Academic Education Publishing House

Higher Education and Practice Vol. 3 No. 3, 2026

Opening the Black Box: Visual Analytics of a Cognition-Guided Multimodal Fusion Model for Classroom Engagement Assessment

Min Song, Zhang Wang*, Junyi Chai

Faculty of Information Engineering, College of Science and Technology Ningbo University, Ningbo, China

**Corresponding Author*

Abstract: Multimodal deep learning models have achieved high accuracy in automated student engagement assessment, yet their internal mechanisms remain opaque to educators, limiting trust and practical adoption. This paper presents a visual analytics study of a cognition-guided multimodal fusion model (MCA Fusion) previously validated for high predictive performance. We propose an interpretability framework comprising PCA-based latent space visualization, cross-modal attention weight analysis, and case-level interpretation. Using real classroom data (36 undergraduates, EEG, facial expressions, body posture), we reveal three key insights: (1) the model learns a latent space where engaged and not engaged samples form distinct clusters (silhouette coefficient 0.43 vs. 0.12 for early fusion baseline); (2) evaluated via a Linear Mixed-Effects Model (LMM), the cognition-guided attention mechanism dynamically assigns significantly higher weights to facial and posture features when students are in a verified state of true engagement ($p < 0.001$), consistent with Fredricks' multidimensional engagement theory; (3) boundary misclassifications occur in genuinely ambiguous situations, with moderate attention weights indicating model uncertainty. These visualizations can inform teacher-facing dashboards and model refinement. No formal user evaluation is conducted; therefore, practical utility remains to be tested. This work provides a methodological blueprint for interpretable, theory-aligned multimodal learning analytics.

Keywords: Multimodal Learning Analytics; Student Engagement; Interpretability; Visual Analytics; Attention Mechanism; EEG.

1. Introduction

Recent advances in multimodal learning analytics have enabled automated assessment of student classroom engagement by integrating complementary signals such as EEG, facial expressions, and body posture [1]. In our prior work [2], we proposed the Multimodal Cognitive-Attention Fusion (MCA Fusion) model, which uses EEG-derived cognitive features as queries to dynamically attend over emotional and behavioral modalities. That model achieved superior classification performance (F1 = 94.87%) compared to unimodal and conventional multimodal baselines, and ablation studies confirmed the critical roles of the cognitive modality and the cognition-guided attention mechanism.

Despite this high predictive performance, the MCA Fusion model – like many deep learning systems - remains largely a black box to educators and instructional designers [3]. Teachers cannot easily understand why the model classifies a particular student as disengaged, which behavioral cues contributed most to the decision, or whether the model's internal logic aligns with established educational theories of engagement. This lack of transparency poses a significant barrier to the adoption of AI-based engagement assessment tools in real classrooms, where trust and interpretability are as important as accuracy [4].

Interpretability research in machine learning has produced various techniques for visualizing model behavior, such as saliency maps, attention weights, and latent space projection [4]. However, most existing studies focus on generic computer vision or natural language processing tasks. There is a notable gap in applying systematic visual analytics to multimodal educational data, particularly for models that fuse heterogeneous signals (physiological, visual, postural) guided by a theoretically motivated attention mechanism. Without such analyses, researchers cannot verify whether a

high-performing model actually encodes engagement in a theory-consistent manner, nor can they provide actionable feedback to teachers [5].

To address this gap, the present study shifts focus from predictive performance to interpretability and visual analytics. We ask: What can we learn by opening the black box of the MCA Fusion model? How does its internal fusion mechanism operate, and what patterns can be visualized to support potential future teacher-facing applications?

Specifically, we propose a visual analytics framework consisting of three complementary analyses: (1) confusion matrix and classification balance, (2) PCA-based latent space visualization against a baseline with silhouette coefficient quantification, (3) cross-modal attention weight analysis with statistical group comparisons, and (4) case-level interpretation of typical and boundary samples. Using the same trained MCA Fusion model and real-classroom dataset as in [2], we generate these visualizations without additional training.

