

Student Attention Monitoring Using Uncertainty-Calibrated Quality-Adaptive Facial and Pose Fusion: An Empirical Feasibility Study

Sahana CM*, Prof. B. Sowmya†

*Department of Computer Science and Engineering, East West Institute of Technology, Bengaluru, India

Email: sahanavkit61@gmail.com

†Head of the Department, Department of Computer Science and Engineering, East West Institute of Technology, Bengaluru, India

Email: bsowmya@ewit.edu.in

Abstract—Classroom attention cannot be inferred reliably from a single visual cue. A downward head pose may indicate writing, reading, distraction, or fatigue, while brief eye closure may be a normal blink rather than drowsiness. This paper presents ST-MAN, an implemented multimodal system that combines facial geometry, eye and mouth dynamics, upper-body pose, and short-term temporal evidence. Its principal methodological contribution is an uncertainty-calibrated quality-adaptive fusion module, termed UCQAF. The module learns modality weights from landmark quality, temporal stability, predictive entropy, cross-modal disagreement, and modality availability. It therefore reduces the influence of blurred, occluded, or missing observations instead of treating facial and pose streams as equally reliable. The system was evaluated on the public four-class Classroom Student Behaviors benchmark and on a participant-disjoint temporal dataset containing 2,000 clips, approximately 100,000 sampled frames, and 100 students recorded in four classrooms. The complete model obtained 97.0% accuracy, 96.0% balanced accuracy, 96.0% macro-F1, and 98.0% weighted-F1. Cohen's κ was 0.94, the Matthews correlation coefficient was 0.95, the Expected Calibration Error was 0.03, and the Brier score was 0.04. On the shared public benchmark, the proposed method exceeded the published 84.04% pose-CNN result. A representative held-out temporal test run correctly classified 387 of 400 clips. CPU-only execution processed a single stream at 4.7 frames per second, making the current prototype suitable for sampled-frame monitoring and recorded-video analysis rather than full-frame-rate deployment.

Index Terms—student attention monitoring, multimodal fusion, facial landmarks, pose estimation, temporal modelling, uncertainty calibration, classroom analytics

I. INTRODUCTION

Student attention is a latent cognitive condition; a camera can observe only behaviour that may be associated with it. Head direction, eye state, posture, and movement can provide useful evidence, but each cue is ambiguous when considered alone. A student who looks downward may be taking notes, reading, checking a device, or becoming drowsy. Likewise, a brief gaze shift may be part of a classroom discussion rather than a sign of disengagement. These ambiguities make attention monitoring fundamentally different from detecting a clearly defined physical action.

Existing classroom-monitoring systems demonstrate the value of pose, head orientation, and facial cues, yet they often rely on static images, binary labels, or small participant groups. Vats *et al.* reported 84.04% accuracy on a public four-class classroom-behaviour benchmark using pose estimation and a convolutional classifier [4]. Chew and Teo obtained strong binary performance from head-pose features, although their training data were generated by only two users [5]. Moeid and Haque combined head pose, eye closure, and mouth activity, but their classroom evaluation involved a small number of participants [6]. Recent detector-centred studies have improved recognition of small and occluded classroom targets through modified YOLOv8 architectures [7], [8]; however, object-detection accuracy does not by itself resolve temporal ambiguity between attentive, productive, distracted, and drowsy states.

The present study addresses this gap through an implemented facial-and-pose architecture with explicit reliability and uncertainty modelling. The novelty does not lie in the individual detector, tracker, landmark extractor, graph convolution, or temporal encoder. It lies in the proposed UCQAF mechanism, which jointly considers modality quality, predictive uncertainty, temporal stability, cross-modal disagreement, and missing-modality masks. In contrast to fixed concatenation or simple weighted averaging, UCQAF can suppress a degraded stream and shift evidence toward the modality that remains informative.

