

PAPER

An Efficient Multimodal Causal Attention Architecture for Real-Time Mobile Educational Content Interpretation

Jin Wang ,
Inam Ullah  

City University Malaysia,
Selangor, Malaysia

inam.ullah@city.edu.my

ABSTRACT

The increasing adoption of mobile learning requires efficient multimodal models capable of real-time educational content analysis under resource constraints. However, existing multimodal architectures often involve high inference latency and memory requirements, limiting their deployment on mobile and edge devices. This study proposes the Efficient Multimodal Causal Attention Architecture (EMCAA), integrating speech, lecture transcripts, presentation slides, and contextual learner interactions. EMCAA combines Temporal Directed Acyclic Graph (T-DAG) Causal Attention to prioritize temporally relevant dependencies, Cross-Modal Layer Fusion (CMLF) to integrate complementary multimodal representations, and Causal Context-Aware Quantization (CCQ) to reduce computational and memory requirements through adaptive precision allocation. Experiments on the Multimodal Student Attention Dataset showed that EMCAA achieved 91.4% accuracy, 91.0% precision, 90.8% recall, and 90.9% F1-score. Compared with MC-MIFA, EMCAA improved accuracy by 1.2 percentage points while reducing inference latency from 57 to 43 ms (24.6%) and memory consumption from 401 to 329 MB (18.0%). Ablation analysis further demonstrated the contribution of the proposed components. These findings indicate the potential of EMCAA for efficient multimodal educational content processing in resource-constrained mobile and edge environments.

KEYWORDS

multimodal educational content, causal attention, T-DAG, cross-modal layer fusion, causal context-aware quantisation

1 INTRODUCTION

Mobile computing, artificial intelligence (AI), and learning technologies have revolutionised the conventional forms of education and turned them into interactive and personalised learning experiences. Mobile learning platforms allow learners to have instant access to lectures, presentations, tutorials, digital books, and exercises

Wang, J., Ullah, I. (2026). An Efficient Multimodal Causal Attention Architecture for Real-Time Mobile Educational Content Interpretation. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(19), pp. 31–44. <https://doi.org/10.3991/ijim.v20i19.63224>

Article submitted 2026-05-14. Revision uploaded 2026-07-13. Final acceptance 2026-07-19.

© 2026 by the authors of this article. Published under CC-BY.

at any time and place [1]. Such resources produce heterogeneous data, such as voice, transcripts, presentation slides, images, and learner interaction. Therefore, intelligent solutions that are able to learn from multiple modalities in real time become an important tool for automating lecture analysis, personalization of learning, summarizing of educational content, and educational assistance [2]. Multimodal AI and Transformer-based models have proven themselves to be powerful solutions for learning the dependencies between texts, visuals, and audios [3]. At the same time, many multimodal models have a lot of parameters and perform complex attention operations that make the deployment of such models difficult for smartphones, tablets, and other small devices [4]. It is possible to use cloud-based computation to decrease the computational load for mobile devices, but then the network becomes an issue for performance and data security. Thus, developing multimodal models that can perform inferences in real-time directly from mobile or edge devices is a research challenge [5]. The other disadvantage associated with the use of existing attention models is their inability to explicitly differentiate between causally and temporally relevant information and redundant information [6]. There are usually temporal dependencies in the education data, where basic concepts are introduced before their explanation and examples. Information from speech, slides, transcripts, and learner interaction may also be useful but vary over time [7]. Modelling these temporal dependencies may help in identifying important concepts in the education data without unnecessary computations [8].

In order to overcome the above-mentioned difficulties, this study suggests an Efficient Multimodal Causal Attention Architecture (EMCAA) that provides real-time analysis of mobile educational content. The framework consists of speech, lecture transcripts, presentation slides, and interactional contextual learning processes [9]. To analyze the time-based relations, we design a Temporal Directed Acyclic Graph (T-DAG) Causal Attention model. To integrate multimodal data from different sources, a Cross-Modal Layer Fusion (CMLF) scheme is applied. To reduce memory and computing resource utilisation while retaining context representations, we propose a Causal Context-Aware Quantisation (CCQ) method. The suggested architecture focuses on finding a balance between the interpretation quality and efficiency of the framework usage on mobile devices. The experimental assessment covers the following factors: interpretation accuracy, inference delay, memory consumption, concept identification, contextual comprehension, and content summarization [11]. Accordingly, this study aims to develop an efficient multimodal architecture for real-time educational content interpretation on mobile and edge devices. Specifically, it seeks to model temporally directed and contextually relevant dependencies among multimodal educational representations through T-DAG attention, selectively integrate complementary information across modalities and representation layers using CMLF, and reduce computational and memory requirements through CCQ-based adaptive precision allocation. Finally, the study evaluates the proposed EMCAA framework in terms of predictive performance, inference latency, memory consumption, and the contribution of its individual components through ablation analysis.

