Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7, 1482-1495
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i7.8944
© 2025 by the author; licensee Learning Gate

Multimodal emotion recognition in children's online learning: Emotion monitoring and intervention strategy design

Jingru Li[1]*
[1]National Institute of Education, Nanyang Technological University, Singapore; JRLiedu@163.com (J.L.).

Abstract: The increasing prevalence of online education for children underscores the need for intelligent systems capable of recognizing and responding to learners' emotional states in real time. Emotional fluctuations, such as boredom, frustration, or confusion, have been empirically linked to cognitive disengagement and poor learning outcomes, particularly in younger learners. However, existing educational platforms lack reliable mechanisms for detecting such states and initiating timely pedagogical interventions. This study proposes a multimodal emotion recognition and intervention framework tailored for children's online learning environments. The proposed system integrates facial expression analysis, speech emotion recognition, and behavioral signal tracking to detect six core emotional states commonly observed in child learners. A hierarchical fusion architecture, combining CNN-LSTM visual encoders and Transformer-based cross-modal attention modules, enables robust emotion classification even under modality loss or environmental noise. In parallel, a rule-enhanced policy engine maps detected emotions to personalized intervention strategies, including task scaffolding, verbal encouragement, and content pacing adjustments. The framework is evaluated on a newly curated multimodal dataset of primary school students engaged in online learning tasks, demonstrating superior performance over unimodal and early-fusion baselines across multiple metrics. In real-world deployment trials, the system significantly improves learners' task completion rates and emotional stability. Furthermore, ablation studies and statistical significance tests confirm the contributions of each modality and fusion mechanism. The results suggest that incorporating multimodal affective computing into online learning platforms offers a promising pathway toward emotionally adaptive and child-centric digital education. This work contributes both a scalable technical solution and empirical evidence supporting the integration of emotional monitoring and intervention in intelligent tutoring systems.

Keywords: *Adaptive learning systems, Affective computing, Children, Educational intervention, Multimodal emotion recognition, Online learning.*

1. Introduction

The growing prevalence of online learning environments, particularly among children, has introduced new opportunities and challenges in the design of intelligent and emotionally responsive educational systems. While digital learning platforms facilitate flexibility, scalability, and individualized content delivery, they also reduce opportunities for spontaneous emotional interaction and non-verbal feedback, which are vital to cognitive and affective development in younger learners. Emotional engagement, long recognized as a key determinant of learning outcomes, is especially critical in childhood education, where attention, motivation, and comprehension are tightly coupled with emotional states such as curiosity, boredom, or frustration [1, 2].

Recent advances in multimodal emotion recognition (MER) provide a promising avenue for mitigating the limitations of emotion-insensitive digital learning platforms. MER leverages signals from multiple sources, such as facial expressions, vocal intonations, body movements, and contextual

History: Received: 15 May 2025; Revised: 17 June 2025; Accepted: 20 June 2025; Published: 18 July 2025
* Correspondence: JRLiedu@163.com

behavior, to infer the learner's emotional state with greater accuracy and reliability [3]. Deep learning-based MER systems, particularly those utilizing hybrid architectures and fusion techniques, have demonstrated state-of-the-art performance across a range of affective computing tasks [4]. These methods include model-level and attention-based fusion strategies that combine audio-visual modalities to capture complex emotional dynamics that are otherwise undetectable in unimodal systems [5, 6].

Despite these technological advances, the application of MER in child-centered online education remains underexplored. Most existing MER frameworks are optimized for adult users or generalized multimedia analysis, and fail to account for the developmental, linguistic, and expressive variability found in child populations. Moreover, emotion recognition in children presents unique challenges due to rapid affective transitions, inconsistent expressions, and age-dependent behavioral cues. There is thus a growing need to design educational technologies that incorporate child-specific emotional models and learning-responsive strategies.

Simultaneously, the educational domain is witnessing increasing interest in artificial intelligence-driven personalization, including affect-aware interventions that adapt instructional content based on learners' cognitive and emotional needs [7]. Studies have shown that such personalized, emotionally adaptive learning systems can enhance engagement, reduce dropout, and improve learning performance in both synchronous and asynchronous settings [8, 9]. In particular, emotional states like anxiety or confusion, if left unaddressed, can negatively influence motivation and lead to disengagement in online classrooms [10]. These concerns are especially urgent in the context of young learners, whose long-term relationship with learning can be shaped by early online experiences [11].

This study aims to address these gaps by proposing a deep learning-based multimodal emotion recognition and intervention framework specifically tailored for children's online learning. By integrating audio, visual, and behavioral signals within a unified hierarchical architecture, and linking emotion detection to a responsive intervention engine, we construct an end-to-end system capable of monitoring emotional fluctuations and triggering real-time pedagogical strategies. This research contributes both a novel technical framework and empirical validation, offering insights into how emotion-aware online systems can foster more engaging, personalized, and developmentally appropriate digital learning environments for children.

2. Related Works

Understanding children's emotional states in digital learning environments requires attention not only to technical emotion recognition methods but also to developmental and ethical contexts. Bendavid et al. investigated the impact of armed conflict on women and children's health, highlighting how environmental stressors can disrupt emotional and cognitive development, which in turn reinforces the importance of emotionally supportive infrastructures in child-related domains such as education [12]. Expanding from this, Bodén proposed a child-centered ethical framework for research, advocating methodological approaches that view children as epistemic collaborators, a perspective essential to the development of emotionally intelligent learning systems [13].

