weiweitesting@hotmail.com

How To Cite: Li, J., Wang, N., Zhang, J., Yao, X., & Wei, W. (2025). Real-Time Classroom Behavior Detection and Visualization System Based on an Improved YOLOv11. *Journal of Educational Technology and Innovation*, *7*(4), 1–13. https://doi.org/10.61414/mwad4j38

Received: 2 September 2025
Revised: 11 November 2025
Accepted: 24 November 2025
Published: 31 December 2025

Abstract: Automatic analysis of student behavior in classrooms has gained importance with the rise of smart education and vision technologies. However, the limited real-time accuracy of existing methods severely constrains their practical classroom deployment. To address this issue of low accuracy, we propose an improved YOLOv11-based detector that integrates CARAFE upsampling, DySnakeConv, DyHead, and SMFA fusion modules. This new model for real-time classroom behavior detection captures fine-grained student behaviors with low latency. Additionally, we have developed a visualization system that presents data through intuitive dashboards. This system enables teachers to dynamically grasp classroom engagement by tracking student participation and involvement. The enhanced YOLOv11 model achieves an mAP@0.5 of 87.2% on the evaluated datasets, surpassing baseline models. This significance lies in two aspects. First, it provides a practical technical route for deployable live classroom behavior monitoring and engagement feedback systems. Second, by integrating this proposed system, educators could make data-informed and fine-grained teaching decisions, ultimately improving instructional quality and learning outcomes.

Keywords: classroom behavior detection; real-time object detection; student engagement; visualization dashboard; AI in education

1. Introduction

Classroom behavior is an important indicator of students' engagement and attentiveness. It can also serve as an indirect measure of teaching effectiveness. Having timely behavioral data allows teachers to detect issues in their teaching and take focused remedial actions (Fredricks et al., 2004; Smith et al., 2013). Traditional data collection methods rely mainly on manual observation or on reviewing recorded classroom videos. These methods require substantial time and effort and cannot provide immediate feedback (Wang et al., 2023). With the rise of smart education, real-time classroom monitoring has become a key trend in the future development of education. Teachers can comprehensively monitor teaching dynamics by tracking student behaviors and learning progress in real time, combined with visual analytics technology. Such data not only provide an evidence base for designing personalized instructional strategies but also help improve student performance and overall teaching quality (Wu et al., 2022; Zheng et al., 2021). Against this backdrop, developing an automated tool capable of reliably analyzing classroom activities has become an important goal in educational technology. In recent years, there have been rapid advances in artificial intelligence. These advances have driven the widespread adoption of computer vision techniques. These techniques are now being used in educational settings. As a key component of computer vision, object detection enables the localization and classification of targets in images or videos. This capability provides the technical foundation for automated recognition and analysis of classroom behavior.

However, despite its improved architecture and speed, YOLOv11 still exhibits several practical limitations that are salient for classroom analytics. First, Real-classroom benchmarks highlight occlusions, pose variation, and inconsistent scales, which make subtle behaviors (e.g., hand-raises, phone use, writing) easy to miss. Recent classroom-specific work built on YOLOv11 explicitly reports remaining difficulties in densely populated frames, inter-person occlusions, and small targets, and therefore adds a high-resolution detection head and large-kernel attention to recover fine details (Hou & Huang, 2025). Second, repeated down-sampling and low-resolution high-level maps weaken tiny-target cues, so YOLOv11 variants in high-resolution remote sensing add shallow heads and stronger cross-scale fusion to curb feature loss and improve detection under cluttered backgrounds (Wu et al., 2025). Third, the speed–accuracy trade-off remains hardware-sensitive in crowded, occluded settings: UAV studies observe that YOLOv11-s/m often miss densely distributed small targets in complex scenes; modules like C2PSA can raise parameter counts, and lightweight separable stacks may become memory-bound on GPUs, so alternative heads and stabilized box losses are adopted to recover throughput and localization (Zhong et al., 2025). These gaps—small, subtle gestures; frequent occlusions; and real-time constraints—mirror classroom conditions, motivating our tailored improvements to enhance fine-grained behavior detection while maintaining real-time performance.

2. Related Work

2.1. Evolution of Object-Detection Models

From the perspective of model evolution, object detection methods can be broadly divided into one-stage and two-stage approaches. Faster R-CNN is a two-stage detector that first generates candidate regions then classifies and regresses the features within these regions. The R-CNN family achieves strong accuracy. However, proposal generation and multi-stage processing introduce substantial computational overhead. As a result, its applicability to classroom behavior recognition where real-time performance is essential and limited. In contrast, one-stage detectors discard explicit proposal generation. The design achieves much higher inference speed by performing parallel bounding-box regression and classification directly on feature maps.

Since its introduction in 2015, the YOLO series has undergone continuous refinement that preserves very high speed while steadily improving accuracy. This combination makes the YOLO family a natural choice for large-scale deployment in smart-classroom systems on standard GPUs. For example, YOLOv11 was released by Ultralytics in 2024 as an optimized successor to YOLOv8 and it maintains detection accuracy while significantly improving computational efficiency and model compactness. It replaces C2f with the more expressive C3k2 in the feature-fusion network and introduces C2PSA in the backbone which further enhances feature representation capacity (Khanam & Hussain 2024).

2.2. Behavior Recognition in Classroom Settings

Research on classroom behavior recognition has developed along three main technical lines. These tasks are posing estimation video action recognition and object detection (Chen et al. 2022; Yin et al. 2022; Li & Peng 2025). Early studies frequently used skeletal pose information as the primary cue. On the ActRec-Classroom dataset (Yang, 2023), for example, Faster R-CNN is first applied to detect students. OpenPose then extracts body joints and facial key points. A CNN classifier finally predicts behavior labels based on these pose features. This pipeline couple's detection with pose analysis and it performs well for canonical actions such as hand raising or leaning. However, its robustness decreases in crowded classrooms where severe occlusion is common. Thus, pose based methods face challenges in realistic high density teaching environments.

Video based action recognition focuses on modeling temporal dynamics and Li et al. use an improved SlowFast model to analyze student behaviors (Li et al., 2022). These 3D convolutional networks process consecutive frames and capture complex motion patterns over time. Such 3D convolutional networks can effectively process consecutive frames and capture complex temporal patterns but they require substantial annotated data and incur heavy computation. This requirement hinders their real-time deployment in ordinary classrooms.

