Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Fine-Grained Classroom Activity Recognition via Detection, Head-Pose, and Appearance Fusion

Yaya Wihardi[1*], Raffi Ardhi Naufal[2], Meutia Jasmine Annisa Herawan[3], Rani Megasari[4]
[1,2,3,4]Universitas Pendidikan Indonesia, Indonesia
[1*]yayawihardi@upi.edu, [2]ardhiraffi12@upi.edu, [3]meutia.ja@upi.edu, [4]megasari@upi.edu

open access

***Corresponding Author**

Article History:
Submitted: 30-07-2026
Accepted: 10-08-2026
Published: 18-08-2026

Keywords:
classroom analytics; fine-grained classroom activity recognition; head-pose estimation; multi-cue fusion; smartphone use.

ABSTRACT

Automatic analysis of student behavior in classroom environments is an important prerequisite for smart learning systems that move beyond coarse engagement indicators toward behavior-specific and pedagogically interpretable feedback. Fine-grained classroom activities, however, remain difficult to recognize because several target behaviors exhibit similar visual patterns, are partially occluded, or involve subtle head and upper-body motion. This study addresses five classroom activities commonly observed in authentic instructional settings: nodding, hand-raising, smartphone use, head-supporting, and looking downward. We propose a student-centered framework that first constructs person-consistent 16-frame tracklets and then integrates three complementary sources of evidence: RGB appearance from person crops, head-pose estimation, and an explicit smartphone-related cue. This design addresses two major ambiguity clusters in classroom scenes, namely downward-looking versus smartphone use and downward-looking versus head-supporting, while preserving temporal sensitivity for nodding. The dataset comprises labeled student-centered clips extracted from multi-view classroom recordings. Evaluation follows a subject-separated and session-separated protocol using different acquisition sessions, classroom settings, and participant groups. Under this protocol, the proposed framework achieves 91.13% accuracy and 91.14% macro F1. Compared with a visual-only baseline, the full fusion model improves macro F1 by 6.21 percentage points, while outperforming the strongest non-fusion baseline by 2.83 percentage points. The confusion analysis further indicates that head-pose information and explicit smartphone cues effectively separate visually adjacent downward-oriented behaviors. These findings support multi-cue fusion as an effective strategy for fine-grained classroom activity recognition and classroom analytics under the present held-out evaluation protocol, while broader deployment and cross-session generalization should remain bounded by the current evidence.

INTRODUCTION

The growing interest in smart classrooms has expanded the role of educational technology from content delivery to in-situ observation of how instruction and student response unfold in real time. Within this broader shift, classroom video analysis has emerged as a particularly attractive modality because it preserves spatial context, interaction patterns, posture changes, and short-term behavioral responses that are difficult to recover from coarse logs or aggregate engagement scores alone (Dimitriadou & Lanitis, 2023; Hu et al., 2024; Jiang & Fu, 2024). Recent studies have therefore applied computer vision to classroom behavior detection, attention monitoring, engagement analysis, and behavior-aware teaching support (Abozeid et al., 2026; Q. Liu et al., 2024, 2025).

Despite this progress, much of the existing classroom analytics literature still operates at a relatively broad semantic level. A substantial portion of prior work focuses on global states such as engagement, attentiveness, or generic classroom behavior categories (Consoli et al., 2024; Elbawab & Henriques, 2023; Ikram et al., 2023; Kassab et al., 2023; Q. Xu et al., 2023). These formulations are useful for high-level monitoring, but they are less informative when the practical objective is to distinguish concrete classroom activities that carry different behavioral interpretations. In classroom observation, subtle distinctions matter. A student who is looking downward may be reading course material, using a smartphone, or disengaging from the instructional flow. Similarly, head-supporting and transient downward gaze can share a similar global posture while reflecting different momentary states.

This difficulty is amplified by the visual characteristics of real classroom scenes. Multi-student layouts introduce partial occlusion, variation in scale, overlapping body configurations, and non-uniform illumination. Small but meaningful objects, such as smartphones, may be only partially visible. Motion-sensitive behaviors such as nodding may be expressed with short duration and limited amplitude. These factors make it difficult to rely on a single cue family, especially when the target categories are visually adjacent (Algabri et al., 2024; Asperti & Filippini, 2023; J. Liu

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

et al., 2023; Vermander et al., 2024; L. Xu et al., 2025).

Motivated by these limitations, this study focuses on five pedagogically relevant student activities: nodding, hand-raising, smartphone use, head-supporting, and looking downward. These classes were selected because they recur frequently in real instructional settings and collectively span both active response and potentially disengaged behavior. Hand-raising and nodding can indicate participation or acknowledgment, whereas smartphone use, sustained head-supporting, and prolonged downward orientation are often associated with distraction, fatigue, or reduced attentional engagement (Alkabbany et al., 2023; Hu et al., 2024; Ikram et al., 2023). Recognizing these five behaviors offers a more concrete and operationally useful account of classroom dynamics than broad engagement labels alone.

The central premise of this work is that fine-grained classroom behavior recognition should be treated as a multi-cue problem. Student detection is needed to localize individuals within crowded scenes. Head pose estimation is needed to capture orientation changes and posture-related cues that are poorly represented by appearance features alone. Appearance-based classification remains necessary because pose is insufficient to capture object interaction and contextual visual detail. We therefore propose a multi-branch fusion framework that combines detection, head pose, and visual appearance at the final classification stage. This design is consistent with recent developments in classroom behavior recognition, head pose estimation, lightweight detection, and human activity recognition more broadly (Algabri et al., 2024; Edozie et al., 2025; J. Liu et al., 2023; Sheng et al., 2025; J. Zhang et al., 2025; Y. Zhang & Wang, 2025).

This study addresses three gaps in the current literature. First, relatively few classroom studies explicitly formulate the five target activities considered here as one fine-grained recognition problem with clear operational label boundaries (Alruwais & Zakariah, 2024; Q. Liu et al., 2025; Sheng et al., 2025; Trabelsi et al., 2023). Second, prior classroom pipelines do not consistently combine person-centered appearance, head-orientation dynamics, and explicit device-related evidence within a single decision process for subtle activity discrimination. Third, although real classroom scenes are dominated by occlusion, viewpoint variation, and visually adjacent downward-oriented states, these difficulties are often treated as qualitative caveats rather than as the core reason a stricter evaluation protocol is needed.

The main contribution of this study is therefore not a generic combination of existing modules, but a task-specific formulation that uses person-centered tracklets and explicit cue disambiguation for five confusable classroom micro-activities. Concretely, this study contributes three elements. First, it introduces a classroom activity dataset tailored to five operationally defined student behaviors and organized as labeled student-centered clips for training, validation, and testing under separated collection conditions. Second, it proposes a tracklet-level multi-branch fusion framework that integrates student detection, head pose, appearance cues, and an explicit smartphone cue to resolve inter-class ambiguity. Third, it evaluates the contribution of each component through baseline comparisons and ablation analysis under the current subject-separated and session-separated train-validation-test protocol. Taken together, these contributions aim to move classroom video understanding toward more interpretable student behavior analytics under the present held-out classroom benchmark.

LITERATURE REVIEW

Research on classroom behavior analysis has advanced rapidly with the maturation of deep object detection and visual recognition models. Recent classroom-oriented studies have relied heavily on YOLO-style detectors because they offer a strong practical compromise between accuracy, latency, and implementation simplicity in crowded educational scenes (Cao et al., 2025; Chen et al., 2023; Q. Liu et al., 2024; Sheng et al., 2025). Subsequent refinements have incorporated multi-scale feature fusion, attention modules, super-resolution, and efficiency-oriented architectural changes to improve recognition under classroom-specific conditions such as small targets, viewpoint variability, and real-time constraints (Shou et al., 2024; Wang et al., 2023, 2024; H. Zhang et al., 2025; J. Zhang et al., 2025; X. Zhang et al., 2024; Zhou et al., 2025).

