EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Student Behavior Detection in Classroom Environments Using Deep Learning Models

Thi-Nguyen Nguyen[1], Dinh-Thai Kim[2,*], Anh-Phuong Pham[2]

[1]Faculty of Electrical and Electronics Engineering, Vietnam-Hungary Industrial University, 100000, Vietnam
[2]International School, Vietnam National University, Hanoi, 100000, Vietnam

Abstract

Observing student behavior in classroom environments provides valuable insights into attention and engagement. However, manual monitoring is labor-intensive and often inconsistent, particularly in large lecture settings. This study proposes a deep learning–based framework for automatic student behavior analysis in real classroom scenarios. A dedicated dataset was constructed, consisting of 3,373 images annotated with five attention-related behaviors—"Focused", "Raising Hand", "Distracted", "Sleeping", and "Using Phone"—totaling 9,659 labeled instances. The dataset captures diverse real-world conditions, including variations in classroom layout, camera viewpoints, and occlusion. We systematically evaluated state-of-the-art object detection models, including YOLOv8–YOLOv12 and the Real-Time Detection Transformer (RT-DETR). Experimental results show that YOLOv8m achieved the highest localization accuracy (mAP@0.5 = 0.920), YOLOv11s/m demonstrated the best overall performance (mAP@0.5:0.95 = 0.726), and RT-DETR-X achieved the highest F1-score (0.886). Notably, larger model size does not necessarily translate into better performance. In addition to accuracy, inference speed was evaluated to assess real-time applicability. Lightweight models such as YOLOv11s achieved a favorable balance between performance and efficiency, enabling real-time processing on resource-constrained hardware. Furthermore, YOLOv11s—with high Precision (0.890) and only 9.4M parameters—was integrated with ByteTrackV2 to perform behavior tracking and temporal analysis in classroom environments. This enables the generation of behavior distribution charts that provide interpretable insights into student engagement over time. These findings demonstrate the potential of automated behavior recognition systems for classroom analytics and data-driven teaching improvement.

Received on 27 November 2025; accepted on 28 March 2026; published on 07April 2026

Keywords: Computer Vision, Student Behavior Detection, YOLO, ByteTrack, RT-DETR

doi:10.4108/airo.11141

1. Introduction

Understanding student behavior in classroom environments is fundamentally essential for developing data-driven educational systems and enhancing pedagogical effectiveness. Student behavior reflects not only their level of concentration but also their initiative and engagement in the learning process—two factors that directly influence the efficacy of knowledge acquisition. However, in most classrooms today, the assessment of these behaviors relies heavily on direct teacher observation. While this method can yield reliable results when applied to small groups where teachers can easily monitor individuals, manual observation inherently becomes inconsistent, subjective, and difficult to sustain over time as class sizes grow or teaching durations extend [1]. Consequently, despite an increasing interest in educational computer vision, there remains a critical lack of systematic and real-world validated frameworks capable of reliably analyzing attention and engagement. This prevailing gap severely limits the practical applicability of existing approaches in large-scale educational settings.

*Corresponding author. Email: thaikd@vnu.edu.vn

To overcome these limitations, the rapid development of computer vision and artificial intelligence (AI) has opened new avenues for analyzing student behavior in classroom environments. Advances in object detection, action recognition, and contextual video analysis have encouraged researchers to develop automated classroom monitoring systems capable of identifying and classifying students' actions directly from real-time video streams [2]. Although these systems hold tremendous potential for supporting teaching evaluation, they continue to face significant practical challenges: (1) the diversity of training data remains limited, hindering model generalization across varying classroom layouts and camera perspectives; (2) severe occlusion in crowded classrooms degrades recognition accuracy; and (3) the high computational cost of processing real-time video remains a major deployment obstacle on standard, resource-constrained educational hardware [3, 4].

To address detection accuracy, deep learning models—particularly object detection frameworks—have become the dominant paradigm for classroom behavior analysis, providing a unified technical foundation for advancements in intelligent education systems [5]. Convolutional neural network (CNN)–based models, such as the YOLO (You Only Look Once) family, have demonstrated fast and accurate detection performance in real-world environments. Recent studies have continually extended traditional YOLO architectures to improve sensitivity to small, partially occluded, or low-light objects. For example, Yang et al. [6] proposed a Bi-level Routing Attention (BRA) block in YOLOv7 to enhance contextual feature representation. Meanwhile, Chen et al. [7] and Sheng et al. [8] refined YOLOv8 by incorporating large-kernel convolutions, enabling the model to expand its receptive field and improve multi-scale recognition capability.

More recently, the field has explored advanced architectures beyond conventional CNNs to handle complex scenes and class imbalances. Abozeid et al. [9] introduced a meta-learning framework combining Vision Transformers with contrastive learning to improve feature discrimination. Similarly, transformer-based detectors like the Real-Time Detection Transformer (RT-DETR) [3] have emerged as a powerful approach for maintaining high accuracy in complex classroom scenes. In parallel, hybrid architectures, such as the CBHA-DETR developed by Li et al. [10], integrate multi-kernel attention and deformable cross-scale fusion to specifically tackle occlusion and multi-scale interactions. Beyond object detection, studies have also utilized OpenPose for skeleton-based recognition [11] and 3D CNNs to exploit spatiotemporal motion sequences [2].

Although these efforts have significantly improved detection accuracy, most existing datasets still focus primarily on easily observable physical actions—such as raising hands, reading books, or turning heads. These actions represent only a superficial aspect of the learning process, while subtle cues related to cognitive engagement are often overlooked. Addressing this limitation is imperative for advancing intelligent systems aimed at learning analytics; without it, systems recognize mere movements but fail to distinguish deeper learning-related states. Moreover, the scarcity of studies conducted in uncontrolled, real classroom settings makes it difficult to evaluate true model generalization, highlighting an urgent need for a specialized dataset reflecting authentic educational environments.

To address these limitations, our study introduces a new dataset specifically designed for lecture-style classrooms, where the instructor teaches from the podium and students are expected to stay focused, face the board, or raise their hands when participating. The dataset focuses on five representative behaviors that reflect foundational levels of learning engagement: "Focused," "Distracted," "Sleeping," "Using Phone," and "Raising Hand." By targeting these foundational states, we establish a reliable and unambiguous baseline for attention analysis. While modern education encompasses nuanced activities like collaborative group discussions, accurately detecting these core cognitive states remains the critical prerequisite for building intelligent classroom systems. In total, the dataset contains 3,373 images and 9,659 annotated instances collected from real classroom recordings under diverse lighting conditions, camera angles, and occlusion levels.