Beyond purely vision-based pipelines, sensor-based and multimodal classroom behavior recognition has also attracted considerable attention. Some studies use depth cameras (e.g., Microsoft Kinect) to obtain 2D–3D skeletal data of students, including head pose, gaze direction, and body posture, and then apply machine learning models to estimate attention or engagement levels over time (Zaletelj & Košir, 2017; Wang et al., 2023). These approaches can infer high-level states such as “attentive”, “distracted”, or “off-task” by fusing facial and postural cues, and have reported promising accuracy in small-scale university classrooms (Sümer et al., 2021). Other work leverages wearable or ambient motion sensors to detect typical classroom actions such as standing up, turning around, or leaning on the desk. For example, multisource sensing systems based on inertial measurement units and environmental sensors have achieved very high recognition rates for predefined classroom behaviors under controlled conditions (Yin et al., 2022).

However, sensor-based solutions typically require additional hardware at the student or classroom level, increasing deployment costs and potentially causing usability or acceptance issues. In contrast, object-detection-based methods build on commodity RGB cameras that are already widely installed in smart classrooms. By directly detecting posture- and behavior-related visual patterns from video streams, these methods are easier to scale across multiple rooms and campuses. At the same time, object detection can be integrated with other modalities at the system level, forming a flexible, multimodal analytic framework. Against this background, our work adopts an improved YOLOv11-based detector as the core vision engine, and focuses on enhancing fine-grained classroom behavior recognition under occlusions and dense student layouts while retaining real-time performance.

2.3. Student Engagement Quantification Methods

Quantifying classroom engagement is more critical than merely classifying behaviors. Compared with subjective, low-frequency measures, computer vision leverages head pose, gaze, and facial expressions to generate time series of attention/engagement at both individual and class levels. For example, prior work in real classrooms models engagement by combining face detection, head pose, and facial features, and aggregates results into class-level indicators (Sümer et al., 2021); other real-time systems use commodity cameras to infer each student's attentive/inattentive state and visualize class-wide attention dynamics over time (Renawi et al., 2022); earlier studies approximate on-task vs. off-task states via gaze targets (teacher/board, notebook, elsewhere) (Zaletelj & Košir, 2017). Common challenges include robust feature selection, multimodal fusion, and generalization to authentic classroom conditions.

3. Experimental Design

3.1. Datasets

In this study, we employ two publicly available classroom-behavior datasets—SCB-Dataset3 (Yang & Wang, 2023) and POCO (Li et al., 2024)— to train and evaluate our model. We selected SCB-Dataset3 as our primary benchmark for classroom-behavior detection. It contains 5686 images and 45,578 bounding-box annotations across six behavior categories: hand-raising, reading, writing, phone use, head-lowering, and leaning over the desk. Table 1 summarizes the number of classes, images and annotations for both SCB-Dataset3 and POCO.

Table 1. Comparison of Public Classroom Behavior Datasets.

Model		Classes	Images	Annotations
SCB3	SCB3-S	3	5015	671
	SCB3-U	6	25810	19768
POCO		10	1903	137960

Within SCB-Dataset3, we further analyse class-level distributions as shown in Table 2. Notably, SCB3-S emphasizes core classroom behaviors (hand-raising, reading, writing) in a structured teaching setting, while SCB3-U captures more diverse and natural student postures—such as phone use, head-lowering and leaning over the desk—in a university-level scenario.

Table 2. Class Distribution in SCB-Dataset3.

Model	Classes	
Classes	SCB3	
	SCB3-S	SCB3-U
Hand raising	11207	6
Reading	10841	7826
Writing	3762	2984
Using phone	/	6976
Bowing the head	/	947
Leaning over the table	/	1029

In addition, we exploit the POCO dataset to broaden our analysis of student behavior. It provides ten fine-grained classroom states, encompassing head posture, phone/book usage, body posture and writing poses. The dataset consists of 1903 images and 137960 bounding-box annotations. Table 3 details each behavior code, its class label and description.

Table 3. POCO Dataset: Behavior Codes, Classes, and Descriptions.

Codes	Classes	Descriptions
front_face	front face	Raise head and focus on the teacher or blackboard. Indicates that the student is interested and interacting with the teacher.
bowed_head	bowed head	Head and eyes down. Indicates that the student is negative about the class.
side_face	side face	Head turns significantly to one side. Indicates that the student is switching attention.
phone	phone	Phone on desk without hands on it. Indicates a neutral state; nothing significant is happening.
phone_hands	phone with hands	Phone with hands on it. Indicates that the student may be focusing on the phone.
book	book	Book on desk without hands on it. Indicates a neutral state; nothing significant is happening.
book_hands	book with hands	Hands on the book. Indicates that the student may be focusing on the book.
head_arms	head in arms	Head in arms or on desk. Indicates that the student is negative about the class.
upright	upright body	Upper body without bending. Indicates that the student may be focusing on the teacher or the course.
body_desk	body on the desk	Upper body is lying on the desk. Indicates that the student is negative about the class.

3.2. Evaluation Metrics

(1) Precision: Precision quantifies the number of true positive predictions divided by the total number of positive predictions made by the model. It effectively measures the model's accuracy in identifying only relevant instances.

$$Precision = \frac{True\,Positives(TP)}{True\,Positives(TP) + False\,Positives(FP)} \tag{1}$$

(2) Recall (True Positive Rate): Recall measures the proportion of actual positives that are correctly identified by the model. It is crucial for determining the model's capability to detect all relevant instances without missing any.

$$Recall = \frac{True\ Positives(TP)}{True\ Positives(TP) + False\ Negatives(FN)} \tag{2}$$

(3) F1 Score: The F1 Score is the harmonic mean of precision and recall. It is a single metric that balances both the precision and the recall, providing a comprehensive measure of a model's accuracy, especially when the class distribution is uneven.

$$F_{1SCORE} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

(4) Average Precision (AP): Average Precision summarizes the precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. It provides a single-figure measure of quality across recall levels, which is particularly useful for evaluating models where predictions must be ranked.

$$\mathrm{AP} = \int_0^1 Precision(Recall)dRecall \tag{4}$$

(5) Mean Average Precision (mAP): Mean Average Precision extends AP to multi-class problems. It first computes the AP for each class and then averages these values. Thus, mAP reflects the overall detection quality across all categories.

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i \tag{5}$$

In Equations (1)–(5), TP denotes true positives correctly identified as positive cases. FP denotes false positives, where negative cases are incorrectly labeled as positive. FN denotes false negatives, where positive cases are missed by the model. N is the total number of classes in the dataset. APk is the Average Precision for the k-th class. For instance, AP50 uses an IoU threshold of 0.50, whereas AP75 uses 0.75. These definitions enable consistent comparison of models across classes and IoU thresholds.

3.3 Experimental

All experiments were conducted on Ubuntu 22.04 using the PyTorch deep learning framework. The software/hardware environment was: Python 3.9, Intel Core i9-13900H CPU, NVIDIA GeForce RTX 3090 GPU, and 24 GB system memory. Unless otherwise stated, the input resolution was fixed at 640 × 640. Mosaic augmentation was applied throughout training except for the last 10 epochs; batch size = 8, epochs = 600.