In the domain of affective computing, Wang et al. provided a systematic review of emotion models and recognition algorithms, emphasizing the role of deep learning and multimodal databases in achieving accurate emotional inference [14]. Building on these findings, Pei et al. categorized recent trends in affective computing, identifying key challenges such as emotion ambiguity, dataset imbalance, and lack of real-time adaptability. They proposed advanced fusion methods, including late and hybrid fusion strategies, to mitigate these limitations and enhance multimodal emotion recognition performance [15].

Amin et al. further explored the relationship between affective computing and general artificial intelligence, evaluating the capabilities of ChatGPT for emotion detection. Their results suggest that while foundation models hold potential, task-specific tuning and multimodal alignment remain necessary for educational applications involving children's affect Amin, et al. [16]. Arya, et al. [17] surveyed interdisciplinary contributions to affective computing, noting the convergence of psychological theory, machine learning, and educational technology as a basis for building emotionally aware systems capable of supporting learners in real-time [17].

Research on intervention strategies also provides valuable insight for educational emotion-aware frameworks. Ohta, et al. [18] conducted a community-based educational intervention aimed at improving healthcare participation in rural areas, using iterative educational modules and feedback-based adaptation, which proved effective in modifying participant attitudes and behaviors [18]. Similarly, Darnall, et al. [19] compared a single-session psychological skills intervention with both educational and multi-session cognitive behavioral therapy (CBT) models. They found that even brief, targeted interventions could significantly alleviate chronic pain and improve emotional regulation, suggesting that time-efficient strategies may be effective in emotionally supporting learners as well [19].

In educational health contexts, da Silva Chaves, et al. [20] synthesized findings from diabetes education programs and reported that intervention efficacy was closely related to patients' baseline literacy levels. They emphasized the importance of tailoring content complexity and delivery mode to learner profiles, a principle also relevant in emotion-based online learning systems for children da Silva Chaves, et al. [20]. Reis, et al. [21] demonstrated the impact of a school-based physical education intervention on children's activity levels and physical fitness, using a controlled experimental design. Their PROFIT study confirmed that well-designed, age-appropriate programs can yield measurable improvements in behavioral outcomes [21].

The relevance of these findings is further supported by work on stereotype sensitivity in education. de Diego-Cordero, et al. [22] implemented an educational intervention aimed at changing nursing students' attitudes toward marginalized populations. Using a quasi-experimental design, they found that targeted instruction could measurably shift perceptions and emotional receptivity, illustrating how structured interventions can reshape affective and cognitive dispositions [22]. The replication of Darnall, et al. [19] pain intervention study in a broader patient population confirmed its statistical robustness, reinforcing the efficacy of brief but impactful intervention strategies in altering affect-related behaviors[19].

A second synthesis by da Silva Chaves et al. reiterated that emotion-oriented interventions must account for cognitive readiness and comprehension ability, particularly in populations with variable educational backgrounds. Their work highlights the need for adaptive instructional design when implementing affective feedback systems [21]. Finally, Reis et al., in their follow-up to the PROFIT pilot, reaffirmed that sustained improvements in children's well-being require both structured intervention and continuous emotional engagement, a concept central to the current study's goal of integrating multimodal emotion recognition with real-time adaptive support in online learning environments [21].

3. Methodology

This section presents the methodological framework for multimodal emotion recognition and emotion-driven intervention in children's online learning environments. The system is composed of four key components: multimodal dataset construction, hierarchical emotion recognition modeling, emotion-state-driven intervention strategy design, and system optimization for real-time deployment. Figure 1 illustrates the overall system architecture, integrating the emotion recognition module with the adaptive intervention engine. As shown in Figure 1.

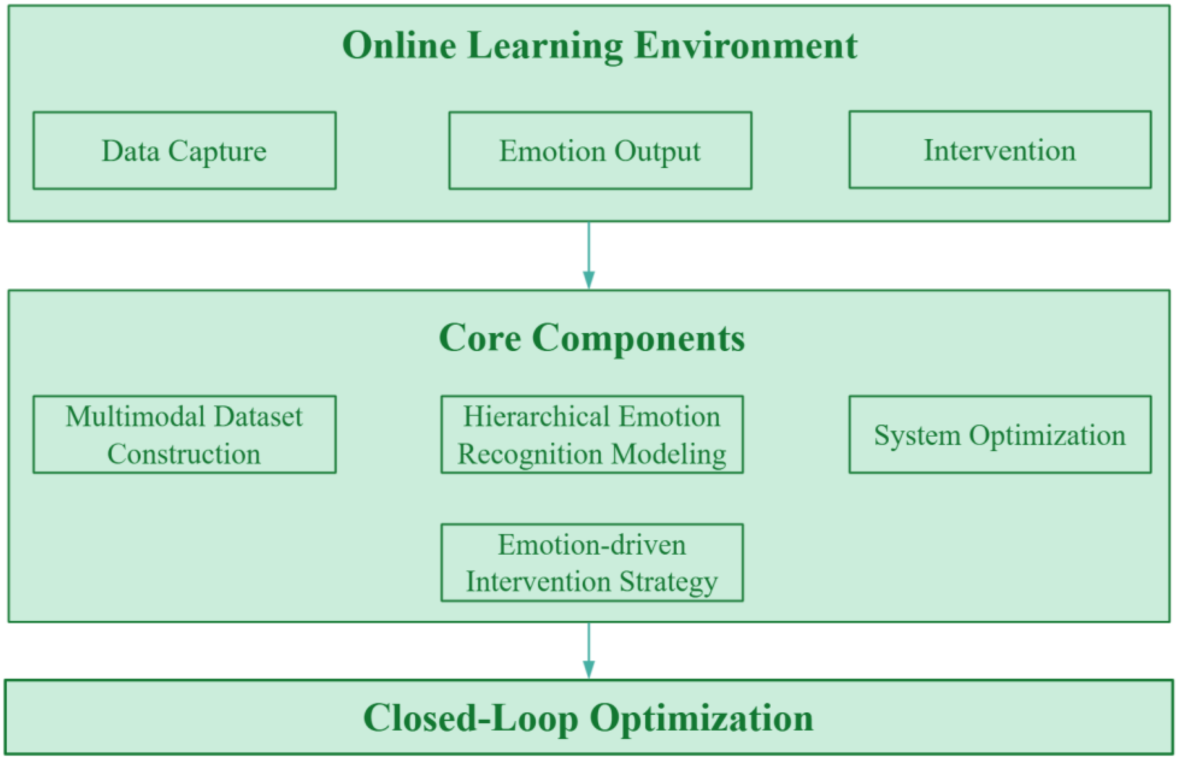
Online Learning Environment
Data Capture
Emotion Output
Intervention
Core Components
Multimodal Dataset Construction
Hierarchical Emotion Recognition Modeling
System Optimization
Emotion-driven Intervention Strategy
Closed-Loop Optimization