The annotation process was designed to prioritize behaviors associated with cognitive engagement rather than merely observable physical appearances, enabling more meaningful interpretation of detection results. Furthermore, rather than proposing a custom network architecture, this study contributes a rigorous, systematic benchmark of state-of-the-art detection models—including YOLOv8 through YOLOv12 [12–16] and RT-DETR [17]—under a unified experimental setup. This comprehensive evaluation assesses both real-world viability and computational efficiency, providing essential deployment guidance for resource-constrained hardware.

Finally, to demonstrate practical applicability, the best-performing model is integrated with the ByteTrackV2 algorithm and a lightweight counting module. Domain charts were generated to visualize engagement dynamics throughout the lecture, providing a quantitative and data-driven perspective on students' classroom interaction levels. Yielding an end-to-end pipeline, this system serves as a foundational practical tool for teaching evaluation and paves the way for

future domain-specific data association strategies that can handle severe classroom occlusion.

The remainder of this paper is organized as follows. Section 2 presents the proposed framework, including dataset construction, behavior annotation, model configuration, and the detection–tracking pipeline. Section 3 reports the experimental results and quantitative evaluations. Section 4 provides a detailed analysis and discussion of the findings. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Material and Method

This section describes the materials and methods used for detecting student behaviors in classroom settings, focusing on dataset construction and model evaluation.

2.1. Dataset

In this study, we introduce a dataset specifically designed for lecture-style classrooms, in which the instructor teaches from the podium and students are expected to face the board and the teacher, maintain focus, or raise their hands when participating—rather than engaging in off-task activities. This context provides a stable camera angle and a consistent scene layout, facilitating objective and systematic analysis of behaviors that reflect students' learning engagement.

Existing public datasets mainly focus on easily observable physical actions, such as reading books or turning heads. Although these behaviors represent movement, they do not fully capture the cognitive aspects of attention and engagement in learning. To bridge this gap, the proposed dataset pivots from purely physical action recognition to attention-centric behavior modeling. By focusing on states that directly reflect cognitive engagement—such as active focus versus digital distraction—this dataset enhances the interpretability of detection results and strictly aligns with the practical requirements of educational analytics.

Our dataset consists of images collected from real classroom settings and annotated with five behavior classes: "Focused," "Raising Hand," "Distracted," "Sleeping," and "Using Phone." These behaviors are defined as follows: "Focused" refers to students sitting upright and looking toward the instructor or the board; "Raising Hand" describes students lifting one arm toward the teacher; "Distracted" indicates students looking or turning away from the lesson; "Sleeping" is characterized by resting the head on a hand or desk; and "Using Phone" describes students visibly holding and looking at a mobile device.

This dataset leverages a hybrid sourcing strategy: approximately half of the images are sourced from existing open educational datasets, namely SCB-Dataset and SCB-Dataset3 [19, 20], while the remaining images are extracted directly from real classroom video recordings. Before annotation, all frames were strictly anonymized to protect students' privacy, in accordance with emerging best practices in privacy-preserving computer vision and educational data governance [21, 22]. In practical deployment, the proposed framework can be configured t o perform on-device inference without storing raw video data, ensuring full compliance with stringent data privacy regulations in educational environments.

The annotation process was conducted on the Roboflow p latform u sing a u nified la beling convention to ensure consistency across subsets. To resolve any potential annotation discrepancies and ensure high-quality ground truths, a rigorous post-annotation filtering phase was applied to remove highly ambiguous or severely truncated bounding boxes. Following this quality control step, the finalized d a taset contains exactly 3,373 images with 9,659 reliably labeled instances. The data are randomly divided into three subsets: training (75%, 2,535 images), validation (15%, 507 images), and testing (10%, 331 images).

Table 1 presents the detailed distribution of annotation counts per behavior class, providing a quantitative overview of the dataset's structure. Notably, the dataset exhibits a natural long-tailed class imbalance, particularly for transient or less frequent behaviors such as "Raising Hand." However, no explicit reweighting or synthetic data augmentation strategies were applied during the dataset preparation phase. Instead, all benchmarked models were trained under a unified, u n modified ex pe rimental se tt ing. Th is deliberate design choice allows us to evaluate the intrinsic algorithmic robustness of each architecture when handling imbalanced data, thereby providing a highly realistic assessment of their baseline performance under authentic classroom conditions.

Table 1. Number of instances per behavior class across the training, validation, and test subsets

Behavior Class	Train	Validation	Test
Focused	2413	472	337
Raising Hand	367	141	91
Distracted	1894	304	185
Sleeping	915	205	159
Using Phone	1502	401	273
Total Instances	7091	1523	1045
Total Images	2535	507	331

All images were resized to 640×640 pixels to match the input format of real-time object detectors such as YOLO and RT-DETR. Overall, this dataset provides a practical and reproducible benchmark for evaluating vision-based models under authentic

classroom conditions and offers a foundation for future research on automated engagement analysis.

2.2. Student Behavior Recognition

Recent advances in computer vision have made it feasible to analyze classroom activities directly from video data. In this study, we evaluate two representative families of object detection models for recognizing student behaviors: the YOLOv8–YOLOv12 series [12–16] and the Real-Time Detection Transformer (RT-DETR) [17]. Each model is tested under multiple configurations—nano, small, medium, large, and extra-large—corresponding to their publicly released variants.

The YOLO series represents a one-stage detection architecture based on convolutional neural networks (CNNs), utilizing cross-stage partial connections and multi-scale feature aggregation to optimize the balance between speed and accuracy. In contrast, RT-DETR adopts a transformer-based encoder–decoder architecture, where detection is performed through set prediction without the need for non-maximum suppression (NMS). All models are trained and evaluated under a unified experimental setup to ensure a fair performance comparison.

Figure 1 illustrates the overall proposed system framework, which comprises three main stages: detection, tracking, and quantification. In the first stage, the selected detection model identifies each student's location within a frame, assigning a bounding box and confidence score to one of five behaviors: "Focused," "Distracted," "Using Phone," "Sleeping," or "Raising Hand." This step provides frame-level recognition of multiple individuals simultaneously within real classroom environments, even under crowded or partially occluded conditions.

