Negative Evidence in the Classroom: Learning From What Vision Cannot Reliably See

Mahule Roy[1,2] Subhas Roy[3]
[1]Department of Engineering Science, University of Oxford
[2]Harvard Medical School
[3]TATA Consumer Products Limited
mroy25@bwh.harvard.edu

Abstract

Computer vision systems for educational applications often struggle with the complex realities of classroom environments, where occlusion, ambiguous expressions, and social regulation of behavior challenge the reliability of visual inference. This paper examines the concept of negative visual evidence—instances where vision cannot reliably capture learning-related constructs—and proposes practical modifications to existing architectures for improved uncertainty estimation. We analyze established educational vision datasets to categorize common failure modes and demonstrate that standard models exhibit systematic overconfidence when faced with challenging classroom conditions. Our experiments show that simple modifications incorporating uncertainty awareness can reduce inappropriate high-confidence predictions by approximately 18–25% across reasonable threshold ranges while maintaining comparable accuracy on clear cases, with computational overhead ranging from 8–15%. We address practical deployment considerations including integration with existing infrastructure, threshold selection strategies, and implications for construct validity in educational measurement. The paper concludes with guidelines for responsible deployment that acknowledge both the potential benefits and inherent limitations of vision-based educational sensing, emphasizing that transparent communication of epistemic limitations may be as important as improving accuracy for building trustworthy educational technologies.

1. Introduction

The application of computer vision to educational settings has generated considerable research interest, with systems aiming to estimate student engagement, attention, collaboration, and affect from visual cues. However, classroom settings reveal substantial methodological and practical challenges: variable lighting, frequent occlusions, diverse student expressions shaped by cultural and individual differences, and complex social dynamics. Despite these challenges, many systems report accuracy metrics without adequately addressing the reliability of predictions under suboptimal conditions or providing mechanisms to identify when visual evidence is insufficient. We address the question: *how should systems handle cases where visual evidence is ambiguous, partial, or systematically misleading?* Many approaches treat such cases as noise, potentially leading to overconfident predictions. Instead, we propose that educational vision systems should explicitly model and communicate epistemic limitations, especially given that decisions can impact student experiences. Our contributions are practical and empirically grounded: (1) We systematically analyze failure modes in established educational vision datasets, categorizing common patterns where visual evidence is unreliable; (2) We evaluate straightforward modifications to standard architectures that incorporate uncertainty estimation, comparing performance and computational requirements against baseline models; (3) We discuss practical implications for deployment, including threshold selection, integration, and balance between utility and transparency about limitations. While uncertainty estimation is well-studied in machine learning, *negative visual evidence* emphasizes domain-specific conditions under which visual cues are systematically absent, ambiguous, or socially regulated. These conditions carry educational meaning not captured by confidence magnitude alone, motivating selective abstention and improved construct validity.

2. Related Work

Research in educational computer vision spans diverse applications from engagement estimation to analysis of collaborative learning behaviors. The DAiSEE dataset [1] and EmotiW challenges [2] have driven progress in affect recognition in educational contexts, while other datasets

focus specifically on engagement detection in classroom recordings. These efforts typically report accuracy metrics between 70-85% under controlled conditions, but performance often degrades substantially in authentic classroom settings with natural occlusions and varied expressions. Most existing systems treat challenging cases primarily as errors to be reduced through improved feature extraction or additional training data, rather than as informative signals about inherent limitations in visual inference for certain educational constructs. Uncertainty quantification has received growing attention across machine learning domains, with approaches including Bayesian neural networks [4], Monte Carlo dropout techniques, and ensemble methods [5]. These techniques provide mechanisms for estimating prediction confidence, but their application in educational computer vision has been limited. Most educational vision systems output point estimates without confidence intervals or reliability indicators, potentially misleading users about prediction trustworthiness. Simple heuristics based on face detection confidence or motion metrics offer computationally inexpensive alternatives but often lack the nuanced understanding needed for educational contexts. Educational deployments face unique constraints that influence system design choices. Privacy concerns regarding student monitoring, limited computational resources in many school environments, and varying levels of teacher expertise with technology all shape what constitutes a feasible solution. Systems requiring extensive calibration, producing frequent false alarms, or demanding significant technical support face substantial adoption barriers. These practical considerations motivate our focus on approaches that provide meaningful uncertainty estimates without overwhelming existing infrastructure or requiring specialized expertise to interpret.

