Systematic Evaluation of YOLO in Student Behavior Detection - A New Dataset

Anh-Tu Tran*, Thai-Khanh Nguyen*, Trung-Thanh Nguyen*, Anh-Ha Tuan*, Trung-Hieu Le*

* Faculty of Information Technology, Dainam University, Hanoi, Vietnam
Email: hieult@dainam.edu.vn

***Abstract*—Detecting student behavior is critical for assessing engagement and enabling adaptive instruction. Manual observation, while informative, lacks scalability in large university classrooms. This paper presents a YOLO-based, frame-level detection framework tailored to real-world settings. Key contributions include: (1) a new dataset from actual classrooms using a single ceiling-mounted camera, annotated with multiple behavior and object classes; (2) a benchmark of YOLOv7, YOLOv8, YOLOv12, and Faster R-CNN under challenging conditions; and (3) a lightweight browser-based system enabling real-time inference without cloud support. Results confirm the feasibility of automated classroom behavior monitoring for educational AI applications.**

***Index Terms*—Student Behavior Detection, YOLO, Real-Time Analytics, Dataset**

I. INTRODUCTION

Understanding student behavior is fundamental to assessing engagement, attention, and learning effectiveness in educational settings [1], [2]. Visual cues such as body posture, gaze direction, or interactions with personal devices often indicate motivation, distraction, or confusion [3]. However, consistently monitoring such behaviors in real-world classrooms especially in large or dynamic environments remains a significant challenge due to occlusions, frequent motion, and the inherent limitations of teacher observation [4], [5]. These limitations often lead to missed opportunities for timely intervention and personalized support.

In response, computer vision–based behavior detection systems have emerged as scalable, automated alternatives. Real-time detection models can help instructors identify disengaged students, adjust instruction dynamically, and enhance classroom awareness. Unlike recognition methods that require temporal modeling, frame-level detection offers a simpler yet effective approach by localizing observable actions from static image frames. This enables practical deployment using standard single-camera setups, offering real-time support without excessive computational demands. Deep learning–based object detectors, particularly those in the YOLO family, have demonstrated strong performance in behavior recognition tasks across domains such as surveillance and healthcare [6]. Their speed and ability to detect small, occluded actions make them well-suited for classroom scenes with multiple students. Prior classroom studies often rely on facial cues or temporal clips to infer engagement [7], whereas our method focuses on frame-level detection of diverse behaviors.

Yet the effectiveness of such systems is often limited by the lack of diverse, well-annotated datasets collected in realistic classroom environments especially for infrequent or overlapping behaviors. For instance, the SCB-Dataset3 benchmark [8] provides bounding-box annotated classroom images across six predefined student behavior categories and serves as a useful reference for evaluating visual detection models. However, it presents limitations in behavioral diversity and annotation design, as it adopts a single-label-per-instance scheme and omits contextual object information. Additionally, it does not address deployment-related aspects such as real-time inference performance or integration with live classroom monitoring systems.

To address these challenges, we propose a YOLO-based detection approach for student behavior analysis at the frame level. Specifically, this study contributes in three aspects: First, we introduce a newly collected dataset annotated with nine categories,including seven student behaviors (using phone, using computer, sleeping, turning left, turning right, raising hand, and writing) and two object types (phone and computer). The data was recorded using a single ceiling-mounted camera in real classroom environments. Second, we benchmark three YOLO-based architectures (YOLOv7, YOLOv8, YOLOv12) and evaluate their detection accuracy across all categories. Third, we demonstrate the feasibility of real-time deployment through a browser-based prototype system that achieves 17 FPS on a standard GPU-equipped machine.

The remainder of this paper is organized as follows: Section II reviews related work in behavior detection and summarizes relevant datasets. Section III details the data collection setup, annotation methodology, and dataset characteristics. Section IV presents experimental results and implementation of the detection models. Section V concludes the study and outlines future research directions.

