MOSAIC-Engage: Uncertainty-Aware Multimodal Deep Learning for Robust Student Engagement Estimation in Online Education

[1]Charanjeet Kaur, [2]Dr. Rajinder Kumar
[1]Research Scholar, Department of Computer Application, Guru Kashi University, Talwandi Sabo, Punjab, India
[2]Assosiate Professor, Department of Computer Application, Guru Kashi University,
Email Id: [1] charanbuttar439@gmal.com, [2] drrajinder1983@gmail.com

Accepted:20-09-.026 Published:12-10 -2026 Page No. 26-35 DOI: 10.5281/zenodo.24388495

Abstract—Web-based teaching makes students less visible to instructors, but camera-facing behavior offers only a partial account of engagement. Looking away may signal distraction, reading, or reflection; equally, silent participation is not necessarily disengagement. Existing recognition studies also differ in labels, class balance, sensor inputs, and evaluation protocols, limiting the meaning of cross-paper accuracy comparisons. This article develops MOSAIC-Engage, a research framework for assessing ordinal engagement indicators from permitted video, audio, and optional interaction features. The proposed system uses independently measured signal-quality cues and availability masks to regulate multimodal fusion, followed by temporal encoding, coherent ordinal classification, and a calibrated option to withhold uncertain predictions. We distinguish two evaluation tracks: visual-only experiments on DAiSEE and genuinely multimodal experiments on CMOSE. A reproducible assessment plan specifies participant-disjoint splits, matched baselines, macro-F1 and ordinal metrics, modality-failure tests, participant-cluster uncertainty intervals, and end-to-end latency measurement. Published DAiSEE scores and CMOSE findings are discussed as secondary evidence, not as results of the proposed model. No MOSAIC-Engage training or deployment experiment is claimed. The principal contribution is a falsifiable architecture and an evaluation protocol addressing reliability, uncertainty, and educational validity without equating observable behavior with learning achievement.

Keywords—component, formatting, style, styling, ins online learning, student engagement estimation, multimodal deep learning, ordinal classification, reliability-aware fusion, selective prediction, uncertainty calibration.

I. INTRODUCTION

Online classes give learners flexibility, although they also make participation harder to observe. A camera feed may show whether a learner is visible, not whether that learner is reasoning through a problem. Engagement includes behavioural participation, emotional involvement, and cognitive effort [1]. Confusion can accompany productive learning, and averted gaze may reflect note-taking rather than inattention [2], [3]. Treating such cues as direct measurements of an internal learning state would therefore overstate what the evidence supports.

Computer-vision studies have used facial expressions, gaze, pose, and temporal dynamics; other investigations incorporate audio or physiological signals [4]–[8]. Yet an accuracy figure alone rarely establishes whether a method works across courses or learners. Label definitions, participant overlap, class prevalence, and modality availability can all change the reported value. The contrast between published DAiSEE studies [9], [10] illustrates this problem: both concern online engagement, but their results should not be

interpreted as a controlled architectural comparison.

We treat engagement estimation as a constrained prediction problem: infer a benchmark-defined ordinal label from observable cues, and acknowledge uncertainty when those cues are weak or unavailable. MOSAIC-Engage brings together quality-conditioned fusion, availability-aware temporal modelling, and selective prediction. The contribution is the joint, testable design and validation protocol, not a claim that any of these mechanisms was invented here. Identifying an individual student or diagnosing emotion is outside the scope of this work.

II. PRIOR WORK AND RESEARCH GAP

A. EDUCATIONAL CONSTRUCT AND OBSERVABLE PROXIES

Fredricks et al. distinguish behavioural, emotional, and cognitive engagement [1]. D'Mello and Graesser documented temporal transitions among engagement, confusion, frustration and boredom in learning situations [2], [3]. A student's looking away can reflect note-taking or reflection; therefore, gaze and expression are noisy indicators, not ground truth. Validation should combine video behavior with opt-in self-report, learning activity and task context. This requirement is central to avoiding over interpretation.

B. DEEP MODELS AND MULTIMODAL DATASETS

The DAiSEE family of studies estimates four ordinal levels of engagement or affect from short video segments. EfficientNetB7 with temporal convolutions or recurrent networks has been evaluated on DAiSEE and VRESEE [9]. A separate ensemble approach reported substantially higher DAiSEE classification accuracy [10]. CMOSE offers multiple modalities and labels designed to address ordinal structure and quality limitations [11]. Other work combines gaze, facial action units, emotion recognition and audio [4], [5], [12]. Sümer et al. additionally investigated attention versus affect features and limited personalization in classrooms [13]. A broad multimodal taxonomy recommends treating representation, alignment, fusion and missingness as distinct issues [8].