At the same time, a parallel line of work has emphasized that student behavior cannot always be inferred reliably from spatial appearance alone. Studies on classroom engagement, attentiveness, and gaze-aware analysis indicate that head orientation, posture, and motion dynamics provide important complementary evidence. This observation is consistent with the broader literature on head pose estimation and posture analysis, where yaw, pitch, roll, and upper-body configuration are treated as informative signals for interpreting attention-related and posture-related states (Algabri et al., 2024; Asperti & Filippini, 2023; Ding et al., 2024; Nadeem et al., 2024; Vermander et al., 2024).

Beyond the classroom domain, the wider human activity recognition literature has increasingly moved toward architectures that preserve both spatial and temporal dependencies through attention mechanisms, transformers, graph-based modeling, and multimodal feature fusion (Elnady & Abdelmunim, 2025; Lamaakal et al., 2025; J. Liu et al., 2023; Qureshi et al., 2025; Y. Zhang & Wang, 2025). These developments are especially relevant when the target behaviors are subtle, partially overlapping, or only weakly separable in single frames. For the present problem, nodding is inherently temporal, smartphone use benefits from explicit object cues, and head-supporting depends on the relation between head orientation and arm placement. A fusion-based design is therefore well aligned with the methodological direction of the field.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Efficiency remains another important consideration. Classroom analytics systems are valuable only to the extent that they remain operational under real deployment constraints, which often include moderate hardware budgets and near-real-time processing requirements. Recent reviews on lightweight detection models and classroom behavior systems have highlighted the need to balance discriminative performance against computational cost, particularly when multiple students appear simultaneously in each frame (Edozie et al., 2025; Q. Liu et al., 2024; Wang et al., 2024).

Taken together, the literature suggests three clear implications for the present work. First, YOLO-based detection remains a reasonable starting point for student localization. Second, head pose and posture-aware signals are likely to be important for distinguishing downward-facing or support-related behaviors. Third, fine-grained classroom activity recognition is better approached through integrated multi-cue reasoning than through visual-only classification or direct one-stage detection alone (Algabri et al., 2024; Edozie et al., 2025; J. Liu et al., 2023; Wang et al., 2024). These observations motivate the proposed architecture.

The closest prior studies generally fall into three groups: direct classroom behavior detectors, engagement-oriented or attentiveness-oriented formulations, and pose-aware or attention-enhanced recognition pipelines. The proposed approach differs from these lines in three respects. Unlike one-stage classroom detectors that directly predict behavior labels from full-frame spatial evidence, the proposed pipeline first forms person-consistent tracklets and then performs tracklet-level classification, thereby preserving identity-consistent temporal and contextual cues for each student. Unlike broad engagement formulations, it focuses on five operationally defined micro-activities with distinct behavioral semantics and explicit label boundaries, rather than collapsing them into wider attentional categories. Unlike pose-aware or attention-enhanced classroom models that emphasize visual or spatial refinement alone, it explicitly treats smartphone-related evidence as a disambiguation cue for downward-oriented behaviors that would otherwise remain confusable under pose and appearance information alone (Q. Liu et al., 2024; Sheng et al., 2025; Wang et al., 2024; J. Zhang et al., 2025). The intended contribution should therefore be understood not as a claim of inventing new standalone components, but as a more tightly specified problem formulation and cue-integration strategy for fine-grained classroom behavior recognition under a stricter held-out evaluation design.

METHOD

Research Design

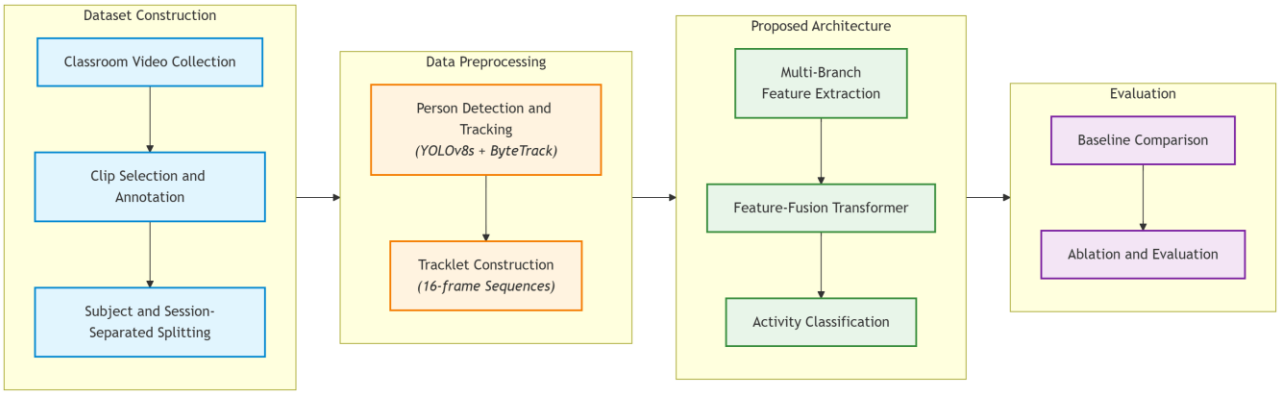

Figure 1. Methodological workflow for fine-grained classroom activity recognition, covering dataset construction, data preprocessing, model development, baseline comparison, and evaluation.

This study follows a quantitative experimental design for fine-grained student activity recognition from classroom video. The methodological objective is to examine whether combining person detection, appearance representation, head-pose dynamics, and smartphone-related cues improves discrimination among visually similar classroom behaviors. The full pipeline consists of dataset construction, data preprocessing, model development, and evaluation. The design therefore treats the problem as a controlled recognition study in which each methodological component is addresses a specific source of ambiguity observed in classroom scenes, including viewpoint variation, partial occlusion, and overlap between downward-oriented behaviors (Algabri et al., 2024; Hu et al., 2024; J. Liu et al., 2023; Sheng et al., 2025; Y. Zhang & Wang, 2025). To clarify this workflow, the main stages of the study are summarized in Figure 1, which organizes the method as a staged recognition pipeline rather than a single-step full-frame prediction scheme.

Dataset

The dataset used in this study was constructed specifically for fine-grained classroom activity recognition and consists of short labeled classroom-video clips representing five target activities: nodding, hand-raising, smartphone use, head-supporting, and looking downward. The data are organized into training, validation, and test partitions with split-specific metadata files. To support a stricter held-out evaluation protocol, the training, validation, and test

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

partitions were drawn from different acquisition sessions, different classroom settings, and different participant groups. Across the full split design, the data were therefore separated by collection time, location, and participant identity rather than being redistributed from the same classroom event. The recordings also include multiple camera viewpoints, including side views and CCTV-style views.

The dataset contains 2,160 labeled clips in total, comprising 1,757 training clips, 200 validation clips, and 203 test clips. Across the three acquisition sessions, the dataset involves 41 unique students. Annotation was performed by three annotators using the same operational labeling protocol. The class distribution is relatively balanced across the five behaviors, as summarized in Table 1.