II. RELATED WORK

A. Datasets for Classroom Student Behavior Detection

The integration of artificial intelligence (AI) in education has increased the demand for automatic student behavior detection. However, existing datasets often remain limited in scale, behavioral diversity, or applicability to real-world classroom settings. SCB-Dataset5 [9] addressed this issue by providing 7,428 images with 106,830 bounding boxes across 20 behavior categories involving both students and teachers.

979-8-3315-9444-2/25/$31.00 ©2025 IEEE

To further enhance dataset diversity and scale, the StudentAct dataset [10] contributed 31,046 multi-view classroom images with over 596,000 annotations spanning five common student behaviors. Similarly, Xu et al. [11] proposed the ARIC dataset, comprising 36,453 images annotated with 32 behavior types. This dataset was captured using multiple camera angles and enriched with textual and audio modalities to support multi-modal behavior analysis. In another contribution, Sharma et al. [12] developed EduNet, a large-scale dataset consisting of 7,851 classroom video clips representing 20 distinct behaviors. A baseline I3D-ResNet-50 model trained on this dataset achieved an accuracy of 72.3%.

B. Student Behavior Detection Methods

Student behavior detection in real-world classrooms presents challenges such as occlusion, inter-class similarity. Recent studies have reframed this task as a real-time object detection problem, leveraging YOLO and its variants to simultaneously localize and classify student behaviors from static images. Chen et al. [6] proposed an enhanced YOLOv8 model integrating C2f_Res2block, Multi-Head Self-Attention (MHSA), and Efficient Multi-scale Attention (EMA). Their model achieved a 4.2% increase in mAP@0.5 over the YOLOv8 baseline, with improved performance under occlusion and small-object conditions. Han et al. [13] introduced WAD-YOLOv8, incorporating CA-C2f, 2DPE-MHA, and Dysample modules for robust detection in wide-angle and crowded classrooms. On the SCB-U dataset, it achieved 90.1% mAP@0.5 and 77.1% mAP@0.5:0.95, outperforming baseline YOLOv8 by 18.7% and 14.8%, respectively, while maintaining real-time speed at 49.8 FPS. Liang et al. [14] developed a YOLOv5s-based framework enhanced with wavelet transforms, Harr downsampling, and MPDIoU loss. The model improved detection robustness, increasing precision by 8.6% and recall by 6.0% compared to the baseline. Ding et al. [15] proposed an improved YOLOv8 detector using SimAM attention, CARAFE upsampling, and MPDIoU regression loss. Evaluated across multiple datasets, their model yielded mAP improvements ranging from 5.0% to 7.4%, demonstrating strong resilience in complex scenes. Tang et al. [16] introduced a YOLOv5-based system combining BiFPN feature fusion and CBAM attention. On a custom dataset of four behavior categories, it achieved 89.8% classification accuracy and 90.4% recall, with real-time inference at up to 45 FPS. Peng et al. [17] presented YOLO-CBD, an improved YOLOv10s-based framework incorporating BiFormer attention and behavior feature aggregation. Their model demonstrated superior mAP and recall over baseline YOLO models and was designed for real-time classroom deployment, though detailed FPS and mAP figures were not disclosed.

Collectively, these studies confirm the effectiveness of YOLO-based detectors for accurate, efficient, and real-time student behavior detection under realistic and challenging classroom conditions.

III. T-Students FITDNU - A Real-World Dataset for Student Behavior Detection

Previous datasets have mainly captured prominent classroom actions like raising hands or standing, often neglecting subtle negative behaviors such as phone usage or dozing off.

While classroom dynamics may vary, we focus on key actions: (1) using phone, (2) using computer, (3) sleeping, (4) turning left, (5) turning right, (6) raising hand, (7) writing, and two objects: (8) phone and (9) computer.

A. Motivation and Collection Environment

The dataset was collected in three real-world university classrooms using a HIKVISION DS-2CD1323G0-IUF IP dome camera mounted centrally on the ceiling (Figure 1). This fixed-angle setup enabled full spatial coverage with minimal intrusion.