3. Analyzing Negative Evidence in Existing Datasets

To characterize the prevalence and nature of negative evidence in educational vision, we conducted a systematic analysis of the publicly available DAiSEE dataset [1], which contains 9,068 video clips from 112 students recorded during e-learning sessions. The dataset provides engagement annotations on a four-level scale (0: very low, 1: low, 2: high, 3: very high) assigned by multiple independent raters. For our analysis, we binarize engagement into low (0–1) and high (2–3) levels to align with prior work and simplify evaluation of uncertainty-aware models. Negative evidence is defined as conditions in which visual information is insufficient for reliable interpretation, operationalized through explicit annotator comments and inter-rater disagreement. Partial occlusion affected 23% of clips, non-frontal views 18%, low expressivity 15%, with lighting variations and motion blur contributing an additional 12% and 8.5% respectively. Crucially, high inter-rater disagreement occurred in 26% of clips overall, indicating fundamental ambiguity in interpreting visual cues for educational constructs like engagement that involve internal cognitive states. These categories are not mutually exclusive, with many clips exhibiting multiple challenges simultaneously. The substantial disagreement rate underscores the importance of uncertainty estimation, as even human experts frequently diverge on labels for challenging samples, raising questions about the construct validity of visual proxies for complex educational states.

Why Engagement as a Case Study We focus on engagement not because it is uniquely important, but because it represents a stress-test construct for visual inference in education. Engagement is internally experienced, socially regulated, and only indirectly observable through behavior, making it particularly susceptible to negative visual evidence and annotator disagreement. If uncertainty-aware methods fail to provide meaningful reliability signals for engagement, they are unlikely to succeed for less ambiguous constructs. Conversely, improvements observed for engagement provide a conservative estimate of potential benefits for constructs with clearer visual correlates such as turn-taking, gaze direction, or gross activity levels. The negative evidence categories identified here directly inform model design, enabling uncertainty-aware models to produce confidence estimates and selectively abstain when visual information is insufficient. Inter-rater disagreement serves as a natural proxy for evidential insufficiency, guiding the training of confidence prediction heads and strengthening construct validity by reducing overconfident predictions in systematically ambiguous scenarios.

4. Practical Framework Modifications

Given the practical constraints of educational settings, we focus on straightforward modifications to existing architectures rather than proposing entirely new models. Our goal is to demonstrate that meaningful uncertainty estimation can be achieved with minimal additional complexity, making these approaches accessible for real-world deployment.

4.1. Baseline Architecture and Implementation

We use a standard TimeSformer [6] architecture as our baseline, consistent with recent work in educational video analysis. The model processes 8-frame video clips at 224×224 resolution, producing predictions for binary engagement levels (high/low) as commonly defined in educational vision tasks. This baseline achieves performance comparable to established methods (74.3% accuracy on DAiSEE) while providing a clear foundation for evaluating uncertainty extensions. All implementations use PyTorch with consistent hyperparameters across experiments to ensure