2 RELATED WORK

Multimodal artificial intelligence has gained increasing attention in educational applications because learning environments generate heterogeneous information from visual, textual, audio, and behavioural sources. Existing studies have applied deep learning to student engagement detection in e-learning environments [21],

multimodal facial cues for engagement recognition [9], and hybrid CNN-ViT architectures for classroom-based engagement analysis [13]. Self-attention combined with 3D DenseNet has also demonstrated the value of attention mechanisms for learning behavioural representations [12], while stability-oriented multimodal modelling has emphasized the importance of reliable prediction beyond peak classification accuracy [14]. More broadly, recent reviews confirm the increasing use of heterogeneous multimodal data for educational analysis [18].

Recent research has increasingly focused on improving the efficiency, temporal modelling, and cross-modal integration of such systems. State-space architectures have emerged as computationally efficient alternatives to conventional Transformer-based models for sequential and visual information processing [5]. Attention-based multimodal alignment [10], question-guided cross-modal attention [11], and dynamic multimodal attention [15] further demonstrate the value of selectively integrating complementary information across heterogeneous modalities. Within educational applications, EfficientNetV2 combined with recurrent models has been used for student engagement detection, illustrating the benefits of efficient feature extraction and temporal modelling [19]. Multimodal AI has also been applied to real-time student engagement detection and adaptive feedback [16], while attention-based multimodal processing has supported real-time listening comprehension assessment [22].

More recently, causal-aware modelling has emerged as a promising direction for multimodal educational analysis [17]. The MC-MIFA framework integrates causal-aware hybrid state-space modelling for robust multimodal student engagement analysis [20], indicating the potential value of incorporating structured dependencies into multimodal representations. However, existing approaches generally address multimodal fusion, temporal modelling, computational efficiency, or causal-aware representation separately. Limited research has jointly investigated temporally directed attention, cross-modal layer fusion, and adaptive quantisation within a unified architecture designed specifically for resource-constrained mobile and edge educational environments.

To address this gap, the present study proposes the Efficient Multimodal Causal Attention Architecture (EMCAA). The framework integrates Temporal Directed Acyclic Graph (T-DAG) Causal Attention for modelling temporally directed dependencies, Cross-Modal Layer Fusion (CMLF) for selective multimodal integration, and Causal Context-Aware Quantisation (CCQ) for adaptive precision allocation. This integration is designed to improve multimodal educational content interpretation while reducing inference latency and memory requirements for mobile and edge deployment.

3 METHODOLOGY

3.1 Problem formulation

The objective of this study is to develop a multimodal framework capable of interpreting educational content in real time while maintaining computational efficiency for mobile and edge deployment. Let the multimodal educational input at time (t) be represented as

$$X_t = \{X_t^s, X_t^l, X_t^v, X_t^c\} \quad (1)$$

where (X_t^s) denotes speech/audio information, (X_t^l) represents lecture transcripts or textual content, (X_t^v) represents presentation slides and visual information, and

(X_t^c) represents contextual learning interactions. Each modality is transformed into a feature representation using an appropriate encoder:

$$F_t^m = E_m(X_t^m), m \in \{s, l, v, c\} \quad (2)$$

where (E_m) denotes the modality-specific feature extraction function and (F_t^m) is the resulting representation. The objective is to combine these heterogeneous representations while preserving their temporal and semantic relationships. Conventional multimodal attention can be expressed as

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (3)$$

where Q, K , and V are the query, key, and value matrices, respectively, and d represents the feature dimensions. However, traditional attention mechanisms could allocate computing power to relationships that are not relevant or redundant. In education, where knowledge tends to be acquired in a chronological sequence, there is an advantage in paying more attention to knowledge that is temporally and causally related. Hence, the proposed problem formulation is

$$H_t = f_\theta(F_t^s, F_t^l, F_t^v, F_t^c, G_t) \quad (4)$$

where (f_θ) represents the proposed EMCAA model and (G_t) represents the temporal directed acyclic graph describing relevant dependencies among multimodal educational concepts. The system produces an interpretation output