As illustrated in Figure 1, we augment YOLOv11 with three targeted modules to tackle three classroom-specific bottlenecks—loss of small-object details, poor depiction of slender limbs, and confusion under heavy occlusion. First, CARAFE replaces conventional interpolation with content-aware upsampling (Wang et al., 2019), providing a larger effective receptive field and instance-adaptive kernel generation to better reconstruct fine details (e.g., writing, head-down) with negligible overhead. Second, DySnakeConv (Figure 2) introduces dynamic sampling along slender or curved contours in the backbone under a topological connectivity constraint, markedly improving recall and boundary sensitivity for elongated shapes such as raised arms (Qi et al., 2023). Third, DyHead injects spatial–channel–task–multi-scale attentions into the detection head, jointly strengthening classification and regression, thereby separating adjacent targets more reliably and improving localization consistency in crowded/occluded scenes. SMFA (in Figure 3) performs self-modulated fusion after feature concatenation, preserving both global context and local textures to reduce false positives in complex backgrounds (Zheng et al., 2024). Together they target detail reconstruction, shape modeling, and discriminative representation, forming a complementary bundle that enhances features for small/elongated structures, boosts robustness of discrimination and localization in crowded views, and preserves lightweight, real-time efficiency.

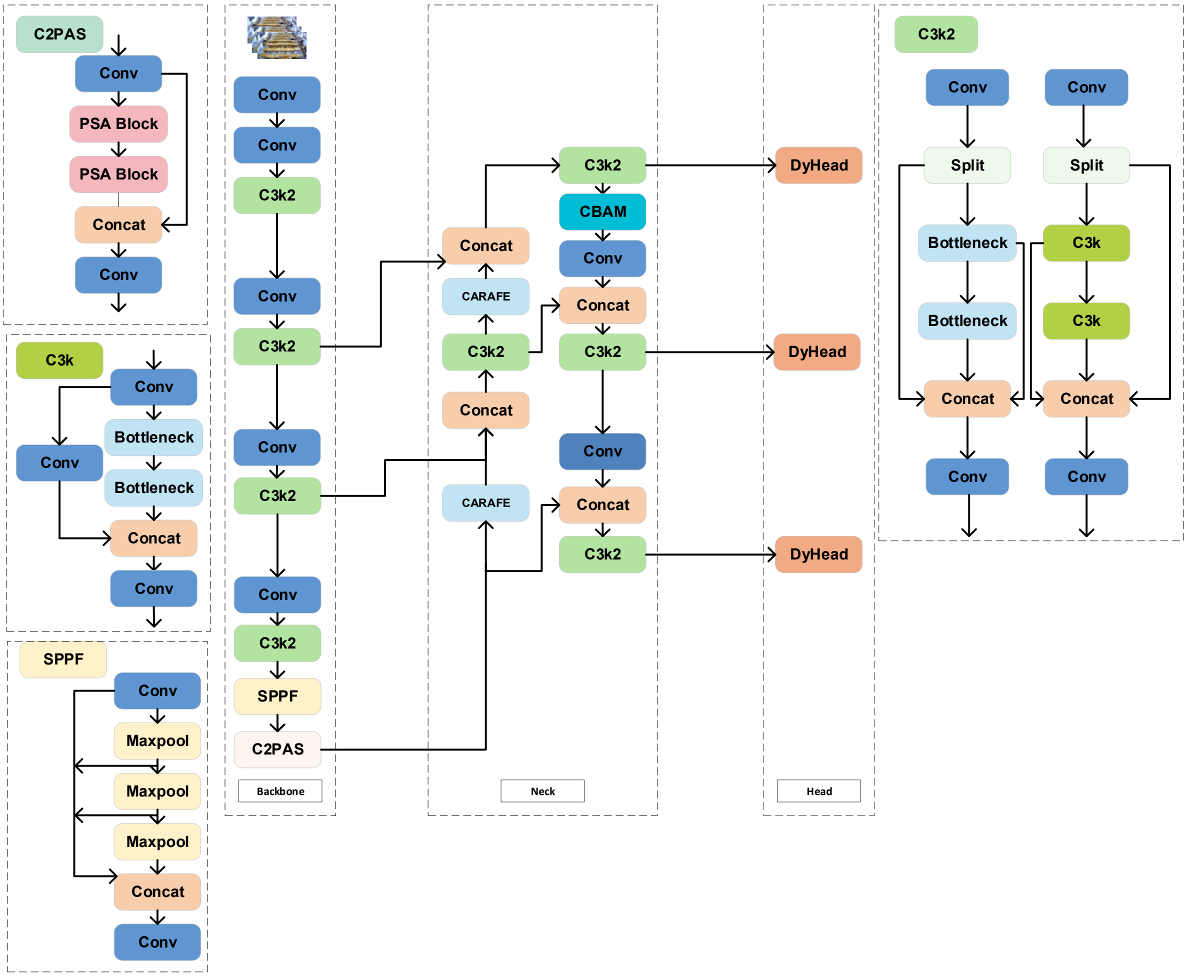


Figure 1. Improved YOLOv11 Network Architecture Diagram.

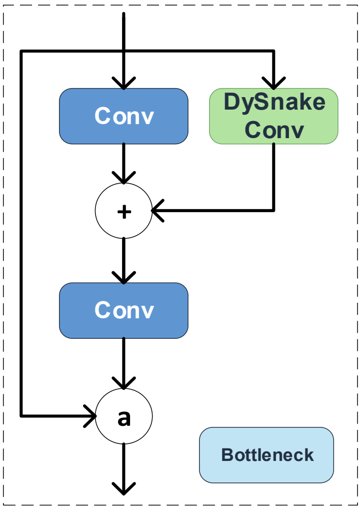


Figure 2. DySnakeConv Module.

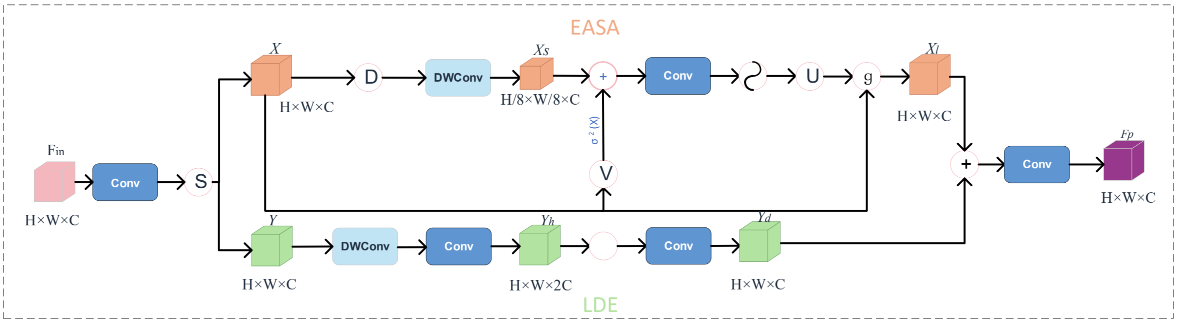


Figure 3. SMFA Module.