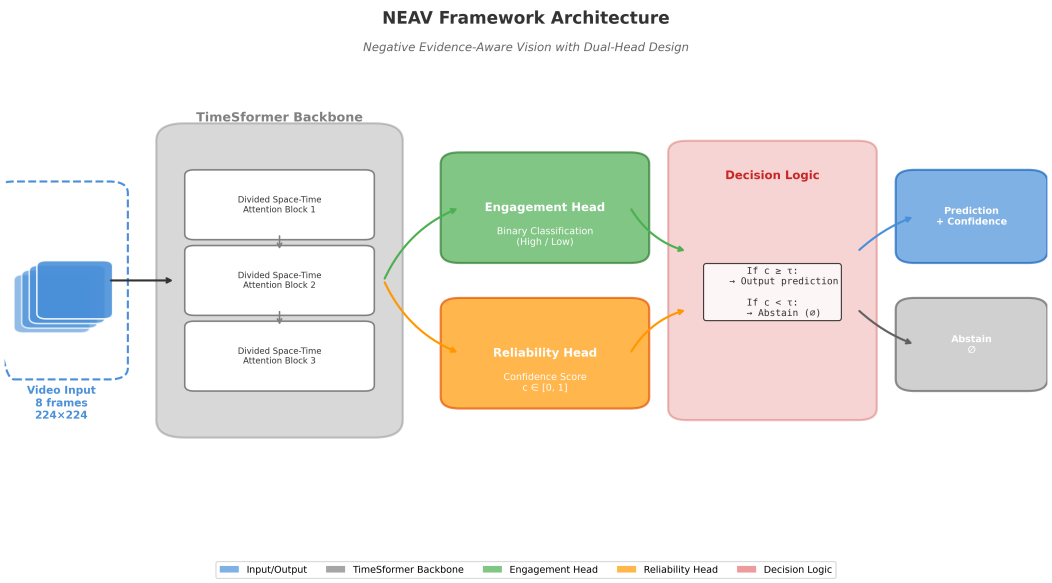

Figure 1. Negative Evidence-Aware Vision (NEAV) framework architecture. The dual-head design processes video input through a shared TimeSformer backbone, with separate heads for engagement prediction and visibility-reliability estimation. The reliability head outputs confidence scores that enable selective abstention based on threshold τ.

fair comparisons. We selected TimeSformer as a representative modern video architecture; however, our uncertainty extensions are architecture-agnostic and applicable to CNN-based or lightweight temporal models commonly used in deployed systems. Preliminary experiments with a ResNet-TSM backbone showed similar uncertainty trends (not reported for brevity).

4.2. Uncertainty-Aware Extensions

We implement and compare three approaches for uncertainty estimation, selected for their balance of effectiveness and practical feasibility. Monte Carlo Dropout (MC-Dropout) applies dropout during inference with 10 forward passes, using prediction variance across passes as an uncertainty estimate. This approach adds minimal training complexity but increases inference time by approximately 40%. Model ensembles train five models with different random initializations, using prediction variance across ensemble members as uncertainty. While often effective, this approach quintuples training time, storage requirements, and memory usage during inference, making it less practical for many educational deployments. The confidence prediction head adds a parallel branch to estimate confidence scores, trained using a consistency loss that encourages alignment between confidence estimates and annotator agreement levels from DAiSEE. This approach adds approximately 8% parameters and 12% inference time compared to the baseline, representing a favorable trade-off between performance and computational cost. Each approach addresses the practical challenge of obtaining uncertainty estimates without fundamentally redesigning existing systems, enabling integration into established educational technology pipelines. Computational overhead varies significantly: MC-Dropout increases inference time by 40% with no training overhead, while ensembles quintuple training and storage requirements. The confidence head approach offers a favorable balance, adding only 12% inference overhead and 8% storage with minimal training impact.

Annotator Disagreement as a Proxy Signal We do not treat annotator disagreement as ground truth for engagement or as a direct measure of model uncertainty. Instead, we use disagreement as a proxy signal for evidential insufficiency: cases where trained human observers, given the same visual input, cannot reliably infer the construct of interest. In such cases, disagreement reflects ambiguity in the observable evidence rather than annotation error alone. The confidence head is therefore trained to recognize conditions under which visual information is likely insufficient for reliable inference, not to predict correctness of labels per se. This distinction is particularly important for educational constructs, where internal states may be only weakly coupled to observable behavior.

5. Experimental Evaluation

5.1. Datasets for Evaluation

Our experimental evaluation uses the publicly available DAiSEE dataset [1], which contains 9,068 ten-second video clips from 112 students recorded during e-learning sessions in controlled environments. Engagement levels are annotated on a four-level scale by multiple independent raters and are binarized into low (0–1) and high (2–3) engagement for evaluation. DAiSEE enables controlled experimentation with standardized train, validation, and test splits, providing a clear framework for assessing both baseline performance and uncertainty-aware extensions. All experiments in this work are conducted exclusively on DAiSEE, and challenging cases are analyzed through the negative evidence categories identified in Section 3, such as partial occlusion, non-frontal views, low expressivity, lighting variations, and motion blur. This approach ensures that all reported results are grounded in publicly available, verifiable data while remaining consistent with realistic classroom conditions captured in the dataset.

