World Journal of Educational Studies
Print ISSN: 2959-9989
Online ISSN: 2959-9997
DOI: https://doi.org/10.61784/wjes3106

CLASSROOM VIDEO BEHAVIOUR PROPOSAL MODEL BASED ON MULTIMODAL ATTENTION MECHANISMS AND ADAPTIVE SEARCH

Ji Li[1], Jin Lu[2*], MaoLi Wang[3]
[1]*Research Management Office, Shenzhen Polytechnic University, Shenzhen 518000, Guangdong, China.*
[2]*Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen Polytechnic University, Shenzhen 518000, Guangdong, China.*
[3]*Institute for Technical and Vocational Education, Shenzhen Polytechnic University, Shenzhen 518000, Guangdong, China.*
Corresponding Author: Jin Lu

Abstract: The analysis of teacher-student behaviour within classroom settings forms the bedrock of smart education research and application. However, existing general-purpose behaviour detection models often exhibit suboptimal accuracy and efficiency when processing extended classroom videos. This stems primarily from their inability to effectively address four key challenges: variable behaviour duration, complex semantic layers, heterogeneous multimodal information, and high background redundancy. To address these challenges, this paper proposes a novel classroom video behaviour proposal model. Its core innovation lies in the synergistic utilisation of multimodal attention mechanisms and adaptive search strategies. First, a robust multimodal feature extraction backbone network is constructed to extract highly discriminative features from video, audio, and automatic speech recognition (ASR) transcribed text. Subsequently, a hierarchical multimodal attention fusion module is designed. This module dynamically captures and integrates behaviour-related key visual segments, audio events, and semantic keywords through two-stage computations: intra-modal attention and cross-modal attention. Building upon this foundation, we innovatively propose an adaptive boundary search algorithm inspired by reinforcement learning principles. This algorithm dynamically adjusts search stride and direction based on the contextual semantics and behavioural confidence of the current video segment, enabling efficient and precise boundary localisation for action proposals within lengthy video sequences. To validate model performance, we constructed a large-scale classroom behaviour dataset, 'Edu-Action'. Comprehensive experimental results demonstrate that our model achieves significant improvements in the core evaluation metric for action proposal tasks, average recall at action number (AR@AN). At a tIoU threshold of 0.5, recall reaches 68.7%, comprehensively outperforming multiple advanced baseline models. Extensive ablation studies further validate the effectiveness and necessity of each component within the model. This paper presents an effective solution for fine-grained action localisation in long-duration video environments, holding significant theoretical implications and broad practical application prospects.
Keywords: Behavioural proposal generation; Multimodal learning; Attention mechanisms; Adaptive search; Classroom video analysis; Smart education; Deep learning

1 INTRODUCTION

The intrinsic demand for enhancing quality and efficiency within the context of educational informatisation has made the digital and intelligent analysis of classroom teaching processes a research hotspot in the field of education [1]. The vast volume of classroom video recordings generated and stored constitutes a valuable educational big data goldmine. Automatically identifying, locating, and understanding teaching behaviours such as 'teacher board writing', 'student raising hands to speak', and 'group collaborative inquiry' holds revolutionary significance for achieving objective classroom teaching evaluation, precise teaching reflection, personalised learning situation analysis, and deep mining of educational big data [2,3]. Time-based behaviour proposals serve as the bridge connecting low-level video features with high-level behavioural understanding, representing the primary and critical component within the behavioural analysis pipeline [4]. The task objective is to precisely locate all potential start and end timepoints for behaviours of interest within an unedited, uncropped video sequence, without pre-assigned behavioural category labels, and to generate confidence scores for these locations [5]. However, the unique characteristics of classroom settings render this task exceptionally complex.

Firstly, the extreme variability in behavioural duration, with classroom actions exhibiting an exceptionally broad distribution range [6]. Instantaneous, atomic behaviours such as 'a pupil raising their hand' or 'a teacher pointing at the screen' may last merely 1 or 2 seconds [7]. Conversely, complex, high-level teaching activities like 'group project collaboration' or 'classroom debates' may persist for several minutes or even an entire lesson. This vast scale disparity poses a formidable challenge to a model's multiscale perception capabilities. Secondly, the hierarchical and nested nature of behavioural semantics. Classroom activities do not exist in isolation but form a complex hierarchical structure [8]. For instance, a macro-level behaviour like 'teacher explaining a new concept' may internally embed multiple