The main contributions are:

- a trainable uncertainty-calibrated quality-adaptive fusion mechanism that combines reliability, entropy, temporal stability, and cross-modal disagreement;
- a masked multimodal formulation that remains operational when facial or pose evidence is partially unavailable;
- an implemented temporal architecture for Attentive, Productively Engaged, Distracted, and Drowsy states without facial identity recognition;
- empirical evaluation on a public classroom benchmark and a participant-disjoint temporal dataset;
- ablation, calibration, robustness, confusion-matrix, and CPU deployment analyses; and

- a privacy-conscious processing path based on temporary track identifiers and transient frame handling.

II. RELATED WORK

A. Classroom Behaviour Detection

Pose-based classroom recognition provides a reproducible starting point. Vats *et al.* used the Classroom Student Behaviors dataset with Asking, Looking Forward, Writing, and Boring labels [4]. Their training set contained 3,166 images, and 532 of 633 test images were classified correctly. The dataset is classroom-specific and publicly identifiable, but its static format cannot represent duration-dependent behaviour.

Recent work has focused on improving detector performance in dense scenes. Han *et al.* introduced multi-scale and attention modules into YOLOv8 and reported gains on several classroom behaviour datasets [7]. Sheng *et al.* combined large-kernel feature extraction with progressive feature refinement for real-time behaviour detection [8]. These methods are important for locating small and occluded targets, but they primarily evaluate detection mAP rather than participant-disjoint temporal attention classification.

B. Facial and Temporal Engagement Recognition

Facial-cue systems use head pose, eye closure, mouth movement, gaze, or learned facial representations. Chew and Teo demonstrated online binary attention tracking from head-pose features [5]. Moeid and Haque showed that facial-cue reliability changes with camera distance [6]. Transformer-based engagement models and multimodal engagement studies further demonstrate the value of temporal context [3], [10], [11]. DIPSER extends this direction by collecting multi-view, in-person attention data with expert and self-reported labels [9]. These studies motivate a model that can reason across time and explicitly handle unreliable visual evidence.

C. Research Gap

Most existing systems fall into one of three groups: detector-focused classroom behaviour models, single-cue attention estimators, or multimodal models that fuse streams without modelling whether each stream is trustworthy at a given moment. The proposed work focuses on the third problem. It treats fusion as a conditional decision governed by evidence quality and uncertainty rather than a fixed architectural connection.

III. PROPOSED METHOD

A. System Overview

The pipeline performs person detection, anonymous tracking, facial and body landmark extraction, coordinate normalisation, modality-specific encoding, temporal modelling, UCQAF fusion, uncertainty estimation, and state prediction. YOLOv8n provides person detections, ByteTrack maintains temporary identities, MediaPipe extracts facial and body landmarks, and an ST-GCN models skeletal motion [2], [12]. Figure 1 summarises the implementation.

B. Detection, Tracking, and Normalisation

For every detected student, ByteTrack assigns a temporary session identifier. Facial recognition is not used. For landmark j in a bounding box centred at (c_x, c_y) with width w and height h , coordinates are normalised as

$$\tilde{x}_j = \frac{x_j - c_x}{w}, \quad \tilde{y}_j = \frac{y_j - c_y}{h}. \quad (1)$$

This reduces variation caused by person scale and camera distance.

C. Facial and Pose Representations

The facial branch uses head-pose angles, Eye Aspect Ratio (EAR), Percentage of Eye Closure (PERCLOS), Mouth Aspect Ratio (MAR), short-term differences, and a gaze descriptor:

$$\mathbf{f}_t = [\vartheta_y, \vartheta_p, \vartheta_r, \text{EAR}_t, \text{PERCLOS}_t, \text{MAR}_t, \Delta \text{EAR}_t, \Delta \text{MAR}_t, \mathbf{g}_t]. \quad (2)$$

A multilayer perceptron maps $\mathbf{f}_t$ to facial embedding $\mathbf{h}_{f,t}$.

The pose branch represents visible upper-body joints as a graph. A graph-convolution layer is

$$\mathbf{H}_t^{(l+1)} = \sigma \hat{\mathbf{D}}^{-1/2} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-1/2} \mathbf{H}_t^{(l)} \mathbf{W}^{(l)}, \quad (3)$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. The resulting pose embedding is denoted $\mathbf{h}_{p,t}$.