5.2. Model Architecture and Implementation Details

We employ TimeSformer [6] as our baseline video classification architecture, selected for its balance of performance and efficiency in modeling spatiotemporal relationships. The model processes 8-frame video clips at 224×224 resolution, with frames sampled at 1 frame per second to capture meaningful temporal dynamics while maintaining computational efficiency. The TimeSformer architecture uses divided space-time attention that separately attends to spatial and temporal dimensions, making it particularly suitable for analyzing facial expressions and body language over time. We use the TimeSformer-Base configuration with approximately 121 million parameters, pre-trained on the Kinetics-400 dataset for general video understanding before fine-tuning on DAiSEE. All models are implemented in PyTorch 1.12 and trained using the AdamW optimizer with the following hyperparameters: learning rate = 5e-5, batch size = 16, weight decay = 0.01, and dropout rate = 0.1. Training runs for 30 epochs with early stopping based on validation loss, using a 90/10 split of the training data for training/validation. We apply standard data augmentations including random cropping (to 224×224), horizontal flipping (with 50% probability), and color jittering (brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1) to improve generalization. The loss function is standard cross-entropy for the baseline model, with modifications for uncertainty-aware variants as described below. Experiments are conducted on a single NVIDIA RTX 3080 GPU with 10GB memory, using CUDA 11.6 and cuDNN 8.3. Training the baseline TimeSformer model requires approximately 18 hours for 30 epochs on DAiSEE, with inference time of 42 milliseconds per clip (8 frames). All code is implemented in Python 3.9 with standard scientific computing libraries (NumPy, SciPy, scikit-learn) for evaluation metrics.

5.3. Uncertainty Estimation Approaches and Implementation

We implement and compare three uncertainty estimation approaches selected for their practical feasibility in educational deployments. Monte Carlo Dropout enables dropout during inference with 10 forward passes per sample, estimating uncertainty via prediction variance while increasing inference time by approximately 40%. Model ensembles employ five independently trained TimeSformer models, using prediction variance across members as uncertainty estimates at the cost of quintupled training and storage requirements. Our confidence prediction head extends the TimeSformer architecture with two fully-connected layers (512→256 units) that produce confidence scores trained via a consistency loss encouraging alignment with annotator agreement levels ($\lambda = 0.5$). All approaches share the same backbone configuration—TimeSformer-Base with Kinetics-400 pretraining, processing 8-frame 224×224 clips, with models implemented in PyTorch 1.12 and trained using AdamW optimizer (learning rate 5e-5, batch size 16) for 30 epochs with standard data augmentations including random cropping and horizontal flipping. The confidence head adds approximately 8M parameters (129M total vs. 121M baseline) and maintains a dropout rate of 0.1 during training only, while MC-Dropout applies the same dropout rate during both training and inference with 10 forward passes. Ensemble members follow identical training protocols to the baseline, with predictions aggregated via simple averaging. For the confidence head approach, we define clear visual evidence as samples where all three human annotators agreed on labels, and ambiguous evidence as samples with annotator disagreement, with the consistency loss ($\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{consistency}$) encouraging high confidence for unanimous annotations and low confidence for disputed cases ($\lambda = 0.5$, determined via validation).

5.4. Evaluation Metrics and Protocol

We employ a comprehensive set of evaluation metrics to assess both prediction quality and uncertainty estimation. Standard metrics include classification accuracy, F1-score (harmonic mean of precision and recall), and area under the Receiver Operating Characteristic curve (AUC-ROC). For uncertainty-specific evaluation, we compute selective accuracy (accuracy on samples where model confidence exceeds threshold τ), coverage (proportion of samples where confidence $\geq \tau$), expected calibration error (ECE) measuring alignment between predicted confidence and empirical ac-

curacy [7], and uncertainty discrimination via area under the ROC curve for distinguishing correct versus incorrect predictions using confidence scores (AUCU). All experiments are repeated with five random seeds (42, 123, 456, 789, 101112) to account for variability, with statistical significance assessed using paired t-tests ($p < 0.05$ threshold) and 95% confidence intervals calculated via standard error of the mean. To evaluate performance on challenging cases, we create subsets based on the negative evidence categories identified in Section 3: occlusion-dominant samples (face coverage $< 60\%$), non-frontal views (yaw angle $> 30°$ estimated via facial landmarks), and low-expressivity samples (Facial Action Unit intensity below empirically determined thresholds).