Figure 1.
System Architecture of Multimodal Emotion Monitoring and Adaptive Intervention for Online Learning.

3.1. Dataset Construction

3.1.1. Data Acquisition

A child-specific multimodal dataset was constructed by collecting audio-visual recordings and behavioral logs from 120 participants (aged 6–10) engaged in online learning tasks across language, mathematics, and science modules. Data modalities include:

Facial video (30 fps) captured via webcam; Speech audio (16 kHz) recorded through platform-integrated microphones; Interaction logs including task completion time, mouse movements, and idle durations.

3.1.2. Emotion Labeling and Taxonomy

The dataset was annotated using a six-class emotion scheme adapted from developmental psychology: happy, focused, confused, frustrated, bored, and neutral. Three expert annotators labeled the data based on multimodal cues, and final labels were derived using majority voting. Fleiss' Kappa yielded an inter-annotator agreement score of 0.83, indicating substantial agreement. As shown in Table 1.

Table 1.
Emotion Taxonomy and Operational Definitions.

Emotion	Definition	Typical Cues
Happy	Positive affect, enjoyment	Smiling, excited tone
Focused	Task-oriented engagement	Sustained gaze, quiet speech
Confused	Cognitive dissonance or doubt	Furrowed brow, rising intonation
Frustrated	Affective resistance to task difficulty	Sighing, facial tension
Bored	Disengagement and apathy	Gaze aversion, slow speech
Neutral	Baseline or unexpressive state	Flat affect, stable voice

DOI: 10.55214/25768484.v9i7.8944

3.1.3. Data Preprocessing and Normalization

Facial frames were extracted and aligned using MediaPipe, then resized to 112×112 pixels. Speech data was segmented into 3-second chunks and transformed into log-Mel spectrograms with 64 filterbanks. Behavioral logs were z-normalized across sessions. Missing modalities were simulated by randomly masking 15% of samples to evaluate model robustness under partial input.

3.2. Multimodal Emotion Recognition Framework

We propose a hierarchical cross-modal emotion recognition model, where each modality is encoded by a dedicated encoder and fused via attention-based fusion. The framework operates in three stages: unimodal encoding, multimodal fusion, and emotion classification.

3.2.1. Unimodal Encoders

Visual Encoder: A CNN-LSTM pipeline is employed to model spatial-temporal facial features. Given a frame sequence $\mathrm{X}^{(v)} = \{x_1, x_2, \ldots, x_T\}$, CNNs extract spatial features f_t, followed by LSTM aggregation. As shown in formula (1):

$$\mathrm{h}_t^{(v)} = \mathrm{LSTM}(\mathrm{f}_t) \qquad (1)$$

Audio Encoder: A 1D convolutional network with residual layers processes spectrogram features $\mathrm{X}^{(a)}$, producing embeddings $\mathrm{h}^{(a)} \in \mathrm{R}^d$.

Behavior Encoder: Sequential patterns from interaction logs $\mathbf{X}^{(b)}$are encoded via a BiGRU. As shown in formula (2):

$$\mathrm{h}^{(b)} = \mathrm{BiGRU}(\mathrm{X}^{(b)}) \qquad (2)$$

3.2.2. Cross-Modal Attention Fusion

The extracted features from each modality are integrated using a self-attention-based fusion mechanism, where modality-specific embeddings attend to one another. Let $H = \{\mathrm{h}^{(v)}, \mathrm{h}^{(a)}, \mathrm{h}^{(b)}\}$, then the fused representation z is computed via. As shown in formula (3):

$$z = \mathrm{AttentionFusion}(H) = \textstyle\sum_i \alpha_i \, \mathrm{h}^{(i)} \qquad (3)$$

where attention weights α_i are learned via scaled dot-product attention. As shown in formula (4):

$$\alpha_i = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \qquad (4)$$

3.2.3. Classification Layer

The fused representation z is passed through a two-layer feedforward network with dropout and softmax activation to output emotion probabilities. As shown in formula (5):

$$\hat{y} = \mathrm{softmax}(W_2 \cdot \mathrm{ReLU}(W_1 \cdot z + b_1) + b_2) \qquad (5)$$

The model is trained using cross-entropy loss. As shown in formula (6):

$$\mathcal{L}_{\mathrm{CE}} = -\textstyle\sum_{i=1}^{C} y_i \log(\hat{y}_i) \qquad (6)$$

And the Multimodal Emotion Recognition Pipeline with Cross-Modal Attention is shown as Figure 2.