In the second stage, the ByteTrackV2 algorithm [18] is applied to maintain temporal consistency across frames. The algorithm assigns a fixed tracking ID to each detected student through a two-step data association process. First, high-confidence detections (confidence > 0.5) are matched based on motion estimation from a Kalman filter and Intersection over Union (IoU) overlap. Then, lower-confidence detections—often caused by occlusion or partial visibility—are re-associated using appearance features. This approach reduces identity switches and ensures stable, continuous tracking for each student throughout the video. In this study, the tracking component is not only used for identity preservation but also serves as a crucial temporal bridge. Specifically, by preserving tracklets even when a student is temporarily obscured by a peer or moving out of the optimal camera angle, the tracking module effectively mitigates false negatives (missed detections) that typically plague frame-independent detection models.

Finally, in the third stage, frame-level detections are aggregated over time to estimate the proportion of students performing each behavior. The counting module computes the percentage of each behavior relative to the total number of valid students per frame. For example, if 10 valid students are detected, with 6 "Focused", 2 "Distracted," and 2 "Raising Hand", the corresponding distribution is 60%, 20%, and 20%. These quantitative sequences are visualized as behavior domain charts, reflecting engagement dynamics over time and providing a data-driven perspective on classroom interaction levels.

Unlike conventional approaches that treat detection and tracking as independent components, the proposed framework explicitly integrates these stages with temporal behavior quantification to produce continuous engagement indicators. By converting discrete bounding boxes into continuous temporal proportions, this module abstracts raw visual data into high-level pedagogical metrics, allowing educators to observe macro-trends in classroom focus rather than isolated instances.

Algorithm 1 summarizes this computation process. The algorithm iterates through all frames, filters out low-confidence detections (ensuring that ambiguous actions or background artifacts do not skew the final distribution), accumulates the counts for each behavior class, normalizes the ratios per frame, and outputs a time series used to plot engagement curves representing learning participation trends.

Algorithm 1 Student Behavior Counting Algorithm

Input: Frames $F = \{f_1, f_2, \ldots, f_T\}$; detections per frame $D_t = \{(s_i, b_i, c_i)\}_{i=1}^{N_t}$, where s_i is the track ID, $b_i \in B = \{$five behaviors: Focused, Distracted, Using Phone, Sleeping, Raising Hand$\}$, and c_i is the confidence score.
Output: Behavior distributions $P = \{p_t\}_{t=1}^{T}$, where $p_t(b)$ is the percentage of behavior b at time t.
Initialize $P \leftarrow \emptyset$.
for $t = 1$ to T **do**
  Filter D_t to retain only detections with $c_i \geq 0.5$.
  $N_t \leftarrow$ total number of valid detections in frame f_t.
  Initialize $C(b) \leftarrow 0$ for all $b \in B$.
  for each valid detection (s_i, b_i, c_i) in D_t **do**
    $C(b_i) \leftarrow C(b_i) + 1$
  end for
  if $N_t > 0$ **then**
    Compute $p_t(b) \leftarrow \frac{C(b)}{N_t} \times 100$ for all $b \in B$.
  else
    $p_t(b) \leftarrow 0$ for all $b \in B$.
  end if
  Append p_t to P.
end for
Return: P (used for visualization in the behavior domain chart over time).

Figure 1. Overview of the student behavior recognition system, illustrating the pipeline from input video to behavior detection, tracking with ByteTrack, counting, and visualization in a behavior domain chart

2.3. Evaluation Metrics

Model performance was quantitatively evaluated using standard object detection metrics to assess both localization accuracy and classification reliability across the five behavior categories: "Focused", "Raising Hand", "Distracted", "Sleeping", and "Using Phone".

We denote the Intersection over Union (IoU) threshold by τ. A predicted box is counted as a true positive (TP) if its class matches a ground-truth instance and IoU $\geq \tau$. A predicted box that does not match any ground-truth instance at τ is a false positive (FP). Any ground-truth instance that is not matched by a prediction at τ is a false negative (FN). True negatives (TN) are not used in object detection scoring because the background is not explicitly enumerated.

The IoU metric measures the spatial overlap between predicted and ground-truth bounding boxes:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}. \tag{1}$$

Detection accuracy was assessed using Precision and Recall:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

The F1-score provides a single measure that balances Precision and Recall:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Overall detection performance was summarized by mean Average Precision (mAP), which is widely adopted as the primary evaluation metric for object detection tasks and is commonly used in benchmark datasets such as PASCAL VOC and MS COCO [23]. Average Precision (AP) is computed from the area under the precision–recall curve for a given class at a specified IoU threshold τ. We report mAP@0.5 at $\tau = 0.5$ and mAP@0.5:0.95, which averages AP over $\tau \in \{0.50, 0.55, \ldots, 0.95\}$ in steps of 0.05:

$$\text{mAP@0.5:0.95} = \frac{1}{10} \sum_{\tau \in \{0.50, 0.55, \ldots, 0.95\}} \text{mAP@}\tau. \tag{4}$$

These metrics together provide a comprehensive basis for comparing models in terms of localization precision and behavior recognition under the same evaluation protocol.

3. Results

3.1. Experimental Setup

To ensure fair comparison and reproducibility, all experiments were conducted under a unified training and evaluation setup. Specifically, we benchmarked six state-of-the-art object detection models for classroom behavior recognition, including YOLOv8 through YOLOv12 (each tested with nano, small, medium, large, and extra-large variants) and the Real-Time Detection Transformer (RT-DETR) in both large and extra-large versions. The YOLO family models were trained using the Ultralytics framework, while RT-DETR was implemented based on the official PyTorch source code to ensure compatibility and fidelity of results.

Training was performed in the Google Colab Pro environment using NVIDIA Tesla T4 or P100 GPUs, each equipped with 16 GB of VRAM. To maintain consistency across models, all hyperparameters were kept identical throughout the experiments. The number of training epochs was set to 100, with a batch size of 16 and an input resolution of 640×640 pixels. The Stochastic Gradient Descent (SGD) optimizer was used with an initial learning rate of 0.01. All input images were resized and normalized to 640×640 pixels to match the architectural requirements of the detectors and to ensure consistency in the training process.