$$Y_t = \{C_t, S_t, R_t\} \quad (5)$$

where (C_t) represents extracted educational concepts, (S_t) denotes contextual content summarization, and (R_t) represents the interpreted semantic relationships among the input modalities. Let (L) denote inference latency, (M) denote memory consumption, and (C) denote computational complexity. The optimisation problem can therefore be expressed as

$$\max A(Y, \hat{Y}); L \leq L_{\max}; M \leq M_{\max}; C \leq C_{\max} \quad (6)$$

where (A) represents the interpretation accuracy, while $(L_{\max})$, $(M_{\max})$, and $(C_{\max})$ represent the maximum allowable latency, memory, and computational resources of the target mobile device. Overall, the optimization objective is to maximize interpretation accuracy while satisfying the latency, memory, and computational constraints of the target mobile device.

3.2 Architecture of the overall framework

The proposed Efficient Multimodal Causal Attention Architecture (EMCAA) is designed for real-time educational content interpretation on mobile and edge devices. As shown in Figure 1, EMCAA consists of five stages: multimodal input acquisition, modality-specific feature extraction, T-DAG Causal Attention, Cross-Modal Layer Fusion (CMLF), and Causal Context-Aware Quantisation (CCQ).

Speech, lecture transcripts, presentation slides, and learner interactions are first transformed into modality-specific feature representations and temporally aligned.

T-DAG identifies temporally directed and contextually relevant dependencies, while CMLF integrates complementary information across modalities. CCQ then applies adaptive precision to reduce memory and computational requirements while preserving important representations. The resulting representation supports key concept extraction, contextual interpretation, and educational content summarisation.

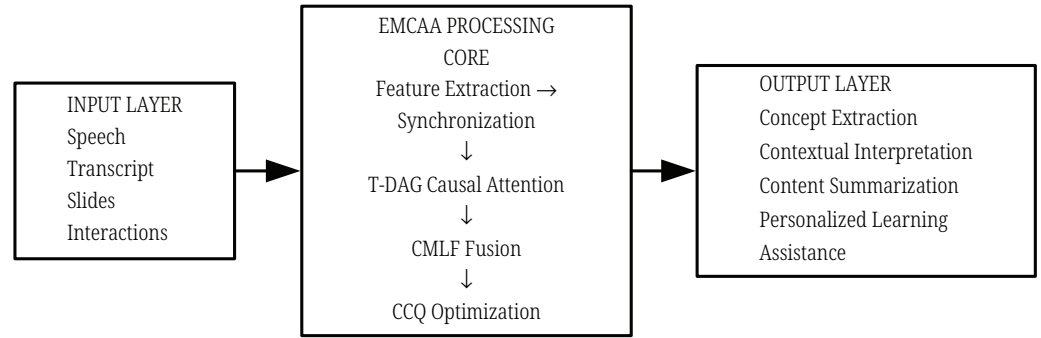


Fig. 1. Block diagram of proposed efficient multimodal causal attention

3.3 Multimodal input processing

The EMCAA model utilises four different educational modes that complement each other: speech, lecture notes, slides and contextual learning experiences. Given any learning session, the multimodal data can be captured as an array of educational events:

$$X = \{x_1, x_2, \dots, x_T\} \quad (7)$$

where (T) denotes the number of temporal learning segments. Each event contains information from the available modalities:

$$x_t = (x_t^a, x_t^q, x_t^i, x_t^r) \quad (8)$$

where (a) , (q) , (i) , and (r) correspond to audio, textual, visual, and interaction information, respectively. The quality of an audio segment is evaluated using a normalized representation:

$$\tilde{x}_t^a = \frac{x_t^a - \mu_a}{\sigma_a + \epsilon} \quad (9)$$

where (μ_a) and (σ_a) represent the mean and standard deviation of the audio features, and (ϵ) is a small constant used for numerical stability. The importance of a textual token can be represented using a contextual weighting function:

where (k) defines the contextual observation window. The temporal correspondence between two modality events can be expressed through a similarity function:

$$S_{ij}^{temp} = \exp\left(-\frac{|t_i - t_j|}{\tau}\right) \quad (10)$$

where (t_i) and (t_j) are the timestamps of two multimodal events and (τ) controls the temporal tolerance.