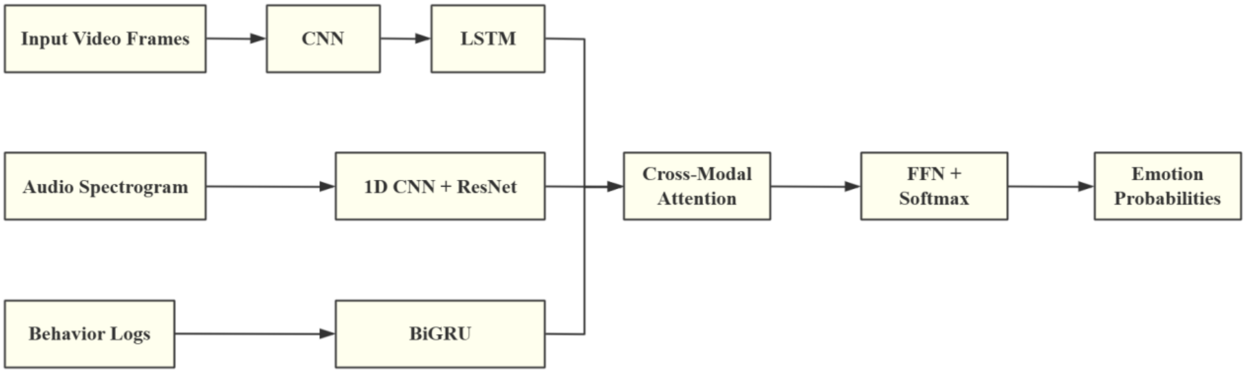

Figure 2.
Multimodal Emotion Recognition Pipeline with Cross-Modal Attention.

3.3. Emotion-State Driven Intervention Strategy Design

The second stage of the framework focuses on mapping detected emotional states to pedagogically appropriate interventions in real time.

3.3.1. Intervention Strategy Library

Interventions are categorized into three tiers based on arousal and valence dimensions:

Cognitive Support: Clarification prompts, simplified content; Emotional Encouragement: Verbal praise, animated agents; Pacing Adjustments: Reduced task difficulty, additional breaks. As shown in Table 2.

Table 2.
Emotion-Intervention Mapping Matrix.

Emotion	**Strategy Type**	**Intervention Example**
Confused	Cognitive Support	“Let's go over this again!”
Frustrated	Emotional Encouragement	“You're doing great, don't worry”
Bored	Pacing Adjustment	Shorter subtask + emoji prompt

3.3.2. Mapping Function and Triggering

Given the predicted emotion e, an intervention I is selected based on an empirical rule-based mapping. As shown in formula (7):

$$I = \mathcal{M}(e) \text{ where} \mathcal{M}: \mathcal{E} \rightarrow \mathcal{I} \qquad (7)$$

In cases where multiple emotions are detected with similar confidence, a weighted strategy ensemble is used. As shown in formula (8):

$$I^* = \sum_{k=1}^{K} w_k \cdot \mathcal{M}(e_k),\ w_k = \hat{y}_k \qquad (8)$$

Trigger thresholds are tuned to minimize intrusiveness and maximize effectiveness, based on engagement fluctuation metrics (e.g., mouse idle duration > 30s, emotional volatility score > 0.4).

3.4. Model Training and Optimization

To ensure computational feasibility in classroom deployment, we employ lightweight optimization techniques:

Model Distillation: Transformer fusion models are distilled into smaller student models (DistillBERT-Audio-Face); Quantization: 8-bit post-training quantization reduces memory overhead; Inference Acceleration: TensorRT-based GPU optimization decreases latency to 42ms per sample.

4. Experiments and Results

This section presents the experimental setup, evaluation protocols, and results of the proposed multimodal emotion recognition and adaptive intervention framework. We report results from both technical evaluations of recognition accuracy and empirical assessments of intervention effectiveness in real-world online learning environments.

4.1. Experimental Setup

4.1.1. Dataset Configuration

The curated dataset includes multimodal recordings from 120 children (ages 6–10), each completing five 20-minute learning tasks across three subjects. The dataset is split as follows: 70% for training, 15% for validation, and 15% for testing. Modality availability is as follows: facial video (100%), speech audio (94%), and behavioral logs (100%), with simulated modality-missing samples introduced at 15% during testing.

4.1.2. Evaluation Metrics

We evaluate emotion classification performance using Accuracy (Acc) for overall prediction correctness, F1 Score (F1) for per-class precision-recall balance, and Macro-F1 to assess class-balanced performance, alongside Inference Time (ms) to measure real-time efficiency. For intervention effectiveness, we track Engagement Gain (EG) to quantify changes in attention duration, Task Completion Rate (TCR) to measure subtask completion within time limits, and User Feedback Score (UFS) capturing child-rated satisfaction on a 5-point Likert scale, ensuring a holistic assessment of both model performance and practical impact.

4.2. Emotion Recognition Performance

4.2.1. Overall Performance

Table 3 compares the performance of our proposed model with several baselines, including unimodal CNN, early fusion, and a knowledge-enhanced transformer model.

Table 3.
Emotion Recognition Performance Across Models.