4. Experimental Results and Analysis

4.1. System Design

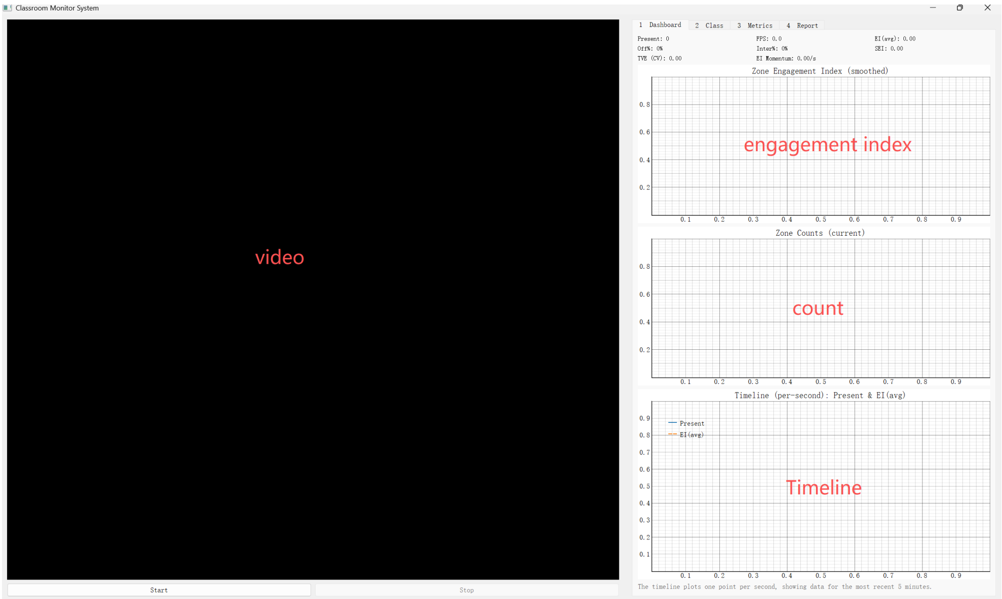
The system employs a PyTorch-based YOLOv11 detector to perform real-time recognition of student postures and behaviors in classroom video frames. For data processing, OpenCV and NumPy are used for image pre-processing and vectorized computation, while exponential moving averages and per-second aggregation enable real-time estimation of multi-region metrics, including Engagement Index, temporal variability estimate, and momentum. The visualization interface is built with PyQt5, using pyqtgraph for interactive bar and line charts and embedding Matplotlib pie charts for composition analysis.

Finally, the system supports CSV export via the Python standard library, local PDF generation with ReportLab, and optional professional report.

Figure 4 shows the initial interface of the Classroom Monitor System. The system consists of a video panel on the left and a data-visualization panel on the right. The home page presents three metrics: Engagement Index, Counts, and Timeline. The Engagement Index divides the classroom view into a 3 × 3 grid (front–middle–back × left–center–right), and computes an engagement score for each region at every frame. The behavior labels are grouped as follows: On-Task = upright, look_book, book, bow_head, turn_head, bend; Interact = raise_head; Off-Task = phone, play_phone, sleep. EI aggregates the proportions of On-Task, Interact, and Off-Task behaviors and normalizes them by the number of people, defined as follows:

$$\mathrm{EI}_z(t)=\frac{\alpha\cdot\mathrm{On}_z(t)+\beta\cdot\mathrm{Inter}_z(t)-\gamma\cdot\mathrm{Off}_z(t)}{\max\left(\mathrm{Count}_z(t),1\right)}\in[0,1] \tag{6}$$

Let $\mathrm{On}_z(t)$, $\mathrm{Inter}_z(t)$, $\mathrm{Off}_z(t)$ denote, for that region *r* at *t*, the numbers of students in the On-Task, Interact, and Off-Task categories, respectively. Let $\alpha,\beta,\gamma \geqslant 0$ ≥ 0 be the corresponding weights, with $\beta>\alpha$ to treat Interact as stronger engagement, and γ as the penalty for Off-Task.

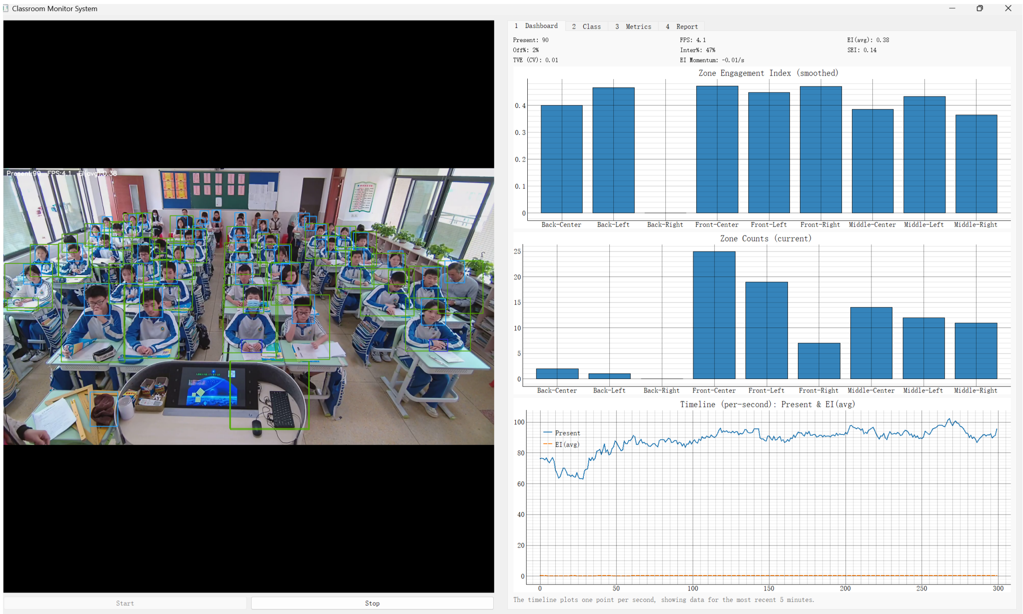


Present: 0
Off%: 0%
TVE (CV): 0.00
FPS: 0.0
Inter%: 0%
EI Momentum: 0.00/s
EI(avg): 0.00
SEI: 0.00

The timeline plots one point per second, showing data for the most recent 5 minutes.