3.4 Temporal Directed Acyclic Graph (T-DAG) Causal Attention

The Temporal Directed Acyclic Graph (T-DAG) Causal Attention mechanism is the central reasoning component of EMCAA. For a sequence of educational representations, the graph can be represented as

$$G = (V, E) \quad (11)$$

where (V) represents the set of multimodal educational nodes and (E) represents directed dependency relationships between nodes. For two educational nodes (v_i) and (v_j) , a directed relationship is established when the earlier node provides relevant contextual information for interpreting the later node. The temporal dependency strength can be estimated using

$$d_{ij} = \exp(-\gamma \Delta t_{ij}) \quad (12)$$

where (Δt_{ij}) represents the temporal distance between the two educational events and (γ) controls the rate at which temporal relevance decreases. The attention score between two nodes is defined as

$$e_{ij} = \frac{q_i k_j^T}{\sqrt{d}} + \beta d_{ij} \quad (13)$$

where (q_i) and (k_j) denote the query and key representations, (d) is the feature dimension, and (β) controls the contribution of temporal dependency. The normalised causal attention weights are calculated as

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum \exp(e_{ik})} \quad (14)$$

The resulting causal representation is obtained through the weighted aggregation of relevant multimodal features:

$$h_i^{casual} = \sum \alpha_{ij} v_j \quad (15)$$

where (v_j) represents the value associated with the preceding educational node. This representation captures the information that is most relevant for understanding the current concept.

3.5 Cross-Modal Layer Fusion (CMLF)

The Cross-Modal Layer Fusion (CMLF) paradigm has been created in order to enable an efficient merging of the complementary information extracted from speech, transcripts, slides, and interactions during learning. Let us denote the causal representations created for the four modalities in layer (l) by (H_l^a) , (H_l^q) , (H_l^i) , and (H_l^r) . CMLF first evaluates the contribution of each modality using a learned fusion coefficient:

$$w_l^m = \frac{\exp(g(H_l^m))}{\sum \exp(g(H_l^n))} \quad (16)$$

where $(g(\cdot))$ represents a learnable importance function, (M) is the number of modalities, and (w_l^m) indicates the relative contribution of modality (m) at layer (l) . The fused representation at each selected layer is then obtained as

$$F_l = \sum w_l^m H_l^m \quad (17)$$

where (F_l) represents the cross-modal representation generated at layer (l) . Instead of discarding the information generated at earlier layers, CMLF selectively combines representations from different depths:

$$F^* = \sum \rho_l F_l \quad (18)$$

where (L) represents the selected fusion layers and (ρ_l) denotes the importance assigned to each layer. A residual connection is incorporated to preserve important information that may otherwise be weakened during fusion:

$$H_{out} = F^* + H_{casual} \quad (19)$$

where (H_{casual}) represents the causal representation obtained from the T-DAG module.

3.6 Causal Context-Aware Quantisation (CCQ)

The Causal Context-Aware Quantisation (CCQ) mechanism is introduced to reduce the computational and memory requirements of EMCAA during deployment on mobile and edge devices. For an intermediate representation (h_i) , a causal-context importance score can be defined as

$$r_i = \lambda_c c_i + \lambda_p f_i + \lambda_t t_i \quad (20)$$

where (c_i) represents causal relevance obtained from the T-DAG module, (f_i) represents cross-modal importance obtained from CMLF, and (t_i) represents temporal relevance. The coefficients (λ_c) , (λ_p) , and (λ_t) control the contribution of these factors. The quantised representation can be expressed as

$$\hat{h}_i = Q(h_i, b_i) \quad (21)$$

where $(Q(\cdot))$ denotes the quantisation operation and (b_i) represents the precision level selected according to the importance of (h_i) . A simple uniform quantisation operation can be represented as

$$\hat{h}_i = \text{round}\left(\frac{h_i - z}{s}\right)s + z \quad (22)$$

where (s) is the quantisation scale and (z) represents the zero-point. CCQ differs from conventional uniform quantisation because the scale and precision can be adjusted according to the contextual importance of the representation. The adaptive precision allocation can be represented as

3.7 Algorithm for EMCAA for real-time educational content interpretation

Algorithm 1: EMCAA for Real-Time Educational Content Interpretation

Input: Speech (X^a), transcript (X^q), slides (X^s), contextual interactions (X^r)
Output: Educational concepts (C), interpretation (I), and summary (S)