TABLE I: Dataset suitability and anticipated limitations.

Dataset	Available signal / task	Required caution
DAiSEE	Video clips; ordinal engagement/affective ratings	Verify release, class counts, agreement, licenses and subject splits.
CMOSE	Multimodal features with engagement annotations	Confirm which samples contain audio/speech and identitydisjoint grouping.
SEA repository	GitHub-hosted supplementary research data [17]	Check data dictionary, consent and comparability before use.

C. COMPARABILITY AND NOVELTY BOUNDARY

The literature supports multimodal fusion and ordinal engagement modeling, but it leaves a practical comparison unresolved: when sensor reliability varies, does quality-conditioned fusion outperform a capacity-matched alternative on participants excluded from training, and can it do so without unacceptable calibration or latency costs? We formulate this as a testable question rather than claiming a proven state-of-the-art model. The required evidence is a common-participant evaluation with explicit missingness tests, not an accuracy ranking assembled from unrelated papers.

III. DATASETS, PROVENANCE, AND LICENSING.

Benchmark Track A (DAiSEE): train and evaluate video-only temporal and ordinal models on an authoritative licensed version; the dataset must not be represented as audio-equipped without evidence. Benchmark Track B (CMOSE): evaluate video, audio/speech and permitted feature combinations with MocoRank and matched fusion comparators [11]. Track C (external validation) is conditional on a new cohort with compatible label semantics and

ethical permissions; simple mapping between DAiSEE and CMOSE labels is not assumed. Official CMOSE project: https://jasonwuchi.github.io/CMOSE/; dataset: https://huggingface.co/datasets/cwuau/CMOSE. The listed independent DAiSEE GitHub reproduction does not establish access to the official dataset. Preserve version identifiers, licenses, participant manifests, class histograms and SHA-256 checksums; exclude unverified Kaggle mirrors. No dataset files have been downloaded for this manuscript.

IV.PROPOSED MOSAIC-ENGAGE METHOD

A. MODALITY-SPECIFIC REPRESENTATION

For each consented segment of length T, the visual encoder extracts frame-level embeddings $h_{v,t} = f_v(I_t)$; an acoustic encoder, if speech/audio is permitted, extracts $h_{a,t} = f_a(A_t)$; and an optional interaction encoder maps coarse non-content telemetry to $h_{i,t} = f_i(X_t)$. Interaction telemetry should not record raw keystrokes or message contents. Modalities are time-aligned using timestamps; padded observations are masked. Each encoder may be initialized from publicly documented pretraining but must be fine-tuned exclusively with the training portion of each benchmark.

B. RELIABILITY-AWARE GATED FUSION

Let $m_{k,t} \in \{0,1\}$ indicate presence of modality k at time t, and let $q_{k,t} \in [0,1]$ represent a quality measurement made without reference to engagement labels. Candidate video-quality variables are frame sharpness (variance of Laplacian, normalized on training data), detection continuity and occlusion; candidate audio-quality variables are valid-speech fraction and speech-to-noise ratio. After projecting $h_{k,t}$ to a shared dimension, compute $s_{k,t} = w^T \tanh(U\tilde{h}_{k,t} + Vq_{k,t} + b)$. Equation (2) defines the masked softmax coefficients, with normalization evaluated only when the denominator is positive. Then $z_t = \Sigma_k \alpha_{k,t} \tilde{h}_{k,t}$. Measurement units, normalization ranges and the handling of unusable frames must be fixed before testing. A quality gate can learn shortcuts, so conditional-on-quality and group-specific rejection statistics are necessary.

$$s_{k,t} = w^T \tanh(U\tilde{h}_{k,t} + Vq_{k,t} + b). \quad (1)$$

$$\alpha_{k,t} = m_{k,t} \exp(s_{k,t}) / \sum_j m_{j,t} \exp(s_{j,t}); \quad z_t = \sum_k \alpha_{k,t} \tilde{h}_{k,t}. \quad (2)$$