micro-behaviours such as ‘teacher writing on the board,’ ‘teacher posing questions,’ or ‘playing instructional videos.’ This phenomenon of ‘behaviours within behaviours’ makes it exceedingly difficult to clearly and accurately delineate behavioural boundaries. Thirdly, the strong multimodal dependency of behavioural identification. Defining classroom behaviours often cannot rely solely on visual information. Auditory cues—including shifts in the teacher's intonation, students' choral responses, sudden quietness—and linguistic information—such as specific phrasing in teacher questions or core conceptual terms mentioned during explanations—are crucial clues for identifying behavioural onset and transitions [9]. For instance, the initiation of a ‘teacher posing a question’ behaviour may be jointly signalled by visual cues such as a ‘teacher's pause’, auditory cues like an ‘upward inflection in tone’, and textual cues such as the presence of ‘interrogative words’. Effectively aligning and integrating these heterogeneous modal information streams represents a core challenge. Fourthly, the high redundancy and intra-class variability within video backgrounds. Extensive segments unrelated to target behaviours exist within lengthy classroom videos, such as student self-study periods, classroom silences, and camera transitions [10]. Traditional sliding window or dense anchor methods generate numerous invalid proposals in these regions, resulting in substantial computational resource wastage and reduced recall rates. Concurrently, the visual and auditory manifestations of the same behaviour may exhibit significant variations across different classes and subjects, demanding robust generalisation capabilities from the model.

Existing behavioural proposal methods, such as the anchor-based SSN[11] and boundary-matching BMN[12], have achieved success on general datasets. However, their original design did not sufficiently account for the aforementioned particularities of classroom scenarios. Most rely on a single visual modality or perform simple post-fusion of multimodal information, failing to fully exploit the deep interconnections between modalities. Furthermore, they commonly employ predefined, fixed-scale anchors or sliding windows, rendering them ill-suited to accommodate the extreme temporal variability inherent in classroom behaviours. This limitation creates bottlenecks in both accuracy and efficiency. To address these issues, this paper proposes an end-to-end classroom behaviour proposal model. Its core contribution lies in designing a hierarchical multimodal attention fusion module that dynamically and efficiently integrates visual, auditory, and linguistic information while focusing on behaviour-relevant key cues. Concurrently, an adaptive boundary search algorithm is introduced. By simulating human browsing and focusing behaviours, it dynamically adjusts search strategies, significantly enhancing efficiency in long-video analysis while maintaining recall rates. Notably, a large-scale, high-quality classroom behaviour dataset, ‘Edu-Action’, has been constructed to advance research in this domain.

2 RELATED RESEARCH

2.1 Generation of Timed Behaviour Proposals

Time-based action proposal generation constitutes a foundational task within video understanding, with its specific architecture illustrated in Figure 1. Early approaches such as S-CNN[13] and SST[14] primarily relied on sliding windows of varying scales across video sequences to generate candidate segments, a method characterised by high computational demands and limited flexibility. Subsequently, the BMN model, based on boundary matching principles, achieved high-quality proposal generation by evaluating all candidate intervals between start-end point pairs, becoming a landmark work in the field. Later efforts such as DBG[15] and RTD-Net[16] further optimised the precision of boundary localisation. However, these approaches were primarily designed for short video clips or sports events. Their fixed anchor scales or matching mechanisms often prove inadequate when confronted with behaviours spanning vast durations within lengthy classroom videos. Moreover, the substantial volume of negative sample proposals they generate severely impacts training efficiency and final performance.

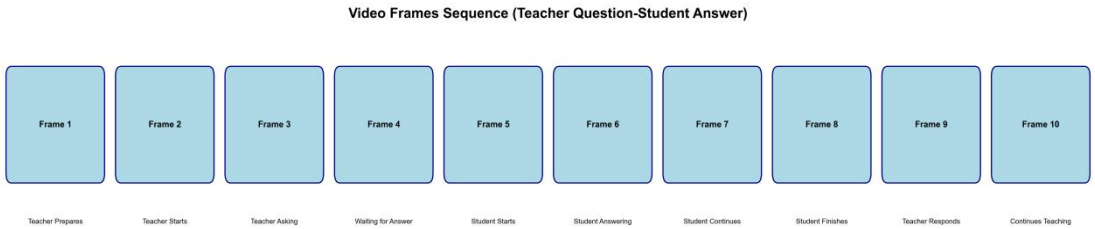

Figure 1 Temporal Behavior Proposal Generation Structure

2.2 Multimodal Video Understanding

Video is inherently a natural amalgamation of visual, auditory, and textual information. Effectively integrating these heterogeneous modalities constitutes the core challenge of multimodal learning [17]. Early fusion approaches included simple feature concatenation, max/mean pooling, and similar techniques. Subsequently, tensor-based fusion methods and bilinear pooling were proposed to capture more complex interactions between modalities, though these often entailed substantial computational overhead [18]. In recent years, attention mechanisms—particularly self-attention within Transformer architectures and cross-modal attention—have become mainstream techniques for multimodal fusion [19]. These enable models to dynamically compute importance weights across different modalities and within the