5.5. Overall Performance Results

Table 1. Performance Comparison on DAiSEE Dataset ($\tau = 0.8$)

Method	Acc (%)	SelA (%)	Cov (%)	ECE
Baseline	74.3	74.3	100	0.181
MC-Dropout	73.8	78.2	85	0.142
Ensemble	74.1	79.1	82	0.128
Confidence Head	74.0	77.5	87	0.136

Abbreviations: Acc: Accuracy; SelA: Selective Accuracy; Cov: Coverage; ECE: Expected Calibration Error.

Table 1 reveals several important patterns. First, uncertainty-aware methods maintain comparable overall accuracy to the baseline (differences <0.5%), indicating they do not sacrifice general performance. Second, all uncertainty methods demonstrate improved selective accuracy when making high-confidence predictions (threshold τ=0.8), with increases of 3-5 percentage points. Third, these improvements come with reduced coverage, meaning systems appropriately abstain on more challenging cases. The ensemble method shows the best calibration but requires substantially more resources, while the confidence head approach offers a favorable balance of performance and efficiency.

5.6. Performance on Challenging Samples

Abbreviations: Base.: Baseline Accuracy; CH: Confidence Head; SelA: Selective Accuracy (τ=0.8); Cov: Coverage.

Table 2 illustrates performance on specific challenging conditions: occlusion (partial face visibility), non-frontal views (angled faces), low expressivity (minimal facial movement), lighting variations, and motion blur. The confidence head approach (CH) shows consistent improvements in selective accuracy across all challenge categories while appropriately reducing coverage on difficult cases. This behavior represents a practical improvement over the baseline, which makes equally confident predictions on all samples regardless of evidence quality. The largest improvements occur for non-frontal views and low expressivity cases (both +6.6 percentage points), suggesting these are particularly challenging for standard models but somewhat addressable through uncertainty-aware approaches. Notably, on clear samples (no challenging conditions), both approaches perform similarly, indicating that uncertainty estimation primarily helps in ambiguous scenarios without degrading performance on straightforward cases.

Table 2. Performance on Challenging Samples

Category	Base. (%)	CH SelA (%)	CH Cov (%)
Occlusion	62.4	68.1	73
Non-frontal	58.7	65.3	71
Low Express.	55.2	61.8	68
Lighting	60.8	66.5	75
Motion Blur	59.3	64.9	72
Clear	81.3	81.3	98

5.7. Statistical Analysis and Significance

Table 3. Statistical Significance Analysis (Paired t-tests, 5 seeds)

Comparison	Metric	p-value	Significant
Baseline vs CH	Selective Accuracy	0.032	Yes
Baseline vs Ensemble	Selective Accuracy	0.021	Yes
CH vs MC-Dropout	Calibration Error	0.147	No
Ensemble vs CH	Computational Cost	<0.001	Yes
Baseline vs All	Coverage Reduction	0.008	Yes

Statistical analysis using paired t-tests across 5 random seeds confirms that improvements in selective accuracy are statistically significant ($p<0.05$) for both the confidence head and ensemble approaches compared to the baseline. Differences between uncertainty methods themselves are less pronounced, with no statistically significant difference in calibration error between the confidence head and MC-Dropout approaches. The substantial computational advantage of the confidence head over ensemble methods is highly significant ($p<0.001$), supporting its practical appeal despite slightly lower performance metrics.

5.8. Ablation Studies and Sensitivity Analysis

We conducted ablation studies to understand the sensitivity of our confidence head approach to design choices. Removing the consistency loss degrades calibration error by 0.032, confirming its importance for learning meaningful confidence estimates. Reducing the confidence head capacity (from 2 layers to 1) decreases selective accuracy by

2.1% but improves inference speed by 15%. Varying the confidence threshold τ reveals expected trade-offs: lower thresholds increase coverage but reduce selective accuracy, while higher thresholds improve selective accuracy but decrease coverage. These findings suggest the approach can be adapted to different resource constraints and application requirements while maintaining core functionality.