When every modality is unavailable, the system abstains; it must not interpret an all-zero input as a genuine observation. The fused sequence passes through a temporal encoder g_θ to give representation r. For ordered labels $Y \in \{0,1,2,3\}$, define $P(Y>c|r) = \sigma(a(r) - b_c)$, $c \in \{0,1,2\}$, where $b_0 \le b_1 \le b_2$. Enforce these thresholds by writing $b_1 = b_0 + \text{softplus}(\delta_1)$ and $b_2 = b_1 + \text{softplus}(\delta_2)$. Class probabilities are obtained by differencing adjacent cumulative probabilities, which preserves non-negativity and a total probability of one.

$$L_{train} = (1/3) \sum_{(c=0)^2} \text{BCE}(\mathbb{1}[y > c], \sigma(a(r) - b_)). \quad (3)$$

Equation (3) is the prespecified primary training objective. Class-weighted variants or a consistency penalty may be evaluated as separate ablations, with coefficients tuned only on training/validation data. Confidence calibration is performed after training on a disjoint calibration partition; expected calibration error (ECE) is an evaluation metric rather than an unspecified differentiable training loss. This separation prevents the manuscript from implying a calibration objective that has not been defined or tested.

C. SELECTIVE INFERENCE AND PRIVACY CONTROL

A prediction is shown only when consent is active, at least one permitted modality passes minimum signal-quality checks, and its validation-calibrated confidence meets a prespecified acceptance rule. Otherwise, the interface reports 'insufficient signal.' The design does not require face identity matching or student-name recognition. Instructor feedback should emphasize aggregated trends and

uncertainty, not individual punitive rankings. The confidence rule and coverage threshold must be fixed before testing.

V.EXPERIMENTAL PROTOCOL FOR ORIGINAL VALIDATION

A. SPLITS AND TRAINING

The proposed protocol is: (1) obtain authoritative dataset versions and permissions; (2) compute manifests and SHA-256 checksums; (3) build participant-disjoint train/validation/test partitions, or disclose when subject IDs prohibit them; (4) fit pre-processing using the training partition only; (5) train visual-only, audio-only, early-fusion, late-fusion and reliability-gated baselines under matching tuning budgets; (6) repeat across five specified random seeds; (7) disclose per-class support, paired uncertainty and missingness sensitivity; and (8) evaluate transfer only when label constructs are demonstrably compatible. Each split and model selection decision must be made without access to held-out test labels.

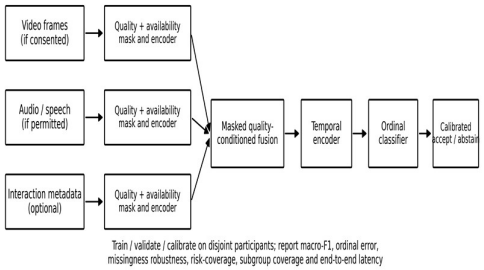


Fig. 1: MOSAIC-Engage proposed processing and evaluation architecture. The design diagram is original; it is not evidence of implementation or measured.

B. METRICS AND INFERENCE MEASUREMENTS.

The primary confirmatory endpoint is four-level macro-F1 on unseen participants with a prespecified minimum no abstention coverage. Report the same participant split for all matching baselines and additionally publish quadratic weighted kappa (QWK), ordinal mean absolute error, balanced accuracy, ECE, per-class precision/recall, risk–coverage curves, size, and p50/p95/p99 glass-to-decision latency on named hardware. Participant-cluster bootstrap intervals preserve within-student dependence. Class counts and chance/majority-label baselines must accompany percentage accuracy. Visual-only and multimodal tracks are reported separately.

C. PRE-SPECIFIED ABLATIONS

The ablation matrix covers: excluding audio; removing reliability quality scores; replacing learned gates with mean concatenation; omitting temporal modeling; dropping ordinal constraints; varying dropout from 0 to 75%; changing input clip windows; and calibrating versus not calibrating. Fair ablation requires the same seeds, group splits, and compute budgets. For educational validation, assess association with contemporaneous self-reported engagement and blinded teacher ratings without treating either as infallible. Stratify performance by lighting, bandwidth, camera use and accommodation status where ethically and lawfully measurable.

VI. PUBLISHED EVIDENCE: QUANTITATIVE SYNTHESIS

A. REPORTED BENCHMARK EVIDENCE

Table II and Fig. 2 report published observations [9], [10], not novel experimental measurements. EfficientNetB7–LSTM achieves 67.48% within the 2022 comparison [9]; the 2024 bagging hybrid reports 94.25% in another study [10]. Differences across publications are not claims of architecture superiority because the model features, splits, labels and validation conditions may differ. Only within-study contrasts under the authors' described protocols are tabulated as published evidence.