Figure 4. Initial interface of the Classroom Monitor System.

Figure 5 Initial interface of the Classroom Monitor System. After the video feed is connected, the left panel shows the live video and the right panel shows the analytics dashboard. The top row summarizes key performance indicators (Present, FPS, EI (avg)). The two bar charts in the middle visualize zone-level EI and zone-level headcounts, respectively. The bottom Timeline is a line plot with one point per second, tracking recent trends in class size and mean engagement.



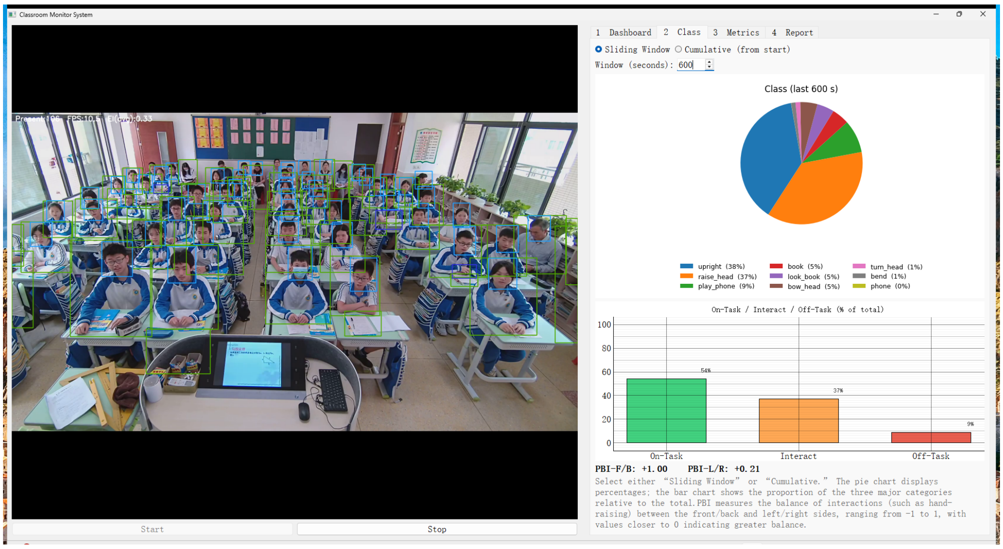
Present: 90
Off%: 2%
TVE (CV): 0.01
FPS: 4.1
Inter%: 47%
EI Momentum: -0.01/s
EI(avg): 0.38
SEI: 0.14

The timeline plots one point per second, showing data for the most recent 5 minutes.

Figure 5. Dashboard module.

Figures 6 and 7 Class view (sliding-window mode). A user-defined time window (e.g., 60–120 s) is applied. The pie chart reports the percentage shares of fine-grained behaviors. The three-bar chart aggregates them into On-Task/Interact/Off-Task, plotted as percentages summing to 100%, which quickly indicates the class status within the window.

The system not only presents data through interactive visual charts, but also supports report export (Figure 8). The report includes: the analysis mode (sliding-window duration or cumulative), keywords, core metrics (Present, FPS, EI (avg), SEI, TVE, EI Momentum, percentages of On-Task /Interact/Off-Task, PBI-F/B, PBI-L/R), and per-region EI tables, enabling both overall and spatial analyses of classroom engagement. On the Metrics page, users can inspect the data and export a CSV with one click for further processing; on the PDF page, they can generate a neatly formatted report to help teachers quickly review class dynamics, archive sessions, and share findings.



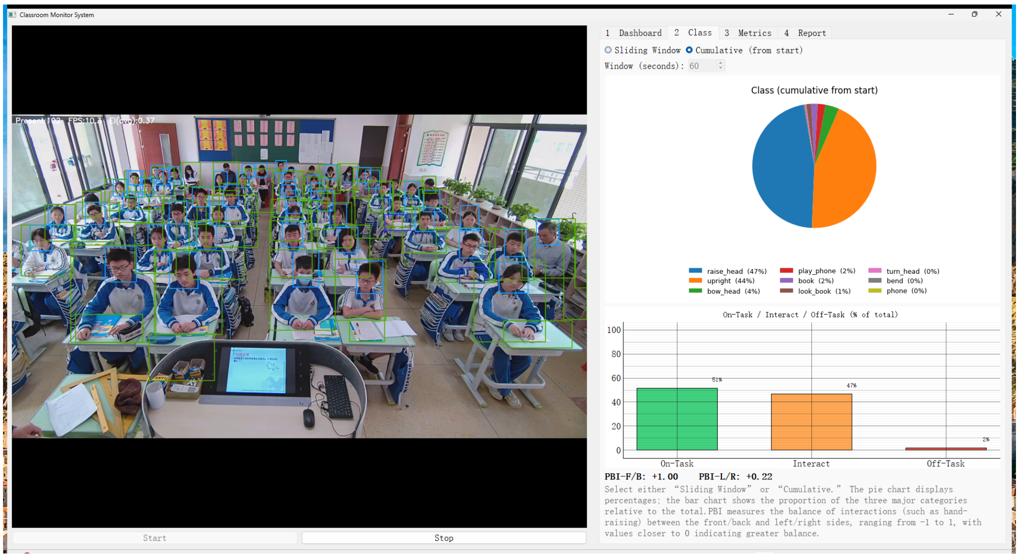
● Sliding Window ○ Cumulative (from start)

Window (seconds): 600

PBI-F/B: +1.00 PBI-L/R: +0.21

Select either “Sliding Window” or “Cumulative.” The pie chart displays percentages; the bar chart shows the proportion of the three major categories relative to the total. PBI measures the balance of interactions (such as hand-raising) between the front/back and left/right sides, ranging from -1 to 1, with values closer to 0 indicating greater balance.

Figure 6. Class module.