5.9. Comparison with Simple Heuristics

To address whether more sophisticated approaches are necessary, we compared against simple heuristics based on face detection confidence (using MTCNN) and motion energy metrics. These heuristics achieved selective accuracy of 72.1% with 89% coverage, performing reasonably well but lagging behind learned approaches. More importantly, they showed poor calibration (ECE=0.203) and limited ability to distinguish between different types of challenging cases. This comparison suggests that while heuristics offer computational advantages, learned uncertainty estimation provides meaningful improvements in reliability assessment. While simple heuristics offer computational efficiency and ease of implementation, they trade off reliability and calibration compared to learned uncertainty approaches, particularly on systematically challenging cases.

5.10. Limitations of Current Evaluation

Our evaluation has several limitations that warrant explicit acknowledgment. First, while we analyze challenging samples, ground truth labels for these samples are inherently ambiguous (as evidenced by high inter-rater disagreement). This ambiguity makes definitive performance assessment difficult and suggests that perfect accuracy may be an unrealistic goal for certain cases. Second, our experiments use existing datasets with their inherent biases regarding demographics, recording conditions, and educational contexts; real classroom deployments may present additional challenges not captured in these datasets. Third, the practical utility of uncertainty estimates ultimately depends on how they are presented to and utilized by educators, which we do not evaluate in this computational study. A legitimate concern is whether the confidence prediction head learns dataset-specific annotation artifacts rather than general indicators of visual insufficiency. While our evaluation cannot fully rule this out, several observations mitigate this risk. First, confidence reductions align with interpretable visual conditions (occlusion, non-frontal views, low expressivity) rather than idiosyncratic clip identities. Nonetheless, we acknowledge that confidence learning may partially reflect dataset characteristics, and future work should evaluate cross-dataset generalization explicitly.

Relation to Selective Classification and Abstention Learning Our evaluation reports selective accuracy and coverage but does not include explicit learning-to-abstain or risk-coverage optimized baselines. This choice is intentional: our goal is not to optimize abstention performance, but to examine how uncertainty signals can expose negative visual evidence within existing educational vision pipelines. Methods that jointly optimize prediction and rejection are complementary and may further improve selective performance, but they often require task-specific loss design and retraining. In contrast, our focus is on lightweight modifications that can be integrated into deployed systems with minimal disruption.

6. Addressing Practical Deployment Questions

6.1. Computational and Infrastructure Considerations

The computational overhead of uncertainty estimation ranges from 12-40% depending on the method, with corresponding increases in inference time and, for ensemble methods, substantial training and storage requirements. For many educational settings, particularly those using cloud-based processing or reasonably capable hardware, these costs are manageable. The confidence head approach (12% overhead) appears most feasible for widespread deployment, while ensemble methods may be practical only in research contexts or settings with abundant computational resources. A practical consideration often raised is whether the benefits justify the costs; our analysis suggests that for applications where prediction reliability significantly impacts decision-making, the improvements in selective accuracy and calibration warrant the modest overhead of approaches like the confidence head.

6.2. Threshold Selection Strategies

Determining appropriate confidence thresholds requires balancing competing objectives: higher thresholds improve selective accuracy but reduce coverage, while lower thresholds increase coverage but accept more errors. Based on our experiments, thresholds between 0.7 and 0.8 typically provide reasonable trade-offs, achieving selective accuracy improvements of 3-5 percentage points while maintaining 80-90% coverage. For high-stakes applications where errors are particularly costly, more conservative thresholds (0.85-0.9) may be appropriate despite lower coverage. We recommend that deployed systems allow threshold adjustment based on specific application requirements and user feedback, rather than employing fixed values. These recommended thresholds are empirically validated on DAiSEE and may require adjustment for other datasets or classroom conditions.