Despite using a single camera, the system captured diverse classroom layouts, varying student densities, and spontaneous behaviors. Videos were recorded in 1920×1080 resolution at 25 fps across multiple sessions held at different times of day to capture variability in both engagement and lighting. Class sizes ranged from 23 to 35 students, typical of higher education settings. Each session lasted about 50 minutes, but only representative segments were extracted to ensure behavior diversity while keeping the dataset size manageable.

TABLE I
Technical Specifications of the Collected Classroom Video Dataset

Parameter	**Description**
Max. number of students	Approx. 50 students
Camera setup	Single camera, centrally positioned
Camera resolution	1920×1080 pixels (Full HD)
Recording speed	25 frames per second (fps)
Collection sessions	Three real classes
Students per class	23–35 students
Collection times	Morning and afternoon sessions

B. Annotation Process and Class Definitions

We adopted a three-stage annotation pipeline using Roboflow to ensure scalable and high-quality labeling:

- **Stage 1 – Manual Labeling:** An initial seed set of 300 images was manually annotated by four trained annotators, with each image verified by at least two others. The process achieved over 95% inter-annotator agreement, following established object detection practices [18].
- **Stage 2 – Semi-Automated Labeling:** A YOLOv12 model trained on the seed set (mAP@0.5 = 72%) was used to pre-label the remaining images. Annotators then corrected errors, especially under occlusion, class overlap, and small-object confusion.
- **Stage 3 – Augmentation:** To boost small-object performance, we applied brightness and saturation shifts

Fig. 1. Corresponding classroom camera image

($\pm$20%, $\pm$15%) and MixUp [19] ($\alpha = 0.5$), improving phone detection mAP from 10.7% to 12.8%.

The dataset comprises **9 annotated classes**, including:

- **Behavior classes:** writing, raising hand, turning left, turning right, sleeping, using phone, using computer.
- **Object classes:** phone, computer.

C. Dataset Statistics and Visual Challenges

The finalized T-Students FITDNU dataset comprises 3,351 classroom images and 126,429 bounding boxes annotated across nine classes, including seven student behaviors and two object types. These annotations reflect real-world classroom dynamics captured under unconstrained conditions using a single fixed-view camera.

Table II shows the distribution of bounding boxes per class. Class imbalance is evident, with raising_hand containing only 318 instances, while phone appears in over 45,000 bounding boxes. This natural distribution was preserved without artificial balancing to maintain the authenticity of student activity frequency.

TABLE II
DISTRIBUTION OF BOUNDING BOXES BY CLASS

No.	Class	Bounding Boxes
1	Computer	20100
2	Phone	45910
3	Raising Hand	318
4	Sleeping	4826
5	Turning Left	9720
6	Turning Right	9222
7	Using Computer	8412
8	Using Phone	23709
9	Writing	4212

As illustrated in Figure 2, sample frames reveal a wide range of visual challenges, including overlapping actions, pose variation, and substantial differences in object scale. The average bounding box dimensions (in width × height pixels) vary significantly across classes, reflecting both behavioral diversity and object complexity. Specifically: using_computer (46 × 101), using_phone (38 × 82), raising_hand (20 × 63), writing (38 × 83), turning_left (33 × 58), turning_right (31 × 56), sleeping (38 × 55), computer (34 × 55), and phone (11 × 17). Such scale variance complicates small-object detection under crowded scenes.