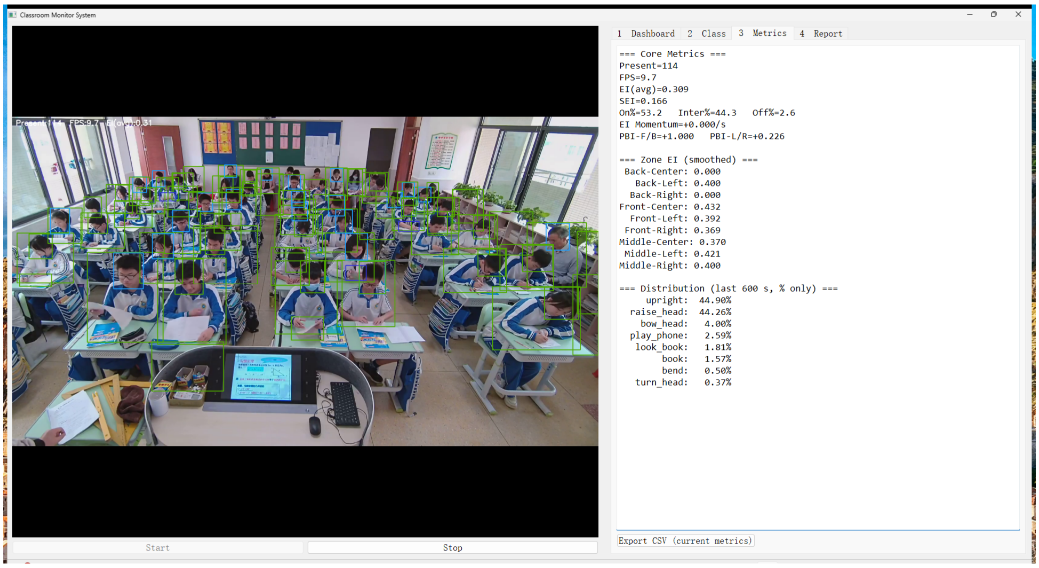
○ Sliding Window ● Cumulative (from start)

Window (seconds): 60

PBI-F/B: +1.00 PBI-L/R: +0.22

Select either “Sliding Window” or “Cumulative.” The pie chart displays percentages; the bar chart shows the proportion of the three major categories relative to the total. PBI measures the balance of interactions (such as hand-raising) between the front/back and left/right sides, ranging from -1 to 1, with values closer to 0 indicating greater balance.

Figure 7. Metrics module.



```
=== Core Metrics ===
Present=114
FPS=9.7
EI(avg)=0.309
SEI=0.166
On%=53.2   Inter%=44.3   Off%=2.6
EI Momentum=+0.000/s
PBI-F/B=+1.000   PBI-L/R=+0.226

=== Zone EI (smoothed) ===
 Back-Center: 0.000
   Back-Left: 0.400
  Back-Right: 0.000
Front-Center: 0.432
  Front-Left: 0.392
 Front-Right: 0.369
Middle-Center: 0.370
 Middle-Left: 0.421
Middle-Right: 0.400

=== Distribution (last 600 s, % only) ===
     upright:  44.90%
  raise_head:  44.26%
    bow_head:   4.00%
  play_phone:   2.59%
   look_book:   1.81%
        book:   1.57%
        bend:   0.50%
   turn_head:   0.37%
```

Figure 8. Report module.

4.2. Effectiveness Analysis of the Proposed Module and Ablation Experiments Analysis

This study utilized CARAFE for replacing all upsamplings, integrating DySnakeConv into the Backbone, substituting DyHead for detection heads, and replacing C3k2 with SMFA after Concat for achieving self-modulation fusion. These modifications led to a substantial enhancement in detection accuracy and real-time performance in classroom scenarios with small targets or occlusions.

To assess the effectiveness of the introduced modules, we perform stepwise ablations on the YOLOv11 baseline the outcomes are presented in Table 4 (P 84.7, R 78.4, mAP@0.5 83.3, mAP@0.5:0.95 57.2). Replacing all upsampling layers with CARAFE yields P85.6, R78.8, mAP@0.5 84.1, and mAP@0.5:0.95 57.9. Adding DySnakeConv to the backbone (CARAFE + DySnakeConv) produces P 85.4, R 80.6, mAP@0.5 84.8, mAP@0.5:0.95 59.1. With DyHead as the detection head (CARAFE + DySnakeConv + DyHead), the results reach P 86.8, R 80.4, mAP@0.5 86.9, mAP@0.5:0.95 59.7. Finally, replacing C3k2 with SMFA after concatenation for self-modulated fusion (full stack) yields P 87.6, R 81.8, mAP@0.5 87.2, and mAP@0.5:0.95 61.0. Overall, CARAFE improves small-object detail preservation; DySnakeConv better models elongated and contour-like structures and strengthens high-IoU performance; DyHead enhances discrimination in crowded scenes; and SMFA consolidates accuracy through global–local coordination, leading to clear synergistic gains on small and occluded targets. Quantitatively, CARAFE increases mAP@0.5 by 0.8 percentage points compared with the baseline, DySnakeConv brings an additional +0.7 gain, DyHead further improves +2.1, and SMFA contributes +0.3 to reach the overall +3.9 improvement.

Table 4. Comparison of ablation experiment results.

Model	Upsample	DySnakeConv	DyHead	SMFA	Precision	Recall	mAP@0.5	mAP@0.5:0.95
A					84.7	78.4	83.3	57.2
B	✓				85.6	78.8	84.1	57.9
C	✓	✓			85.4	80.6	84.8	59.1
D	✓	✓	✓		86.8	80.4	86.9	59.7
E	✓	✓	✓	✓	87.6	81.8	87.2	61.0

Note: ✓ indicates that the corresponding component is used.

4.3. Comparative Experiments

Using a standardized data split and a consistent training/inference protocol, we comparatively evaluated YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and our proposed model, as summarized in Table 5. Successive iterations of YOLO demonstrate incremental improvements across all evaluated metrics from YOLOv5 to YOLOv11. In comparison to the most robust baseline, YOLOv11, our model enhances Precision, Recall, and mAP@.5 by +2.9, +3.4, and +3.9 percentage points, respectively. Experimental results demonstrate that jointly integrating CARAFE, DySnakeConv, DyHead, and SMFA not only consistently improves the model's recall and discriminative capability, but also maintains a high level of stability during both training and inference.

Table 5. performance comparison of different models.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)
YOLOv5n	81.0	73.9	78.6
YOLOv8n	82.2	75.5	80.3
YOLOv9n	83.6	76.6	81.2
YOLOv10n	84.2	77.4	82.1
YOLOv11n	84.7	78.4	83.3
Our model	87.6	81.8	87.2

5. **Implications and Ethical Considerations**

5.1. Practical Implications for Teaching and Learning

A real-time behavior detection system continuously analyzes students' engagement data in class. It provides teachers with an immediate evidence base for refining their instructional strategies. Once the system detects that students are becoming distracted or confused, it promptly alerts the teacher. Based on this feedback, teachers can flexibly adjust their teaching approach or incorporate more engaging content. This will help them to guide students back into an effective learning state in a timely manner (Xu et al., 2023). The feasibility of this model has been supported by prior research. For example, Renawi et al. (2022) showed that a simple vision-based attention monitoring tool not only received positive feedback from teachers but also had the potential to enhance instructional effectiveness. Building on such findings, teachers can rely on real-time data provided by the system to implement precise instructional interventions. For example, they can adapt the pace of instruction or initiate timely interactive questioning to maintain a high level of classroom engagement.