1. Initialize EMCAA, T-DAG, CMLF, and CCQ parameters.
2. Receive multimodal educational input:
 $X = \{X^a, X^q, X^s, X^r\}$
3. Preprocess each modality and extract:
 $F_a \leftarrow$ Speech features
 $F_q \leftarrow$ Transcript features
 $F_i \leftarrow$ Slide features
 $F_r \leftarrow$ Interaction features
4. Align all modality features according to their temporal information.
5. Construct the T-DAG:
FOR each educational representation
Calculate temporal and semantic dependency.
IF dependency > threshold THEN
Create a directed causal connection.
ELSE
Ignore the connection.
END IF
END FOR
6. Apply T-DAG Causal Attention to assign higher weights to causally relevant information and suppress redundant features.
7. Apply CMLF:
Calculate modality importance.
IF modality is highly relevant THEN
Increase its fusion weight.
ELSE
Reduce its fusion weight.
END IF
Generate the fused multimodal representation.
8. Apply CCQ:
FOR each fused representation
Calculate causal-context importance.
IF importance is high THEN
Use higher precision.
ELSE
Apply lower-precision quantization.
END IF
END FOR
9. IF inference latency exceeds the target threshold THEN
Increase compression of low-importance features.
ELSE
Maintain the current configuration.
END IF
10. Generate:
 $C \leftarrow$ Key educational concepts
 $I \leftarrow$ Contextual interpretation
 $S \leftarrow$ Content summary
11. Return C, I, and S.
12. End Algorithm

4 RESULTS AND DISCUSSION

4.1 Dataset

The Multimodal Student Attention Dataset, publicly available on Kaggle, was used to evaluate the proposed EMCAA framework. The dataset contains multimodal information related to student attention and learning activities and was organised into synchronised learning segments. The data were divided into 70% training, 15% validation, and 15% testing sets. The training set was used for model learning, the validation set for parameter selection and optimisation, and the testing set for final performance evaluation. The same data split was applied to EMCAA and all baseline models to ensure consistent comparison.

4.2 Pre-processing

Each modality was pre-processed separately before temporal alignment. Visual frames were resized to 224×224 pixels and normalised. Audio signals were normalised and segmented into 2-s overlapping windows to preserve temporal information. Lecture transcripts were cleaned and limited to a maximum sequence length of 128 tokens. A temporal tolerance of 1 s was used to synchronise information across modalities, while numerical features were normalised before being processed by the T-DAG module.

4.3 Experimental setup

The EMCAA framework was implemented using Python and PyTorch. Model training employed the Adam optimiser, with a learning rate of (1×10^{-4}) , a batch size of 32, and a maximum of 50 epochs. Early stopping was applied when validation performance showed no improvement for eight consecutive epochs. CCQ assigned 8-bit precision to high-importance representations and 4-bit precision to low-importance representations. The complete experimental configuration is summarised in Table 1.

Table 1. Experimental configuration

Parameter	Setting
Dataset	Multimodal Student Attention Dataset
Data source	Kaggle
Training/Validation/Testing	70%/15%/15%
Image/frame resolution	224×224 pixels
Audio segment duration	2 s
Temporal synchronization tolerance	1 s
Maximum text sequence length	128 tokens
Optimizer	Adam
Learning rate	(1×10^{-4})
Batch size	32
Maximum training epochs	50
Early stopping patience	8 epochs
High-importance quantization	8-bit
Low-importance quantization	4-bit

4.4 Performance evaluation

EMCAA was evaluated using accuracy, precision, recall, F1-score, inference latency, and memory consumption. Performance was compared with CNN-LSTM, 3D-DenseNet-SA, CNN-ViT, EfficientNetV2-RNN, and MC-MIFA. An ablation study was also conducted to assess the individual contributions of T-DAG, CMLF, and CCQ.

Overall performance comparison. Table 2 compares EMCAA with the baseline models. EMCAA achieved the highest performance, with 91.4% accuracy, 91.0% precision, 90.8% recall, and 90.9% F1-score. Compared with MC-MIFA, the strongest baseline, EMCAA improved accuracy by 1.2 percentage points and F1-score by 1.4 percentage points, indicating improved predictive performance.

Table 2. Overall performance comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-LSTM	84.8	83.9	82.7	83.3
3D-DenseNet-SA	86.7	85.9	85.1	85.5
CNN-ViT	88.5	87.8	87.1	87.4
EfficientNetV2-RNN	89.1	88.6	88.0	88.3
MC-MIFA	90.2	89.7	89.3	89.5
EMCAA	91.4	91.0	90.8	90.9

Computational efficiency. Table 3 compares the computational efficiency of EMCAA with the baseline models. EMCAA achieved the lowest inference latency (43 ms) and memory consumption (329 MB). Compared with MC-MIFA, this represents reductions of 24.6% in inference latency and 18.0% in memory consumption, indicating improved efficiency for resource-constrained deployment.