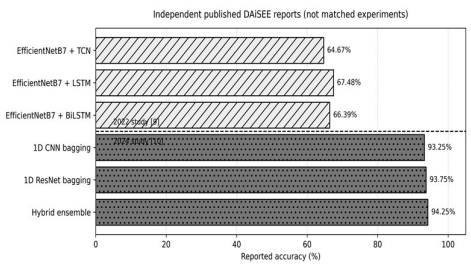


Fig. 2: Chart of previously published DAiSEE accuracy reports [9], [10], redrawn by the authors of this manuscript. It illustrates variation between studies, not improvement achieved by MOSAIC-Engage.

Within [10], hybrid bagging exceeds its reported CNN-bagging and ResNet-bagging comparators by 1.00 and 0.50 percentage points. The 2022 and 2024 figures should not be joined with an arrow labeled as an improvement: neither paper supplies the paired same-participant experiment needed to attribute the numerical difference to the proposed MOSAIC-Engage architecture. Fig. 2 is therefore a provenance-labeled literature overview only.

B. MULTIMODAL EVIDENCE AND MISSINGNESS

In the CMOSE primary paper [11], MocoRank improves overall accuracy by 1.32 percentage points and average accuracy by 5.05 points relative to the authors' own comparators; the reported gain should not be transferred to MOSAIC-Engage. Multimodal systems can fail when speech availability, compression, recording quality or label definitions shift. Their benefits must therefore be evaluated against equal-capacity single-modality controls under deliberate dropout and corruption.

C. RESULTS NOT AVAILABLE IN THIS MANUSCRIPT

The published-study comparison is the only quantitative evidence reported here. MOSAIC-Engage has not been trained or tested in this manuscript. Accordingly, original macro-F1, confusion matrices, ablation effects, calibration curves, confidence intervals, and measured inference times are not available. They are prospective evaluation outcomes, not omitted positive findings.

TABLE II: Published numerical evidence: independent study-level results, not same-protocol head-to-head testing.

Published study	Dataset / model	Reported outcome	Interpretation and comparability limitation
Al-Sakkaf et al. [9], IEEE Access 2022	DAiSEE /EfficientNetB7 +TCN	64.67% accuracy	Within-paper reported four-level engagement task; details must match before replication.
Al-Sakkaf et al. [9], IEEE Access 2022	DAiSEE/EfficientNetB7+LSTM	67.48% accuracy	Same publication; strongest of its three DAiSEE numbers.
Al-Sakkaf et al. [9], IEEE Access 2022	DAiSEE/ EfficientNetB7 +BiLSTM	66.39% accuracy	Not comparable to other-publisher values without consistent evaluation.
Santoni et al. [10], IEEE Access 2024	DAiSEE / 1D CNN bagging	93.25% accuracy	Reported model-level result; inspect features and resampling before comparison.
Santoni et al. [10], IEEE Access 2024	DAiSEE / 1D ResNet bagging	93.75% accuracy	Same study family, but not paired with [9].
Santoni et al. [10], IEEE Access 2024	DAiSEE / hybrid ensemble	94.25% accuracy	Highest published number in this table, *not a proven universal SOTA*.
Wu et al. [11], CVPRW 2024	CMOSE / MocoRank	+1.32 pp overall; +5.05 pp average accuracy	Improvements against authors' own comparators; absolute score not extracted here.
Sümer et al. [13], IEEE TAFFC 2023	School-classroom face video / attention+affect	AUC .620 (grade 8), .720 (grade 12)	Classroom, labels and AUC metric differ from DAiSEE.

Numerical rows are taken from the cited primary publications. No values are generated by the present manuscript. "pp" means percentage points. Accuracy and AUC must not be placed on a single ranking scale.

Claim/end point	Status
Dataset access, hashes, exact split	Not yet executed
Subject-disjoint macro-F1	Not measured
Gating versus late fusion (paired)	Not measured
25/50/75% modality dropout	Not measured
Calibrated risk–coverage	Not measured
Batch-1 end-to-end latency	Not measured
Privacy-safe classroom evaluation	Not conducted
Published comparisons, Table III	Source-grounded

Table III identifies the evidence still needed to assess the proposal. In particular, the reported performance of previously published methods cannot substitute for a controlled test of quality-aware fusion.