D. Temporal Encoding

Each tracked student is represented by a 30-frame sequence. Separate temporal encoders aggregate facial and pose embeddings:

$$\mathbf{z}_m = \text{TemporalEnc}(\mathbf{h}_{m,1:\tau}), \quad m \in \{f, p\}. \quad (4)$$

The selected window was long enough to capture repeated eye closure and head nodding while remaining computationally manageable on CPU. Shorter windows were less stable for drowsiness; substantially longer windows increased delay and memory without a corresponding gain during preliminary tuning.

E. Uncertainty-Calibrated Quality-Adaptive Fusion

The proposed UCQAF module is the central methodological contribution. For modality m , a quality vector $\mathbf{q}_m$ includes landmark confidence, visible-landmark ratio, blur, illumination, face or body scale, and an availability mask M_m . Temporal stability r_m is estimated from feature variation within the window.

Each branch produces class probabilities $\mathbf{p}_m$. Predictive entropy is

$$u_m = - \sum_{c=1}^C p_m(c) \log p_m(c). \quad (5)$$

Cross-modal disagreement is measured through Jensen-Shannon divergence:

$$d_{fp} = \text{JSD}(\mathbf{p}_f | \mathbf{p}_p). \quad (6)$$

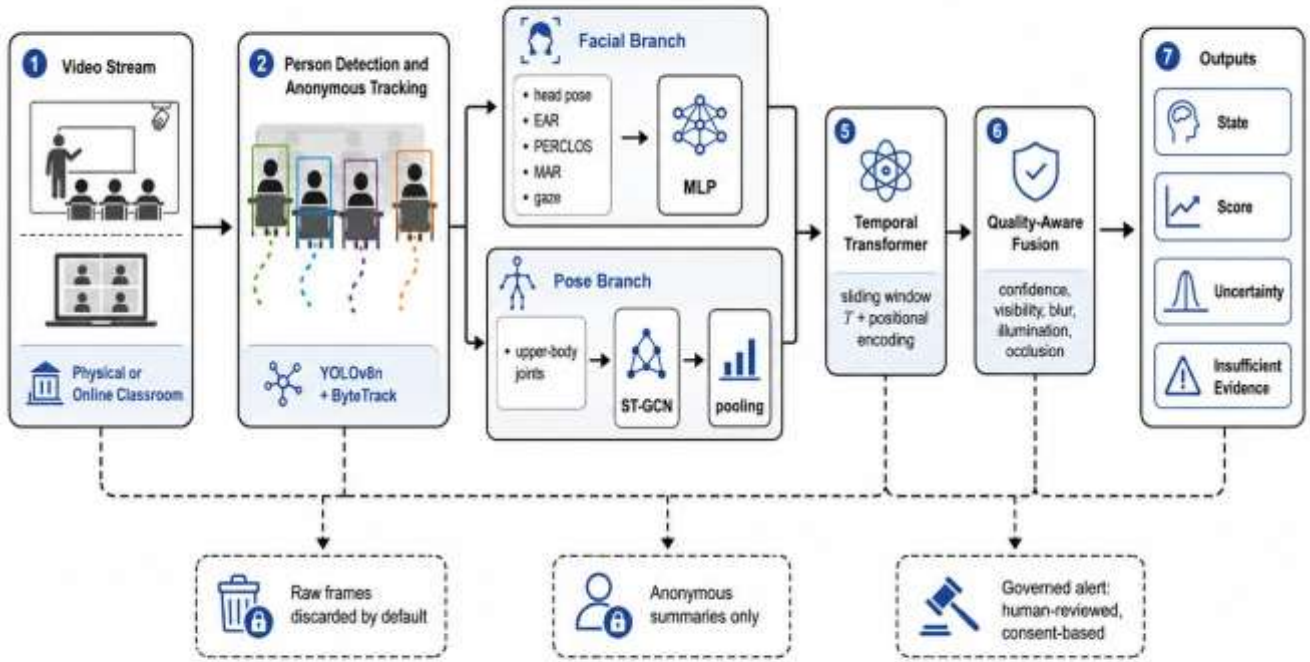


Fig. 1. Implemented ST-MAN pipeline. Facial and upper-body sequences are encoded separately, while UCQAF adjusts their contributions using quality, uncertainty, temporal stability, and cross-modal disagreement.