same modality at distinct temporal steps. In classroom settings, Sameer et al. attempted to utilise audio event detection to augment behaviour recognition, yet failed to achieve end-to-end deep fusion [20]. Our work draws inspiration from this approach but introduces a more refined hierarchical attention structure designed to capture cross-modal temporal alignment relationships in classroom behaviours with greater precision.

2.3 Efficient Video Analysis and Search Strategies

To address the inefficiency of long video analysis, researchers have proposed various strategies. Some approaches employ a two-stage strategy, involving coarse screening followed by fine-tuning. Others attempt to learn search strategies through reinforcement learning, intelligently skipping irrelevant frames. In recent years, the state space model Mamba has garnered attention for its efficiency in modelling long sequences [21]. Our adaptive search module shares the underlying philosophy with such approaches, but innovatively links the search stride directly to the local contextual information and behavioural confidence of the current segment. This achieves a data-driven, content-aware dynamic search mechanism better suited to the uneven distribution of classroom behaviours.

3 METHODS

In the method proposed in this study, we construct an end-to-end framework whose core process commences with deep feature extraction from visual, audio, and transcribed textual components of classroom videos. Subsequently, through a hierarchical multimodal attention module, it dynamically calculates intra-modal and cross-modal attention weights to adaptively fuse the most semantically relevant visual segments, audio events, and textual keywords associated with behavioural semantics. This ultimately drives an innovative adaptive search algorithm. This algorithm intelligently adjusts the search stride and direction based on the contextual semantics and behavioural confidence of the current segment. Consequently, it efficiently and accurately locates the start and end boundaries of potential behavioural segments within lengthy video sequences. The specific architecture is illustrated in Figure 2. The overall architecture comprises three core components: multimodal feature extraction, a hierarchical multimodal attention fusion module, and an adaptive proposal search module.

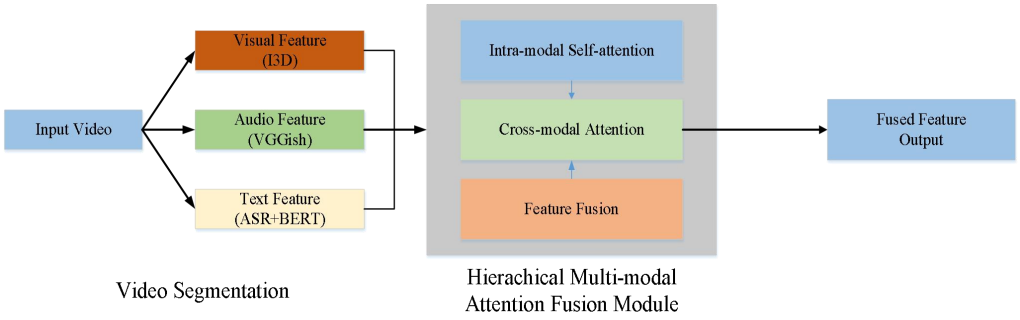

Figure 2 Model Architecture Diagram

3.1 Multimodal Feature Extraction

As illustrated in Figure 1, given a long classroom video clip V, we first uniformly partition it into non-overlapping segments of length L. For each segment t, we concurrently extract features from three modalities, as detailed in Figure 3.

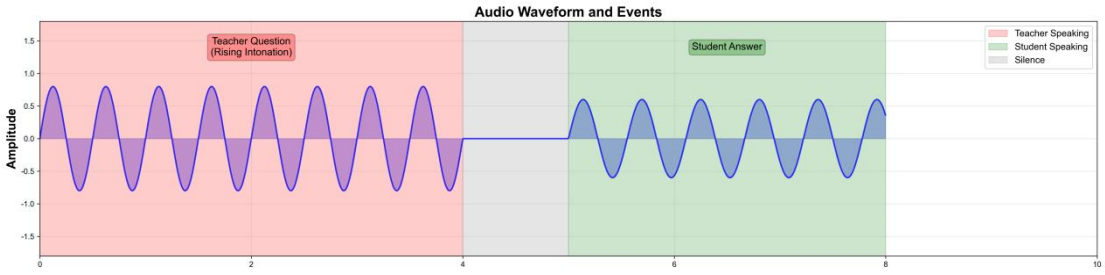

Figure 3 Multimodal Features

Visual features $F_t^v \in R^{d_v}$. To capture appearance and motion information, we employ the I3D model pre-trained on the large-scale action recognition dataset Kinetics-400 [22] as the backbone network [23]. For each clip, we extract its RGB frames and corresponding optical flow frames, pass them through the I3D network respectively, and concatenate the features obtained before the final fully-connected layer to yield the final visual feature vector F_t^v.