Table 1. Summary of the classroom activity dataset and split composition

Item	Train	Validation	Test	Total
Nodding	365	40	40	445
Hand-raising	367	38	39	444
Smartphone use	337	42	43	422
Head-supporting	359	42	43	444
Looking downward	329	38	38	405
Total clips	1,757	200	203	2,160

Table 1 highlights the fixed split composition used throughout the study. The training, validation, and test partitions come from different acquisition sessions with different participants, locations, and collection times. No participant overlap was allowed across splits, and clips were not randomly redistributed from the same classroom event into multiple partitions. The train, validation, and test memberships were fixed by predefined dataset partitions and their accompanying metadata files and were kept unchanged throughout model development and comparison. The benchmark should therefore be interpreted as a subject-separated and session-separated train-validation-test protocol, not as a conventional random clip-level split. Although some CCTV-style camera placements appear visually similar across sessions because they reflect a similar classroom monitoring configuration, the underlying recordings still come from independent acquisition settings. This design reduces leakage between splits and provides a more demanding test of generalization than mixing clips from a single acquisition context.

Figure 2 provides representative examples of the five target classes and illustrates the visual similarity between some downward-oriented behaviors. Clip-level labels were assigned using operational definitions fixed before model training. Nodding was defined as visible vertical head motion sustained for at least 6 consecutive frames. Hand-raising was defined as one or both hands lifted above shoulder level. Smartphone use was assigned when a phone was directly visible or when hand posture and gaze direction jointly supported a stable phone-interaction interpretation. Head-supporting referred to sustained cases in which the head rested on the hand or forearm. Looking downward was defined as sustained downward head orientation without dominant physical support from the hand or arm.

To improve annotation consistency, dominant labels were assigned only to intervals in which the behavior was visually stable enough to support confident class assignment. Because several classes partially overlap in pose and gaze pattern, annotation followed an explicit precedence rule rather than unconstrained visual judgment. Hand-raising took precedence whenever one or both hands were clearly elevated above shoulder level. Smartphone use required either visible device evidence or a stable interaction pattern in which hand configuration and gaze direction jointly favored phone use over generic downward attention. Head-supporting was assigned only when sustained physical support from the hand or forearm was visually dominant. Looking downward served as the residual downward-orientation class only when neither smartphone interaction nor dominant head support was sufficiently supported by the visual evidence. Nodding was assigned only when the minimum motion criterion was satisfied over the required temporal interval. This rule-based annotation hierarchy was applied by three annotators under a shared operational protocol and is important because it makes the semantic boundary between the most confusable classes explicit rather than leaving it to ad hoc frame-by-frame judgment.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Figure 2. Representative samples from the classroom activity dataset for the five target classes. From left to right and top to bottom, the panels correspond to nodding, hand-raising, smartphone use, head-supporting, and looking downward.

Ambiguous transitions, severe occlusion, and intervals in which the target activity could not be interpreted reliably were excluded from the final tracklet set. This decision favors label precision over raw sample count and is particularly important for separating looking downward, smartphone use, and head-supporting. In borderline cases, such as partial head support during downward gaze, temporary hand-to-face contact without stable support, or suspected smartphone interaction without adequate object or posture evidence, samples were excluded rather than forced into a noisy class. Consequently, the remaining confusions are more likely to reflect genuine model difficulty than avoidable annotation inconsistency.

Data Preprocessing

After annotation, each clip was processed in three stages. First, student instances were localized in each frame to produce person-centered crops. Second, framewise detections were associated across time to form short identity-consistent tracklets of 16 frames, which served as the fixed-length prediction unit throughout training and evaluation. Third, auxiliary signals were extracted from each tracklet, including head-pose estimates and upper-body keypoints. This pipeline reflects the broader methodological trend of combining detection, tracking, pose estimation, and sequence-aware recognition for behavior analysis in complex visual scenes (Algabri et al., 2024; Ding et al., 2024; Elnady & Abdelmunim, 2025; J. Liu et al., 2023; Sharma et al., 2024).

Input crops for the appearance branch were resized to 224 x 224 pixels. Spatial augmentation consisted of horizontal flipping, random brightness-contrast adjustment, mild Gaussian noise, and random scaling. Lightweight temporal augmentation was applied through frame dropping and temporal position jitter, meaning that frame positions within the 16-frame window were slightly perturbed during training to reduce sensitivity to imperfect tracking, short-term occlusion, and missing-frame effects. These augmentations were intended to improve robustness at the tracklet level rather than to simulate entirely new viewing conditions.

Each tracklet, rather than each frame, is treated as the basic unit of classification. For a tracklet $T = \{x_1, x_2, \ldots, x_{16}\}$, the appearance branch extracts RGB features from person crops, the pose branch estimates framewise yaw, pitch, and roll, and the object branch records smartphone-related evidence observed within the same temporal interval. A single activity label is predicted for the full tracklet. This formulation reduces frame-level instability, limits sensitivity to isolated noisy detections, and better matches the temporal structure of behaviors such as nodding and sustained downward orientation.

Upper-body keypoints are used as auxiliary preprocessing signals rather than as a standalone prediction branch. Their main role is to support region interpretation around the head, shoulder, elbow, and wrist configuration during tracklet construction, smartphone-cue generation, and annotation review, especially for hand-raising and head-supporting. The final classifier therefore uses RGB appearance, head-pose dynamics, and smartphone-related evidence as its principal learned inputs, while keypoints remain a supporting signal for sample preparation and ambiguity screening.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Proposed Architecture

The proposed framework is a student-centric, tracklet-level multi-branch classifier composed of person localization, short-horizon tracklet construction, appearance encoding, head-pose encoding, and smartphone-related cue integration. The architecture is motivated by the observation that the five target classes are not separable by a single cue family alone. In particular, hand-raising is strongly posture-dominant, smartphone use depends on object-level evidence, and nodding requires short-term motion sensitivity rather than purely static appearance. A late-fusion design is therefore more appropriate than direct full-frame behavior prediction because it preserves person-level attribution while allowing complementary cues to contribute only when they are behaviorally informative.

Student localization and tracklet construction. A YOLOv8s detector restricted to the person class localizes student instances in each frame, after which temporally adjacent detections are associated by ByteTrack to form identity-consistent 16-frame tracklets (Jocher et al., 2026; Y. Zhang et al., 2022). This design choice is motivated by the practical need to preserve person-level localization accuracy while maintaining computational feasibility in crowded classroom scenes (Cao et al., 2025; Chen et al., 2023; Q. Liu et al., 2024; Sheng et al., 2025; Wang et al., 2024). It also prevents the downstream classifier from mixing cues from different students in the same classroom view.

Appearance branch. An EfficientNet-B2 encoder extracts visual embeddings from person-centered crops (Tan & Le, 2019). This branch captures clothing configuration, arm posture, hand visibility, body support configuration, and other appearance cues that remain informative even when motion cues are weak or partially occluded.

Head-pose branch. This branch uses SixDRepNet to estimate yaw, pitch, and roll and then converts their framewise sequence into a compact tracklet descriptor that captures both static orientation and short-term directional change across the 16-frame window (Hempel et al., 2022). The objective is to capture orientation-sensitive evidence that is especially relevant for nodding, head-supporting, and looking downward (Algabri et al., 2024; Asperti & Filippini, 2023; L. Xu et al., 2025). In methodological terms, this branch supplies a structured motion-and-orientation descriptor that is more behavior-specific than RGB appearance alone.

Smartphone cue branch. A person-level smartphone-related cue is computed for the same temporal interval as the target tracklet. Smartphone evidence is obtained from a YOLOv8-based detector using the cellphone class, while wrist or hand coordinates are derived from the upper-body keypoint signal (Ding et al., 2024; Jocher et al., 2026). For a detected phone instance, the overlap relation between the wrist/hand region and the phone bounding box is evaluated using Intersection over Union (IoU); when the IoU reaches at least 10%, the binary cue is set to $\mathrm{PhoneCue} = 1$. This threshold should be interpreted as a permissive contact cue rather than as a precise localization criterion, because the purpose of this branch is to preserve plausible phone-interaction evidence under partial visibility. This branch is intended to preserve explicit evidence for device interaction when downward gaze alone is insufficient to separate smartphone use from other downward-facing behaviors. It therefore functions as a disambiguation cue rather than as the sole basis for the class decision.