3.2. Student Behavior Detection Results

This section presents the performance evaluation results of the behavior detection models, including YOLOv8–YOLOv12 and the Real-Time Detection Transformer (RT-DETR). All models were trained and tested under the same experimental setup described in Section 3.1, using the same dataset and hyperparameter configuration. The models were evaluated on five representative classroom behaviors: "Focused," "Raising Hand," "Distracted," "Sleeping," and "Using Phone." The aggregated results are summarized in Table 2, while visualizations are illustrated in Figures 2–3. In addition to accuracy metrics, inference speed (measured in frames per second, FPS) is reported to evaluate the real-time applicability of each model. The results show that lightweight models such as YOLOv11s achieve a favorable balance between detection performance and computational efficiency, making them highly suitable for real-time classroom deployment on resource-constrained devices. Detailed FPS comparisons are integrated alongside model performance to provide a comprehensive evaluation of both localization accuracy and runtime efficiency.

Overall Performance. Across all experiments, YOLOv11 and RT-DETR achieved the highest overall accuracy while maintaining real-time inference speed. The YOLO family demonstrated stable performance across different model sizes, whereas RT-DETR provided a better balance between Precision and Recall, particularly in crowded or partially occluded classroom scenarios, where the transformer architecture benefits from its ability to model spatial relationships effectively.

Results by Model Family. Within the YOLOv8 series, the medium variant achieved outstanding accuracy with mAP@0.5 = 0.920 and mAP@0.5:0.95 = 0.709, while the small model yielded the highest Precision (0.889) and F1-score (0.872). These results suggest that medium-scale architectures offer the optimal trade-off between representational capacity and computational cost.

For YOLOv9, the medium and compact variants performed best, with mAP@0.5 $\approx$ 0.894 and F1-score $\approx$ 0.854. In contrast, the tiny version (YOLOv9t) performed significantly worse due to limited feature extraction depth, despite having the fewest parameters (2.0 M).

The YOLOv10 series showed strong stability across behaviors, with the small variant achieving Precision = 0.869, Recall = 0.867, mAP@0.5 = 0.902, and F1-score = 0.868, representing the best balance within this family.

YOLOv11 further improved overall performance, reaching the highest Recall (0.883) and F1-score (0.879) with the large variant. Notably, YOLOv11s achieved the highest Precision (0.890), and both the small and medium versions shared mAP@0.5:0.95 = 0.726, indicating consistent localization performance even under strict IoU thresholds.

The YOLOv12 models exhibited a slight decline in accuracy compared to YOLOv11. Although the medium variant achieved Precision = 0.876, its Recall (0.794) was lower, reflecting reduced robustness under complex conditions. Nevertheless, the results remained competitive, suggesting that recent architectural modifications did not significantly compromise overall detection accuracy.

Transformer-based Model Performance.. With its transformer architecture, RT-DETR-L and RT-DETR-X achieved the highest F1-scores (0.885–0.886), along with Precision $\approx$ 0.894 and Recall $\approx$ 0.877. Although their mAP@0.5 (0.860–0.866) was slightly lower than that of YOLOv11, RT-DETR demonstrated exceptional consistency across all behavior classes—particularly for visually similar behaviors such as "Distracted" and "Using Phone," which often cause confusion in conventional CNN-based detectors.

Performance Trends by Model Size. As illustrated in Figure 2, scaling up the model size does not consistently guarantee superior accuracy. Specifically, Figure 2(a) demonstrates that mid-sized architectures, such as YOLOv8m and YOLOv11s, achieve peak mAP@0.5 scores, frequently outperforming their heavier counterparts. This trend is corroborated by Figure 2(b) under the stricter mAP@0.5:0.95 metric, where the YOLOv11 family maintains a clear dominance. Furthermore, Figure 2(c) explicitly delineates the computational trade-offs: while lightweight models (under 15M parameters) comfortably exceed 70 FPS—making them ideal for edge deployment—models exceeding 40M parameters experience a sharp decline in inference speed without a proportional enhancement in detection reliability.

Impact of Class Imbalance on Minority Classes. Table 3 details the per-class performance of YOLOv11s, offering critical insights into how the dataset's class imbalance impacts detection capabilities. Notably, despite the lack of explicit data reweighting strategies during training, the model maintained robust performance for minority classes. For instance, "Raising Hand" achieved a high mAP@0.5 of 0.931 and the highest mAP@0.5:0.95 of 0.770, heavily benefiting from its distinct spatial characteristics that prevent misclassification. Conversely, the "Distracted" class exhibited the lowest Precision (0.820) and Recall (0.790). This reduction reveals that class imbalance, when coupled with subtle inter-class visual similarities (e.g., distinguishing a distracted student from one reading or casually using a phone under the desk), remains a primary challenge for object detection frameworks.

Detection Robustness in Diverse Real-World Scenarios. Figure 3 illustrates the model's robust detection capabilities across varying population densities, camera perspectives, and lighting environments. The top row demonstrates performance in sparsely populated classroom settings, successfully identifying behaviors from (a) an oblique camera angle, (b) a close-up viewpoint that captures intricate behavioral details, and (c) a straight, long-distance perspective where the model maintains accuracy despite the extended range. Conversely, the bottom row highlights the system's resilience in heavily crowded classrooms under challenging visual constraints. Specifically, the detector effectively localizes multi-person behaviors under (d) a high-angle viewpoint illuminated by bright natural outdoor lighting, (e) a relatively lower camera angle within a darker environment reliant on artificial electric lighting, and (f) an extreme high-mounted camera placement in a dimly lit classroom. Across these complex scenarios, the model consistently overcomes severe occlusion, scale variations, and illumination fluctuations to maintain precise behavioral classification.

In summary, both YOLOv11 and RT-DETR demonstrated strong performance in recognizing classroom behaviors. YOLOv11 proved more effective for precise localization, while RT-DETR maintained reliable real-time tracking through its balanced Precision and Recall. Together, these results suggest that combining transformer-based global reasoning with refined spatial attention mechanisms could further enhance detection accuracy in complex, multi-student environments. These rigorous evaluations provide a reproducible baseline for future studies, offering consistent guidelines for deploying detection models across diverse and attention-focused educational datasets.

3.3. Domain Chart Analysis

Behavior domain chart analysis enables the visualization of students' learning engagement dynamics over time by quantifying the frequency and distribution of detected behaviors across different phases of a lecture. In this study, we employ the YOLOv11s model—a variant achieving mAP@0.5:0.95 = 0.726 with only 9.4 million parameters—which demonstrates an optimal balance between accuracy and computational efficiency, to analyze video sequences from three representative classroom sessions, each lasting approximately 10–12 minutes.