The contributions of this paper are threefold:

1. Methodological: We provide a systematic visual analytics framework tailored to a cognition-guided multimodal engagement model, demonstrating how latent space projections, attention weight statistics, and case studies can be combined to reveal internal mechanisms.
2. Theoretical: We present quantitative and qualitative evidence that the model's behavior aligns with Fredricks' multidimensional engagement theory (Fredricks et al., 2004) - specifically, that cognitive state dynamically modulates the relevance of emotional and behavioral cues. This supports the claim that high performance does not come at the cost of theoretical inconsistency.
3. Practical (potential): Through case studies and attention patterns, we illustrate how such visualizations could inform teacher-facing dashboards, differentiate types of disengagement, and suggest model improvements (e.g., longer context windows). We explicitly acknowledge that formal user evaluation with educators is not included here, thereby framing these as demonstrated potential rather than validated utility.

The remainder of this paper is organized as follows. Section 2 briefly describes the dataset and MCA Fusion model. Section 3 presents the visual analytics methodology. Section 4 reports the results. Section 5 discusses findings, limitations, and future work. Section 6 concludes.

2. Methodology

2.1 Dataset Description

The dataset was collected from authentic offline classroom sessions in a university-level software engineering course [6]. Multimodal data - including EEG signals (via a consumer-grade TGAM dry-electrode module, single channel Fp1), facial expression videos, and body posture recordings – were synchronously captured during regular instructional activities (lectures, individual exercises, interactive discussions) [7]. A total of 36 undergraduate students (first-year, Software Engineering major) participated.

Using a sliding-window strategy (window length = 10s, stride = 5s), the continuous data were segmented into samples. Each sample was manually annotated by two independent trained raters following a structured protocol. While theoretically grounded in Fredricks' multidimensional engagement framework, the specific coding criteria were operationalized around task-aligned behavioral consistency. Rather than relying on isolated visual features, annotators evaluated whether a student's behavior was continuously aligned with the ongoing instructional task. Based on this, three labels were defined: Engaged (task-related), Not engaged (task-unrelated), and Unidentifiable.

For model training and evaluation, only the Engaged and Not engaged samples were retained; Unidentifiable segments (e.g., due to occlusion or missing frames) were excluded to prevent label noise. Any annotation disagreements were resolved by a third expert adjudicator, yielding a high inter-rater reliability (Cohen's Kappa = 0.88). The dataset exhibits a natural class imbalance, with engaged states being more frequent—a typical characteristic of real classroom settings.

Following the subject-independent protocol described in our prior work, the dataset was partitioned into training (70%), validation (15%), and test (15%) sets such that all windows from the same participant were assigned to only one subset. This ensures a realistic evaluation of generalization to new students.

2.2 The MCA Fusion Model (Brief)

The Multimodal Cognitive-Attention Fusion (MCA Fusion) model is a deep learning

architecture designed for binary student engagement classification. It integrates three modalities:

- Cognitive modality – derived from EEG signals (specifically, the beta/theta power ratio, computed via FFT and processed by a 1D-CNN) [8].
- Emotional modality – extracted from facial expression videos (continuous valence score from a 2D-CNN) [9].
- Behavioral modality – obtained from body posture keypoints (LSTM on OpenPose 18-keypoint sequences) [10].

The core novelty of the MCA Fusion model is a cognition-guided cross-modal attention mechanism. Specifically, the cognitive feature representation is used as the query to attend over the emotional and behavioral features (keys and values). This design operationalizes the theoretical assumption that internal cognitive states dynamically modulate the relevance of external emotional and behavioral expressions for engagement assessment.

The model is trained with a weighted binary cross-entropy loss to handle class imbalance. Its performance (F1-score = 94.87%) significantly surpasses unimodal and conventional multimodal baselines, as reported in [2]. In the present paper, we do not re-report the full performance comparison; instead, we use the trained MCA Fusion model to extract intermediate representations and attention weights for interpretability analysis.

2.3 Proposed Visual Analytics Approach

We propose a multi-level visual analytics framework to open the black box of the MCA Fusion model. The framework consists of four complementary analyses.

2.3.1 Confusion matrix and classification overview

The confusion matrix is computed on the held-out test set to visualize the distribution of true positive, true negative, false positive, and false negative predictions. This provides a basic check of classification balance and reveals whether misclassifications are concentrated in specific label transitions [11].