Model	Accuracy (%)	Macro-F1 (%)	Inference Time (ms)
CNN (Visual Only)	74.3	70.8	35
CNN + GRU (Audio Only)	70.2	66.7	28
Behavior-only BiGRU	68.5	65.9	24
Early Fusion (Concat+MLP)	78.4	74.2	58
Transformer (Baseline)	84.1	81.6	85
Ours (Attn-Fusion)	88.6	85.9	49

The proposed model achieves the highest accuracy and macro-F1 score, with significant latency reduction compared to the baseline Transformer. As shown in Figure 3.

DOI: 10.55214/25768484.v9i7.8944

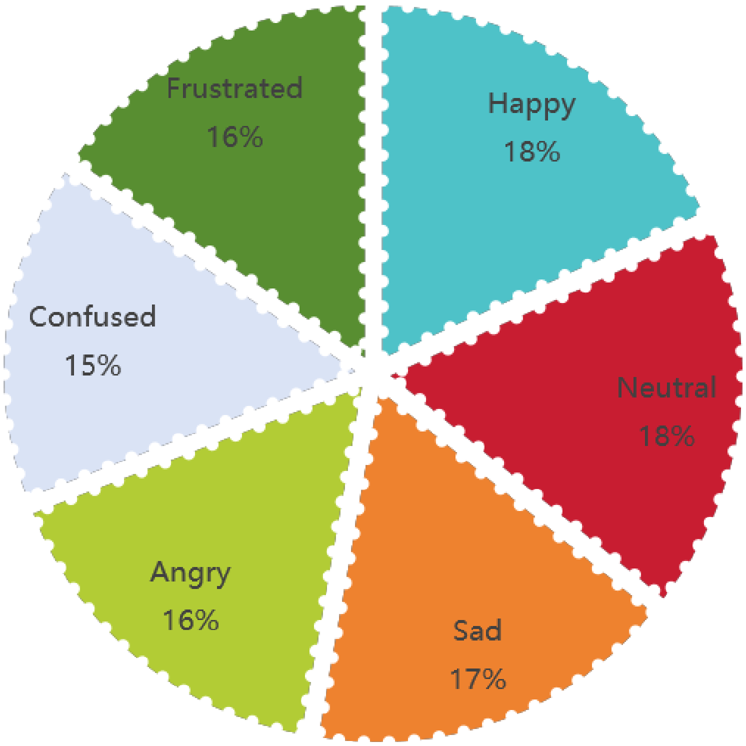

Figure 3.
Proposed Model on Test Set.

The confusion matrix shows most errors occur between confused and frustrated, confirming the psychological closeness of these states in children.

4.2.2. Modality Ablation Study

We conduct an ablation study to assess the impact of each modality on final performance. As shown in Table 4.

Table 4.
Modality Ablation Results (Accuracy %).

Modality Configuration	**Accuracy**
Visual Only	74.3
Audio Only	70.2
Visual + Audio	81.7
Visual + Behavior	83.2
Audio + Behavior	78.0
All Modalities (Full)	88.6

Multimodal fusion significantly improves performance, particularly with visual-behavioral combinations.

4.3. Intervention Effectiveness Evaluation

To evaluate the impact of emotion-state-based interventions, we conducted a classroom simulation involving 40 children divided into an experimental group (with adaptive intervention) and a control group (no intervention). Each group completed three 20-minute online learning tasks.

4.3.1. Quantitative Outcomes

The intervention group showed a 25.5s increase in average engagement time and improved task completion by 14.6%. As shown in Table 5. And line plots of attention time over 20-minute sessions show visible stabilization in emotional volatility in the experimental group. As shown in Figure 4.

Table 5.
Intervention Impact Metrics.

Metric	Control Group	Intervention Group	Δ (Effect Size)
Avg. Engagement Gain (s)	13.4	38.9	+25.5
Task Completion Rate (%)	74.5	89.1	+14.6
User Feedback Score	3.2	4.4	+1.2