VII. MECHANISTIC HYPOTHESES AND BENCHMARK GOVERNANCE

A. WHY RELIABILITY CONDITIONING COULD WORK

Quality-aware fusion is motivated by heteroscedastic observation noise: when a stream becomes less informative because of blur, packet loss or absent audio, a masked gate may reduce its contribution to the fused representation. This is not a guarantee of improved classification. In particular, missingness itself may correlate with access constraints or pedagogy, inducing shortcut learning or unfair rejection. Therefore, the

evaluation must include missing-not-at-random stress tests and group-stratified coverage, not only average accuracy.

B. LEAKAGE-RESISTANT DATASET GOVERNANCE

The unit of statistical independence is a participant, not a video frame. All records sharing a participant identifier are allocated to the same split. If participant identifiers are missing or ambiguous, the benchmark cannot be described as participant-independent without an auditable reconstruction protocol. Clip selection, feature extraction, temporal sampling, annotation harmonization and any feature scaling must occur within the training fold. Duplicate and near-duplicate videos should be identified using hashes and perceptual similarity. A held-out external dataset is used only to estimate domain transportability, not for hyper parameter selection. Because CMOSE and DAiSEE may use no identical constructs, no label-space conversion is assumed without a documented mapping and agreement study.

C. CLASS IMBALANCE, ANNOTATION QUALITY AND BASELINES

Before training, report per-class prevalence and chance/majority-class benchmarks. Optimize class-balanced cross-entropy with a fixed ordinal auxiliary term selected only by validation, and report both macro-F1 and macro recall. Accuracy alone should never determine the preferred model. Where annotator-level labels exist, measure weighted agreement and adjudicate systematic disagreements; when only consensus labels exist, discuss their unmeasured subjectivity. Reproduce MocoRank or its published reference implementation as a CMOSE-specific baseline when implementation and licensing permit [11]. Compare at equal encoder capacities and input sampling rates. A calibration model fitted to test logits would invalidate selective-risk conclusions.

D. TRAINING AND REPORTING RECIPE

Predeclare 16 temporally stratified RGB frames per clip as a starting configuration, adjusting it only within training/validation; document the underlying clip length and true sampling rate. For CMOSE, respect which audio and video streams are genuinely supplied under its license. Match shared encoder backbones, input budget and search effort across late fusion, concatenation and gated models. Use five fixed random seeds, participant-grouped partitions, test-blind tuning, class weighting learned only on training data, and a calibration subgroup that does not share participants with the final evaluation. Code, hardware versions, random seeds and participant-level results should be archived in a permitted repository.

TABLE IV: Prospective controlled experiments: independent variable, control, outcome, and decision rule (not observed results)

Test	Intervention and matched comparator	Primary outcome (held-out participants)	Prespecified interpretation
H1: quality-aware fusion	Reliability gate versus embedding concatenation with matched encoders, seeds, and compute budget	Difference in four-class macro-F1; paired participant-bootstrap 95% CI	Support only if interval excludes zero and ordinal MAE does not materially worsen
H2: modality dropout	Randomly blank one permitted stream at 25%, 50%, and 75% test missingness; compare masked versus unmasked fusion	Macro-F1 as a function of missingness; worst-condition macro-F1	Report slope and absolute change; no claim of robustness from full-modality data alone
H3: selective prediction	Temperature calibration and threshold fixed on validation split versus uncalibrated model	Risk–coverage curve, area under risk–coverage curve, ECE	Improvement only at same predeclared coverage; rejection cost disclosed
H4: real-time feasibility	Batch-1 inference at fixed frame resolution on declared CPU/mobile GPU	Median/p95 end-to-end latency, throughput, peak memory and power	Real-time claim only if full acquisition-tooutput path meets target budget
H5: transportability	In-dataset evaluation versus untouched external cohort with compatible labels	Macro-F1degradation and subgroup calibration spread	Explicitly report negative transfer; no pooled cross-dataset rankings

ordinal MAE, macro-F1, model size and throughput in the same result table. The confidence interval must be computed by participant-block resampling rather than frame resampling; bootstrapping random frames would produce misleadingly narrow uncertainty intervals.

Separately publish calibration curves stratified by lighting and audio availability, conditional on adequate participant counts.