Table 3. Computational efficiency comparison

Model	Inference Latency (ms)	Memory Consumption (MB)	Relative Efficiency
CNN-LSTM	82	512	Low
3D-DenseNet-SA	74	478	Low
CNN-ViT	68	451	Moderate
EfficientNetV2-RNN	61	426	Moderate
MC-MIFA	57	401	High
EMCAA	43	329	Very High

Compared with MC-MIFA, EMCAA reduced inference latency from 57 ms to 43 ms (24.6%) and memory consumption from 401 MB to 329 MB (18.0%), while improving accuracy by 1.2 percentage points and F1-score by 1.4 percentage points. These results indicate that EMCAA improves computational efficiency without compromising predictive performance. The decrease in latency is due to selective causal attention and cross-modal fusion, which minimise any interaction of irrelevant features.

The CCQ module also minimises memory usage by decreasing the numerical precision of less critical features.

Ablation study. Table 4 indicates an ablation study was carried out to evaluate the role of the three core components of EMCAA. It involved comparing the full version with versions where either T-DAG, CMLF, or CCQ was left out. Removing T-DAG produced the largest decline in accuracy, from 91.4% to 88.9%, indicating its importance for predictive performance.

Table 4. Ablation analysis of EMCAA

Model Variant	Accuracy (%)	F1-Score (%)	Latency (ms)	Memory (MB)
Without T-DAG	88.9	88.3	51	352
Without CMLF	89.7	89.1	48	341
Without CCQ	90.5	90.0	46	331
Full EMCAA	91.4	90.9	43	329

Figure 2 Removing CMLF reduced accuracy to 89.7%, while removing CCQ produced a smaller accuracy reduction of 90.5%. Overall, the ablation results show that the three components contribute differently to the performance and efficiency of EMCAA.