5.2. Multimodal Extensions and Feedback Loops

In future system design and optimization, a key direction is to integrate multimodal data and build more sophisticated feedback mechanisms to comprehensively enhance the learning experience. For example, (1) audio analysis can capture speaking intensity and questioning frequency in the classroom. It can provide quantitative evidence for teacher–student interaction (Jia et al., 2025); (2) facial expression and gaze-tracking techniques can reveal students' emotional states and levels of attention (Falcon et al., 2023); and (3) physiological signal monitoring, using wearable devices to collect indicators such as heart-rate variability and electrodermal activity. It can objectively reflect students' cognitive load and emotional arousal (Bustos-López et al., 2022). Integrating these multimodal signals will lay the technical foundation for more accurate assessment of student engagement. With such feedback mechanisms, future classroom systems will not only present behavioral data but may also draw on the ideas of intelligent tutoring systems to provide teachers with deep-learning-based, personalized instructional recommendations (Zhao, 2024). However, these recommendations should enhance teachers' professional autonomy and creativity, rather than replacing their pedagogical judgement (Zhu et al., 2025).

Another important development direction is to leverage longitudinal data for personalized learning analytics. This approach can be combined with the multimodal feedback described above. The system can regularly generate reports on students' engagement and academic performance. These reports support reflective dialogue between teachers and students on how to improve engagement and adjust learning strategies (Yang et al., 2024). With a multimodal and adaptive design, the system can gradually evolve into a comprehensive intelligent classroom assistant. It will not only recognize classroom behavior but also support ongoing improvement of both teaching and learning processes. These advances will help turn smart classrooms from theory into practice and make sure that technology truly supports the fundamental goals of education.

5.3. Ethical and Privacy Considerations for Real-Time Classroom Analytics

Although deploying AI-based monitoring technologies in classrooms holds considerable application potential. Any use of such systems must be very careful. The goal is to make sure the technology is used in a positive way and students' rights are protected. A recent review by Zhu et al. (2025) categorized the ethical risks of AI in education into three dimensions: technology, education, and society. At the technological level, key risks include privacy breaches, model opacity, and potential bias in algorithmic decision-making. In education, there are concerns about over-reliance on AI. It may lead to homogeneous teaching, weaken the role of teachers, and harm teacher–student relationships as well as students' psychological well-being. In addition, a recent systematic review on AI and critical thinking has shown that AI technologies can function either as facilitators or as barriers to the development of students' critical thinking skills, depending on how they are designed and embedded in instruction (Lin et al., 2025). For the societal aspect, AI systems may worsen the digital divide. They also raise difficult questions about who should be held accountable when the system makes an error.

AI-based classroom behavior systems must prioritize transparency, informed consent, and fairness in both their design and deployment. The above analysis shows that excellent technical performance alone is not sufficient to guarantee responsible use. In practical terms, schools should use video-based monitoring only after students and parents have been fully informed and have given their consent. All monitoring activities must strictly comply with the relevant data privacy regulations. All collected data should be securely stored and anonymized to minimize the risk of misuse. The feedback from the system should only play a supportive and advisory role. It is designed to empower teachers, not to replace their professional judgment. In this way, teachers can still keep their main authority and leading role in the classroom. This aligns with the core principle of "responsible AI in education": technology should empower educators instead of undermining their authority and creativity (Song & Gong, 2024).

Looking ahead, building trustworthy intelligent classroom systems is a systemic endeavor in which ethical safeguards must take priority. For the technological issues, regular audits for algorithmic bias are needed to ensure fairness, alongside continuous reinforcement of privacy-preserving techniques such as differential privacy and federated learning. At the level of teaching practice, system design and iteration should incorporate the perspectives of teachers, students, and administrators, ensuring that technology genuinely supports teachers' professional judgment and classroom creativity. At the governance level, transparent and accountable rules and oversight mechanisms should be put in place. These rules need to state clearly how data can be used and who is responsible.

They should also provide channels for informed consent and appeal for teachers, students, and parents. Only by strictly following a responsible AIED framework and systematically addressing these ethical challenges can real-time behavior monitoring technologies return to their original educational mission (Zhu et al., 2025; Miao et al., 2021).

6. Conclusions

This study proposes a real-time student behavior detection and visualization system for smart classrooms, which is of significant value for both educational practice and research. Built on an improved YOLOv11 model, the system provides real-time feedback and interactive functions, effectively addressing several limitations of existing work. Through this system, teachers can monitor classroom dynamics in real time, promptly identify students who are not paying attention. By observing classroom activity, teachers can flexibly adjust their teaching pace and strategies to enhance instructional precision and optimizing classroom management.

Educational administrators and researchers can also draw on the objective behavioral data generated by the system to rigorously evaluate teaching methods and levels of student participation, thereby supporting evidence-based educational decision-making. Experiments on public classroom-behavior datasets show that the proposed model achieves higher mean Average Precision (mAP) and recall than YOLOv5, YOLOv7, YOLOv8, YOLOv9, and YOLOv10.

At the same time, it maintains a higher frame rate, which enables smooth online deployment in real classroom environments. Ablation studies further verify the effectiveness of each individual module and their combined contribution. This research effectively integrates computer vision techniques with authentic educational settings and promotes the transition of smart classroom systems from theoretical validation to practical application. At the same time, we acknowledge several limitations of the current system. The set of behavior categories covered remains relatively limited, and its adaptability to different subjects, grade levels, and complex classroom scenarios still needs further validation. Future work will focus on expanding the behavior classification system and incorporating multimodal data to capture students' learning states in a more comprehensive way, thereby further enhancing the model's representational capacity and robustness in real-world educational contexts.

Author Contributions

Conceptualization, J.L.; methodology, J.L. and N.W.; software, J.L and X.Y.; validation, J.L. and X.Y.; investigation, J.L.; resources, J.Z.; data curation, J.L. and J.Z.; writing—original draft preparation, J.L.; writing—review and editing, J.L. and W.W.; supervision, W.W.; visualization, J.L. project administration, W.W. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

This study used only publicly released open datasets (including POCO, SCB, and other open resources) and public online videos for system development and testing, with no new participant recruitment or on-site data collection. No identity inference is performed, and any exemplar frames are de-identified when necessary. This study used only publicly released open datasets (including POCO, SCB, and other open resources) and public online videos for system development and testing, with no new participant recruitment or on-site data collection.

Informed Consent Statement

Participant consent was waived because this study involved only secondary analysis of publicly available open datasets (e.g., POCO, SCB, and other open resources) and public online videos, with no new participant recruitment or on-site data collection; the work reports only behavior recognition outputs and aggregated statistics and makes no attempt to identify individuals. Written informed consent for publication is not applicable because the manuscript does not include identifiable personal information or identifiable images/frames of any individuals.