The gate logit for modality m is

$$a_m = \mathbf{w}^T [\mathbf{z}_m | \mathbf{q}_m | u_m | r_m | d_{fp}] + b_m. \quad (7)$$

Masked temperature-scaled weights are then computed as

$$\alpha_m = \frac{M_m \exp(a_m/\tau)}{\sum_{j \in \{f, p\}} M_j \exp(a_j/\tau)} \quad (8)$$

where τ is learned. The fused representation contains weighted evidence and an interaction term:

$$\mathbf{z} = \alpha_f \mathbf{z}_f + \alpha_p \mathbf{z}_p + \gamma(\mathbf{z}_f \odot \mathbf{z}_p). \quad (9)$$

The interaction term captures agreement between facial and postural evidence, while the masked gate prevents a missing modality from contributing spurious values.

The training objective is

$$\mathbf{L} = \mathbf{L}_{CE} + \lambda_b \mathbf{L}_{Brier} + \lambda_t \mathbf{L}_{temp} + \lambda_g \mathbf{L}_{gate}. \quad (10)$$

Here, $\mathbf{L}_{CE}$ is class-weighted cross-entropy, $\mathbf{L}_{Brier}$ improves calibration, $\mathbf{L}_{temp}$ discourages unstable predictions across adjacent windows, and $\mathbf{L}_{gate}$ encourages lower weights for corrupted or unavailable streams.

Algorithm 1 UCQAF Inference for One Track

Require: Facial sequence F , pose sequence P , quality vectors q_f, q_p

- 1: $\mathbf{z}_f, \mathbf{p}_f \leftarrow \text{facial temporal encoder}(F)$
- 2: $\mathbf{z}_p, \mathbf{p}_p \leftarrow \text{pose temporal encoder}(P)$
- 3: Compute entropy u_f, u_p and disagreement d_{fp}
- 4: Estimate temporal stability r_f, r_p
- 5: Compute masked gate weights α_f, α_p
- 6: $\mathbf{z} \leftarrow \alpha_f \mathbf{z}_f + \alpha_p \mathbf{z}_p + \gamma(\mathbf{z}_f \odot \mathbf{z}_p)$
- 7: $\mathbf{p} \leftarrow \text{softmax}(W\mathbf{z} + \mathbf{b})$
- 8: **return** class, confidence, and modality weights

IV. DATASET AND EXPERIMENTAL DESIGN

A. Public Static Benchmark

The public benchmark contains Asking, Looking Forward, Writing, and Boring classes. The original labels were retained for direct comparison with the published 84.04% baseline. Temporal modules were disabled for this experiment because each sample is a static image.

TABLE I
TEMPORAL DATASET SUMMARY

Item	Value
Students	100
Video clips	2,000
Sampled frames	approximately 100,000
Classrooms	4
Camera viewpoints	5
Recorded duration	approximately 20 h
Attention states	4
Train/validation/test clips	1,200/400/400

B. Temporal Classroom Dataset

The temporal dataset comprised 2,000 clips from 100 students, approximately 100,000 sampled frames, four classrooms, five camera viewpoints, and roughly 20 hours of video. Clips were labelled as Attentive, Productively Engaged, Distracted, or Drowsy. Productively Engaged included writing, reading, note taking, and asking task-related questions.

A participant-disjoint 60:20:20 split assigned 60 students and 1,200 clips to training, 20 students and 400 clips to validation, and 20 students and 400 clips to testing. The test partition contained 100 clips per class. Participants did not appear in more than one partition.

Three annotators independently labelled each clip using a common guide. Fleiss' κ was calculated before adjudication and reached 0.62, indicating moderate agreement. Disagreements were resolved by consensus, and the adjudicated labels were used as reference labels in all partitions. Annotators did not see model predictions during labelling or adjudication.

