Proceedings of Machine Learning Research 339:160–169, 2026 IRAISE Workshop at the Festival of Learning 2026

# Privacy-Preserving Region-Level Classroom Engagement Analytics for Teacher Reflection

**Zhiyuan Xiang** julian.xiang1123@gmail.com
*Duke University, Pratt School of Engineering, Department of Mechanical Engineering*

**Qiao Lin*** sherrylin417@outlook.com
*University of Pennsylvania, Graduate School of Education*

**Xueyan Jia** JXYZON@163.com
*Hainan University, School of Information and Communication Engineering*

**Shengkai Zheng** shengkaizheng@outlook.com
*North China University of Science and Technology, School of Materials Science and Engineering*

## Abstract

Monitoring classroom-wide students' visible behavioral engagement is important for teacher reflection and instructional adjustment, yet video-based classroom analytics raise substantial privacy concerns when students may be identifiable from visual data. To address this tension, this paper presents a teacher-facing, region-level classroom engagement analytics framework that converts classroom video into temporal and spatial summaries while reducing reliance on identity-sensitive visual information. The framework combines stochastic head obfuscation, adversarial identity decoupling through a gradient reversal layer, and differentially private release of aggregated engagement summaries through Laplace noise injection. Using 21 short-duration classroom video segments from the Dataset for Classroom Group Engagement Recognition and a 13-minute extended instructional sequence, we examine what interpretable temporal and spatial signals the system can generate for classroom reflection across short-duration and extended analyses. We also conduct a professional evaluation with four middle school teachers to examine whether these outputs are understandable and aligned with teachers' professional observations. The generated outputs and teacher evaluation suggest that region-level timelines and heatmaps can help teachers review moments and classroom areas that warrant contextual interpretation.

**Keywords:** Engagement Analysis; Classroom Analytics; Privacy.

## 1. Introduction

Student engagement is closely related to students' conceptual understanding and academic achievement (Reeve et al., 2004), and it provides teachers with important cues for adjusting instruction and reflecting on classroom interaction (Virtanen et al., 2013). In live classrooms, however, teachers often need to monitor many students at once while simultaneously delivering instruction, managing activities, and responding to classroom dynamics. As a result, they may rely on partial and uneven cues, such as eye contact, visible participation, or responses from a small number of vocal students. These cues are valuable but incomplete, making it difficult for teachers to notice classroom-wide temporal changes, spatial participation gaps, or quiet forms of disengagement during instruction.

Automated classroom analytics can help surface temporal and spatial engagement patterns that are difficult to observe in real time (Liu et al., 2025). However, classroom sensing systems also raise concerns about student privacy, surveillance, and the interpretation of engagement indicators. When analytics are reported at the individual-student level, visible behaviors may be misread as judgments of motivation, conduct, or teaching quality. To address this tension, our framework reorients classroom engagement analytics toward teacher reflection by shifting the analytic unit from individual students to predefined classroom regions. The system treats classroom video as a spatial signal of collective activity and shows how visible engagement patterns vary across regions and over time. This region-level framing supports teacher interpretation while reducing reliance on identifiable student information and individual-level behavioral judgments.

This paper presents the design rationale of the framework and examines how it translates classroom video into teacher-facing temporal, spatial, and disparity-sensitive summaries for reflective interpretation. Using 21 short-duration classroom video segments from the Dataset for Classroom Group Engagement Recognition (Lu et al., 2025), together with a 13-minute extended classroom recording constructed by combining selected segments, we examine what temporal, spatial, and disparity-sensitive visible behavioral engagement signals the system can generate for classroom reflection across short-duration and extended analyses. We also conduct an exploratory professional evaluation with four middle school teachers to examine whether these outputs are understandable, aligned with teachers' professional observations, and potentially useful in practice. This study is guided by two research questions: *RQ1.* What temporal, spatial, and disparity-sensitive visible behavioral engagement signals can a privacy-aware classroom analytics tool generate from classroom video data? *RQ2.* How do teachers interpret these signals, and how useful do they perceive them to be for classroom reflection?