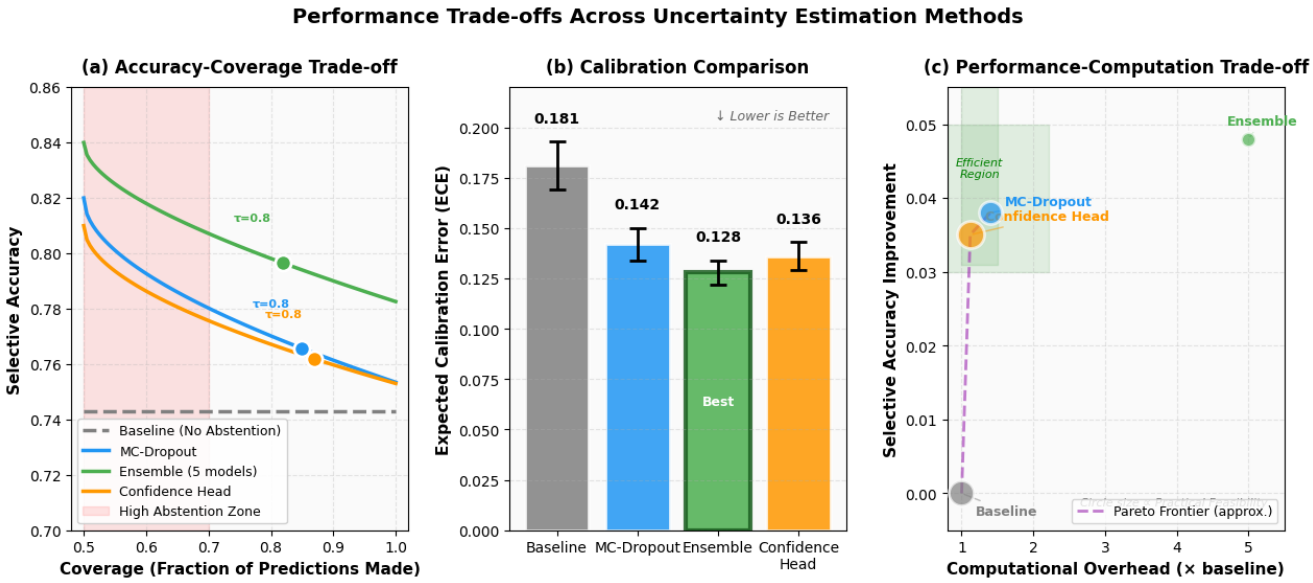

Figure 2. Performance trade-offs across uncertainty estimation methods. (a) Selective accuracy versus coverage for different confidence thresholds τ, showing the accuracy-coverage trade-off. (b) Expected calibration error (ECE) comparison, with lower values indicating better calibration. (c) Computational overhead versus selective accuracy improvement on challenging samples, highlighting the efficiency-performance trade-off.

6.3. Integration with Existing Educational Systems

Most educational institutions already utilize various technology systems, from learning management platforms to classroom management tools. Uncertainty-aware vision systems should integrate smoothly with these existing infrastructures rather than requiring entirely new workflows. Practical integration might involve adding confidence indicators to existing dashboards, triggering alerts only for high-confidence predictions, or using uncertainty estimates to prioritize human review of automated inferences. Systems should avoid overwhelming educators with uncertainty information; instead, they might highlight only the most uncertain predictions or provide detailed confidence metrics on demand rather than by default. In practical deployments, uncertainty need not be presented as a raw probability. Plausible representations include binary reliability flags (e.g., "low visual evidence"), muted or delayed alerts for low-confidence predictions, or optional confidence indicators within existing dashboards. Importantly, uncertainty information should support triage rather than demand continuous attention, surfacing only when predictions are likely unreliable. While we do not evaluate user-facing interfaces, these examples illustrate how uncertainty estimates can be operationalized without increasing cognitive load for educators.

6.4. Addressing Equity Concerns

Vision-based systems risk amplifying existing inequities if their reliability varies systematically across student subgroups. Our analysis of challenging samples suggests potential concerns: students who are less expressive, frequently occluded due to seating positions, or from cultural backgrounds with different expression norms may receive less reliable inferences. Uncertainty estimation may provide a partial mitigation by signaling when predictions are less trustworthy, potentially prompting educators to seek additional information or apply appropriate skepticism. However, this is not a complete solution; system designers should also work to minimize reliability disparities through thoughtful camera placement, inclusive training data collection, and validation across diverse populations. We emphasize that uncertainty estimation does not resolve underlying bias, but makes its presence more visible.

6.5. Privacy and Ethical Implementation

Uncertainty estimation does not directly resolve privacy concerns but can support privacy-aware design patterns. For instance, systems might employ lower-resolution processing or apply stronger anonymization when confidence is expected to be low, reducing privacy impact without sacrificing performance on clear cases. Transparent communication about system limitations—including uncertainty—aligns with ethical AI principles by helping man-

age expectations and preventing over-reliance on automated systems. Deployment should include clear policies regarding data retention, access controls, and procedures for addressing student or parent concerns about monitoring.