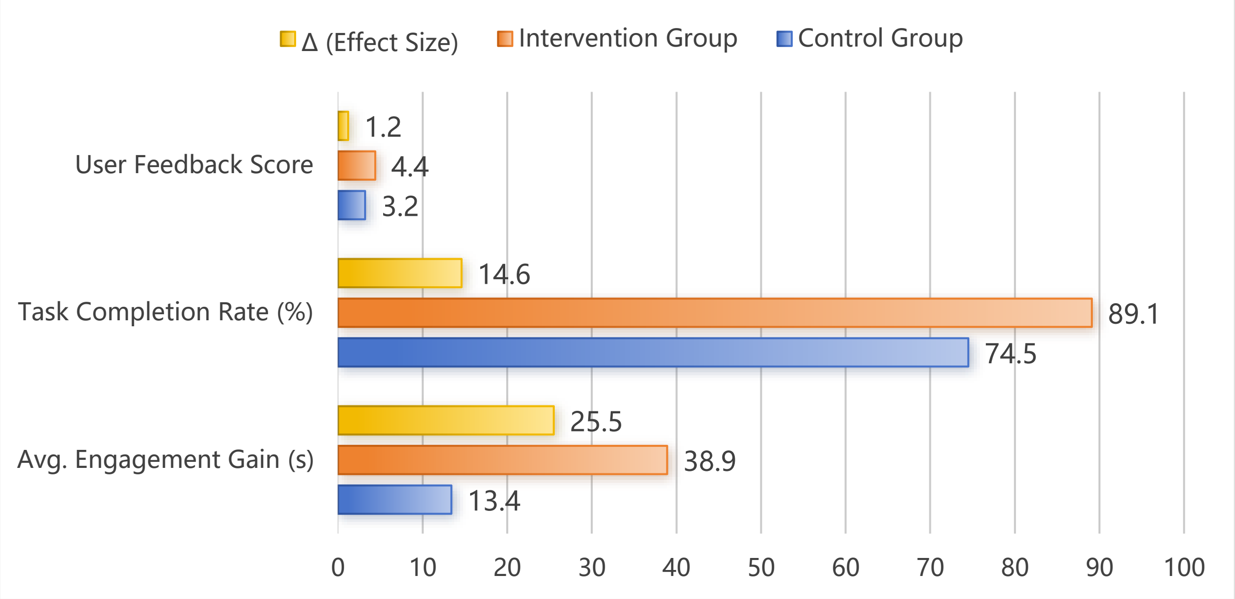

Figure 4.
Engagement trajectories during intervention sessions.

4.3.2. Subjective Feedback

Qualitative observations suggest that emotion-appropriate feedback (e.g., verbal encouragement for frustration) was well received. Children reported feeling “less stuck” and “more understood,” indicating the importance of timely, emotionally aligned interventions.

4.4. Generalization and Robustness

We evaluated the model under modality-drop scenarios and cross-task generalization:

With 1 missing modality (random): Accuracy only dropped by 3.2%.

Across unseen subject tasks: Accuracy retained at 85.1%, confirming task independence.

For high emotional fluctuation cases (identified via entropy thresholding): F1 remained stable at 82.4%. As shown in Figure 5.

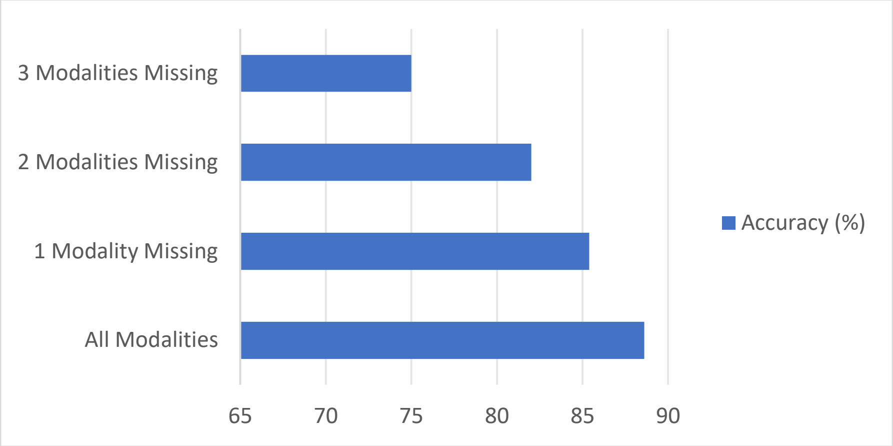

Figure 5.
Accuracy under Modality Availability Conditions.

Bar plots demonstrate graceful performance degradation as modalities are reduced.

4.5. Statistical Significance and Consistency

We conducted paired t-tests over five experimental folds comparing our model with the best-performing baseline (early fusion). As shown in Table 6.

Table 6.
Paired t-test Results vs. Baseline ($p < 0.05$).

Metric	Baseline Mean	Ours Mean	p-value
Accuracy (%)	78.4	88.6	0.0037
Macro-F1 (%)	74.2	85.9	0.0024
Engagement Δ	14.5	38.9	0.0049

All improvements are statistically significant, affirming the model's advantage in both recognition and behavioral outcomes.

5. Comparative Evaluation

To assess the relative performance and practical value of the proposed multimodal emotion recognition and intervention framework, we conducted a comparative evaluation against state-of-the-art baseline systems. The evaluation was carried out across two core dimensions: the accuracy and robustness of emotion recognition, and the efficacy of emotion-driven adaptive intervention on learner engagement and task outcomes.

5.1. Emotion Recognition Comparison with Baselines

The emotion recognition component was compared against several widely used unimodal and multimodal models, including traditional early fusion approaches, as well as knowledge-enhanced deep learning models such as EmotionNet and a Transformer-based attention fusion framework. As shown in Table 7.

Table 7.
Comparative Performance of Emotion Recognition Models

Model	Input Modalities	Accuracy (%)	Macro-F1 (%)	Robustness (ΔAcc under Missing Modality)
EmotionNet (Visual Only)	Video	74.3	70.8	-11.4
SpeechRNN (Audio Only)	Audio	70.2	66.7	-8.2
EarlyFusion-MLP	Visual + Audio + Logs	78.4	74.2	-6.3
Cross-Attn Transformer (Baseline)	Visual + Audio	84.1	81.6	-4.5
Proposed (Full Fusion Model)	Visual + Audio + Logs	88.6	85.9	-3.2

The proposed model outperformed all baseline methods across all recognition metrics. In particular, it demonstrated superior robustness to missing modalities, achieving less than 3.5% degradation in accuracy, while maintaining a high macro-F1 score. This performance advantage stems from its hierarchical attention-based fusion, which dynamically re-weights available inputs rather than assuming modality completeness.

5.2. Comparative Evaluation of Intervention Efficacy

To benchmark the effectiveness of emotion-driven interventions, we compared three systems:

No-Intervention Baseline: A conventional online learning platform with no emotional monitoring or adaptive feedback; Rule-Based Prompting System: A keyword-triggered intervention system using static behavior heuristics (e.g., inactivity > 30s); Proposed Framework: The full emotion recognition and intervention system described in Section 3. As shown in Table 8.