## 2. Related Works

Recent advances in classroom sensing and video analysis create new possibilities for capturing fine-grained engagement patterns in authentic educational settings (Liu et al., 2025). Bidwell and Fuchs (2011) used color-based markers and Kinect depth sensing to connect classroom observation with automated engagement measurement. Raca and Dillenbourg (2013) later developed a classroom attention monitoring system that estimated students' attention from motion information and behavioral synchronization among neighboring students. Raca (2015) further examined camera-based estimation of students' attention through manually extracted visual features, including head pose, classroom position, periods of stillness, and head movement distance. However, much of this earlier work focused on estimating attention or engagement from low-level behavioral features, with less attention to how analytics should be reported to teachers in a privacy-sensitive and pedagogically interpretable form. For teacher reflection, the central question is not only whether a model can detect behavior, but how its outputs can help teachers notice classroom patterns without turning students into individually monitored subjects.

Beyond questions of analytic utility and pedagogical interpretation, systems that analyze or monitor students through classroom video also raise ethical privacy risks and technical privacy-utility challenges. Classroom videos may expose students' faces, identities, attendance patterns, and behavioral traces, and visual representations can create risks of identity inference when models retain or reconstruct sensitive facial information (Khosravy et al., 2022). These risks are especially important in educational contexts because students may not have equal power to opt out of data collection, and because privacy concerns shape students' experiences of technology-enhanced learning environments (Blackmon and Major, 2023). In response to these risks, prior privacy-preserving video analytics research has explored technical approaches for reducing identity exposure while preserving analytic usefulness. Ren et al. (2018) proposed a video face anonymizer for privacy-preserving action detection, using adversarial training to remove privacy-sensitive face identity information while maintaining action detection performance. In their framework, the anonymizer modifies face pixels, while the discriminator attempts to recover identity-sensitive information from the anonymized video. Wang et al. (2019) recognized that conventional video sanitization often detects and blurs regions of faces, license plates, locations, but does not formally quantify or bound privacy leakage, so differentially private video analytics were proposed.

Although prior work has explored privacy-preserving video analytics, these approaches do not fully address the specific privacy risks of classroom engagement analysis, which still focus on individual action recognition. For example, Ren et al. (2018) reduce facial identity information for privacy-preserving action detection, but the downstream output, fine-grained descriptions of individual actions or behavioral patterns, may still be linked to particular students. Inspire by Wang et al. (2019) 's emphasis on privacy-preserving video analytics, our framework avoids reporting what each student is doing nor does it assign engagement judgments to individual students. Students' visible behaviors are reflected through the collective behavior of the regions in which they are located, forming region-level temporal and spatial summaries of classroom engagement. This design helps teachers focus on whole-class patterns, thereby reducing the risk of student monitoring while preserving useful signals for teacher reflection.

## 3. Privacy Preserving Classroom Engagement Analytics Framework

The proposed tool estimates classroom-level visible behavioral engagement proxies through aggregated, region-level analytics designed for teacher reflection. Given an input video stream $V = \{I_1, I_2, \dots, I_t\}$, where $I_t$ denotes the frame at time step $t$, the system produces a set of regional engagement outputs $E_t = \{e_{t,1}, e_{t,2}, \dots, e_{t,Z}\}$ over $Z$ pre-defined classroom zones. The framework maps classroom video to temporal and spatial engagement summaries that can support interpretable whole-class reflection.

### 3.1. Input Privacy Enhancement Module

To reduce dependence on facial information, the model applies stochastic head obfuscation during training. Unlike post-hoc anonymization, which removes sensitive information only after inference, this strategy limits the model's opportunity to learn identifiable facial representations in the first place. For each annotated bounding box $b_i = (x_1, y_1, x_2, y_2)$, the head region is defined as the top $r\%$ of the box. With probability $p$, the cropped head region $I_{crop}$ is randomly transformed as follows:

$$I'_{crop} = \begin{cases} \text{Gaussian}(I_{crop}, \sigma), & \text{if random} < 0.5 \\ \text{Blackout}(I_{crop}), & \text{otherwise} \end{cases}. \tag{1}$$

This training strategy encourages the model to rely more on body posture and movement cues than on facial details, thereby reducing identity leakage at the input stage.

### 3.2. Feature Identity Decoupling Module

To further reduce identity-related information from the learned representations, we introduce an adversarial learning branch composed of a feature extractor $E$ and an identity discriminator $D_{id}$. The discriminator is trained to predict student identity labels from the extracted features $F$, while the feature extractor is optimized to produce representations that are invariant to identity information. The intermediate feature representations $F$ implicitly encode identity-related cues. This is because visual features extracted from human subjects often contain subtle but distinctive patterns, such as body shape, habitual posture, motion dynamics, and appearance-related characteristics. These cues can be unintentionally used by a sufficiently expressive model to distinguish between individuals, even in the absence of explicit facial features. This adversarial mechanism is applicable in scenarios where identity labels or video-based tracking identifiers are available during training. These labels are used solely for regularization purposes to suppress identity-sensitive features and are not involved in any downstream inference or teacher-facing outputs. The adversarial objective is defined as:

$$\mathcal{L}_{adv} = \sum_{i} \mathcal{L}_{CE}\big(D_{id}(\mathrm{GRL}(F_i)), y_{id}\big), \tag{2}$$

where $\mathcal{L}_{CE}$ denotes cross-entropy loss, GRL denotes the gradient reversal layer, and $y_{id}$ denotes the identity label or video-local track identifier used during adversarial training. During forward propagation, the GRL acts as an identity mapping. However, during back-propagation, it reverses the gradient by multiplying it with a negative scalar.

### 3.3. Multi-Task Prediction Module

The system uses a lightweight YOLOv8-C2f (Sapkota et al., 2025) multi-task architecture to capture classroom behavioral cues at multiple levels. A detection head localizes students in the scene. An orientation head estimates body direction as a coarse signal of attentional focus, and a pose head captures fine-grained body movements such as hand-raising or note-taking. Together, these components provide the behavioral basis for temporal and spatial engagement estimation at the classroom level. The multi-task objective is defined as:

$$\mathcal{L}_{task} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{orient}\mathcal{L}_{orient} + \lambda_{pose}\mathcal{L}_{pose}, \tag{3}$$

where $\mathcal{L}_{box}$, $\mathcal{L}_{cls}$, $\mathcal{L}_{orient}$, and $\mathcal{L}_{pose}$ denote the losses for bounding box regression, classification, orientation estimation, and pose regression, respectively. The weighting coefficients $\lambda$ balance the contribution of each task during training.

### 3.4. Output Differential Privacy Module

To support privacy-preserving reporting, the classroom is divided into $Z$ pre-defined zones and engagement is summarized at the group level. For each zone, the final reported engagement score is computed as:

$$y_z = \hat{y}_z + \eta_z, \quad \eta_z \sim \mathrm{Laplace}\left(0, \frac{\Delta f}{\epsilon}\right), \tag{4}$$

where $\hat{y}_z$ is the raw aggregated score and $\eta_z$ is Laplace noise controlled by the privacy budget $\epsilon$. This mechanism reduces the influence of any single student on the released outputs and helps limit the risk of inferring individual attendance or behavior from regional summaries.

### 3.5. Comprehensive Training Objective

The framework is trained end-to-end by jointly optimizing behavioral prediction, zone-level consistency, and privacy-related constraints. The overall objective is:

$$L_{total} = L_{task} + \lambda_{zone}L_{zone} + \lambda_{adv}L_{adv} + \lambda_{attn}L_{attn}. \tag{5}$$

Here, $L_{task}$ is the primary multi-task loss, $L_{zone}$ supervises region-level statistical estimation, $L_{adv}$ penalizes identity-predictive features, and $L_{attn}$ reduces sensitivity to high-frequency facial details. These objectives support engagement analysis while constraining the model's dependence on identity-revealing information.

## 4. System-Generated Visible Behavioral Engagement Analytics Outputs

To evaluate whether the system could generate interpretable engagement signals across temporal scales, we applied it to 21 short-duration classroom video segments from the Dataset for Classroom Group Engagement Recognition (Lu et al., 2025). To further illustrate the system's capacity for longer-duration analysis, we constructed a 13-minute extended instructional sequence by combining a subset of short-duration segments that appeared to focus on the same instructional topic and showed comparable classroom conditions. The short-duration segments were used to inspect finer-grained engagement changes, while the 13-minute extended recording was used to examine broader engagement trends over a longer instructional sequence. The analyzed videos contained an average of 8 visible students per frame, with the detected number varying across frames due to occlusion, camera angle, and student movement. For each video, the system produced teacher-facing outputs, including a class-level engagement timeline, a region-by-time engagement heatmap, and summary indicators such as average engagement, equity Gini coefficient, lowest-engagement region, and largest engagement drop. These outputs were presented through an interactive dashboard (Fig. 1) that supports regional inspection, $2 \times 2$ and $3 \times 3$ grid switching, and time-based review.