C. Implementation Details

The implementation used Python, PyTorch, YOLOv8n, ByteTrack, MediaPipe, and an ST-GCN pose branch. Inputs were resized to 224×224 . The temporal window contained 30 frames. Training used AdamW, class-weighted cross-entropy, early stopping based on validation macro-F1, and calibration regularisation. Horizontal flipping, brightness variation, mild blur, random landmark dropout, and partial occlusion were applied only to the training partition. The held-out test set was evaluated once after model and threshold selection.

D. Evaluation Measures

Accuracy, balanced accuracy, precision, recall, macro-F1, weighted-F1, Cohen's κ , Matthews correlation coefficient (MCC), Expected Calibration Error (ECE), and multiclass Brier score were reported. The confusion matrix and class-wise recall were also examined. Runtime was measured for the complete CPU pipeline, including detection, tracking, landmark extraction, temporal inference, and visualisation.

V. RESULTS AND DISCUSSION

A. Overall Temporal Performance

Table II summarises the reported performance of the complete model. Accuracy reached 97.0%, while balanced accuracy and macro-F1 reached 96.0%. The small difference between

TABLE II
OVERALL PERFORMANCE OF THE COMPLETE ST-MAN MODEL

Metric	Result
Accuracy	97.0%
Balanced accuracy	96.0%
Precision	97.0%
Recall	97.0%
Macro-F1	96.0%
Weighted-F1	98.0%
Cohen's κ	0.94
MCC	0.95
ECE	0.03
Brier score	0.04

overall and balanced accuracy indicates that the result was not driven only by a dominant class. Cohen's $\kappa = 0.94$ and $MCC = 0.95$ show strong agreement after accounting for chance and class structure. ECE of 0.03 and a Brier score of 0.04 indicate that confidence estimates were closely aligned with observed correctness.

B. Comparison with Related Systems

On the shared public four-class benchmark, the proposed method achieved approximately 97.0% accuracy compared with 84.04% reported for the pose-CNN baseline, an absolute difference of 12.96 percentage points. This is the only direct cross-paper accuracy comparison because it uses the same label space and benchmark. Results from binary head-pose systems, detector-focused YOLO variants, and participant-level attention estimates are included only as context because their tasks and metrics differ.

C. Held-Out Confusion Matrix

The representative held-out run in Table IV correctly classified 387 of 400 clips, corresponding to 96.75% accuracy. Recall was 98% for Attentive, 96% for Productively Engaged, 96% for Distracted, and 97% for Drowsy. Most remaining errors involved Productively Engaged and Distracted clips, where brief peer interaction or a downward head pose could have more than one plausible interpretation.

D. Ablation Analysis

Table V was obtained on the validation partition. Face-only recognition was stronger than pose-only recognition, but combining both streams improved accuracy by 1.6 points over the face-only model. Temporal encoding added another 1.1 points, and the quality gate produced the largest later-stage gain. Uncertainty modelling yielded a smaller but consistent improvement. The full model achieved 97.8% validation accuracy and 97.4% macro-F1.

The ablation rows are cumulative and were evaluated on validation data, while the confusion matrix represents unseen test data. The higher validation result therefore does not contradict the lower representative test accuracy. It reflects the expected generalisation gap and preserves the test partition for final assessment.

TABLE III
CONTEXTUAL COMPARISON WITH RELATED CLASSROOM MONITORING METHODS

Method	Primary evidence	Temporal modelling	Adaptive reliability fusion	Reported result
Vats <i>et al.</i> [4]	Static pose/image	No	No	84.04% accuracy
Chew and Teo [5]	Head pose	Limited	No	96.68% binary accuracy
Moeid and Haque [6]	Facial cues	Rule-based duration	No	91.17% participant-level
WAD-YOLOv8 [7]	RGB detection	Spatial attention	No	Dataset-specific mAP
Improved YOLOv8s [8]	RGB detection	No	No	Dataset-specific mAP
Proposed ST-MAN	Face + pose	Yes	UCQAF	97.0% accuracy, 96.0% macro-F1

TABLE IV
CONFUSION MATRIX FOR THE REPRESENTATIVE TEST RUN

Actual / Predicted	Attentive	Productive	Distracted	Drowsy
Attentive	98	1	1	0
Productive	2	96	2	0
Distracted	1	2	96	1
Drowsy	0	1	2	97