Let f_v denote the appearance embedding, f_p the head-pose descriptor, and f_s the smartphone-related cue. These person-level inputs are fused at the tracklet level through a Feature-Fusion Transformer (Lamaakal et al., 2025), yielding the representation $\hat{f}$, which is then passed to the final classification head. In other words, the fusion module does not simply concatenate heterogeneous cues; it reweights appearance, pose, and smartphone-related evidence according to their relevance for the current tracklet before the classifier outputs one of the five target activity classes. Throughout the manuscript, this operation is referred to as tracklet-level feature fusion. Its purpose is to prevent any single cue family from dominating the decision in all cases while still allowing pose and object evidence to become decisive when the visual appearance alone is ambiguous.

The head-pose branch captures both static orientation and short-term change. Rather than relying on a single-frame estimate, it summarizes the pose sequence over the full tracklet through descriptors derived from yaw, pitch, and roll trajectories. This design is particularly important for nodding, for which the discriminative signal lies in short-term directional change rather than in a single static posture. More broadly, it aligns the model with state-of-the-art activity-recognition practice in which temporally structured auxiliary cues are used to stabilize decisions for subtle behaviors rather than treating each frame as an independent classification event (Elnady & Abdelmunim, 2025; J. Liu et al., 2023; Qureshi et al., 2025; Y. Zhang & Wang, 2025).

The Feature-Fusion Transformer serves as the final integration module rather than a standalone visual backbone. Its role is to model cross-cue dependencies among appearance, head-pose, and smartphone-related evidence before the final decision layer is applied. In practice, the fusion stage produces one tracklet-level representation that is then mapped to a five-class activity decision. This is methodologically important because the most difficult classes in the present study are not ambiguous within one cue family alone; instead, they are disambiguated by how multiple cues reinforce or contradict each other at the tracklet level.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

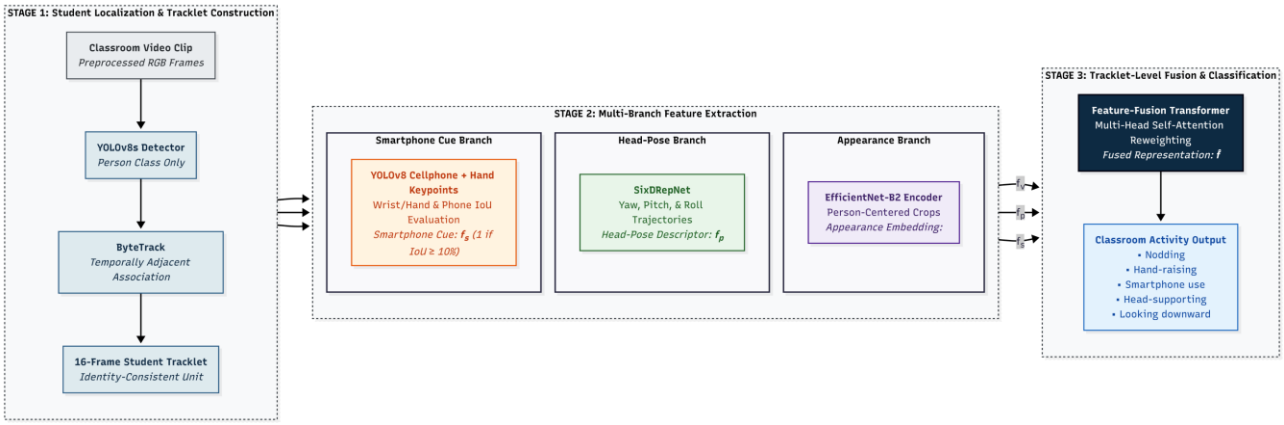

Figure 3. Architecture of the proposed student-centered recognition pipeline.

The stage-wise organization of the proposed architecture is illustrated in Figure 3, which traces the pipeline from person detection and tracklet construction to multi-branch feature extraction, feature fusion, and final activity classification. As shown in the figure, the framework first localizes individual students, then forms identity-consistent tracklets, extracts complementary appearance, head-pose, and smartphone-related cues, and finally integrates these cues at the tracklet level for five-class activity prediction. This explicit stage-wise structure clarifies how person-level localization, cue extraction, and feature fusion interact within the proposed recognition pipeline.

The smartphone cue is integrated at the person level rather than at the full-frame level. This reduces the risk that detections elsewhere in the classroom will influence the label of the current student track and makes the cue more directly attributable to the target person. Because the cue combines cellphone detection with wrist-to-phone overlap consistency, it is also less likely to overfire on nearby devices that do not belong to the target student. The resulting pipeline is therefore behavior-centric at the level of the individual student, not merely scene-centric at the level of the classroom image.

Model training was framed as five-class activity classification under a shared train/validation/test protocol. Optimization was performed with AdamW (Loshchilov & Hutter, n.d.) using an initial learning rate of 1×10^{-4} and a batch size of 32. The tracklet length was fixed at 16 frames for the proposed model, and the same temporal extent was used for all baselines requiring sequence input in order to prevent performance differences from arising solely from unequal temporal context. Hyperparameter selection and early stopping were performed on the validation split only, with validation macro F1 used as the principal model-selection signal. This shared stopping rule is important because it keeps model selection aligned with the class-balanced metric emphasized in the final evaluation. All experiments were executed on a workstation equipped with an NVIDIA RTX 5090 GPU, an Intel Core i9 CPU, and 128 GB RAM.

Baselines

Three baselines are used for comparison. Baseline A: a visual-only classifier based on EfficientNet-B2 without head-pose or smartphone-cue input. Baseline B: a one-stage YOLOv8-based direct behavior detector representing the dominant family of YOLO-style classroom behavior detection pipelines (Cao et al., 2025; Chen et al., 2023; Q. Liu et al., 2024; Sheng et al., 2025; Wang et al., 2024). Baseline C: a CNN-LSTM visual-temporal model without pose-aware fusion, representing a conventional sequence-based recognition formulation (Elnady & Abdelmunim, 2025; Qureshi et al., 2025; Y. Zhang & Wang, 2025).

These baselines were selected to represent three common design families: appearance-only recognition, direct detection-based classification, and explicit temporal modeling without pose-aware fusion. The goal was not to claim exhaustive superiority over every possible classroom-recognition architecture, but to test whether the proposed combination of person-centered appearance, head-pose dynamics, and smartphone evidence adds measurable value over plausible alternative formulations under the same benchmark.

To preserve comparison fairness, all baselines were trained and evaluated on the same subject-separated and session-separated split and under the same label definitions as the proposed model. The baselines were chosen as representative alternatives rather than weak strawman models, so that the comparison isolates the value of multi-cue fusion rather than differences in split design or class taxonomy.

The comparison protocol also follows a common validation-driven tuning principle across methods. Each model is tuned on the same validation split, uses the same stopping principle, and is interpreted under the same class taxonomy. Although architecture-specific implementation details still differ across model families, this protocol makes the comparison substantially more defensible than using heterogeneous splits, inconsistent labels, or unequal evaluation criteria. In particular, the direct detector baseline tests whether one-stage classroom behavior detection is already sufficient, the visual-only baseline tests whether RGB appearance alone is adequate, and the CNN-LSTM baseline tests

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

whether temporal modeling without explicit pose-aware disambiguation can close the remaining gap.