E. REAL-TIME LATENCY DEFINITION

End-to-end latency is $L=t_{display}-t_{latest_capture}$, measured from the timestamp of the newest signal used to the teacher-visible decision. Sampling-window duration must ALSO be reported; a ten-second observation window with 20-ms network execution is not an instantaneous learning signal. Benchmark batch size one after warm-up for at least 1,000 decisions on specified edge CPU/GPU hardware, reporting p50/p95/p99 processing delay, window length, peak RAM, energy where measurable, failures and the percentage of abstained samples. No such values have been measured here.

VIII. STATISTICAL ANALYSIS AND INTERPRETATION GUARDRAILS

The primary effect is Δ=macro-F1(quality-gated fusion) −macro-F1(matched concatenation), calculated on the same held-out participants. A paired participant-cluster bootstrap with at least 2,000 replicates can estimate a two-sided 95% confidence interval for Δ. The test split, principal metric, minimum acceptable prediction coverage, and decision thresholds must be frozen in advance. Secondary comparisons should be labelled exploratory or handled with a prespecified multiplicity correction. The result should also be checked against ordinal error and inference cost, so that a marginal gain in F1 does not disguise a worse system.

Classification validity and educational validity are different questions. A model may predict a benchmark label yet fail to measure meaningful classroom participation. To examine this gap, a prospective opt-in study should compare predictions with short experience-sampling responses, independent teacher judgments, and formative-task evidence, while accounting for course, task, learner, and time effects.

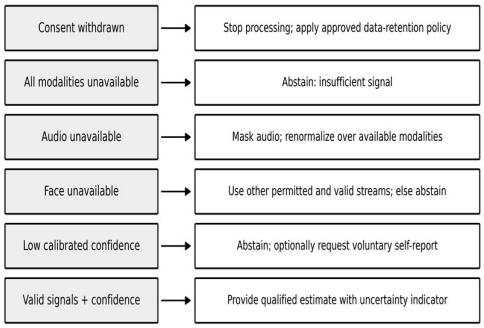


Fig. 3: Proposed operational failure-handling policy. These are design rules, not experimentally verified system behaviours.

Because facial cues and device availability can differ across learner groups, subgroup analyses need appropriate consent, sufficiently large samples, and clearly reported uncertainty. No automatically generated engagement score should be used to grade, discipline, or exclude a student. These safeguards concern the proposed deployment design; they have not yet been evaluated in a classroom trial.

IX. DEPLOYMENT, FAILURES AND THREATS TO VALIDITY

A deployment should avoid forcing a low-confidence prediction when cameras are disabled or bandwidth is restricted (Fig. 3). Model quality can shift across countries, devices, lighting, accessibility needs and teaching styles; therefore, cross-environment performance must be monitored only with consent and without automated disciplinary decisions. Reviews have emphasized that classroom deployment evidence and ethical safeguards lag behind technical development [19]. Face-derived features can also encode demographic differences unrelated to actual learning. Reports should disclose the study population and consider consent withdrawal, data minimization, access control, retention limits and human oversight.

There are five major threats to internal and external validity. Label construct: observable affect is not engagement itself. Leakage: clips from one learner or session can silently occur in both train and test. Imbalance: overall accuracy can hide errors on minority

classes. Measurement mismatch: DAiSEE, CMOSE, classroom face-video and EEG studies use different labels and sensors. Deployment mismatch: a GPU model benchmarked offline does not prove real-time operation on a learner laptop. The pre-specified tests in Section IV address but cannot automatically eliminate these threats.

X. REPRODUCIBILITY BLUEPRINT AND DECISION THRESHOLDS

A. OPERATIONAL RESEARCH QUESTIONS

RQ1: Under participant-disjoint evaluation, does quality-conditioned fusion improve macro-F1 relative to an equal-capacity concatenation baseline?

RQ2: Does prediction quality deteriorate less rapidly under 25%, 50% and 75% modality dropout, including non-random dropout?

RQ3: At a fixed coverage, does post-hoc calibration reduce selective risk? RQ4: Does the full camera/microphone-to-interface pipeline meet a prespecified latency and memory budget on named edge hardware? All four questions remain open until implementation.

B. EVALUATION ESTIMANDS AND UNCERTAINTY

For each class c, let precision P_c and recall R_c defines $F1_c=2P_cR_c/(P_c+R_c)$, with $F1_c=0$ when its denominator is zero. The macro score is $\text{macro-F1}=(1/K)\sum_{c=1}^{K}F1_c$, where K=4. For ordinal predictions, $MAE=(1/N)\sum_{i=1}^{N}|y_i-\hat{y}_i|$. Given accepted cases S_τ at confidence threshold τ, coverage is $C(\tau)=|S_\tau|/N$ and selective classification risk is $R(\tau)=\sum_{i\in S_\tau}1[\hat{y}_i\neq y_i]/|S_\tau|$. Risk is undefined at zero coverage and must not be plotted as zero.