Ablation Study Results

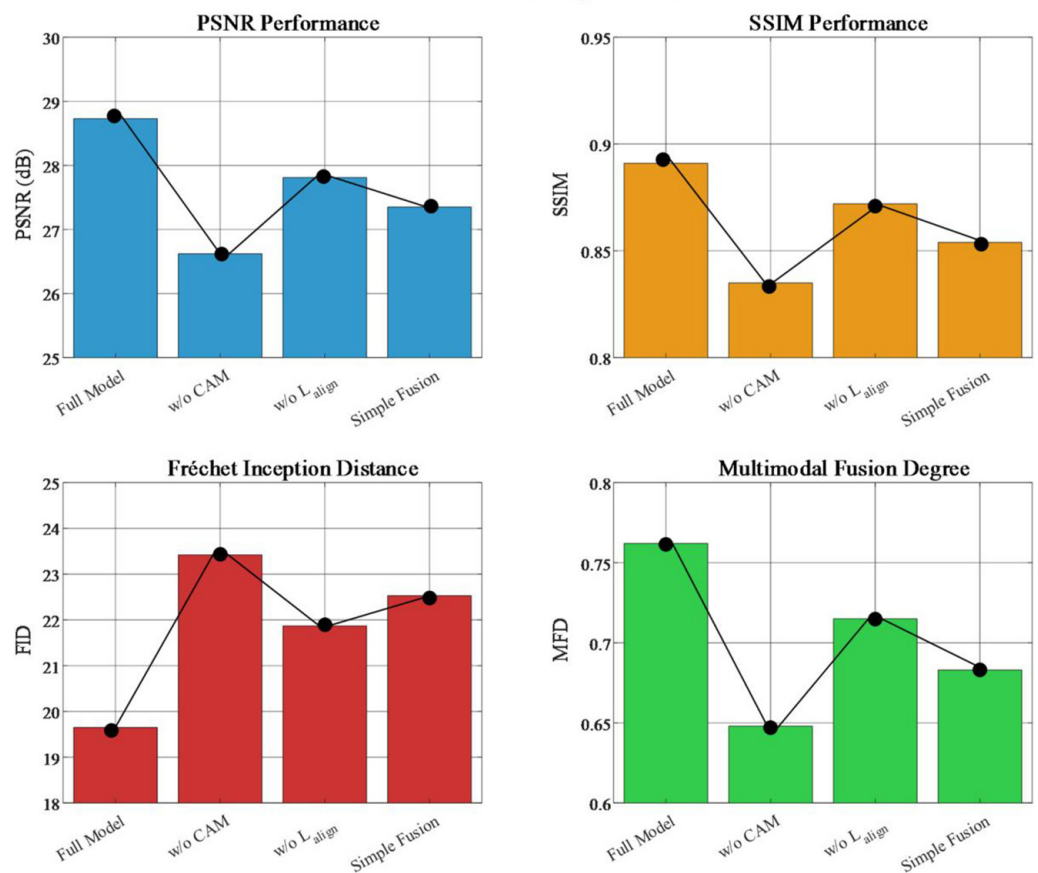


Fig. 2. Ablation analysis of T-DAG, CMLF, and CCQ

5 DISCUSSION

The results demonstrate that EMCAA achieves a favourable balance between predictive performance and computational efficiency. The model achieved 91.4% accuracy and a 90.9% F1-score, while reducing inference latency and memory consumption relative to the baseline models. The ablation analysis further shows that T-DAG contributes most strongly to predictive performance, as its removal produced the largest decline in accuracy. CMLF also improved performance by integrating complementary multimodal representations, while CCQ supported efficient deployment through adaptive precision allocation. Overall, the findings suggest that selectively prioritising temporally and contextually relevant representations can improve multimodal educational content processing without treating all input information equally. This selective processing strategy provides a practical direction for developing efficient multimodal AI systems for resource-constrained mobile and edge learning environments.

6 CONCLUSION AND FUTURE WORK

This study proposed the Efficient Multimodal Causal Attention Architecture (EMCAA) for real-time educational content interpretation on mobile and edge devices. EMCAA integrates T-DAG, CMLF, and CCQ to model temporally relevant dependencies, fuse multimodal representations, and reduce deployment overhead. The model achieved 91.4% accuracy and a 90.9% F1-score, with an inference latency of 43 ms and memory consumption of 329 MB. Compared with MC-MIFA, EMCAA reduced latency by 24.6% and memory consumption by 18.0%, while improving accuracy by 1.2 percentage points. These findings demonstrate the potential of EMCAA for efficient multimodal processing in resource-constrained educational environments. Future research should evaluate EMCAA using larger and more diverse educational datasets, including different subjects, languages, and learning environments. Additional modalities, such as eye tracking and physiological signals, may also be investigated. Further work should explore federated and privacy-preserving learning to strengthen secure deployment across mobile and edge devices.

7 DATA AVAILABILITY STATEMENT

The study used the publicly available Multimodal Student Attention Dataset available on Kaggle: <https://www.kaggle.com/datasets/colabsss/multimodal-student-attention-dataset>.

8 REFERENCES

- [1] I. S. Almuniri *et al.*, “Beyond peak accuracy: A stability-centric framework for reliable multimodal student engagement assessment,” *Scientific Reports*, vol. 16, no. 1, 2026. <https://doi.org/10.1038/s41598-025-31215-7>
- [2] P. Bhardwaj *et al.*, “Application of deep learning on student engagement in e-learning environments,” *Computers & Electrical Engineering*, vol. 93, p. 107277, 2021. <https://doi.org/10.1016/j.compeleceng.2021.107277>