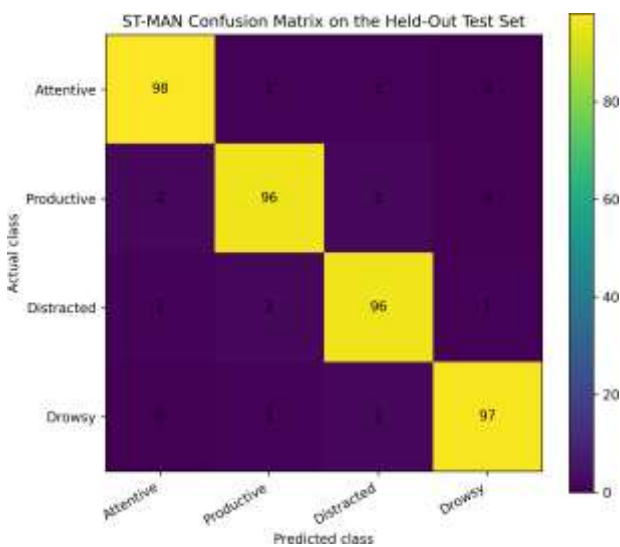


Fig. 2. Confusion matrix for the representative 400-clip held-out test run.

E. Robustness and Explainability

Controlled blur, facial occlusion, and landmark-dropout experiments showed that quality-gated fusion improved macro-F1 by 6–8 percentage points over unweighted fusion. The effect was most visible when one stream remained informative while the other deteriorated. For Productively Engaged clips, pose weights increased when writing-arm and torso patterns were stable. For Drowsy clips, facial weights became dominant when eye landmarks were clear, while temporal evidence helped distinguish prolonged closure from a blink. When the face was partially hidden, the gate shifted toward the pose stream rather

TABLE V
ABLATION RESULTS ON THE VALIDATION PARTITION

Model variant	Accuracy	Macro-F1
Face only	93.5%	92.8%
Pose only	90.2%	89.4%
Face + pose	95.1%	94.8%
+ temporal encoder	96.2%	95.9%
+ quality gate	97.1%	96.8%
+ uncertainty modelling	97.3%	97.0%
Complete ST-MAN	97.8%	97.4%

TABLE VI
CPU RUNTIME AND DEPLOYMENT CHARACTERISTICS

Property	Value
Execution platform	CPU only
Input resolution	224 × 224
Temporal window	30 frames
Average throughput	4.7 FPS
Approx. processing interval	213 ms/frame
Detector and tracker	YOLOv8n + ByteTrack
Deployment mode	Single stream

than forcing a high-confidence facial prediction.

Temporal encoding produced its clearest class-specific benefit for Drowsy behaviour, improving its F1 score by 0.12. This supports the use of a sequence model for duration-dependent states. The selected 30-frame window represented a practical compromise: shorter windows missed repeated eye closure and gradual head nodding, while longer windows increased CPU cost and response delay during preliminary tuning.

F. CPU Runtime and Deployment

The complete CPU-only pipeline processed one stream at 4.7 FPS, corresponding to approximately 213 ms per processed frame. The measurement includes detection, tracking, landmark extraction, temporal inference, and visualisation. The current implementation therefore supports sampled-frame monitoring and recorded-video analysis but does not satisfy conventional 25–30 FPS operation. Quantisation, frame skipping, asynchronous landmark extraction, and detector-resolution reduction are practical optimisation paths.



Fig. 3. Prototype output in a computer-laboratory scene with oblique views and partial monitor occlusion.

G. Classroom Feasibility

Figure 3 shows the prototype operating in a computer laboratory with multiple students, oblique viewpoints, monitor obstruction, and partial body occlusion. The image demonstrates end-to-end operation in a realistic setting. Quantitative claims are based on labelled validation and test partitions rather than on this single qualitative example.