A paired stratified bootstrap should resample participants with replacement, carrying all their correlated segments together. For each of at least 2,000 bootstrap replicates, calculate the difference in macro-F1 between two systems on a common test participant sample. The percentile interval for that paired difference quantifies uncertainty without treating temporally adjacent frames as independent observations. Because threshold tuning can inflate test results, abstention thresholds and any ordinal decision boundaries must be frozen using validation data. A preregistered minimum-coverage requirement prevents a trivial model from achieving low error by rejecting almost every clip.

C. RESOURCES AND IMPLEMENTATION CHECKLIST.

For each dataset, store cryptographic hashes, a machine-readable manifest, licensing confirmation, a table of participants per split, class histograms, pre-processing scripts, hyper parameter search logs, training seed, hardware identifiers, driver/CUDA versions, and wall-clock measurement code. A minimum comparison comprises five named baselines: visual-only, audio-only, early concatenation, probability-level late fusion, and masked reliability-gated fusion. Parameter counts, multiply–accumulate estimates and modality acquisition costs should accompany accuracy because edge-compute constraints are central to the title's real-time claim. Public code must not redistribute any underlying protected student video.

D. INTERPRETATION OF FAILURE MODES

A visual system may mistake reading a printed handout for disengagement; an audio-based system may consider muted students to be inactive; interaction signals may favor learners who type frequently. These are construct-related measurement errors, not merely noise. During informed consent, participants should be able to disable any sensor without educational penalty, and the inference engine should record that a modality was unavailable rather than silently imputing a false observation. Aggregate instructor feedback should be accompanied by an uncertainty indicator and an explicit warning against using predictions for grading or surveillance. Research claims regarding improved student learning require a separate prospective, appropriately controlled educational outcome study.

XI. DISCUSSION AND FUTURE EMPIRICAL WORK.

A reliable engagement estimator must satisfy three separate standards. First, it should generalize to unseen participants under the benchmark's own label definition. Second, its predictions should bear a demonstrable relationship to meaningful educational indicators instead of facial appearance alone. Third, the complete capture-to-feedback system should meet acceptable reliability, calibration, and latency requirements on specified hardware. MOSAIC-Engage provides an evaluation design for these questions but does not yet establish any of the three empirically. Because missing audio and low-quality video may track connectivity, disability-related

accommodation, or teaching practice, coverage and rejection rates deserve as much scrutiny as classification accuracy.

The immediate next study is therefore a preregistered, participant-disjoint CMOSE comparison of quality-conditioned gating with equal-capacity early and late fusion baselines. Its outputs should include per-class errors, an ablation for missing modalities, calibration and risk–coverage plots, and batch-one end-to-end timing. DAiSEE can provide a separate visual-only assessment, but should not be presented as supplying audio. Any later instructor-facing deployment needs independent educational validation and appropriate ethics review.

XII. CONCLUSION

MOSAIC-Engage offers an explicit architecture and a reproducible test plan for uncertainty-aware engagement estimation. Quality-conditioned masked fusion, ordinal classification, and selective prediction form a coherent hypothesis about reliability when classroom signals are incomplete. Existing published results establish the importance—and the difficulty—of the problem, but they do not verify this framework. Whether MOSAIC-Engage improves generalization, calibration, robustness, or practical inference speed remains an empirical question. Reporting that boundary clearly is a prerequisite for credible subsequent experimentation.

RESEARCH INTEGRITY AND DATA AVAILABILITY

Research integrity and data availability—This article specifies a proposed method and interprets independently published findings. Its authors do not report new data collection, model training, test-set performance, or classroom deployment. DAiSEE and CMOSE are external resources subject to their respective access terms. Dataset versions, code, permissions, and ethics approval must be documented before any empirical follow-up. No secondary result should be relabelled as an original MOSAIC-Engage measurement.

REFERENCES

[1]. J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, “School engagement: Potential of the concept, state of the evidence,” Rev. Educ. Res., vol. 74, no. 1, pp. 59–109, 2004, doi: 10.3102/00346543074001059.