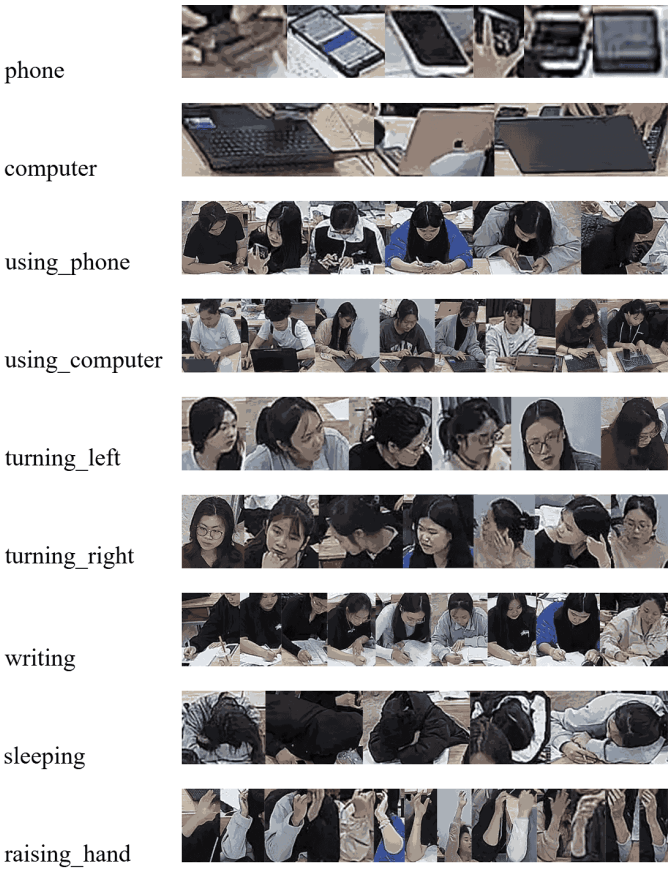

Fig. 2. Sample images showing class diversity, scale variation, and occlusion.

The dataset captures multiple students per frame, leading to an average of 5.6 objects per image. Dense scenes and real-time interaction introduce label overlap between behaviors and objects. Notably, behavior labels often encompass interaction zones, while object labels remain tightly localized posing a unique detection challenge. These attributes make the dataset particularly suitable for evaluating multi-scale, context-aware object detection models under realistic classroom constraints.

To encourage further research in classroom behavior detection, we publicly release the T-Students FITDNU dataset, which consists of real classroom images with detailed annotations. The dataset has been collected with institutional ethics approval and is available for academic and non-

commercial use. For access, please contact us via email at: tu05062005@gmail.com.

D. Comparison with Existing Datasets

To emphasize the unique characteristics and practical challenges of the T-Students FITDNU dataset, we present a comparative analysis against two representative classroom behavior datasets: StudentAct [10] and SCB-Dataset3 [8]. Table III summarizes key attributes across these datasets, including scale, behavior categories, annotation methodology, and deployment settings.

While StudentAct and SCB-Dataset3 have contributed significantly to benchmark development, they differ substantially in their design assumptions and practical applicability. StudentAct was constructed from scripted activities across synchronized multi-camera views, limiting its realism for frame-level, real-time deployment. SCB-Dataset3, though captured in natural classroom scenes, contains fewer behavior categories and does not provide object annotations reducing its suitability for evaluating multi label detection frameworks.

In contrast, T-Students FITDNU was designed for real-world, single camera classroom environments. It captures spontaneous student behaviors under occlusion, pose variation, and varying lighting conditions often underrepresented in existing datasets. Furthermore, it combines both behavior and object annotations in each frame, enabling fine-grained evaluation of detection models under multi-label and small-object constraints. Importantly, all annotations were verified under a semi-automated pipeline to ensure consistency while scaling, and the dataset has been released publicly with institutional approval making it a reproducible and accessible resource for practical classroom behavior detection research.

IV. Proposed Method for Behavior detection

A. Behavior detection models used

Detecting student behaviors in real world classroom environments presents a distinct set of challenges compared to traditional single subject action recognition tasks. In these settings, multiple students may perform different actions simultaneously, often under occlusion, lighting variation, and within high density layouts. Therefore, the detection model must not only localize multiple subjects but also correctly classify subtle and overlapping behaviors in cluttered scenes.