Because regional engagement comparisons can be confounded by differences in student density, we normalized engagement at the region level. A region with fewer detected students may produce a lower aggregate score simply because fewer students are present. Therefore, region-level engagement was computed as the average engagement proxy of all detected students within each predefined region at each time step. Each detected student was assigned to a spatial region based on the center point of the bounding box, projected onto a normalized classroom plane. The spatial boundaries for the $2 \times 2$ and $3 \times 3$ grids were calibrated according to the classroom seating layout. This calibration is intended to make the structural capacity and baseline number of seats across regions as comparable as possible. The engagement score for region $z$ at time $t$ is then computed as:

$$\hat{y}_{z,t} = \frac{1}{N_{z,t} + \delta} \sum_{i \in \mathcal{S}_{z,t}} e_{i,t} \tag{5}$$

Here, $\mathcal{S}*z,t$ refers to the set of students detected in region $z$ at time $t$; $N*z,t$ gives the corresponding number of detected students; $e_{i,t}$ captures the individual engagement proxy inferred from posture and pose-based signals; and $\delta = 10^{-5}$ serves as a small stabilization constant to prevent division by zero when a region contains no detected students. This normalization allows dashboard indicators, including the equity Gini coefficient and lowest-region tracking, to reflect average region-level engagement patterns.

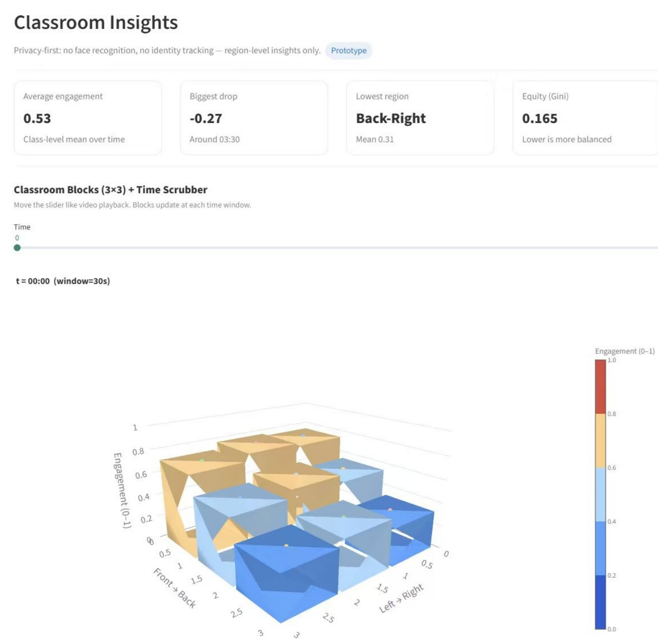
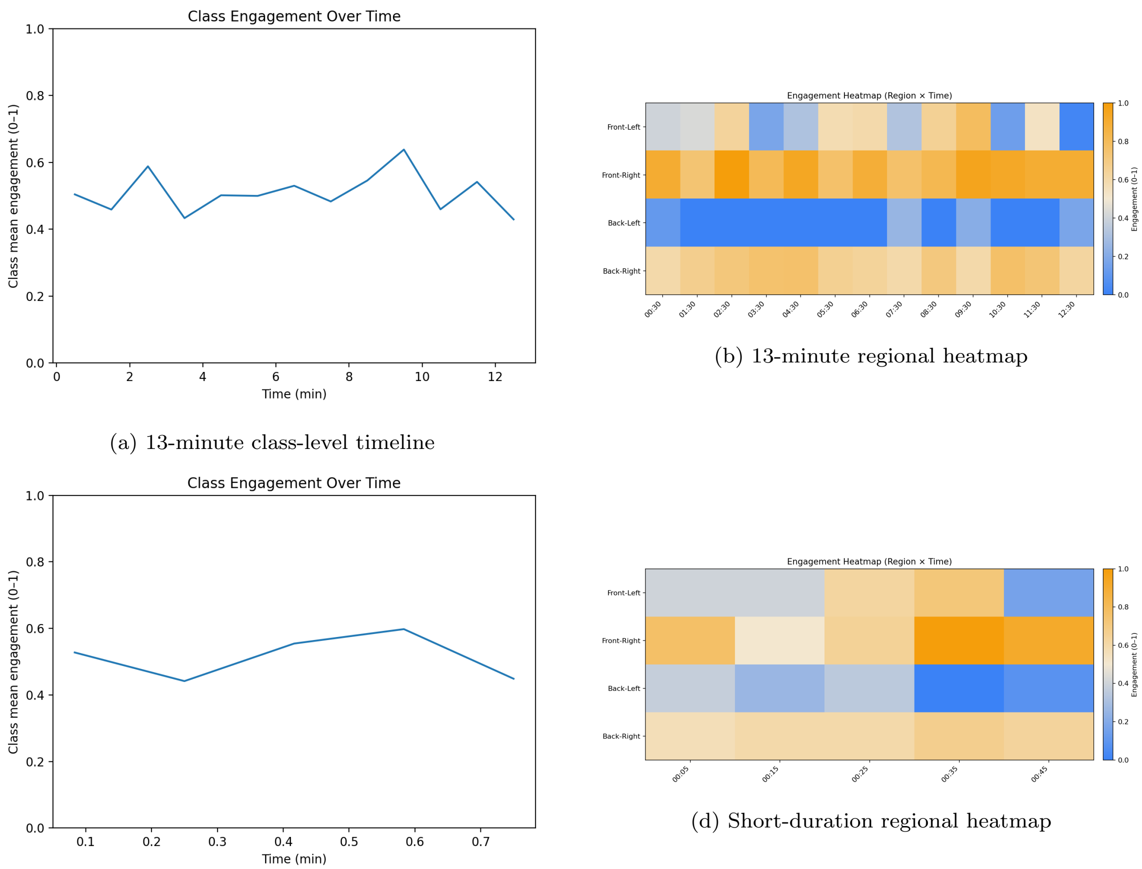