Evaluation Metrics

Performance is evaluated using overall accuracy, per-class precision, recall, and F1-score, together with macro F1-score as the principal class-balanced summary metric. Macro F1 is emphasized because the class frequencies are only approximately balanced and because the recognition difficulty differs across behaviors. Unless otherwise stated, all reported metrics are computed at the tracklet level, which is the prediction unit of the proposed framework. Validation data are used for model selection and early stopping, whereas the test partition is reserved for final reporting only. Inference speed is reported in frames per second (FPS) as an empirical runtime indicator rather than as a hardware-independent deployment guarantee. Repeated-run variance and confidence intervals were not evaluated in the present study and are therefore not claimed.

RESULT

Main Results

Table 2 presents the quantitative comparison between the proposed framework and the three baseline models in terms of accuracy, macro F1, and inference speed. The proposed model achieves the strongest recognition performance, reaching 91.13% accuracy and 91.14% macro F1, whereas the lighter baselines retain a throughput advantage. Among the baseline methods, CNN-LSTM provides the closest competitive result, with 88.18% accuracy and 88.31% macro F1, but it still remains below the proposed framework across both principal recognition metrics. Relative to this strongest non-fusion baseline, the proposed model improves accuracy by 2.95 percentage points and macro F1 by 2.83 percentage points. The gain is especially meaningful in macro F1, which is the more informative metric here because the five target classes are only approximately balanced and differ in visual difficulty. This result suggests a more stable decision boundary across the full label space rather than improvement concentrated only in dominant or easier categories.

Table 2. Quantitative comparison between the proposed framework and the baseline models

Model	Accuracy (%)	Macro F1 (%)	FPS
Baseline A: EfficientNet-B2	84.73	84.93	320.0
Baseline B: YOLOv8 direct behavior detector	86.21	86.42	210.0
Baseline C: CNN-LSTM	88.18	88.31	105.0
Proposed model	91.13	91.14	85.0

Table 2 also clarifies the role of architectural design choices. The lighter baselines remain faster, but their recognition capacity is more limited for visually adjacent activities under the current benchmark. The CNN-LSTM baseline benefits from temporal modeling, yet it still lags behind the proposed approach, suggesting that temporal information alone is insufficient when head orientation and object-level evidence are not explicitly modeled. The performance advantage of the proposed framework is therefore more plausibly associated with complementary cue integration than with merely choosing a single stronger backbone. At the same time, the results should be interpreted as evidence of comparative advantage within the selected benchmark and baseline set, not as proof of universal superiority across all classroom settings. Figure 4 provides a compact visual summary of this trade-off between recognition quality and throughput.

Figure 4 shows that the proposed model occupies the strongest overall position in accuracy and macro F1, whereas the lighter baselines retain a throughput advantage. This visual pattern is consistent with Table 2 and reinforces the interpretation that the main contribution of the proposed framework lies in improved recognition quality rather than in raw inference speed.

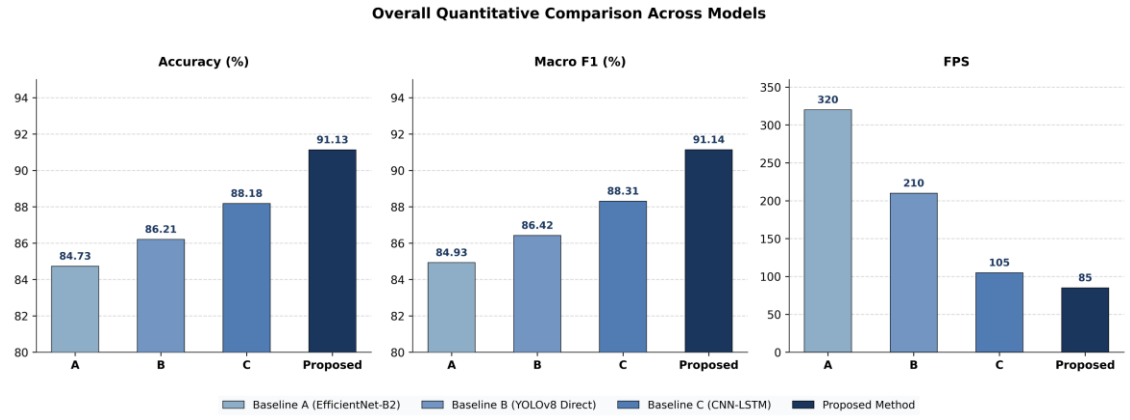

Figure 4. Visual summary of the overall comparison in Table 2

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Per-Class Performance

Table 3 presents the class-wise precision, recall, and F1-scores of the proposed model across the five target activities. Hand-raising achieves the strongest class-wise result, with an F1-score of 96.10%, followed by smartphone use at 93.02% and head-supporting at 92.86%. By contrast, looking downward remains the most difficult category, with an F1-score of 83.95%, while nodding is also comparatively challenging at 89.74%. This pattern is consistent with the task structure: looking downward functions as a residual downward-orientation class after smartphone use and head-supporting have been separated by stronger evidence, whereas nodding depends on short-horizon temporal variation that is inherently harder to preserve than static posture cues.

Table 3. Class-wise precision, recall, and F1-score of the proposed model

Class	Precision (%)	Recall (%)	F1-score (%)
Nodding	92.11	87.50	89.74
Hand-raising	97.37	94.87	96.10
Smartphone use	93.02	93.02	93.02
Head-supporting	95.12	90.70	92.86
Looking downward	79.07	89.47	83.95

Table 3 also helps explain why a multi-cue formulation is beneficial. Smartphone use and looking downward share a similar downward head orientation, but differ in the presence of an explicit object interaction cue. Head-supporting also overlaps with other seated downward-facing states, but its class-wise performance remains comparatively strong when explicit pose and support cues are available. These class-wise results remain broadly consistent with the design rationale of the proposed model and help explain why the complete fusion configuration in Table 4 yields the highest aggregate performance. Importantly, the pattern is evidence-based rather than rhetorical: the model is not equally strong on all classes, but it is comparatively strongest where additional cue disambiguation is most needed.

Confusion Analysis

The confusion pattern, visualized in Figure 5, indicates that the single largest error mode is nodding being classified as looking downward. This is structurally plausible because both behaviors can share a downward-oriented head posture in individual frames, while nodding additionally requires short-term temporal evidence that may be weakened by limited motion amplitude, partial occlusion, or imperfect tracklet sampling. A second cluster of errors involves downward-oriented classes being reassigned to nearby alternatives, including smartphone use being predicted as looking downward and looking downward being predicted as smartphone use. The matrix therefore supports a restrained interpretation: the proposed framework reduces, but does not eliminate, ambiguity among downward-oriented classes.

This confusion matrix further indicates that the strongest residual ambiguity is concentrated in downward-oriented behaviors rather than in head-supporting versus looking-downward alone. The pattern is especially consistent with two ambiguity pairs: nodding versus looking downward, and smartphone use versus looking downward. By contrast, hand-raising remains rarely confused with the other categories because its body configuration is spatially salient and less dependent on subtle orientation changes.