To address these challenges, we evaluated three state-of-the-art object detection models from the YOLO family: YOLOv7, YOLOv8, and YOLOv12. These models were selected for their strong balance between accuracy and inference speed, as well as their proven performance on small-object detection in dense scenes properties critical for behavior detection in classrooms.

1) YOLOv7 – anchor-based detection with repconv optimization: YOLOv7 is an anchor-based object detector that employs the e-elan backbone and repconv modules to balance accuracy and efficiency. its relatively compact architecture makes it suitable for resource-constrained systems, establishing a strong baseline in our experiments.

2) YOLOv8 – anchor-free with decoupled heads and c2f modules: YOLOv8 is a fully anchor free model that integrates c2f modules and decoupled detection heads, improving both classification and localization. this architecture is particularly effective for detecting small-scale or overlapping behaviors under complex classroom conditions.

3) YOLOv12 – encoder–decoder design with attention mechanisms: YOLOv12 adopts a hybrid encoder–decoder architecture enhanced by multi scale attention. its emphasis on semantic representation and spatial reasoning makes it highly robust to occlusion, low contrast, and subtle posture variations critical for fine grained behavior analysis.

Our proposed behavior detection pipeline consists of three main stages: data acquisition, model training, and inference. High resolution classroom images are first collected and annotated using a structured semi automatic labeling process. The annotated dataset is then used to train each YOLO variant using standard detection architecture. During inference, each trained model receives a single frame classroom image and directly outputs bounding boxes with behavior or object class labels. An overview of the entire pipeline is illustrated in Figure 3.

B. Evaluation Metrics and Results

To evaluate behavior detection accuracy, we adopt the standard Intersection over Union (IoU), defined as:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{B_{gt} \cap B_{pred}}{B_{gt} \cup B_{pred}}$$

where B_{gt} and B_{pred} denote the ground truth and predicted bounding boxes, respectively. Although extensions such as Generalized IoU [20] address alignment in distant or occluded cases, we report results using standard IoU for clarity and consistency with prior work.

For performance comparison, we report precision, recall, mAP@0.5, and mAP@[0.5:0.95]. Table IV presents results across nine annotated behavior and object classes using four detection models: YOLOv7, YOLOv8l, YOLOv12s, and Faster R-CNN.

Experimental findings indicate that detection performance is strongly influenced by model architecture. **YOLOv8l** achieves the highest overall accuracy, particularly on subtle behaviors such as Using Computer (83.0% mAP@[0.5:0.95]), Sleeping (82.6%), and Turning Left (80.7%). **YOLOv7** remains competitive on dominant behaviors like Using Phone (96.9% mAP@0.5), but its performance declines on visually ambiguous classes, such as Raising Hand (50.8%). This category is especially challenging due to its very limited data (only 318 samples), small gesture size, and frequent occlusion, which restrict the model's ability to generalize. While **YOLOv12s** incorporates attention mechanisms and encoder–decoder structures, it underperforms due to model complexity and sensitivity to class imbalance, leading to inconsistent results across underrepresented behaviors. In contrast, **Faster R-CNN** yields the lowest scores across all metrics. This underperformance may stem from its two-stage pipeline, which is less effective

TABLE III
COMPARISON OF KEY DATASET CHARACTERISTICS

Dataset	T-Students FITDNU	StudentAct	SCB-Dataset3
Number of images	3,351	31,046	5,686
Number of boxes	126,429	596,371	45,578
Class types	9 (7 behaviors + 2 objects)	5 behaviors	6 behaviors
Camera setup	Single fixed ceiling camera	Five synchronized cameras	Not specified
Annotation method	Manual + semi-auto (YOLOv12)	Manual + interpolation	Fully manual
Collection context	Real university classrooms (non-scripted)	Scripted multi-view sessions	School scenes
Key challenges	Small objects, occlusion, multi-label	View fusion, motion blur	Subtle postures, limited scale