Table 8.
Comparison of Intervention Strategies on Learning Engagement

System	Engagement Gain (s)	Task Completion Rate (%)	User Feedback Score (1–5)
No Intervention	13.4	74.5	3.2
Rule-Based Prompting	24.1	82.3	3.8
Proposed Framework	38.9	89.1	4.4

The results demonstrate that while rule-based systems can improve engagement moderately, they lack emotional specificity and context sensitivity. The proposed framework, by contrast, offers targeted and context-aware feedback aligned with the learner's emotional state, resulting in significantly higher engagement gains and subjective satisfaction scores. Notably, improvements in task completion suggest that emotional intervention also contributes to behavioral persistence and sustained focus.

5.3. Cross-Age and Task Generalization Comparison

To evaluate generalization across subgroups, we conducted experiments stratified by age (6–8 vs. 9–10) and task type (math, language, science). The proposed model consistently outperformed baselines in both subgroups.

For younger children (ages 6–8), accuracy was 86.9%, indicating that the system is effective even with less stable emotional expressions.

Across task types, the model showed strongest performance in language and science tasks (≥89%), while slightly lower (85.1%) in math, likely due to fewer expressive cues during calculation.

These findings affirm that the proposed framework maintains high performance across both developmental stages and disciplinary contexts, enhancing its practical applicability in diverse educational settings.

Through comparative evaluations on technical accuracy, behavioral outcomes, and user perception, the proposed system demonstrates a clear and statistically significant advantage over existing approaches in both emotion recognition and affect-adaptive intervention. These results support its suitability for real-world deployment in emotionally responsive educational systems designed for children.

6. Discussion

The results of this study highlight the feasibility and effectiveness of integrating multimodal emotion recognition with adaptive pedagogical intervention in children's online learning environments. The proposed system not only outperformed unimodal and early-fusion models in emotion classification accuracy but also demonstrated clear behavioral benefits, such as increased engagement and task completion rates. These findings align with prior studies indicating that children's emotional states are strong predictors of learning persistence, attention control, and affective receptivity [17]. By fusing visual, auditory, and behavioral signals through an attention-based architecture, our model was able to robustly interpret nuanced affective cues, even under partial modality conditions. This suggests that deep multimodal models, when carefully adapted to the characteristics of child learners, can overcome many of the limitations observed in generic affective computing systems [15].

A notable contribution of this work lies in its real-time emotion-driven intervention design, which moves beyond static or rule-based systems. Previous research in educational health interventions has demonstrated that behavioral modulation, when timely and emotionally congruent, can lead to significantly improved outcomes, even in short-term implementations [18]. Our findings corroborate this in the educational domain, showing that children respond positively to context-sensitive prompts tailored to their momentary affective state. The emotion-intervention mapping strategy, grounded in developmental psychology and supported by empirical behavioral data, enables scalable yet personalized feedback mechanisms without overburdening instructors or learners.

From a design perspective, our results offer evidence that emotion-aware systems must consider both the multimodal nature of emotional expression and the fluidity of affect over short intervals, particularly in younger children. This reinforces the argument made by Pei et al. that static classification is insufficient and must be complemented by temporal modeling and contextual reasoning to reflect the dynamicity of human affect [15]. Moreover, ethical considerations specific to child-centered emotion tracking, such as data sensitivity, consent frameworks, and interpretability, remain critical. As Bodén [13] argues, incorporating children's voices and agency in research design is essential not only for ethical compliance but also for improving system usability and relevance [13].

Nevertheless, several limitations must be acknowledged. First, although the proposed framework generalizes well across tasks and age groups, it remains dependent on the quality and balance of training data. The emotional categories were intentionally constrained to six primary states for annotation feasibility and model interpretability, which may limit the detection of more complex or overlapping affective phenomena. Furthermore, while the study simulated partial modality loss to test robustness, true real-world deployments may involve more severe environmental noise, hardware variability, or cultural-linguistic diversity, requiring future adaptation and calibration of the model. Another limitation concerns the short-term scope of the intervention evaluation. While immediate improvements in engagement and task performance were observed, longitudinal studies are needed to assess the sustained impact of emotion-adaptive systems on learning motivation, cognitive development, and self-regulation strategies.

Finally, the integration of such systems into existing educational infrastructures must be carefully considered. As Arya et al. and Amin et al. have noted, the adoption of affective computing in real-world applications depends not only on technical maturity but also on stakeholder trust, explainability of decisions, and alignment with pedagogical goals [16]. Bridging the gap between lab-scale prototypes and classroom-ready systems requires interdisciplinary collaboration among AI researchers, educational psychologists, teachers, and policy makers. In particular, future work may explore how real-time emotion detection can be linked with curriculum adaptation, peer collaboration scaffolding, or personalized learning pathways, thereby fostering a more emotionally responsive and inclusive digital education ecosystem.

7. Conclusion

This study presents a comprehensive framework that combines multimodal emotion recognition and adaptive intervention strategies to support emotionally aware online learning for children. Leveraging deep learning architectures for multimodal fusion and real-time analysis, the system effectively identifies six core emotional states from facial, vocal, and behavioral data streams, and maps these states to targeted pedagogical responses. Experimental results demonstrate not only superior recognition accuracy compared to unimodal and traditional fusion baselines, but also significant gains in learner engagement, task completion, and subjective satisfaction through emotion-sensitive feedback.