[2]. S. D’Mello and A. Graesser, “Dynamics of affective states during complex learning,” Learn. Instr., vol. 22, pp. 145–157, 2012, doi: 10.1016/j.learninstruc.2011.10.001.

[3]. S. D’Mello and A. Graesser, “The half-life of cognitive-affective states during complex learning,” Cogn. Emot., vol. 25, no. 7, pp. 1299–1308, 2011, doi: 10.1080/02699931.2011.613668.

[4]. Sukumaran and A. Manoharan, “Multimodal engagement recognition from image traits using deep learning techniques,” IEEE Access, vol. 12, pp. 25228–25244, 2024, doi: 10.1109/ACCESS.2024.3353053.

[5]. “A multimodal approach for real-time engagement monitoring in elearning using machine learning,” 2024 IEEE Frontiers in Education Conf., doi: 10.1109/FIE61694.2024.10892967.

[6]. D. Kayande and S. Kukreja, “Design of an integrated multi-modal machine learning framework for real-time student engagement evaluation and learning outcome optimizations,” MethodsX, vol. 15, art. 103588, 2025, doi: 10.1016/j.mex.2025.103588.

[7]. “Student engagement recognition: Comprehensive analysis through EEG and verification by image traits using deep learning techniques,” IEEE Access, vol. 13, pp. 11639–11662, 2025, doi: 10.1109/ACCESS.2025.3526187.

[8]. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2,pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[9]. “Students engagement level detection in online e-learning using hybrid EfficientNetB7 together with TCN, LSTM, and Bi-LSTM,” IEEE Access, vol. 10, pp. 99573–99583, 2022, doi: 10.1109/ACCESS.2022.3206779.

[10]. M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, “Automatic detection of students’ engagement during online learning: A bagging ensemble deep learning approach,” IEEE Access, vol. 12, pp. 96063–96073, 2024, doi: 10.1109/ACCESS.2024.3425820.

[11]. C.-H. Wu, S.-Y. Liu, X. Huang, X. Wang, R. Zhang, L. Minciullo, W. K. Yiu, K. Kwan, and K.-T. Cheng, “CMOSE: Comprehensive multi-modality online student engagement dataset with high-quality labels,” Proc. IEEE/CVF CVPR Workshops, pp. 4636–4645, 2024, doi: 10.1109/CVPRW63382.2024.00466.

[13]. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," Proc. ICMI, pp. 784–789, 2020, doi: 10.1145/3382507.3417973.

[14]. Ö. Sümer et al., "Multimodal engagement analysis from facial videos in the classroom," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1012– 1027, 2023, doi: 10.1109/TAFFC.2021.3127692.

[15]. "Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures," Future Gener. Comput. Syst., vol. 108, pp. 334–348, 2020, doi: 10.1016/j.future.2020.02.075.

[16]. P. R. Komaravalli and B. Janet, "Detecting academic affective states of learners in online learning environments using deep transfer learning," Scalable Comput. Pract. Exp., 2023, doi: 10.12694/scpe. v24i4.2470.

[17]. "Engagement and attention level detection on the DAiSEE dataset," 2025 Int. Conf. Emerging Technologies and Innovation for Sustainability, doi: 10.1109/EmergIN67762.2025.11450643.

[18]. L. Yan, X. Wu, and Y. Wang, "Student engagement assessment using multimodal deep learning," PLOS ONE, vol. 20, no. 6, art. e0325377, 2025, doi: 10.1371/journal.pone.0325377.

[19]. "An approach of analyzing classroom student engagement in multimodal environment by using deep learning," 2023 IEEE WIECON-ECE, doi: 10.1109/WIECON-ECE60392.2023.10456488.

[20]. "Affective computing in online higher education: systematic literature review," Comput. Educ.: Artif. Intell., vol. 10, art. 100499, 2026, doi: 10.1016/j.caeai.2025.100499.

[21]. Kumar, R. (2025, January). Artificial intelligence's metamorphosis in higher education (HEIs): A transformational study. Social Research Foundation, 9(1). "A multimodal AI framework for real-time student engagement detection and adaptive feedback in higher education," J. King Saud Univ. Comput. Inf. Sci., 2026, doi: 10.1007/s44443-026-00639-0.

[22]. Devasya, A. Abel, Y. Wang, and L. Jia, "Affective state recognition in online learning: A deep learning approach with data augmentation," 2025 IEEE SMAP, pp. 108–113, doi: 10.1109/SMAP66932.2025.00027.