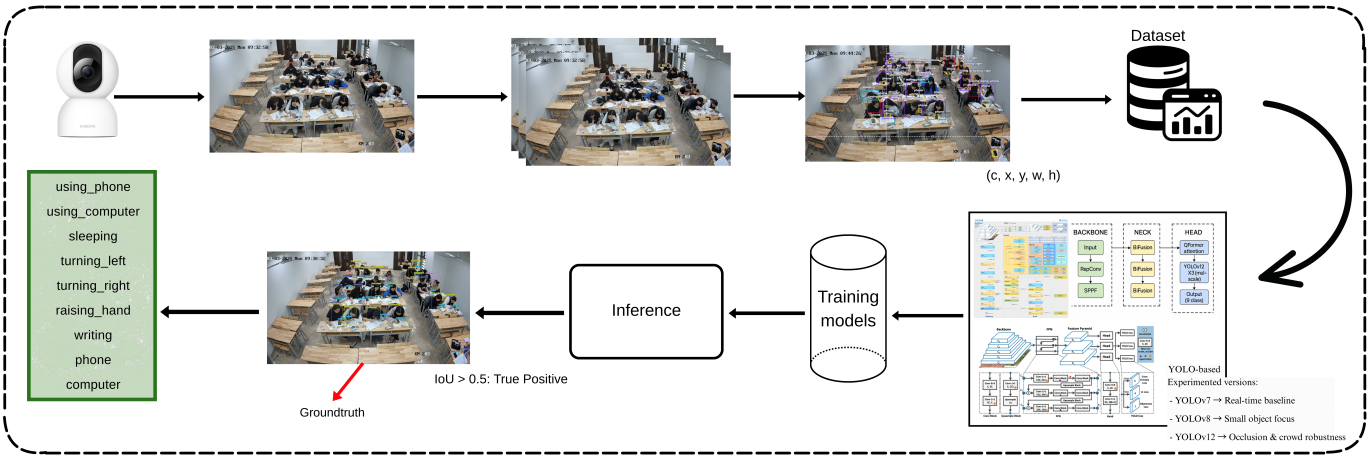

Fig. 3. Overview of the classroom behavior detection system architecture.

in dense, small-object classroom environments. Notably, it only achieves 7.8% mAP@0.5 for Phone and 13.6% for Using Phone, reflecting its limited applicability in real-time educational scenarios.

Overall, **YOLOv8l** demonstrates the most reliable balance between detection accuracy and robustness, offering clear advantages for classroom behavior monitoring in visually complex environments.

C. Real-Time Inference Result

To demonstrate the practical deployment capability of the proposed system, we implemented a real-time student behavior detection module based on the YOLOv8l model. The system receives live RTSP video streams from a ceiling-mounted HIKVISION IP camera and delivers annotated output through a lightweight browser-accessible interface.

The system operates via a multithreaded architecture comprising three main components:

- Frame Capture: A dedicated thread continuously fetches video frames at 25 FPS and resizes them to 640×640 to match the model input size.
- Behavior Inference: A second thread performs detection using the trained YOLOv8l model and overlays bounding boxes and class labels on each frame.
- Web Visualization: Annotated frames are encoded in MJPEG and streamed via HTTP using a Flask server. A live behavior frequency chart is also rendered for classroom analytics.

During deployment in actual classroom sessions, the system achieved a consistent speed of approximately 17 FPS on a local machine equipped with an NVIDIA RTX GPU, an Intel i7 CPU, and 32GB RAM. As shown in Figure 4, the system accurately detects high-frequency actions such as *writing*, *using phone*, and *sleeping*, while more subtle or occluded behaviors like *raising hand* remain challenging.