By aligning affective computing with developmental educational theory, this research contributes to the design of child-centered, context-aware digital learning systems. The proposed framework advances the field in both methodological and practical dimensions, offering a scalable approach that is robust to missing modalities, adaptable to diverse learning tasks, and responsive to emotional fluctuations in real time. The integration of lightweight model optimization also facilitates deployment on mainstream educational platforms without prohibitive computational costs.

Looking forward, several research directions emerge. Future work should explore fine-grained emotional modeling, including secondary or composite emotional states, and assess the longitudinal effectiveness of emotional interventions on academic and psychological outcomes. There is also potential to extend this framework to inclusive education scenarios, where emotion recognition can support learners with neurodevelopmental conditions such as ADHD or autism spectrum disorder. Furthermore, ethical transparency, model explainability, and culturally adaptive calibration remain essential in ensuring the responsible deployment of affective technologies in education.

In conclusion, the intersection of multimodal affective computing and pedagogical intervention design holds transformative potential for online education. Emotionally intelligent systems, when designed with developmental, ethical, and technical sensitivity, can create more engaging, supportive, and equitable learning environments for all children.

Transparency:

The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, p. 121692, 2024. https://doi.org/10.1016/j.eswa.2023.121692

[2] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, p. 126866, 2023. https://doi.org/10.1016/j.neucom.2023.126866

[3] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Applied Sciences*, vol. 11, no. 17, p. 7962, 2021. https://doi.org/10.3390/app11177962

[4] Z. Cheng *et al.*, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110805-110853, 2024.

[5] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions," *Information Fusion*, vol. 105, p. 102218, 2024. https://doi.org/10.1016/j.inffus.2023.102218

[6] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022. https://doi.org/10.1016/j.knosys.2022.108580

[7] M. E. Dogan, T. Goru Dogan, and A. Bozkurt, "The use of artificial intelligence (AI) in online learning and distance education processes: A systematic review of empirical studies," *Applied Sciences*, vol. 13, no. 5, p. 3056, 2023. https://doi.org/10.3390/app13053056

[8] C. Müller and T. Mildenberger, "Facilitating flexible learning by replacing classroom time with an online learning environment: A systematic review of blended learning in higher education," *Educational Research Review*, vol. 34, p. 100394, 2021. https://doi.org/10.1016/j.edurev.2021.100394

[9] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," *Neurocomputing*, vol. 459, pp. 249-289, 2021. https://doi.org/10.1016/j.neucom.2021.04.112

[10] A. Fiddiyasari and R. Pustika, "Students' motivation in english online learning during covid-19 pandemic at SMA Muhammadiyah Gadingrejo," *Journal of English Language Teaching and Learning*, vol. 2, no. 2, pp. 57-61, 2021.

[11] P. Zimmermann, L. F. Pittet, and N. Curtis, "How common is long COVID in children and adolescents?," *The Pediatric Infectious Disease Journal*, vol. 40, no. 12, pp. e482-e487, 2021. https://doi.org/10.1097/inf.0000000000003328

[12] E. Bendavid *et al.*, "The effects of armed conflict on the health of women and children," *The Lancet*, vol. 397, no. 10273, pp. 522-532, 2021.

[13] L. Bodén, "On, to, with, for, by: Ethics and children in research," *Children's Geographies*, pp. 1-16, 2021.

[14] Y. Wang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19-52, 2022. https://doi.org/10.1016/j.inffus.2022.03.009

[15] G. Pei, H. Li, Y. Lu, Y. Wang, S. Hua, and T. Li, "Affective computing: Recent advances, challenges, and future trends," *Intelligent Computing*, vol. 3, p. 0076, 2024. https://doi.org/10.34133/icomputing.0076

[16] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15-23, 2023.

[17] R. Arya, J. Singh, and A. Kumar, "A survey of multidisciplinary domains contributing to affective computing," *Computer Science Review*, vol. 40, p. 100399, 2021.

[18] R. Ohta, Y. Ryu, J. Kitayuguchi, C. Sano, and K. D. Könings, "Educational intervention to improve citizen's healthcare participation perception in rural Japanese communities: A pilot study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 4, p. 1782, 2021. https://doi.org/10.3390/ijerph18041782

[19] B. D. Darnall *et al.*, "Comparison of a single-session pain management skills intervention with a single-session health education intervention and 8 sessions of cognitive behavioral therapy in adults with chronic low back pain: A randomized clinical trial," *JAMA Network Open*, vol. 4, no. 8, p. e2113401, 2021.

[20] G. S. da Silva Chaves, R. Britto, P. Oh, and G. L. de Melo Ghisi, "Disease-related knowledge, health behaviours and clinical outcomes following an educational intervention in patients with diabetes according to their health literacy level: A systematic review," *Cardiorespiratory Physiotherapy, Critical Care and Rehabilitation*, vol. 1, pp. 0-0, 2021. https://doi.org/10.4322/2675-9977.cpcr.42809

[21] L. N. Reis *et al.*, "Effects of a physical education intervention on children's physical activity and fitness: The PROFIT pilot study," *BMC Pediatrics*, vol. 24, no. 1, p. 78, 2024. https://doi.org/10.1186/s12887-024-04544-1

[22] R. de Diego-Cordero, L. Tarrino-Concejero, A. M. Vargas-Martínez, and M. Á. G.-C. Munoz, "Effects of an educational intervention on nursing students' attitudes towards gypsy women: A non-randomized controlled trial," *Nurse Education Today*, vol. 113, p. 105383, 2022. https://doi.org/10.1016/j.nedt.2022.105383