These three sessions were selected to represent distinct learning engagement states. The first condition corresponds to a **Highly Focused Class**, in which most students maintain continuous attention throughout the lecture with very few off-task actions. The second condition represents a **Highly Active Participation Class**, characterized by frequent student interactions, such as raising hands and direct verbal exchanges with the instructor. Finally, the third condition corresponds to a **Low Engagement Class**, where inattentive behaviors such as sleeping, using mobile phones, or side conversations are clearly observed.

Each behavior domain chart depicts the temporal variation of five behavioral categories—Focused, Raising Hand, Distracted, Sleeping, and Using Phone—providing a concise yet comprehensive visualization of students' attention and engagement patterns throughout the lecture (Figure 4).

In the **Highly Focused Classroom** session (10 minutes), the "Focused" behavior dominates with 78.7%, followed by "Raising Hand" at 10.1% and "Distracted" at 11.1%. Behaviors such as "Sleeping" and "Using Phone" are almost absent, accounting for less than 0.2%. This pattern reflects a traditional lecture environment in which students maintain high attention but exhibit limited active participation. From a temporal perspective, Figure 4(a) reveals a highly stable engagement baseline. The purple "Focused" area remains consistently broad across the 600-second timeline, while

Table 2. Performance comparison of YOLOv8–v12 and RT-DETR models for student behavior recognition. Best results are shown in bold

Model	Input	Precision	Recall	mAP@0.5	mAP@0.5:0.95	F1-score	Params (M)	FPS
YOLOv8n[12]	640	0.857	0.856	0.896	0.703	0.856	3.2	146.32
YOLOv8s[12]	640	0.889	0.856	0.915	0.700	0.872	11.2	139.02
YOLOv8m[12]	640	0.869	0.860	**0.920**	0.709	0.864	25.9	88.79
YOLOv8l[12]	640	0.861	0.845	0.893	0.698	0.853	43.7	63.96
YOLOv8x[12]	640	0.877	0.842	0.891	0.704	0.859	68.2	42.57
YOLOv9t[13]	640	0.838	0.840	0.885	0.680	0.839	**2.0**	89.15
YOLOv9s[13]	640	0.861	0.837	0.892	0.694	0.849	7.2	77.29
YOLOv9m[13]	640	0.859	0.849	0.894	0.697	0.854	20.1	58.13
YOLOv9c[13]	640	0.858	0.844	0.895	0.696	0.851	25.5	57.23
YOLOv9e[13]	640	0.853	0.853	0.890	0.697	0.853	58.1	31.28
YOLOv10n[14]	640	0.865	0.844	0.891	0.697	0.854	2.3	112.03
YOLOv10s[14]	640	0.869	0.867	0.902	0.721	0.868	7.2	82.46
YOLOv10m[14]	640	0.832	0.823	0.874	0.671	0.827	15.4	75.65
YOLOv10l[14]	640	0.833	0.848	0.887	0.685	0.840	24.4	66.07
YOLOv10x[14]	640	0.840	0.825	0.878	0.678	0.832	29.5	50.44
YOLOv11n[15]	640	0.878	0.847	0.903	0.707	0.862	2.6	111.14
YOLOv11s[15]	640	0.890	0.865	0.909	**0.726**	0.877	9.4	97.67
YOLOv11m[15]	640	0.885	0.871	0.907	**0.726**	0.878	20.1	89.84
YOLOv11l[15]	640	0.876	**0.883**	0.907	0.724	0.879	25.3	55.49
YOLOv11x[15]	640	0.886	0.869	0.906	0.725	0.877	56.9	39.68
YOLOv12n[16]	640	0.848	0.811	0.888	0.658	0.829	2.6	71.94
YOLOv12s[16]	640	0.843	0.824	0.891	0.677	0.833	9.3	69.94
YOLOv12m[16]	640	0.876	0.794	0.898	0.680	0.833	20.2	55.85
YOLOv12l[16]	640	0.853	0.815	0.896	0.678	0.834	26.4	45.23
YOLOv12x[16]	640	0.860	0.825	0.900	0.684	0.842	59.1	36.30
RT-DETR-L[17]	640	**0.894**	0.876	0.866	0.661	0.885	45.0	15.94
RT-DETR-X[17]	640	0.893	0.877	0.860	0.664	**0.886**	86.0	11.09

Table 3. Per-class performance of YOLOv11s on the testing dataset

Class	Precision	Recall	mAP@0.5	mAP@0.5:0.95
All	0.890	0.865	0.909	0.726
Distracted	0.820	0.790	0.835	0.663
Focused	0.913	**0.928**	0.942	0.729
Raising Hand	0.903	0.873	0.931	**0.770**
Sleep	**0.947**	0.905	**0.945**	0.739
Using Phone	0.868	0.829	0.893	0.730

"Distracted" behaviors manifest only as brief, low-amplitude spikes. These transient fluctuations likely correspond to natural cognitive pauses, momentary fatigue, or minor instructional transitions before students quickly regain focus.

In the **Highly Active Classroom** session, interaction levels increase markedly, with "Raising Hand" reaching 16.9%, while "Focused" remains high at 76.2% and "Distracted" decreases to 6.4%. The "Sleeping" and "Using Phone" behaviors appear only marginally (0.4% and 0.1%, respectively). This pattern characterizes a dynamic learning environment where students actively engage and interact with the instructor. Temporally, Figure 4(b) vividly captures the episodic nature of active learning. The "Raising Hand" behavior is not uniformly distributed; instead, it exhibits sharp, periodic downward peaks penetrating the focused area throughout the session. These cyclical bursts directly correspond to interactive Q&A segments or instructor-prompted discussions. Notably, the blue "Distracted" band remains consistently suppressed at the bottom during these high-interaction peaks, demonstrating that active pedagogical strategies effectively mitigate attention drift.