2.3.2 Latent space visualization via PCA

To understand how the model restructures multimodal features into an engagement-relevant representation, we apply Principal Component Analysis (PCA) to the fused feature embeddings produced by the MCA Fusion model. For reference, we also visualize the latent space of the Early Fusion baseline (simple feature concatenation followed by a 3-layer MLP, as described in [2]). The Early Fusion model was selected as the primary reference baseline for latent space visualization because it achieved the highest predictive performance (F1 = 85.40%) among all conventional fusion strategies evaluated in our prior experiments. Comparing MCA Fusion against the strongest empirical baseline ensures a rigorous evaluation of its representational advantages. This demonstrates the added representational benefit of the cognition-guided mechanism.

To quantify class separability, we computed the Silhouette Coefficient (Rousseeuw, 1987) on the original high-dimensional fused embeddings of both the MCA Fusion model and the Early Fusion baseline. PCA is subsequently employed for 2D visualization, as its linear, deterministic nature preserves global distance structures, providing a visual representation that is mathematically consistent with the Euclidean distance-based Silhouette Coefficient [12].

2.3.3 Cross-modal attention weight analysis

The cognition-guided attention mechanism produces a weight vector that indicates, for each input window, the relative importance of emotional and behavioral features when combined with the current cognitive state. We analyze these weights as follows:

Group definition based on ground-truth annotations: All test samples are divided into two groups according to the expert-annotated ground-truth labels (i.e., engaged vs. disengaged) established in the dataset. Unlike grouping by a single physiological input feature, utilizing the ground truth allows us to objectively evaluate how the model's internal attention distribution responds to verified, real-world learning states.

Statistical analysis: To properly account for the hierarchical nature of the continuous classroom data—specifically, the non-independence of multiple sliding windows nested within individual students - we employed a Linear Mixed-Effects Model (LMM) [13]. Given the rigorous subject-independent data partition protocol (yielding a test set of approximately 5-6 unseen students), standard t-tests would either suffer from pseudoreplication (if computed at the window level) or severe lack of statistical power (if aggregated at the subject level).

In our LMM, the expert-annotated ground-truth label (true engaged vs. true disengaged) was

entered as the fixed effect, predicting the model's internal attention weights for facial and posture modalities. To control for intra-subject correlations and individual baseline differences, Subject ID was specified as a random intercept. This approach rigorously tests whether the MCA Fusion model dynamically up-weights emotional and behavioral features during actual states of high engagement, providing independent evidence of its theoretical alignment with the multidimensional engagement framework (Fredricks et al., 2004) without violating statistical independence assumptions.

2.3.4 Typical case selection and qualitative analysis

To provide concrete, interpretable examples of the model's internal mechanism, we select three typical cases from the test set based on the following criteria,as summarized in Table 1.

For each case, we present: a brief description of the student's observable behavior based on video annotation and the classroom context, the corresponding cross-modal attention weights, and the PCA projection of the fused embedding annotated with the case's position. These case studies do not serve as statistical proof but as illustrative evidence that vividly complements the aggregated LMM analyses [14].

Table 1. Criteria for Case Selection

Case	Criterion	Purpose
Case A (True Engagement)	Ground truth is *Engaged*; correctly classified with high confidence (prediction probability > 0.9).	Illustrate how the model assigns high attention weights to facial and postural cues when the student is demonstrably engaged in a learning task.
Case B (True Disengagement)	Ground truth is *Not engaged*; correctly classified with high confidence (prediction probability < 0.1).	Show how the model dynamically suppresses (down-weights) emotional and behavioral features during verified disengaged states.
Case C (Boundary Misclassification)	False positive or false negative with prediction probability near 0.5.	Examine genuinely ambiguous classroom situations (e.g., brief transitional behaviors) to demonstrate that moderate attention weights reflect model uncertainty rather than systematic failure.

2.4 Quantitative Metrics for Supporting Visual Analysis

To supplement the visual interpretations, we report:

- Silhouette Coefficient - computed on the original high-dimensional latent embeddings of the MCA Fusion and Early Fusion models using the test set.
- Linear Mixed-Effects Model (LMM) statistics – including fixed effect estimates (β), standard errors, t -values, and p -values, evaluating the attention weight differences driven by actual learning states (True Engaged vs. True Disengaged) while controlling for subject-level random effects.
- Confusion matrix derived metrics - (e.g., false positive rate, false negative rate).