Data Availability Statement

The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Bustos-López, M., Cruz-Ramírez, N., Guerra-Hernández, A., Sánchez-Morales, L. N., Cruz-Ramos, N. A., & Alor-Hernández, G. (2022). Wearables for engagement detection in learning environments: A review. *Biosensors*, *12*(7), 509. https://doi.org/10.3390/bios12070509.

Chen, G., Ji, J., & Huang, C. (2022, April 15–17). *Student classroom behavior recognition based on OpenPose and deep learning* [Conference session]. 7th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 576–579), Xi'an, China. https://doi.org/10.1109/ICSP54964.2022.9778501.

Falcon, S., Alonso, J. B., & Leon, J. (2023). Teachers' engaging messages, students' motivation to learn and academic performance: The moderating role of emotional intensity in speech. *Teaching and Teacher Education*, *136*, 104375.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059.

Hou, P., & Huang, S. (2025). BCSM-YOLO: An improved product package recognition algorithm for automated retail stores based on YOLOv11. *IEEE Access*, *13*, 139665–139679.

Jia, L., Sun, H., Jiang, J., & Yang, X. (2025). High-quality classroom dialogue automatic analysis system. *Applied Sciences*, *15*(3), 1613. https://doi.org/10.3390/app15031613.

Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv*. https://arxiv.org/abs/2410.17725.

Li, C., Bao, W., Chen, X., Jing, Y., & Qu, X. (2022, May 20–23). *SlowFast with DropBlock and smooth samples loss for student action recognition* [Conference session]. 14th International Conference on Digital Image Processing (ICDIP) (Vol. 12342, pp. 205–212), Wuhan, China.

Li, C., & Peng, C. (2025, January 17–19). *Student classroom behavior recognition based on improved YOLOv7* [Conference session]. 2nd International Conference on Informatics Education and Computer Technology (pp. 64–69), Kuala Lumpur, Malaysia.

Li, J., Wang, N., & Wei, W. (2024). POCO: Pedagogical objects in college classroom dataset [Data set]. *GitHub*. https://github.com/jwmianzu/POCO-Dataset.

Lin, H., Wei, W., & Lu, H. (2025). Facilitator or barrier? A systematic review on the relationship between artificial intelligence technologies and the development of critical thinking skills. *Journal of Educational Technology and Innovation*, *7*(2), 11–24.

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *AI and education: Guidance for policy-makers*. UNESCO Publishing.

Qi, Y., He, Y., Qi, X., Zhang, Y., & Yang, G. (2023, October 1–6). *Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation* [Conference session]. IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 6070–6079), Paris, France.

Renawi, A., Alnajjar, F., Parambil, M., Trabelsi, Z., Gochoo, M., Khalid, S., & Mubin, O. (2022). A simplified real-time camera-based attention assessment system for classrooms: Pilot study. *Education and Information Technologies*, *27*(4), 4753–4770.

Song, F., & Gong, X. (2024). Substitution or empowerment: The impact and response of artificial intelligence teaching on teachers' teaching rights. *China Distance Education*, *44*(4), 15–27. [In Chinese]

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education*, *12*(4), 618–627. https://doi.org/10.1187/cbe.13-08-0154.

Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2021). Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, *14*(2), 1012–1027.

Wang, H., Gao, C., Fu, H., Ma, C. Z.-H., Wang, Q., He, Z., & Li, M. (2023). Automated student classroom behaviors' perception and identification using motion sensors. *Bioengineering*, *10*(2), 127. https://doi.org/10.3390/bioengineering10020127.

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., & Lin, D. (2019, October 27–28). *CARAFE: Content-aware reassembly of features* [Conference session]. IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3007–3016), Seoul, Republic of Korea.

Wu, Z. Y., Tang, Y., & Liu, H. (2022). Survey of personalized learning recommendation. *Journal of Frontiers of Computer Science and Technology*, *16*(1), 21–40. [In Chinese]

Wu, Z., Zhen, H., Zhang, X., Bai, X., & Li, X. (2025). SEMA-YOLO: Lightweight small object detection in remote sensing image via shallow-layer enhancement and multi-scale adaptation. *Remote Sensing*, *17*(11), 1917. https://doi.org/10.3390/rs17111917.

Xu, X., Dugdale, D. M., Wei, X., & Mi, W. (2023). Leveraging artificial intelligence to predict young learner online learning engagement. *American Journal of Distance Education*, *37*(3), 185–198.

Yang, F. (2023). SCB-dataset: A dataset for detecting student classroom behavior. *arXiv*. https://arxiv.org/abs/2304.02488.

Yang, F., & Wang, T. (2023). SCB-dataset3: A benchmark for detecting student classroom behavior. *arXiv*. https://arxiv.org/abs/2310.02522.

Yang, K. B., Borchers, C., Falhs, A. C., Echeverria, V., Karumbaiah, S., Rummel, N., & Aleven, V. (2024, September 16–20). *Leveraging multimodal classroom data for teacher reflection: Teachers' preferences, practices, and privacy considerations* [Conference session]. 19th European Conference on Technology Enhanced Learning (pp. 498–511), Krems, Austria.

Yin, C. C., Sun, Y., Li, G., Peng, J., Ran, F., Wang, Z., & Zhou, J. (2022). Identifying and monitoring students' classroom learning behavior based on multisource information. *Mobile Information Systems*, *2022*, 9903342.

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing*, *2017*, 80.

Zhao, C. (2024). AI-assisted assessment in higher education: A systematic review. *Journal of Educational Technology and Innovation*, *6*(4), 39–58.

Zheng, M., Sun, L., Dong, J., & Pan, J. (2024, September 29–October 4). *SMFANet: A lightweight self-modulation feature aggregation network for efficient image super-resolution* [Conference session]. 18th European Conference on Computer Vision (ECCV) (pp. 359–375), Milan, Italy.

Zheng, Y. F., Zhao, Y. N., Bai, X., & Fu, Q. (2021). Survey of big data visualization in education. *Journal of Frontiers of Computer Science and Technology*, *15*(3), 403–422. [In Chinese]

Zhong, H., Zhang, Y., Shi, Z., Zhang, Y., & Zhao, L. (2025). PS-YOLO: A lighter and faster network for UAV object detection. *Remote Sensing*, *17*(9), 1641. https://doi.org/10.3390/rs17091641.

Zhu, H., Sun, Y., & Yang, J. (2025). Towards responsible artificial intelligence in education: A systematic review on identifying and mitigating ethical risks. *Humanities and Social Sciences Communications*, *12*(1), 1–14.