- [3] I. Boutabia *et al.*, “Hybrid CNN-ViT model for student engagement detection in open class-room environments,” *SN Computer Science*, vol. 6, no. 6, 2025. <https://doi.org/10.1007/s42979-025-04228-2>
- [4] B. D. Bhavani, R. Shetty, and D. Mahesh, “Enhancing student engagement and personalized learning through AI tools: A comprehensive review,” *Computer Science and Engineering International Journal*, vol. 15, no. 1, pp. 111–130, 2025. <https://doi.org/10.5121/cseij.2025.15113>
- [5] G. Chen *et al.*, “Video Mamba suite: State space model as a versatile alternative for video understanding,” *International Journal of Computer Vision*, vol. 134, no. 1, 2026. <https://doi.org/10.1007/s11263-025-02597-y>
- [6] Z. Dan, Z. Yali, and Z. Tong, “A multimodal AI framework for real-time student engagement detection and adaptive feedback in higher education,” *Journal of King Saud University—Computer and Information Sciences*, vol. 38, 2026. <https://doi.org/10.1007/s44443-026-00639-0>
- [7] A. A. Fraiden, “Anticipatory thinking and AI-driven assessments: A balanced approach to AI integration in education aligned with Saudi Vision 2030,” *African Journal of Biomedical Research*, vol. 27, no. 3, pp. 619–630, 2024. <https://doi.org/10.53555/AJBR.v27i3.2560>
- [8] J. D. T. Guerrero-Sosa, F. P. Romero, V. H. M. Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. Á. Olivas, “A comprehensive review of multimodal analysis in education,” *Applied Sciences*, vol. 15, no. 11, p. 5896, 2025. <https://doi.org/10.3390/app15115896>
- [9] S. Gupta, P. Kumar, and R. Tekchandani, “A multimodal facial cues based engagement detection system in e-learning context using deep learning approach,” *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 28589–28615, 2023. <https://doi.org/10.1007/s11042-023-14392-3>
- [10] Z. Ji and S. Li, “Multimodal alignment and attention-based person search via natural language description,” *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11147–11156, 2020. <https://doi.org/10.1109/JIOT.2020.2995148>
- [11] W. Li, F. Deng, and Z. Li, “Question-guided attention and cross-modal alignment for knowledge-based visual question answering,” *Information Processing & Management*, vol. 63, no. 3, p. 104578, 2026. <https://doi.org/10.1016/j.ipm.2025.104578>
- [12] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, “Three-dimensional densenet self-attention neural network for automatic detection of student’s engagement,” *Applied Intelligence*, vol. 52, no. 12, pp. 13803–13823, 2022. <https://doi.org/10.1007/s10489-022-03200-4>
- [13] H. Chen and Y. Shang, “Construction and Application of a learning resource sharing platform in higher education based on mobile interactive technology,” *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 20, no. 01, pp. 19–33, 2026. <https://doi.org/10.3991/ijim.v20i01.59785>
- [14] F. Naseer and S. Khawaja, “Mitigating conceptual learning gaps in mixed-ability classrooms: A learning analytics-based evaluation of AI-driven adaptive feedback for struggling learners,” *Applied Sciences*, vol. 15, no. 8, p. 4473, 2025. <https://doi.org/10.3390/app15084473>
- [15] P. Thottempudi, V. Kumar, and R. Kumar, “Dynamic multi-modal attention network for robust and real-time through-wall human activity recognition,” *Results in Engineering*, vol. 28, p. 107632, 2025. <https://doi.org/10.1016/j.rineng.2025.107632>
- [16] W. Qu, L. Guo, J. Cui, and X. Jin, “Multimodal attention-based instruction-following part-level affordance grounding,” *Applied Sciences*, vol. 14, no. 11, p. 4696, 2024. <https://doi.org/10.3390/app14114696>
- [17] F. A. Raza, A. D. Singh, J. J. S. Kovilpillai, A. Hamdan, and V. Rajaratnam, “Safeguarding integrity in AI-enhanced education: Stakeholder perspectives on accuracy, validity, and ethics in ASEAN,” *European Journal of STEM Education*, vol. 10, no. 1, p. 22, 2025. <https://doi.org/10.20897/ejsteme/17307>

- [18] M. Roshanaei, H. Olivares, and R. R. Lopez, "Harnessing AI to foster equity in education: Opportunities, challenges, and emerging strategies," *Journal of Intelligent Learning Systems and Applications*, vol. 15, no. 4, pp. 123–140, 2023. <https://doi.org/10.4236/jilsa.2023.154009>
- [19] F. M. Shiri, E. Ahmadi, M. Rezaee, and T. Perumal, "Detection of student engagement in e-learning environments using EfficientNetV2-L together with RNN-based models," *Journal of Artificial Intelligence*, vol. 6, no. 1, pp. 85–103, 2024. <https://doi.org/10.32604/jai.2024.048911>
- [20] C. Wang, S. Zhang, T. Li, and Z. Lu, "MC-MIFA: A causal-aware hybrid state space framework for robust multimodal student engagement analysis," *Scientific Reports*, vol. 16, Art. no. 20467, 2026. <https://doi.org/10.1038/s41598-026-51563-2>
- [21] C. W. Woo *et al.*, "Students' perception of the classroom environment: A comparison between innovative and traditional classrooms," *Journal of School Teaching and Learning*, vol. 22, no. 1, pp. 17–31, 2022. <https://doi.org/10.14434/josotl.v22i1.30735>
- [22] C. Wang, "Mobile AIGC Image Generation Interactive Training Model for Design Education," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 20, no. 02, pp. 63–77, 2026. <https://doi.org/10.3991/ijim.v20i02.60137>

9 AUTHORS

Jin Wang is affiliated with the Faculty of Information Technology, City University Malaysia. The research interests include multimodal financial sentiment analysis, financial communication, machine learning, explainable AI, and causal inference (E-mail: 57213681@qq.com).

Dr. Inam Ullah holds a PhD in Computer Science from Universiti Teknikal Malaysia Melaka (UTeM). His research interests include artificial intelligence, machine learning, deep learning, cybersecurity, computer vision, and medical image analysis. He is currently affiliated with Faculty of Information Technology, City University Malaysia (E-mail: inam.ullah@city.edu.my).