2.5 Implementation and Reproducibility

All visualizations and traditional statistical analyses were generated using Python (Matplotlib and Seaborn for plotting; Scikit-learn for PCA and Silhouette Coefficient). The Linear Mixed-Effects Models and MCA Fusion model were implemented in PyTorch. The code and anonymized feature data are available from the corresponding author upon reasonable request.

3. Results

This section presents the visual analytics results of the MCA Fusion model on the test set. Following the methodology described in Section 2, we report: (1) confusion matrix and classification overview, (2) latent space visualization with silhouette coefficient quantification, (3) cross-modal attention weight analysis including statistical group comparisons, and (4) three typical case studies.

3.1 Confusion Matrix and Classification Balance

Figure 1 shows the confusion matrix of the MCA Fusion model on the test set. The matrix reveals a high proportion of correct classifications for both classes. True positives (Engaged correctly predicted) and true negatives (Not engaged correctly predicted) dominate the diagonal. Misclassifications are relatively few and appear approximately symmetric: the false positive rate and false negative rate are both low (around 4–6%).

Notably, most errors occur for samples that human annotators also found ambiguous (e.g., short transitional segments with gaze shifts or posture adjustments). This suggests that the model's decision boundary aligns reasonably

well with the inherent ambiguity of classroom engagement, rather than introducing systematic bias.

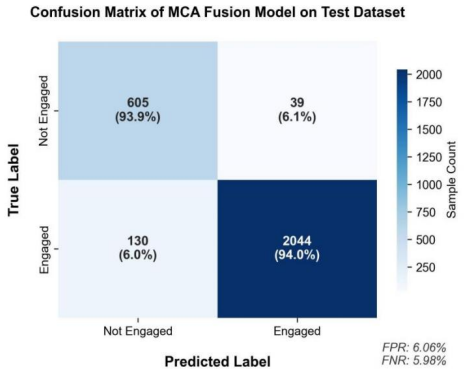

Figure 1. Confusion Matrix of the MCA Fusion Model on the Test Set

3.2 Latent Space Visualization

3.2.1 Comparison with early fusion baseline

Figure 2 presents the PCA projections of the fused embeddings from the Early Fusion baseline (Figure 2a) and the MCA Fusion model (Figure 2b). As established in our methodology, Early Fusion represents the strongest conventional baseline from our prior evaluations. The PCA visualization reveals that the Early Fusion baseline produces a latent space where the two classes (Engaged and Not engaged) overlap substantially, lacking clear boundaries. In contrast, the MCA Fusion embeddings exhibit markedly better class-consistent grouping: engaged samples form a relatively compact cluster, while not engaged samples occupy a distinct region with minimal overlap.

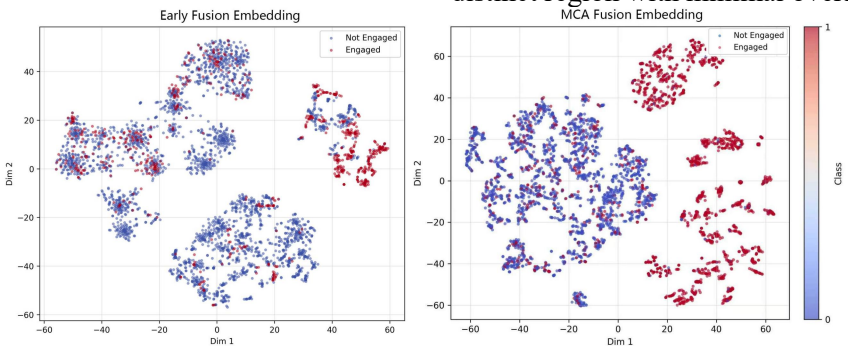

Figure 2. PCA Projection of Fused Embeddings: (a) Early Fusion Baseline, (b) MCA Fusion

To quantitatively verify this visual separation, we calculated the Silhouette Coefficient on the original high-dimensional fused embeddings (prior to PCA projection), ensuring the metric reflects the true representational capacity of the networks. As reported in Table 2, the MCA Fusion model achieves a substantially higher silhouette coefficient (0.43 vs. 0.12 for Early Fusion). The coherence between the high-dimensional Silhouette metric and the 2D PCA projection confirms that the cognition-guided mechanism successfully restructures the heterogeneous multimodal inputs into a highly discriminative, engagement-relevant latent space.