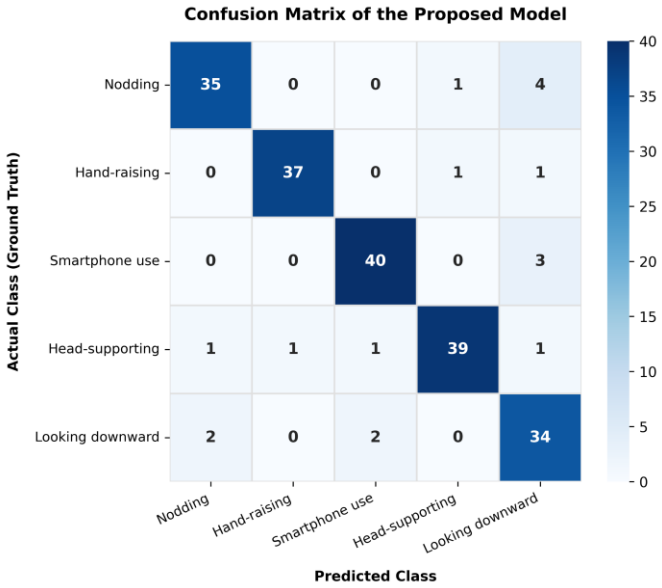

Figure 5. Confusion matrix of the proposed model on the test set

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Figure 5 shows that the most concentrated error region lies between nodding and looking downward, followed by confusion between smartphone use and looking downward. This visual pattern is consistent with the discussion above and confirms that the strongest residual errors remain concentrated in the downward-oriented classes rather than being uniformly distributed across all categories.

Ablation Study

Table 4 presents the ablation results of the proposed framework in terms of accuracy and macro F1 across five configuration settings. The addition of head-pose information produces the largest intermediate gain, increasing performance from 84.73% accuracy and 84.93% macro F1 in the visual-only setting to 89.90% accuracy and 88.80% macro F1. By contrast, adding the smartphone cue alone yields a smaller improvement, from 84.93% to 85.40% macro F1. The complete configuration, which combines visual input, head pose, smartphone cue, and attention-based fusion, achieves the best overall result at 91.13% accuracy and 91.14% macro F1. Under the present benchmark, this pattern indicates that pose-sensitive information contributes more broadly across the label space than the phone cue alone.

The smartphone cue plays a complementary role. Its contribution is smaller at the aggregate level, but it addresses a specific failure mode of downward-facing classes. The final gain introduced by attention-based fusion further suggests that feature integration should not be treated as a simple concatenation problem; the model benefits when it can adaptively reweight cue importance depending on the observed behavioral pattern.

Table 4. Ablation results across the proposed configuration variants

Configuration	Accuracy (%)	Macro F1 (%)
Visual only	84.73	84.93
Visual + phone cue	86.20	85.40
Visual + head pose	89.90	88.80
Visual + head pose + phone cue	91.00	89.80
Visual + head pose + phone cue + attention-based fusion	91.13	91.14

Table 4 also helps explain the role of temporal reasoning in the current design. The gain from adding head-pose dynamics suggests that even a lightweight temporal representation contributes meaningfully to the recognition of motion-sensitive classes. At the same time, the remaining difficulty of nodding indicates that the 16-frame formulation should be viewed as a computationally efficient compromise rather than as the strongest possible temporal encoder. The reported results support the current design as effective under the present benchmark, while still leaving room for additional gains from stronger sequence modeling or tracklet-length sensitivity analysis. Figure 6 visually consolidates the trends from Tables 3 and 4 and makes the relative class difficulty and ablation trajectory easier to inspect.

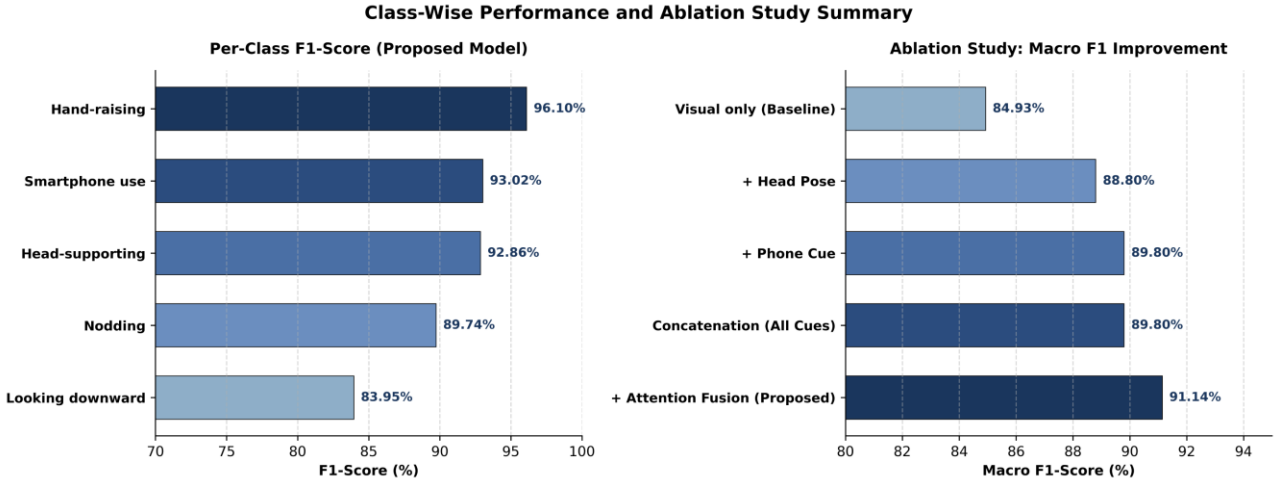

Figure 6. Quantitative visual summary derived from Tables 3 and 4

Figure 6 shows that hand-raising remains the easiest class, whereas looking downward is the most difficult, and it also illustrates the steady performance gain obtained as additional cues and fusion components are incorporated. This visual pattern is consistent with Tables 3 and 4 and reinforces the interpretation that head-pose information and feature-fusion design are the main contributors to the final performance improvement.

Real-Time Analysis

Under the reported workstation configuration, the proposed framework operates at 85 FPS, which places it within a throughput range compatible with real-time use under the reported hardware. For classroom monitoring, the most useful model is not necessarily the fastest one, but the one that preserves sufficient inference speed while

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

improving discrimination for the most behaviorally relevant and visually ambiguous classes.

The reported throughput values should be interpreted as empirical feasibility indicators rather than as universal deployment guarantees. In the present study, runtime measurements were obtained on a system equipped with an NVIDIA RTX 5090 GPU, an Intel Core i9 processor, and 128 GB RAM. End-to-end speed will still depend on tracking stability, crop quality, hardware configuration, and the consistency with which small-object and pose cues remain detectable.

DISCUSSION

The findings reinforce a central argument of this study: fine-grained student activity recognition in classroom video is fundamentally a multi-cue inference problem. Appearance cues alone are insufficient when multiple activities share similar global posture. This is particularly evident for nodding, looking downward, and smartphone-related downward-facing states, where reliable interpretation depends on orientation dynamics and part-level spatial relations rather than on coarse RGB appearance alone. The concentration of residual errors in these downward-oriented classes, as seen in Table 3 and Figure 6, further supports this interpretation, showing that they are the least separable without explicit temporal and object-level cues. In this sense, the proposed formulation is well aligned with prior work that positions head pose, posture, and structured body information as complementary signals for behavior interpretation (Algabri et al., 2024; J. Liu et al., 2023; Nadeem et al., 2024; Vermander et al., 2024).

The improvement associated with the smartphone detector supports a second methodological point: explicit object cues matter when the behavioral label depends on interaction with a small target. A downward head pose does not uniquely imply phone use, and inferring device interaction from gaze direction alone is likely to be brittle under occlusion. Introducing an auxiliary detector therefore improves interpretability as well as recognition logic because the model is no longer forced to treat all downward-facing states as variants of the same visual pattern. The ablation trend in Table 4 is consistent with this reading: the phone cue produces a smaller gain than head pose at the aggregate level, but it contributes to resolving a specific and behaviorally important ambiguity cluster.