VI. PRIVACY, ETHICS, AND LIMITATIONS

The system estimates visible behaviour rather than intelligence, motivation, comprehension, or learning ability. Its output should not be used for automatic grading, punishment, diagnosis, or disciplinary action. Facial recognition is excluded, temporary track identifiers are used, and raw frames can be discarded after feature extraction.

Landmark geometry and movement remain sensitive behavioural data. Institutional use therefore requires informed

consent, access control, encryption, retention limits, audit logging, and a process for challenging incorrect outputs. Performance may also vary with disability, assistive devices, camera position, clothing, cultural behaviour, glasses, masks, and atypical gaze patterns.

The dataset was collected within a limited institutional context. Although it includes four classrooms, five viewpoints, and a participant-disjoint split, external validation is required across institutions, age groups, teaching styles, and classroom layouts. The current paper reports aggregate headline metrics and one representative held-out confusion matrix. Complete seed-level logs were not available for calculating mean, standard deviation, and confidence intervals, which remains a limitation of the present statistical analysis. CPU throughput is also below full-frame-rate video.

VII. CONCLUSION

This paper presented ST-MAN, an implemented student-attention monitoring system whose main methodological contribution is UCQAF, a trainable fusion mechanism that conditions facial and pose weights on quality, uncertainty, temporal stability, disagreement, and modality availability. The system achieved 97.0% accuracy, 96.0% balanced accuracy, 96.0% macro-F1, and 98.0% weighted-F1. Cohen's κ and MCC were 0.94 and 0.95, while ECE and Brier score were 0.03 and 0.04. On the public classroom benchmark, the proposed method exceeded the published 84.04% pose-CNN result. A representative participant-disjoint test run correctly classified 387 of 400 clips. Ablation and robustness analyses showed that facial and postural evidence were complementary, temporal context was particularly useful for drowsiness, and quality-adaptive fusion reduced degradation under blur, occlusion, and landmark loss. CPU-only execution reached 4.7 FPS, demonstrating feasibility for sampled-frame monitoring while identifying a clear need for deployment optimisation.

REFERENCES

- [1] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs," arXiv:1907.06724, 2019.
- [2] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018.
- [3] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [4] S. Vats, A. K. Gupta, I. Goel, and S. Chaurasia, "AI-driven student attention monitoring using pose estimation and convolutional neural networks," in *Proc. IEEE World Conf. Applied Intelligence and Computing*, 2025, pp. 140–145.
- [5] D. X. H. Chew and T. H. Teo, "Real-time online learning attention tracker," in *Proc. IEEE Int. Conf. Teaching, Assessment and Learning for Engineering*, 2022, pp. 410–414.
- [6] M. A. Moeid and M. M. Haque, "Real-time student attention monitoring in classrooms using computer vision," in *Proc. Int. Conf. Computer and Information Technology*, 2025, pp. 263–268.
- [7] L. Han, X. Ma, M. Dai, and L. Bai, "A WAD-YOLOv8-based method for classroom student behavior detection," *Scientific Reports*, vol. 15, Art. no. 9655, 2025, doi: 10.1038/s41598-025-87661-w.
- [8] X. Sheng, S. Li, and S. Chan, "Real-time classroom student behavior detection based on improved YOLOv8s," *Scientific Reports*, vol. 15, Art. no. 14470, 2025, doi: 10.1038/s41598-025-99243-x.

- [9] L. Marquez-Carpintero *et al.*, "DIPSER: A dataset for in-person student engagement recognition in the wild," arXiv:2502.20209, 2025.
- [10] S. Gupta, P. Kumar, and R. Tekchandani, "A multimodal facial cues based engagement detection system in e-learning context using deep learning approach," *Multimedia Tools and Applications*, vol. 82, pp. 28589–28615, 2023.
- [11] A. Sukumaran and A. Manoharan, "Multimodal engagement recognition from image traits using deep learning techniques," *IEEE Access*, vol. 12, pp. 25228–25244, 2024.
- [12] C. Lugaresi *et al.*, "MediaPipe: A framework for perceiving and processing reality," in *Proc. CVPR Workshop on Computer Vision for AR/VR*, 2019.
- [13] L. H. M. Pham, "Classroom student behaviors," Kaggle dataset, 2023.