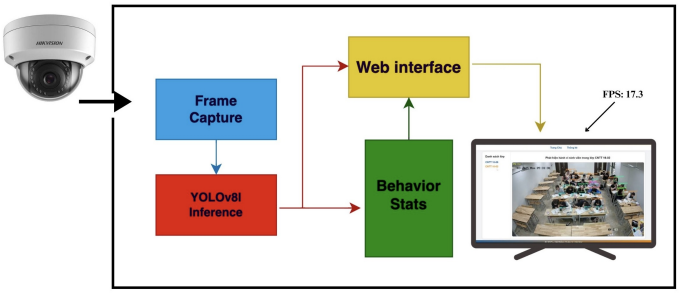

Fig. 4. Architecture of the YOLOv8l-based real-time detection system

This implementation demonstrates that real-time, single-camera classroom behavior recognition is feasible without relying on cloud infrastructure, making it suitable for on-site educational environments.

TABLE IV
PERFORMANCE COMPARISON OF OBJECT DETECTION MODELS ON THE T-STUDENTS FITDNU DATASET

Activity	Model	Precision	Recall	mAP@0.5	mAP@ [0.5:0.95]
Phone	Faster R-CNN	31.4	38.9	7.8	31.4
	YOLOv7	86.8	77.9	83.2	42.1
	YOLOv8l	**89.6**	**78.9**	**87.9**	**55.7**
	YOLOv12s	86.2	62.0	73.3	43.7
Using Phone	Faster R-CNN	55.6	65.2	13.6	55.6
	YOLOv7	92.1	94.9	96.9	67.1
	YOLOv8l	**95.5**	**95.8**	**97.7**	**81.0**
	YOLOv12s	87.8	92.6	93.9	67.1
Computer	Faster R-CNN	58.6	66.9	15.7	58.6
	YOLOv7	94.1	96.6	97.0	66.1
	YOLOv8l	**96.4**	**98.2**	**98.6**	**79.1**
	YOLOv12s	93.6	96.5	96.5	70.5
Turning Left	Faster R-CNN	43.4	54.3	27.6	43.4
	YOLOv7	91.6	91.4	96.2	61.8
	YOLOv8l	**94.8**	**95.7**	**97.5**	**80.7**
	YOLOv12s	84.3	85.3	88.9	58.7
Turning Right	Faster R-CNN	40.3	51.3	28.4	40.3
	YOLOv7	88.2	90.7	94.4	59.9
	YOLOv8l	**94.1**	**93.4**	**96.4**	**78.8**
	YOLOv12s	83.6	83.7	87.6	58.6
Using Computer	Faster R-CNN	54.4	64.2	23.3	54.4
	YOLOv7	87.0	**96.3**	96.3	69.7
	YOLOv8l	**95.2**	94.9	**97.6**	**83.0**
	YOLOv12s	85.2	90.2	92.0	69.1
Sleeping	Faster R-CNN	55.9	63.4	51.0	55.9
	YOLOv7	92.4	96.4	96.1	65.8
	YOLOv8l	**96.5**	**98.8**	**98.9**	**82.6**
	YOLOv12s	91.1	97.2	96.4	69.7
Writing	Faster R-CNN	48.2	57.4	40.7	48.2
	YOLOv7	89.7	92.7	95.5	63.3
	YOLOv8l	**93.8**	**93.1**	**95.6**	**78.2**
	YOLOv12s	83.4	86.9	88.0	64.5
Raising Hand	Faster R-CNN	25.5	33.9	39.9	25.5
	YOLOv7	**90.9**	83.1	87.6	50.8
	YOLOv8l	84.4	**96.2**	**95.7**	**70.3**
	YOLOv12s	61.2	34.6	52.9	39.0

V. CONCLUSION AND FUTURE WORK

This study introduced the T-Students FITDNU dataset, capturing frame-level student behaviors in authentic Vietnamese university classrooms using a single fixed camera. The dataset includes nine annotated categories covering both actions and related objects, constructed via a semi-automated labeling process. Evaluation of three YOLO-based models showed that YOLOv8l consistently outperformed others, especially for subtle or small-scale behaviors, though challenges remain in detecting occluded actions such as hand-raising.