More broadly, the study is consistent with an emerging direction in classroom analytics toward integrated pipelines that combine efficient detection with richer semantic reasoning. Rather than treating classroom monitoring as a purely coarse engagement problem, the proposed formulation emphasizes operationally defined behaviors that can be linked more directly to classroom observation and intervention. This shift is important for smart learning environments in which interpretability and actionable granularity may be more useful than a single generic attentiveness score.

Several limitations should nevertheless be acknowledged. First, recognition quality remains sensitive to tracking errors under severe occlusion or overlapping students. Second, downward-oriented classes, especially looking downward, temporally subtle nodding instances, and phone-related downward attention, still pose the greatest challenge because they require stronger temporal sensitivity and finer cue integration than the current lightweight fusion design may provide. Third, the class distribution is not perfectly balanced, so some residual bias may persist even when macro F1 is used as the primary class-balanced metric. Fourth, although the split design is stronger because train, validation, and test come from independent sessions, broader claims about robustness would still benefit from additional held-out sessions, larger participant diversity, and cross-institution validation. Fifth, reproducibility would be further strengthened by public release of the code, complete hyperparameter tables, and implementation checklists. Sixth, additional robustness analyses such as occlusion-stratified evaluation, cross-view performance comparison, tracklet-length sensitivity, and failure-case visualization would further strengthen the empirical account of where the model succeeds and where it remains brittle. These limitations point to clear opportunities for future work on stronger temporal modeling, improved tracking robustness, and broader cross-classroom validation.

CONCLUSION

This study presents a multi-branch fusion framework for fine-grained recognition of five student activities in classroom video: nodding, hand-raising, smartphone use, head-supporting, and looking downward. By integrating student detection, head-pose information, and appearance cues, the framework is designed to address behaviors that are visually similar yet behaviorally distinct. Within the present held-out evaluation protocol, the results indicate that explicit multi-cue modeling is better suited to this problem than relying exclusively on direct detection or appearance-only classification.

The proposed formulation outperforms the selected baselines and is particularly beneficial for classes dominated by subtle head orientation or small-object interaction. Under the present benchmark, the complete fusion model improves macro F1 from 88.31% to 91.14% relative to the strongest non-fusion baseline and from 84.93% to 91.14% relative to the visual-only baseline. From an application-oriented perspective, the study points toward a more interpretable form of smart classroom analytics that can move beyond global engagement labels toward concrete behavior categories. From a methodological perspective, it highlights the value of combining pose-aware reasoning and object-level evidence under a subject-separated and session-separated train-to-evaluation benchmark. These conclusions should be read as benchmark-supported findings rather than as definitive evidence of full deployment robustness across broader classroom domains.

E-ISSN : 2807-9035
Volume 6, **Number** 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Future work can extend this direction through richer temporal modeling, self-supervised pretraining, occlusion-aware evaluation, tracklet-length sensitivity analysis, and cross-classroom or cross-institution evaluation to strengthen generalization under broader deployment settings.

REFERENCES

Abozeid, A., Alrashdi, I., & Al-Makhlasawy, R. M. (2026). Intelligent recognition of students' behavior for smart learning environments. *Scientific Reports 2026 16:1*, *16*(1), 5674-. https://doi.org/10.1038/s41598-026-36633-9

Algabri, R., Abdu, A., & Lee, S. (2024). Deep learning and machine learning techniques for head pose estimation: a survey. *Artificial Intelligence Review 2024 57:10*, *57*(10), 288-. https://doi.org/10.1007/S10462-024-10936-7

Alkabbany, I., Ali, A. M., Foreman, C., Tretter, T., Hindy, N., & Farag, A. (2023). An Experimental Platform for Real-Time Students Engagement Measurements from Video in STEM Classrooms. *Sensors 2023, Vol. 23, Page 1614*, *23*(3), 1614. https://doi.org/10.3390/S23031614

Alruwais, N. M., & Zakariah, M. (2024). Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education. *IEEE Access*, *12*, 66110–66128. https://doi.org/10.1109/ACCESS.2024.3354981

Asperti, A., & Filippini, D. (2023). Deep Learning for Head Pose Estimation: A Survey. *SN Computer Science*, *4*(4), 349-. https://doi.org/10.1007/S42979-023-01796-Z/FIGURES/17

Cao, Y., Cao, Q., Qian, C., & Chen, D. (2025). YOLO-AMM: A Real-Time Classroom Behavior Detection Algorithm Based on Multi-Dimensional Feature Optimization. *Sensors 2025, Vol. 25, Page 1142*, *25*(4), 1142. https://doi.org/10.3390/S25041142

Chen, H., Zhou, G., & Jiang, H. (2023). Student Behavior Detection in the Classroom Based on Improved YOLOv8. *Sensors 2023, Vol. 23, Page 8385*, *23*(20), 8385. https://doi.org/10.3390/S23208385

Consoli, T., Schmitz, M. L., Antonietti, C., Gonon, P., Cattaneo, A., & Petko, D. (2024). Quality of technology integration matters: Positive associations with students' behavioral engagement and digital competencies for learning. *Education and Information Technologies 2024 30:6*, *30*(6), 7719–7752. https://doi.org/10.1007/S10639-024-13118-8

Dimitriadou, E., & Lanitis, A. (2023). A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments 2023 10:1*, *10*(1), 12-. https://doi.org/10.1186/S40561-023-00231-3

Ding, J., Niu, S., Nie, Z., & Zhu, W. (2024). Research on Human Posture Estimation Algorithm Based on YOLO-Pose. *Sensors 2024, Vol. 24, Page 3036*, *24*(10), 3036. https://doi.org/10.3390/S24103036

Edozie, E., Shuaibu, A. N., John, U. K., & Sadiq, B. O. (2025). Comprehensive review of recent developments in visual object detection based on deep learning. *Artificial Intelligence Review 2025 58:9*, *58*(9), 277-. https://doi.org/10.1007/S10462-025-11284-W

Elbawab, M., & Henriques, R. (2023). Machine Learning applied to student attentiveness detection: Using emotional and non-emotional measures. *Education and Information Technologies 2023 28:12*, *28*(12), 15717–15737. https://doi.org/10.1007/S10639-023-11814-5

Elnady, M., & Abdelmunim, H. E. (2025). A novel YOLO LSTM approach for enhanced human action recognition in video sequences. *Scientific Reports 2025 15:1*, *15*(1), 17036-. https://doi.org/10.1038/s41598-025-01898-z

Hempel, T., Abdelrahman, A. A., & Al-Hamadi, A. (2022). 6D ROTATION REPRESENTATION FOR UNCONSTRAINED HEAD POSE ESTIMATION. *Proceedings - International Conference on Image Processing, ICIP*, 2496–2500. https://doi.org/10.1109/ICIP46576.2022.9897219

Hu, J., Huang, Z., Li, J., Xu, L., & Zou, Y. (2024). Real-Time Classroom Behavior Analysis for Enhanced Engineering Education: An AI-Assisted Approach. *International Journal of Computational Intelligence Systems 2024 17:1*, *17*(1), 167-. https://doi.org/10.1007/S44196-024-00572-Y

Ikram, S., Ahmad, H., Mahmood, N., Faisal, C. M. N., Abbas, Q., Qureshi, I., & Hussain, A. (2023). Recognition of Student Engagement State in a Classroom Environment Using Deep and Efficient Transfer Learning Algorithm. *Applied Sciences 2023, Vol. 13, Page 8637*, *13*(15), 8637. https://doi.org/10.3390/APP13158637