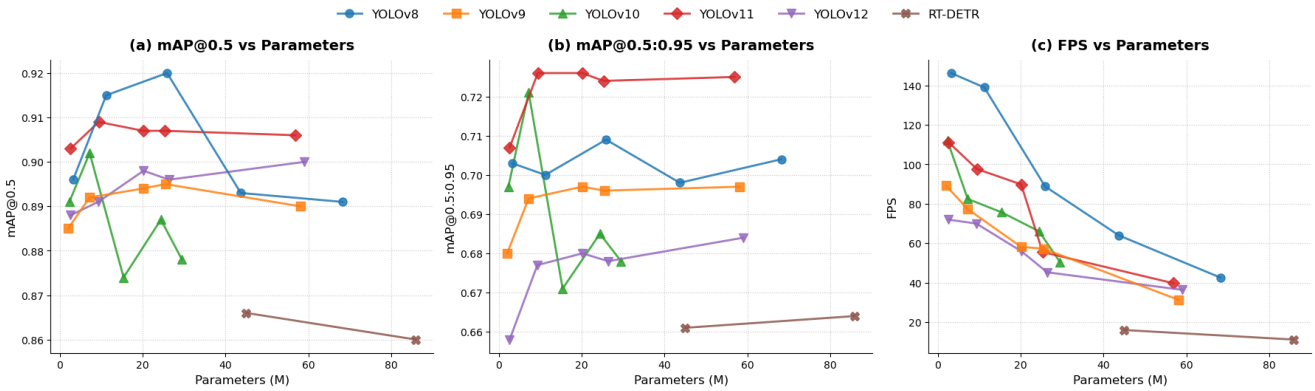
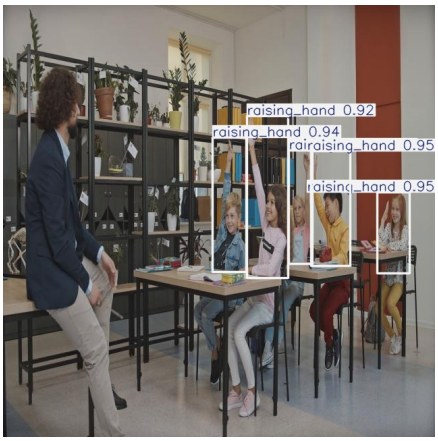
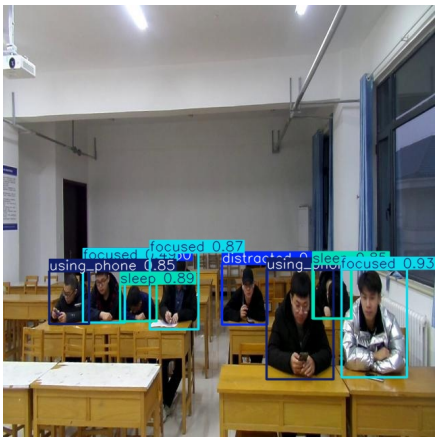


Figure 2. Relationship between model parameter count and detection performance (mAP@0.5, mAP@0.5 : 0.95 and F1-score) for student behavior recognition

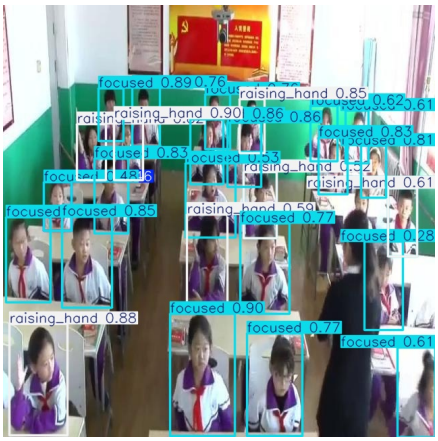


(a) Oblique camera view (sparse)

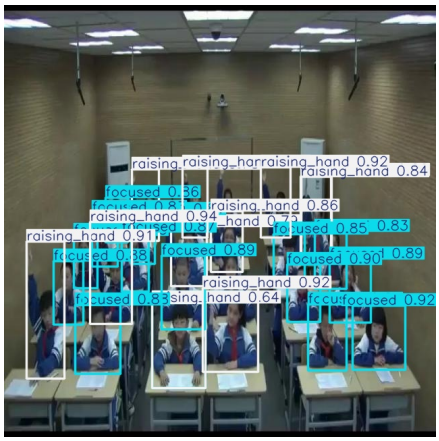
(b) Close-up view (sparse)



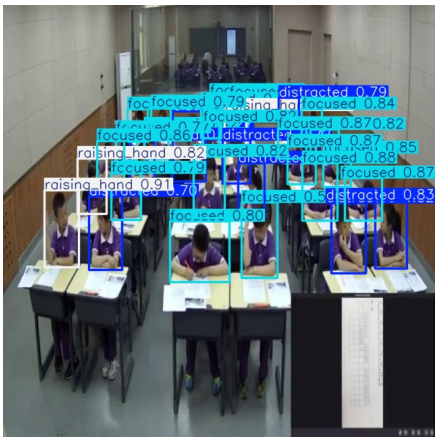
(c) Straight long-distance view (sparse)



(d) High-angle view, bright natural light (crowded)



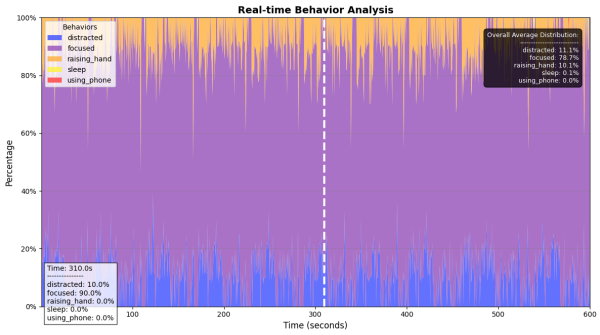
(e) Lower-angle view, artificial light (crowded)



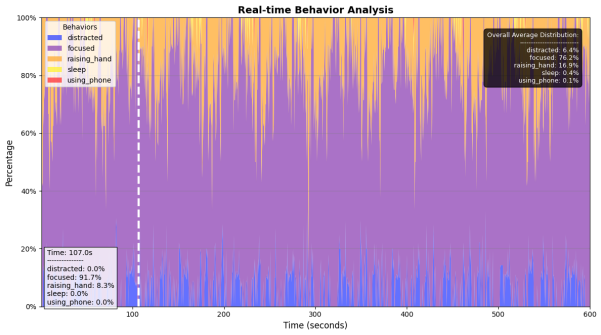
(f) Extreme high-angle view, low light (crowded)

Figure 3. Representative detection results across five student behaviors (*Focused*, *Raising Hand*, *Distracted*, *Sleeping*, and *Using Phone*) under varying population densities, camera perspectives, and lighting conditions