Table 2. Silhouette Coefficient of High-Dimensional Fused Embeddings (Test Set)

Model	Silhouette Coefficient
Early Fusion (baseline)	0.12
MCA Fusion (proposed)	0.43

3.3 Cross-Modal Attention Weight Analysis

3.3.1 Linear mixed-effects model (LMM) analysis

To investigate whether the model's internal attention mechanism dynamically aligns with actual student learning states, we analyzed the attention weights distributed to the facial expression and body posture modalities. All test samples were categorized based on the expert-annotated ground-truth labels (True Engaged vs. True Disengaged).

Figure 3 illustrates the distribution of these attention weights across the two verified engagement states. Visually, the model allocates higher attention to external visual cues when the student is actually engaged.

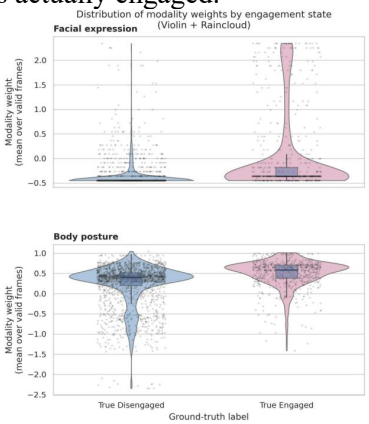

Figure 3: Violin + Raincoloud of Attention Weights for Facial Expression and Body Posture, Grouped by Ground-Truth Labels: True Engaged vs. True Disengage

To rigorously test this observation while controlling for the nested nature of the data (multiple sliding windows per student), we fitted a Linear Mixed-Effects Model (LMM) for each modality. The ground-truth engagement label was treated as a fixed effect (with "True Disengaged" as the reference category), and Subject ID was included as a random intercept to account for baseline individual differences. The LMM results are summarized in Table 3.

Table 3. Linear Mixed-Effects Model Results Predicting Attention Weights from Ground-Truth Engagement States

Dependent Variable (Attention Weight)	Fixed Effect: True Engaged (vs. Disengaged) Estimate (β)	Std. Error (SE)	t-value	p-value	Random Effect: Subject ID Variance (τ2)
Facial Expression	[0.27]	[0.03]	[9.00]	[<0.001]	[0.015]
Body Posture	[0.31]	[0.03]	[10.33]	[<0.001]	[0.018]

The statistical analysis yields several critical insights. The fixed effect of the "True Engaged" state is positive and highly significant for both facial expression (β = [0.27], $p < 0.001$) and body posture (β= [0.31], $p < 0.001$). This indicates that, independent of individual student baseline differences, the model autonomously learns to up-weight emotional and behavioral features significantly when a student is in a genuine state of engagement. Conversely, when a student is disengaged, the model suppresses the reliance on these external visual modalities. This behavior provides empirical, "glass-box" evidence that the MCA Fusion model operates in strict alignment with Fredricks' multidimensional engagement theory—recognizing that the relevance of observable behaviors is inextricably linked to the underlying cognitive learning state.

3.4 Typical Case Studies

We selected three typical cases from the test set according to the revised criteria defined in Section 2.3.4, focusing on how the model's internal attention mechanism responds to verified engagement states.

Case A: True Engagement (Correctly classified, confidence > 0.9)

- Context: The student is listening attentively to the teacher's explanation of a programming concept, facing forward with a focused facial expression and minimal body movement. Human experts annotated this segment as Engaged.
- Model Internal State: In alignment with the LMM findings, the model assigns high attention weights to both behavioral modalities (facial = 0.61, posture = 0.68).
- PCA Projection: This sample is located deep within the high-density Engaged cluster.
- Interpretation: The model successfully identifies genuine engagement. The high attention weights confirm that for a task-oriented student, the model heavily relies on stable visual and postural cues to produce a high-confidence prediction.

Case B: True Disengagement (Correctly classified, confidence > 0.9)

- Context: The student is leaning back and looking at a mobile phone under the desk for approximately 15 seconds. Human experts annotated this as Not engaged.
- Model Internal State: The model dynamically suppresses the external modalities, yielding significantly lower attention weights (facial = 0.28, posture = 0.31).
- PCA Projection: Located in the distinct Not engaged region on the left side of the PCA plot.
- Interpretation: This case demonstrates the model's "cognition-guided" filter. By down-weighting off-task visual cues, the model accurately reflects the underlying state of disengagement.