Future work will focus on expanding the dataset to more diverse classroom settings and applying Focal Loss to address class imbalance. We also plan to transition from static detection to temporal behavior recognition by incorporating sequence-based models, aiming for deeper insight into student engagement and more intelligent classroom monitoring systems.

REFERENCES

[1] W. M. Reinke, K. C. Herman, and C. B. Copeland, "Student engagement: The importance of the classroom context," *Handbook of research on student engagement*, pp. 529–544, 2022.

[2] T. D. Nguyen, M. Cannata, and J. Miller, "Understanding student behavioral engagement: Importance of student interaction with peers and teachers," *The journal of educational research*, vol. 111, no. 2, pp. 163–174, 2018.

[3] E. A. Skinner and M. J. Belmont, "Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year." *Journal of educational psychology*, vol. 85, no. 4, p. 571, 1993.

[4] M. Cents-Boonstra, A. Lichtwarck-Aschoff, E. Denessen, N. Aelterman, and L. Haerens, "Fostering student engagement with motivating teaching: An observation study of teacher and student behaviours," *Research Papers in Education*, vol. 36, no. 6, pp. 754–779, 2021.

[5] T. Havik and E. Westergård, "Do teachers matter? students' perceptions of classroom interactions and student engagement," *Scandinavian journal of educational research*, vol. 64, no. 4, pp. 488–507, 2020.

[6] H. Chen, G. Zhou, and H. Jiang, "Student behavior detection in the classroom based on improved yolov8," *Sensors*, vol. 23, no. 20, p. 8385, 2023.

[7] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*, 2017, pp. 33–40.

[8] F. Yang and T. Wang, "Scb-dataset3: A benchmark for detecting student classroom behavior," *arXiv preprint arXiv:2310.02522*, 2023.

[9] F. Yang, "Scb-dataset: a dataset for detecting student classroom behavior," *arXiv preprint arXiv:2304.02488*, 2023.

[10] P.-D. Nguyen, H.-Q. Nguyen, T.-B. Nguyen, T.-L. Le, T.-H. Tran, H. Vu, and Q. N. Huu, "A new dataset and systematic evaluation of deep learning models for student activity recognition from classroom videos," in *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2022, pp. 1–6.

[11] L. Xu, F. Meng, Q. Wu, L. Pan, H. Qiu, L. Wang, K. Chen, K. Geng, Y. Qian, H. Wang *et al.*, "Aric: An activity recognition dataset in classroom surveillance images," *arXiv preprint arXiv:2410.12337*, 2024.

[12] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Edunet: a new video dataset for understanding human activity in the classroom environment," *Sensors*, vol. 21, no. 17, p. 5699, 2021.

[13] L. Han, X. Ma, M. Dai, and L. Bai, "A wad-yolov8-based method for classroom student behavior detection," *Scientific Reports*, vol. 15, no. 1, p. 9655, 2025.

[14] W. Liang, J. Zhang, J. Wang, and H. Fan, "A yolo-based behavior detection and analysis system for smart classrooms," in *International Conference on Remote Sensing, Mapping, and Image Processing (RSMIP 2025)*, vol. 13650. SPIE, 2025, pp. 340–346.

[15] H. Ding, Y. Zhang, X. Fu, X. En, and M. Cao, "Robust student behavior detection in complex classroom environments using an enhanced yolo-based approach," *Available at SSRN 5297551*.

[16] L. Tang, T. Xie, Y. Yang, and H. Wang, "Classroom behavior detection based on improved yolov5 algorithm combining multi-scale feature fusion and attention mechanism," *Applied Sciences*, vol. 12, no. 13, p. 6790, 2022.

[17] S. Peng, X. Zhang, L. Zhou, and P. Wang, "Yolo-cbd: Classroom behavior detection method based on behavior feature extraction and aggregation," *Sensors*, vol. 25, no. 10, p. 3073, 2025.

[18] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[20] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.