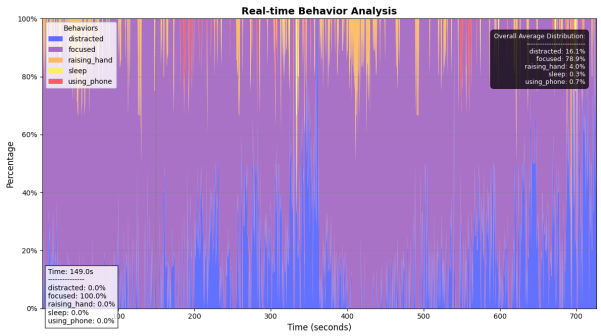
Conversely, in the **Low Engagement Classroom** session (12 minutes), the "Focused" category remains relatively dominant at 78.9%, but "Distracted" increases notably to 16.1%, accompanied by "Using Phone" at 0.7%. Meanwhile, "Raising Hand" drops to 4.0%, and



(a) Focused Classroom



(b) Active Classroom



(c) Low Engagement Classroom

Figure 4. Behavior domain charts illustrating the temporal composition of student behaviors across three engagement states detected by YOLOv11s: (a) Highly Focused, (b) Highly Active Participation, and (c) Low Engagement. The charts quantify not only the overall behavioral distributions but also explicitly highlight the temporal volatility and cyclical fluctuations in attention, participation, and distraction levels throughout the lectures

"Sleeping" remains at 0.3%. This outcome indicates surface-level attention without meaningful participatory behaviors. The temporal dynamics in this session (Figure 4c) starkly contrast with the high-engagement scenarios. Rather than remaining stable, the "Distracted" behavior exhibits significant volatility. Dense, expanding clusters of distraction peak heavily around the mid-point (e.g., between 300 and 400 seconds) and persist in dense blocks towards the end of the session. These widening distraction bands indicate a progressive breakdown of sustained attention and the onset of cognitive fatigue when instructional stimuli lack interactivity.

Overall, the behavior domain charts quantitatively highlight the distinctions among the three classroom states while visually revealing behavioral fluctuations over time. Furthermore, the successful application of the YOLOv11s-ByteTrackV2 pipeline across these three distinct instructional scenarios—without requiring session-specific fine-tuning—demonstrates the system's robust preliminary generalization potential. While the current framework performs reliably across varying engagement levels, future work will explore domain adaptation strategies to further improve scalability and ensure consistent deployment across cross-campus environments with entirely unseen layouts. These findings demonstrate that combining behavior recognition with temporal visualization can provide objective indicators of students' attention and engagement levels, thereby supporting data-driven evaluation of teaching effectiveness and pedagogical optimization.

4. Discussion

4.1. Practical Deployment and Educational Analytics Integration

This study contributes to the broader field of educational computer vision by providing a scalable and objective framework for analyzing student behavior in real classroom environments, effectively addressing the inherent limitations of manual observation and subjective evaluation. It is important to clarify that this

work does not aim to introduce a novel detection architecture. Instead, its primary novelty lies in the seamless integration of object detection, multi-object tracking, and temporal behavior modeling into a unified, end-to-end analytical pipeline tailored for educational settings. This system-oriented design shifts the focus from isolated frame-level recognition to continuous, interpretable analysis of student engagement over time. Such capabilities align directly with recent advancements in intelligent education and learning analytics, where AI-driven systems are increasingly utilized to derive actionable insights and support data-driven pedagogical decision-making [24, 25].

Regarding practical deployment, empirical results demonstrate that lightweight detectors, particularly YOLOv11s, can achieve highly reliable behavior recognition while maintaining minimal computational overhead. By proving that high accuracy can be sustained under real-world classroom constraints without requiring high-end GPUs, this study confirms the feasibility of deploying large-scale behavior analysis systems on resource-constrained edge devices commonly utilized in educational institutions.

Furthermore, the generation of behavior domain charts bridges the critical gap between raw computer vision metrics and actionable educational analytics. Rather than merely logging discrete actions, these modular charts provide educators with a quantitative, interpretable visualization of engagement dynamics. This enables the precise identification of low-engagement or high-distraction periods, thereby directly supporting timely, data-driven adjustments in teaching strategies.

4.2. Limitations and Future Work

Despite its contributions, this study presents several limitations that offer valuable opportunities for future research. First, the selected set of five behaviors serves as a foundational subset specifically tailored for traditional lecture-based settings. Consequently, it does not fully capture the nuanced dynamics of interactive environments, such as detailed note-taking, peer-to-peer interactions, or collaborative group discussions. Second, as highlighted in the experimental results, class imbalance remains an inherent challenge in naturalistic classroom datasets. While the YOLOv11s model demonstrated notable resilience in detecting minority classes such as "Raising Hand," future iterations should explore explicit mitigation strategies—such as focal loss optimization or targeted oversampling—to further stabilize minority class recall. Additionally, differentiating visually similar behaviors (e.g., distinguishing subtle phone usage from reading) under severe multi-person occlusion requires continuous algorithmic refinement.

To address these limitations, future work will prioritize expanding the dataset and extending the evaluation to diverse pedagogical scenarios, including flipped classrooms and active learning environments [26, 27]. A critical next step involves enhancing the system's scalability and cross-scenario generalization. We plan to investigate domain adaptation techniques to ensure robust, out-of-the-box performance across different campus environments with entirely unseen physical layouts and lighting conditions. Ultimately, the modular nature of this framework lays the groundwork for seamless integration with broader Learning Management Systems (LMS), enabling real-time behavioral insights to become a core, automated component of holistic teaching evaluation platforms.

5. Conclusion and Future Work

This study presented a comprehensive, end-to-end framework for classroom behavior analysis, leveraging state-of-the-art object detection and multi-object tracking techniques. To support systematic evaluation, we constructed a specialized real-world dataset comprising 3,373 images and 9,659 annotated instances across five attention-centric behaviors.

Experimental evaluations demonstrated that different model architectures exhibit distinct strengths. Specifically, YOLOv11s achieved an optimal balance between localization accuracy and computational efficiency. Furthermore, the inference speed (FPS) results confirm that such lightweight architectures are highly viable for real-time deployment on resource-constrained edge devices. Conversely, transformer-based models like RT-DETR exhibited exceptional robustness in classification, particularly when distinguishing between visually similar behaviors.

By integrating YOLOv11s with ByteTrackV2 and a temporal aggregation mechanism, the proposed system successfully shifts the analytical focus from isolated frame-level detection to continuous behavioral monitoring. The generation of behavior domain charts provides educators with an interpretable, quantitative visualization of engagement dynamics over time, offering a practical, objective tool for data-driven teaching evaluation.