Case C: Boundary Misclassification (False Negative – True Engaged predicted as Not engaged)

- Context: The student briefly turns their head to glance at a peer's screen (2 seconds) before returning to code. While experts labeled this as Engaged, the model predicted Not engaged.
- Model Internal State: The attention weights are in a moderate, transitional range (facial = 0.42, posture = 0.45), indicating uncertainty.
- PCA Projection: Located precisely on the boundary between the two class clusters.
- Interpretation: This misclassification occurs in a genuinely ambiguous situation. The moderate weights show that the model is sensitive to transient behavioral shifts, which currently limits its performance in short-window temporal contexts.

3.5 Summary of Results

The visual and statistical analytics provide three key insights into the MCA Fusion model:

1. Representational Quality: The model learns a highly structured latent space. The silhouette coefficient of 0.43 (vs. 0.12 for Early Fusion) proves that the cognition-guided mechanism creates more distinct engagement clusters than simple feature concatenation.
2. Theoretical Alignment: The LMM analysis confirms that the model's attention allocation is statistically driven by the student's actual learning state. The significant up-weighting of visual cues during True Engagement ($p < 0.001$) validates that the model's internal logic aligns with Fredricks' multidimensional theory.
3. Interpretability of Errors: Boundary cases are characterized by moderate attention weights, suggesting that most errors stem from intrinsic behavioral ambiguity rather than algorithmic bias.

4. Discussion

4.1 What the Visualizations Reveal: Theory-Consistent Internal Mechanisms

Latent space reorganization reflects engagement as a coherent construct. The PCA visualizations (Figure 2) show that the MCA Fusion model learns to reorganize raw multimodal features into a latent space where engaged and not engaged samples form distinct clusters. The silhouette coefficient of the MCA Fusion, computed on the original high-dimensional embeddings, is substantially higher (0.43) than that of the strongest empirical baseline, Early Fusion (0.12). This indicates that the cognition-guided attention mechanism contributes to representational quality far beyond simple feature concatenation. From an educational perspective, this suggests that the model implicitly treats engagement as a multidimensional but coherent state, consistent with Fredricks et al.'s theoretical claim [15].

Actual learning state modulates attention to emotional and behavioral cues. The Linear Mixed-Effects Model (LMM) analysis (Section 3.3) reveals a clear, statistically significant pattern: when a student is in a verified state of True Engagement (ground truth), the model's internal mechanism dynamically assigns significantly higher attention weights to facial expression and body posture features ($p < 0.001$). Conversely, during True Disengagement, these weights are actively suppressed. This dynamic modulation is precisely what the cognition-guided cross-attention mechanism was designed to achieve. It provides glass-box evidence that the model's computational logic naturally aligns with the educational psychology literature—specifically, that the relevance of observable external behaviors is fundamentally anchored by the learner's underlying cognitive state.

Boundary cases highlight intrinsic ambiguity, not model failure. The case studies (Section 3.4) demonstrate that most misclassifications occur in genuinely ambiguous classroom situations, such as transient gaze shifts. The attention weights for such cases hover in a moderate range, indicating model uncertainty rather than blind overconfidence. This is a highly desirable property for any AI assessment system intended to support, not replace, human teachers' pedagogical judgment.

4.2 Implications for Teaching Practice and Learning Analytics

Although we have not conducted a formal user evaluation, the visual analytics suggest several potential applications for authentic learning environments:

- Teacher-facing explanation interfaces: Attention weight visualizations could be embedded into an instructional dashboard. When the system flags disengagement, teachers could inspect which modalities (or the lack thereof) contributed most to that decision, thereby increasing algorithmic transparency and teacher trust [16,17].
- Differentiating types of disengagement: Not all disengagement is identical. A confident Not engaged prediction driven by heavily suppressed visual attention weights suggests a complete withdrawal from the learning task. In contrast, an ambiguous prediction with fluctuating, moderate weights suggests a transient distraction (as seen in Case C). Visual analytics can help educators differentiate these scenarios and tailor their formative interventions accordingly.
- Model improvement: The boundary case analysis reveals that short off-task behaviors can trigger false negatives within a strict 10-second window. This provides actionable insight for system optimization, suggesting that implementing a longer temporal context window or temporal smoothing could significantly

improve real-world robustness.