Figure 1: Teacher-facing Classroom Insights dashboard prototype.

Users can upload an extended classroom video to the system to observe broader engagement changes over a longer instructional sequence. As shown in Fig. 2a-b, the 13-minute extended recording and regional heatmap captured broad engagement trends, including drops and recoveries around major classroom transitions. Users can also upload short-duration classroom video segments to inspect finer-grained engagement changes. As shown in Fig. 2c-d, the short-duration timeline and heatmap made transient changes across the $2 \times 2$ classroom grid more visible. Across the 21 evaluated short-duration segments, the system surfaced regional participation gaps that may be difficult for teachers to detect during live instruction. The Back-Left region was identified as the lowest-engagement zone in 16 segments, followed by the Back-Right region in 4 segments and the Front-Left region in 1 segment. These spatial gaps were further captured by the disparity-sensitive indicators. The equity Gini coefficient, computed from normalized regional engagement values to assess the balance of participation across classroom regions, had a mean value of 0.351 across the 21 segments, with a 10th-to-90th percentile range of 0.248 to 0.460. Additionally, engagement in the lowest-performing regions averaged 0.095, with a median of 0.073, indicating that the least engaged areas often showed lower participation than the classroom average.




(a) 13-minute class-level timeline

(b) 13-minute regional heatmap

(c) Short-duration class-level timeline

(d) Short-duration regional heatmap

Figure 2: System-generated engagement analytics at two temporal scales. (a) Class-level engagement timeline for the 13-minute extend recording. (b) Region-by-time engagement heatmap for the 13-minute extend recording. (c) Class-level engagement timeline for a short-duration example. (d) Region-by-time engagement heatmap for a short-duration example.

## 5. Professional Evaluation with Teachers

Four middle school teachers participated in the professional evaluation,including two novice teachers with 1 and 2 years of experience and two experienced teachers with 24 and 30 years of experience. The purpose of this review was to examine whether the generated outputs were understandable, pedagogically meaningful, and aligned with teachers' professional observations. Each teacher reviewed selected classroom video clips together with the corresponding engagement timeline, regional heatmap, and summary indicators. Before seeing the system interpretation, teachers were asked to identify moments in the videos where students appeared quiet, less responsive, off-task, or less visibly engaged. These teacher identified moments were then compared with the system detected engagement drops. We calculated a preliminary alignment rate by treating a system detected drop as aligned when it occurred within a five second window around a teacher identified moment. Specifically, 17 out of 23 teacher-identified moments were matched by the system, resulting in a preliminary alignment rate of 73.9%, which indicates a meaningful level of agreement between the system-generated outputs and teachers' professional observations.

Overall, teachers found the engagement timeline and regional heatmap easy to interpret. For timeline, they viewed that it helped them locate moments that deserved closer review, while still requiring them to consider what was instructing. They also praised the regional heatmap because it made spatial participation patterns more visible. For example, the heatmap revealed whether students in the back of the classroom were consistently less active or whether attention decreased after a change in task structure. Since the back of the classroom is often difficult to monitor during live teaching, the heatmap provided a useful visual cue for noticing participation patterns that may otherwise be missed. Teachers also cautioned that low engagement scores should not be interpreted as direct evidence of disengagement or poor instruction. Quiet reading, individual problem solving, note-taking, or activity transitions may all produce lower visible engagement signals. These comments highlight an important limitation: the system can indicate when and where engagement changes occur, but teachers are still needed to interpret why these changes happen.