Despite these advancements, inherent challenges such as dataset class imbalance, severe multi-person occlusion, and the current validation strictly within lecture-style environments remain. Moving forward, future work will focus on three primary directions to address these limitations: First, to improve algorithmic robustness against minority classes and visual ambiguities, we will explore explicit imbalance mitigation strategies (e.g., focal loss, targeted oversampling) and multi-modal feature fusion, such as integrating audio cues or facial expressions. Second, to enhance systemic scalability and cross-campus generalization, we aim to implement domain adaptation techniques and

expand the dataset to encompass more dynamic, interactive pedagogical settings, such as flipped classrooms or group-based activities. Finally, to ensure practical utility, future efforts will prioritize the development of intuitive user interfaces and the seamless integration of this analytical pipeline into broader Learning Management Systems (LMS), empowering educators to leverage AI-driven insights without requiring technical expertise.

Declarations. **Conflict of interest** The authors declare no conflict of interest.
Ethics approval and consent for participate Not applicable.

Acknowledgement. This research was funded by International School, Vietnam National University, Hanoi under project code CS.2024-12.

References

[1] Yu M., Xu J., Zhong J., Liu W. and Cheng W. (2017) *Behavior detection and analysis for learning process in classroom environment*. In 2017 IEEE Frontiers in Education Conference (FIE), pp. 1–4. IEEE.

[2] Yin Albert C.C., Sun Y., Li G., Peng J., Ran F., Wang Z. and Zhou J. (2022) *Identifying and monitoring students' classroom learning behavior based on multisource information*. Mobile Information Systems, 2022(1), 9903342.

[3] Lin L., Yang H., Xu Q., Xue Y. and Li D. (2024) *Research on student classroom behavior detection based on the real-time detection transformer algorithm*. Applied Sciences, 14(14).

[4] Li Y., Qi X., Saudagar A.K.J., Badshah A.M., Muhammad K. and Liu S. (2023) *Student behavior recognition for interaction detection in the classroom environment*. Image and Vision Computing, 136, 104726.

[5] W. Song, T. He, H. Zhang, Z. Shi, H. Chen, C. Shangguan, and C. Hao, "Intelligent recognition and analysis of student behavior in real-classroom scenarios: a comprehensive survey, exploration, and future perspectives," *Journal of King Saud University Computer and Information Sciences*, 2026.

[6] Yang F., Wang T. and Wang X. (2023) *Student classroom behavior detection based on YOLOv7+ BRA and multi-model fusion*. In International Conference on Image and Graphics, pp. 41–52. Springer Nature Switzerland.

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

[8] Sheng X., Li S. and Chan S. (2025) *Real-time classroom student behavior detection based on improved YOLOv8s*. Scientific Reports, 15(1), 14470.

[9] A. Abozeid, I. Alrashdi, and R. M. Al-Makhlasawy, "Intelligent recognition of students' behavior for smart learning environments," *Scientific Reports*, 2026.

[10] T. Li, J. Wang, C. Xu, B. Xu, N. An, and J. Zhang, "CBHA-DETR: multi-kernel attention and deformable fusion network for behavior recognition in classroom monitoring," *Multimedia Systems*, vol. 32, no. 2, p. 112, 2026.

[11] Lin F.C., Ngo H.H., Dow C.R., Lam K.H. and Le H.L. (2021) *Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection*. Sensors, 21(16), 5314.

[12] Jocher G., Chaurasia A. and Qiu J. (2023) *Ultralytics YOLOv8*. Version 8.0.0. Available at: https://github.com/ultralytics/ultralytics

[13] Wang C.-Y., Yeh I.-H. and Liao H.-Y.M. (2024) *YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information*. arXiv:2402.13616. Available at: https://arxiv.org/abs/2402.13616

[14] Wang A., Chen H., Liu L., Chen K., Lin Z., Han J. and Ding G. (2024) *YOLOv10: Real-Time End-to-End Object Detection*. arXiv:2405.14458. Available at: https://arxiv.org/abs/2405.14458

[15] Wang C.-Y. and Liao H.-Y.M. (2024) *YOLOv11: An Overview of the Key Architectural Enhancements*. arXiv:2410.17725. Available at: https://arxiv.org/abs/2410.17725

[16] Tian Y., Ye Q. and Doermann D. (2025) *YOLOv12: Attention-Centric Real-Time Object Detectors*. arXiv:2502.12524. Available at: https://arxiv.org/abs/2502.12524

[17] Zhao Y., Lv W., Xu S., Wei J., Wang G., Dang Q., Liu Y. and Chen J. (2024) *DETRs beat YOLOs on real-time object detection*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974.

[18] Zhang Y., Sun P., Jiang Y., Yu D., Weng F., Yuan Z., Luo P., Liu W. and Wang X. (2022) *ByteTrack: Multi-object tracking by associating every detection box*. In European Conference on Computer Vision, pp. 1–21. Springer Nature Switzerland.

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

[20] Yang F. and Wang T. (2023) *SCB-dataset3: A benchmark for detecting student classroom behavior*. arXiv:2310.02522. Available at: https://arxiv.org/abs/2310.02522

[21] H. Hukkelås and F. Lindseth, "Does image anonymization impact computer vision training?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 140–150, 2023.

[22] Y. Xue, V. Chinapah, and C. Zhu, "A comparative analysis of AI privacy concerns in higher education: News coverage in China and Western countries," *Education Sciences*, vol. 15, no. 6, p. 650, 2025. DOI: 10.3390/educsci15060650.

[23] M. Otani, R. Togashi, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh, "Optimal correction cost for object detection evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21107–21115, 2022.

[24] R. Sajja, Y. Sermet, D. Cwiertny, and I. Demir, "Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education," *Technology, Knowledge and Learning*, pp. 1–31, 2025.

[25] L. Cabral, R. Pinto, and G. Gonçalves, "AI-powered learning analytics dashboards: a systematic review of applications, techniques, and research gaps," *Discover Education*, vol. 4, no. 1, p. 525, 2025.

[26] Q. Liu, X. Jiang, and R. Jiang, "Classroom behavior recognition using computer vision: A systematic review," *Sensors*, vol. 25, no. 2, p. 373, 2025.
[27] R. Yang, T. Tian, and J. Tian, "Versatile teacher: A class-aware teacher–student framework for cross-domain adaptation," *Pattern Recognition*, vol. 158, p. 111024, 2025.