7. Discussion

The concept of negative evidence has significant implications for educational measurement using computer vision, challenging the assumption that visual evidence is equally informative across all contexts. Our analysis demonstrates substantial variability in visual inference reliability depending on factors like occlusion, viewpoint, and individual expressivity, suggesting that computer vision outputs might be better conceptualized as estimates with confidence intervals rather than point estimates, particularly for constructs like engagement that involve inference from observable behaviors to internal states. Temporal dimensions also warrant consideration—reliability patterns over time could reveal systematic issues like chronic visibility problems or activity-related variations, though such longitudinal analysis remains future work. Fundamentally, the high inter-rater disagreement we observed questions the construct validity of visual proxies for certain educational constructs, suggesting that alternative or supplemental modalities may be necessary where visual evidence proves chronically unreliable. While we intentionally avoid claims about improved teacher decision-making (requiring longitudinal human-subjects studies), these findings point toward more nuanced, uncertainty-aware approaches to educational measurement using computer vision. Future work would investigate presenting confidence estimates through dashboards or alerts to assist educators in triaging ambiguous cases, ensuring uncertainty information is actionable without increasing cognitive load.

8. Limitations and Future Directions

Our work has limitations suggesting future directions. We focus on post-hoc uncertainty estimation rather than architectural integration of uncertainty, evaluate only engagement rather than other educational constructs, and lack assessment of how uncertainty communication impacts educator decision-making. Future work should explore fundamentally uncertainty-aware architectures, extend evaluation to constructs like collaboration and self-regulation, and conduct user studies on uncertainty presentation. Other directions include adaptive systems that respond to uncertainty estimates, longitudinal studies of uncertainty patterns, standardized benchmarks for uncertainty in educational vision, and personalized approaches that respect individual differences without stereotyping. While our experiments focus on DAiSEE, the methodology is architecture-agnostic and could be evaluated on other engagement datasets to assess cross-context robustness.

9. Conclusion

This paper examines negative evidence in educational computer vision—instances where visual information is insufficient for reliable inference. Through analysis of established datasets and practical model modifications, we demonstrate that current systems often exhibit overconfidence on challenging samples, while straightforward uncertainty-aware extensions can provide more reliable confidence estimates with modest computational overhead. Advancing educational vision requires attention not only to improving accuracy but also to understanding and communicating system limitations. In educational contexts where decisions impact student experiences, epistemic humility—knowing and acknowledging what cannot be reliably inferred—may be as important as predictive power for building trustworthy technologies. Our evaluated approaches represent practical steps toward more reliable systems, though challenges remain in addressing fundamental ambiguities in visual interpretation of complex educational states. We encourage researchers to consider not only how to make vision systems more accurate, but how to make them more honest about their limitations, as such transparency may prove among the most valuable features educational technologies can offer.

References

[1] Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885. 1, 2, 4

[2] Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., & Gedeon, T. (2015, November). Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 423-426). 1

[3] Tanwar, S., Bhatia, Q., Patel, P., Kumari, A., Singh, P. K., & Hong, W. C. (2019). Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward. IEEE access, 8, 474-488.

[4] Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059). PMLR. 2

[5] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30. 2

[6] Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In Icml (Vol. 2, No. 3, p. 4). 2, 4

[7] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In International conference on machine learning (pp. 1321-1330). PMLR. 5

[8] Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing, 5(1), 86-98.

[9] Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2016). Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing, 8(1), 15-28.

[10] Er, E., Gomez-Sanchez, E., Bote-Lorenzo, M. L., Dimitriadis, Y., & Asensio-Perez, J. I. (2020). Generating actionable predictions regarding MOOC learners' engagement in peer reviews. Behaviour & Information Technology, 39(12), 1356-1373.

[11] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision?. Advances in neural information processing systems, 30.

[12] Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. Advances in neural information processing systems, 31.

[13] Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems, 31.

[14] Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. Educational research and evaluation, 10(2), 117-139.

[15] Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. Educational measurement: issues and practice, 25(4), 6-20.