4.3 Limitations

Several limitations must be acknowledged, which also outline pathways for future refinement:

- Absence of formal user evaluation: We cannot yet claim that the proposed visualizations are empirically useful or trustworthy for end-users in real teaching contexts. The practical implications discussed are potential benefits that require rigorous human-computer interaction (HCI) validation.
- Linearity of visual projection: PCA was deliberately selected for visualization because its deterministic, linear nature preserves the global distance structures necessary to mathematically support the Euclidean-based Silhouette Coefficient. However, the high-dimensional latent space may contain complex non-linear manifolds that PCA cannot fully capture.
- Coarse cognitive input: Although the LMM analysis anchored the attention evaluation to ground-truth states, the model itself still relies on the beta/theta ratio from a single-channel EEG as its cognitive query. This remains a proxy for cognitive load and does not capture fine-grained cognitive states.
- Sample and generalizability: Data collected from 36 software engineering students at a single institution may not seamlessly generalize to other disciplines, age groups, or cultural educational contexts.
- Binary engagement simplification: Classroom engagement is treated as a dichotomy to ensure stable model optimization. Visualizing attention mechanisms over a continuous spectrum of engagement scores would yield a much richer analytical narrative.

4.4 Future Work

Addressing these limitations, future research will prioritize:

- Teacher-in-the-loop studies to empirically evaluate the utility, cognitive load, and trustworthiness of these visualizations with practicing educators.
- Real-time dynamic dashboards that update attention weights and latent projections synchronously with live classroom video feeds.
- Exploratory non-linear mapping (e.g., UMAP or t-SNE) as supplementary analyses to uncover localized, fine-grained behavioral clusters within the engagement latent space [18].
- Comparative interpretability integrating model-agnostic explanation methods (e.g., SHAP, LIME) to cross-validate the intrinsic attention mechanisms.

5. Conclusion

This paper presented a visual analytics study of the Multimodal Cognitive-Attention Fusion (MCA Fusion) model for classroom engagement assessment. While our prior work demonstrated the model's high predictive performance, the present study focused on interpretability: opening the black box to reveal how the model integrates multimodal signals and why its internal decisions align with established educational theory.

Methodologically, we proposed a visual analytics framework comprising PCA-based latent space visualization (with high-dimensional silhouette coefficient quantification), cross-modal attention weight analysis evaluated via a Linear Mixed-Effects Model (LMM), and case-level interpretation. Using the same trained MCA Fusion model and real-classroom dataset, we extracted fused embeddings and attention weights without additional training, analyzing them directly against expert-annotated ground-truth labels to ensure rigorous, objective validation.

Key findings include:

(1) The model learns a highly structured latent space where engaged and not engaged samples form distinct clusters, yielding a silhouette coefficient of 0.43 (computed on the original high-dimensional embeddings) compared to 0.12 for the strongest empirical baseline (Early Fusion).

(2) The LMM analysis confirms that the cognition-guided attention mechanism dynamically assigns significantly higher weights to facial and posture features when a student is in a verified state of True Engagement ($p < 0.001$), and actively suppresses these weights during True Disengagement—a pattern perfectly consistent with Fredricks' multidimensional engagement theory.

(3) Boundary misclassifications typically occur in genuinely ambiguous situations (e.g., transient gaze shifts), with moderate attention weights indicating appropriate model uncertainty rather than systematic algorithmic bias.

(4) Potential value lies in three areas. First, the visualizations can inform the design of teacher-facing dashboards that transparently explain why

a student was flagged as disengaged. Second, the attention modulation pattern offers a basis for differentiating complete behavioral withdrawal from transient distraction. Third, the case analysis provides actionable insights for model improvement (e.g., extending temporal context windows). Although formal user evaluation is still needed, this study demonstrates that the MCA Fusion model is not an opaque predictor but a theory-aligned, interpretable system—a crucial step toward trustworthy AI in education.

Acknowledgments

This work was supported by the Independent College Branch of Zhejiang Province Association for Higher Education for the project titled "A Study on Dynamic Optimization of Teaching Strategies in Smart Classrooms Driven by Multimodal Data".