Jiang, H., & Fu, W. (2024). Computer vision recognition in the teaching classroom: A Review. *EAI Endorsed Transactions on AI and Robotics*, *3*. https://doi.org/10.4108/AIRO.4079

Jocher, G., Qiu, J., Liu, M., Lyu, S., Akyon, F. C., & Kalfaoglu, M. E. (2026). *Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models*. https://doi.org/10.48550/arXiv.2606.03748

Kassab, S. E., Al-Eraky, M., El-Sayed, W., Hamdy, H., & Schmidt, H. (2023). Measurement of student engagement in health professions education: a review of literature. *BMC Medical Education 2023 23:1*, *23*(1), 354-. https://doi.org/10.1186/S12909-023-04344-8

Lamaakal, I., Yahyati, C., Maleh, Y., El Makkaoui, K., Ouahbi, I., El-Latif, A. A. A., Zomorodi, M., & El-Rahiem, B. A. (2025). A tiny inertial transformer for human activity recognition via multimodal knowledge distillation and explainable AI. *Scientific Reports 2025 15:1*, *15*(1), 42335-. https://doi.org/10.1038/s41598-025-26297-2

Liu, J., Mu, X., Liu, Z., & Li, H. (2023). Human skeleton behavior recognition model based on multi-object pose

E-ISSN : 2807-9035
Volume 6, **Number** 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

estimation with spatiotemporal semantics. *Machine Vision and Applications 2023 34:3*, *34*(3), 44-. https://doi.org/10.1007/S00138-023-01396-0

Liu, Q., Jiang, R., Xu, Q., Wang, D., Sang, Z., Jiang, X., & Wu, L. (2024). YOLOv8n-BT: Research on Classroom Learning Behavior Recognition Algorithm Based on Improved YOLOv8n. *IEEE Access*, *12*, 36391–36403. https://doi.org/10.1109/ACCESS.2024.3373536

Liu, Q., Jiang, X., & Jiang, R. (2025). Classroom Behavior Recognition Using Computer Vision: A Systematic Review. *Sensors 2025, Vol. 25, Page 373*, *25*(2), 373. https://doi.org/10.3390/S25020373

Loshchilov, I., & Hutter, F. (n.d.). *Decoupled Weight Decay Regularization*. Retrieved July 28, 2026, from https://github.com/loshchil/AdamW-and-SGDW

Nadeem, M., Elbasi, E., Zreikat, A. I., & Sharsheer, M. (2024). Sitting Posture Recognition Systems: Comprehensive Literature Review and Analysis. *Applied Sciences 2024, Vol. 14, Page 8557*, *14*(18), 8557. https://doi.org/10.3390/APP14188557

Qureshi, T. S., Shahid, M. H., Farhan, A. A., & Alamri, S. (2025). A systematic literature review on human activity recognition using smart devices: advances, challenges, and future directions. *Artificial Intelligence Review 2025 58:9*, *58*(9), 276-. https://doi.org/10.1007/S10462-025-11275-X

Sharma, V., Gupta, M., Kumar, A., & Mishra, D. (2024). STAR-3D: A Holistic Approach for Human Activity Recognition in the Classroom Environment. *Information 2024, Vol. 15, Page 179*, *15*(4), 179. https://doi.org/10.3390/INFO15040179

Sheng, X., Li, S., & Chan, S. (2025). Real-time classroom student behavior detection based on improved YOLOv8s. *Scientific Reports 2025 15:1*, *15*(1), 14470-. https://doi.org/10.1038/s41598-025-99243-x

Shou, Z. ;, Yuan, X. ;, Li, D. ;, Mo, J. ;, Zhang, H. ;, Zhang, J. ;, Wu, Z. A., Shou, Z., Yuan, X., Li, D., Mo, J., Zhang, H., Zhang, J., & Wu, Z. (2024). A Dynamic Position Embedding-Based Model for Student Classroom Complete Meta-Action Recognition. *Sensors 2024, Vol. 24, Page 5371*, *24*(16), 5371. https://doi.org/10.3390/S24165371

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* (pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition. *Big Data and Cognitive Computing 2023, Vol. 7, Page 48*, *7*(1), 48. https://doi.org/10.3390/BDCC7010048

Vermander, P., Mancisidor, A., Cabanes, I., & Perez, N. (2024). Intelligent systems for sitting posture monitoring and anomaly detection: an overview. *Journal of NeuroEngineering and Rehabilitation 2024 21:1*, *21*(1), 28-. https://doi.org/10.1186/S12984-024-01322-Z

Wang, Z., Wang, M., Zeng, C., & Li, L. (2024). SBD-Net: Incorporating Multi-Level Features for an Efficient Detection Network of Student Behavior in Smart Classrooms. *Applied Sciences 2024, Vol. 14, Page 8357*, *14*(18), 8357. https://doi.org/10.3390/APP14188357

Wang, Z. ;, Yao, J. ;, Zeng, C. ;, Fu, S., Yang, B., Zarifis, A., Wang, Z., Yao, J., Zeng, C., Li, L., & Tan, C. (2023). Students' Classroom Behavior Detection System Incorporating Deformable DETR with Swin Transformer and Light-Weight Feature Pyramid Network. *Systems 2023, Vol. 11, Page 372*, *11*(7), 372. https://doi.org/10.3390/SYSTEMS11070372

Xu, L., Li, Z., Gan, Y., & Xia, H. (2025). Soft-label guided stacked dual attention network for head pose estimation and its application to classroom gaze analysis. *Scientific Reports 2025 16:1*, *16*(1), 405-. https://doi.org/10.1038/s41598-025-29814-5

Xu, Q., Wei, Y., Gao, J., Yao, H., & Liu, Q. (2023). ICAPD Framework and simAM-YOLOv8n for Student Cognitive Engagement Detection in Classroom. *IEEE Access*, *11*, 136063–136076. https://doi.org/10.1109/ACCESS.2023.3337435

Zhang, H., Zhou, G., He, W., & Deng, H. (2025). Classroom Behavior Detection Method Based on PLA-YOLO11n. *Sensors 2025, Vol. 25, Page 5386*, *25*(17), 5386. https://doi.org/10.3390/S25175386

Zhang, J., Guo, L., & Wang, X. (2025). Student Classroom Behavior Recognition Based on YOLOv8 and Attention Mechanism. *Information 2025, Vol. 16, Page 934*, *16*(11), 934. https://doi.org/10.3390/INFO16110934

Zhang, X., Nie, J., Wei, S., Zhu, G., Dai, W., & Yang, C. (2024). A Study of Classroom Behavior Recognition Incorporating Super-Resolution and Target Detection. *Sensors 2024, Vol. 24, Page 5640*, *24*(17), 5640. https://doi.org/10.3390/S24175640

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2022). ByteTrack: Multi-object Tracking by Associating Every Detection Box. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13682 LNCS*, 1–21. https://doi.org/10.1007/978-3-031-20047-2_1/SAVE-RESEARCH

Zhang, Y., & Wang, Y. (2025). A comprehensive survey on RGB-D-based human action recognition: algorithms, datasets, and popular applications. *EURASIP Journal on Image and Video Processing 2025 2025:1*, *2025*(1), 15-. https://doi.org/10.1186/S13640-025-00677-0

Zhou, L., Liu, X., Guan, X., & Cheng, Y. (2025). CSSA-YOLO: Cross-Scale Spatiotemporal Attention Network for

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 3, August 2026
https://doi.org/10.47709/brilliance.v6i3.9453

Fine-Grained Behavior Recognition in Classroom Environments. *Sensors 2025, Vol. 25, Page 3132*, *25*(10), 3132. https://doi.org/10.3390/S25103132