## 6. Conclusion and Future Directions

This study set out to develop a privacy-aware approach to classroom video analytics that can generate meaningful engagement insights without assigning engagement judgments to individual students. Building on prior classroom attention analytics that used student localization, body orientation, and pose-based behavioral cues to estimate classroom engagement (Raca and Dillenbourg, 2013; Liu et al., 2025), this work extends that line of research by shifting the analytic focus from individual students to aggregated regional summaries. Inspired by privacy-preserving perspectives in video analytics Wang et al. (2019), the proposed framework transforms student location, body orientation, and pose-based movement signals into region-level engagement timelines, spatial heatmaps, and disparity-sensitive indicators. These teacher-facing outputs provide instructors with a way to reflect on how seating arrangements, task structures, and instructional pacing may shape student participation over time and across different classroom regions.

The current study also points to several directions for improvement. First, future work needs to further examine the privacy-utility tradeoff of the framework. We plan to conduct additional experiments that compare obfuscation strategies and adversarial learning settings to better understand how privacy protection shapes analytic usefulness. Second, the system needs to be evaluated in more authentic classroom contexts. The present evaluation focused on teachers' interpretation of selected video clips and system outputs. During the summer, we plan to conduct an additional classroom observation study in middle school settings. This study will collect further engagement data and examine how teachers use the engagement timeline, regional heatmap, and summary indicators during reflection. We will also investigate how well these outputs align with teachers' classroom observations, what explanations teachers need, and how the interface and indicators can be refined to better support instructional decision-making. This work responds to broader calls for responsible AI in education. As AI systems become more common in classrooms, responsible AI should strengthen teachers' professional noticing, contextual judgment, and equitable responses to students' needs while keeping teachers central in educational decision-making.

## References

Jonathan Bidwell and Henry Fuchs. Classroom analytics: Measuring student engagement with automated gaze tracking. Technical report, University of North Carolina at Chapel Hill, 2011.

Stephanie J. Blackmon and Claire Howell Major. Inclusion or infringement? a systematic research review of students' perspectives on student privacy in technology-enhanced, hybrid and online courses. *British Journal of Educational Technology*, 54(6), July 2023. doi: 10.1111/bjet.13362.

Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi. Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. *IEEE Transactions on Information Forensics and Security*, 17:357–372, 2022.

Qingtang Liu, Xinyu Jiang, and Ruyi Jiang. Classroom behavior recognition using computer vision: A systematic review. *Sensors*, 25(2):373, January 2025. doi: 10.3390/s25020373.

Weigang Lu, Yang Yang, Runfei Song, Ying Chen, and Cunling Bian. A video dataset for classroom group engagement recognition. *Scientific Data*, 12(1):644, April 2025. doi: 10.1038/s41597-025-04987-w.

Mirko Raca. *Camera-Based Estimation of Student's Attention in Class*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2015.

Mirko Raca and Pierre Dillenbourg. System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, pages 265–269. ACM, 2013. doi: 10.1145/2460296.2460351.

Johnmarshall Reeve, Hyungshim Jang, Dan Carrell, Soohyun Jeon, and Jon Barch. Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, 28(2):147–169, June 2004. doi: 10.1023/B:MOEM.0000032312.95499.6f.

Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018. doi: 10.48550/arXiv.1803.11556.

Ranjan Sapkota, Marco Flores-Calero, Rizwan Qureshi, Chetan Badgujar, Upesh Nepal, Alwin Poulose, Peter Zeno, Uday Bhanu Prakash Vaddevolu, Sheheryar Khan, Maged Shoman, et al. Yolo advances to its genesis: A decadal and comprehensive review of the you only look once (yolo) series. *Artificial Intelligence Review*, 58(9):274, 2025. doi: 10.1007/s10462-025-11253-3.

Tuomo Erkki Virtanen, Marja-Kristiina Lerkkanen, Anna-Maija Poikkeus, and Matti Kuorelahti. The relationship between classroom quality and students' engagement in secondary school. *Educational Psychology*, 35(8):963–983, August 2013. doi: 10.1080/01443410.2013.822961.

Han Wang, Shangyu Xie, and Yuan Hong. Videodp: A universal platform for video analytics with differential privacy. *arXiv preprint arXiv:1909.08729*, 2019. doi: 10.48550/arXiv.1909.08729.