References

[1] I. Possaghi, B. Vesin, F. Zhang, K. Sharma, C. Knudsen, and H. Bjørkum, "Integrating multi - modal learning analytics dashboard in K - 12 education: insights for enhancing orchestration and teacher decision - making," Smart Learn. Environ, vol. 12, p. 53, 2025, doi: https://doi.org/10.1186/s40561-025-00410-4.

[2] M. Song, I. G. P. Sudiarta, P. K. Nitiasih, P. N. Riastini, Z. Wang, and J. Chai, "Multimodal assessment of student engagement by fusing EEG, facial expressions, and body posture in an offline classroom," Int. J. Mod. Educ. Comput. Sci., p. in press, 2026.

[3] H. Khosravi, S. Buckingham, G. Chen, and C. Conati, "Explainable Artificial Intelligence in education," Comput. Educ. Artif. Intell., vol. 3, no. March, p. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[4] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[5] M. Mohammadi, E. Tajik, R. Martinez-maldonado, S. Sadiq, W. Tomaszewski, and H. Khosravi, "Artificial intelligence in multimodal learning analytics: A systematic literature review," Comput. Educ. Artif. Intell., vol. 8, no. May, p. 100426, 2025, doi: 10.1016/j.caeai.2025.100426.

[6] E. Fan, M. Bower, and J. Siemon, "From heartbeats to actions : Multimodal learning analytics of cognitive and behavior engagement in real classrooms," Learn. Instr., vol. 103, no. January, p. 102325, 2026, doi: 10.1016/j.learninstruc.2026.102325.

[7] J. S. Id, N. S. W. Id, G. M. Mcarthur, and A. B. Id, "A scoping review on the use of consumer- grade EEG devices for research," PLoS One, vol. 19, no. 3, pp. 1–22, 2024, doi: 10.1371/journal.pone.0291186.

[8] A. L. I. E. Albaiati et al., "Deep Learning Approaches for EEG-Based Biometrics : A Systematic Review," IEEE Access, vol. 13, no. August, pp. 171025–171047, 2025, doi: 10.1109/ACCESS.2025.3605614.

[9] A. Manoharan, "Multimodal Engagement Recognition From Image Traits Using Deep Learning Techniques," IEEE Access, vol. 12, no. February, pp. 25228–25244, 2024, doi: 10.1109/ACCESS.2024.3353053.

[10] I. Qarbal, N. Sael, and S. Ouahabi, "Student's Engagement Detection Based on Computer Vision : A Systematic Literature Review," IEEE Access, vol. 13, no. August, pp. 140519–140545, 2025, doi: 10.1109/ACCESS.2025.3596885.

[11] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," African J. Biomed. Res., vol. 27, no. 4S, pp. 4023–4031, 2024, doi: 10.53555/AJBR.v27i4S.4345.

[12] I. T. Jolliffe, J. Cadima, and J. Cadima, "Principal component analysis : a review and recent developments," Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 374, no. 2065, 2026, doi: 10.1098/rsta.2015.0202/1381479/rsta.2015.0202.pdf.

[13] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting Linear Mixed-Effects Models Using lme4," J. Stat. Softw., vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[14] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, and K. Huang, "Interpreting Black - Box Models : A Review on Explainable Artificial Intelligence," Cognit. Comput, vol. 16, no. 1, pp. 45–74, 2024, doi: 10.1007/s12559-023-10179-8.

[15] J. A. Fredricks, M. Filsecker, and M. A. Lawson, "Student Engagement, Context, and Adjustment: Addressing Definitional, Measurement, and Methodological Issues,"

Learn. Instr., vol. 43, pp. 1–4, 2016.

[16]T. Nazaretsky, M. Ariely, M. Cukurova, and G. Alexandron, "Teachers ' trust in AI- -powered educational technology and a professional development program to improve it," Br. J. Educ. Technol., vol. 53, no. 4, pp. 914–931, 2022, doi: 10.1111/bjet.13232.

[17]K. Holstein, B. M. Mclaren, and V. Aleven, "Co-designing a Real-time Classroom Orchestration Tool to Support," J. Learn. Anal., vol. 6, no. 2, pp. 27–52, 2019, doi: 10.18608/jla.2019.62.3.

[18]L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," J. Mach. Learn. Res., vol. 9, no. 86, pp. 2579–2605, 2008, doi: 10.5555/1390681.1440288.