Audio features $F_t^a \in R^{d_a}$. We employ the VGGish model to extract audio features [24]. This model, pre-trained on a large-scale YouTube audio dataset, captures semantic information of meaningful audio events. We extract the log-Mel spectrogram from the audio waveform aligned with the video clip and feed it into the VGGish network to obtain F_t^a.

Text features $F_t^t \in R^{d_t}$. We first utilise industrial-grade automatic speech recognition (ASR) services (such as Google Cloud Speech-to-Text[25] or Azure Speech Services[26]) to convert the audio stream into a timestamped text transcription. Subsequently, for each video segment t, we aggregate all corresponding transcribed text sentences within its temporal scope. Finally, sentence embedding vectors for this aggregated text are obtained using a pre-trained BERT model[27] as text feature Ftt. This feature encapsulates rich semantic information, such as keywords and interrogative sentences.

Ultimately, we obtained three feature sequences,

$$F^v = \{F_1^v, F_2^v, ..., F_L^v\}, F^a = \{F_1^a, F_2^a, ..., F_L^a\}, F^t = \{F_1^t, F_2^t, ..., F_L^t\}。$$

3.2 Hierarchical Multi-modal Attention Fusion Module

This module is designed to dynamically and selectively fuse information from three modalities, amplifying behaviour-relevant cues while suppressing irrelevant noise. Its architecture, as depicted in Figure 4, comprises two hierarchical levels.

First, we apply a Transformer encoder layer [28] to the feature sequences of each modality, performing intra-modal self-attention calculations. Taking the visual modality as an example, this is illustrated in Equation 1.

$$\tilde{F}^v = Transformer(F^v) \tag{1}$$

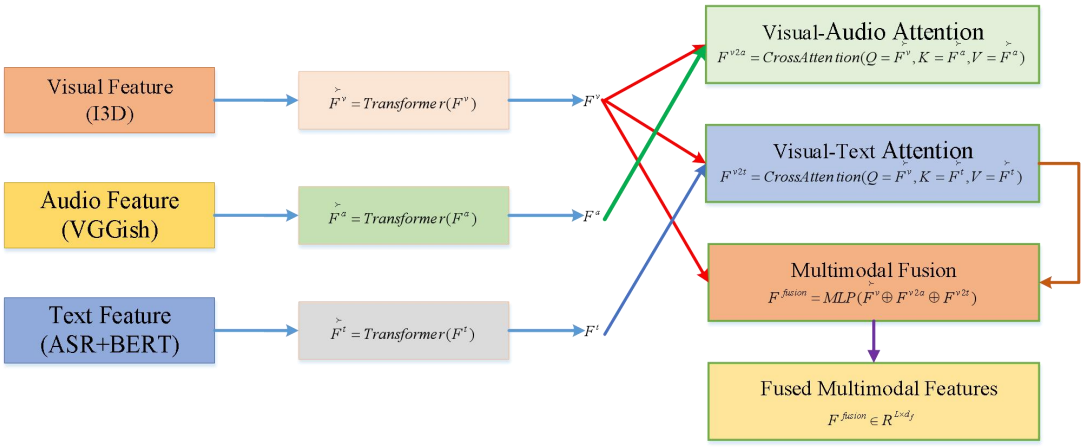

Figure 4 Hierarchical Multi-modal Attention Fusion Module

Here, F^v denotes the sequence of visual features enhanced through modal self-attention. The self-attention mechanism captures long-range temporal dependencies. For instance, it enables the model to recognise that an action such as ‘a pupil standing up’ may correlate with an action like ‘a teacher posing a question’ several seconds prior, even when intervening frames are unrelated. Similarly, the enhanced audio features F^a and text features F^t are obtained as shown in Equations 2 and 3.

$$\tilde{F}^a = Transformer(F^a) \tag{2}$$

$$\tilde{F}^t = Transformer(F^t) \tag{3}$$

Following the extraction of enhanced features across modalities, cross-modal information interaction and fusion are performed. A vision-dominant fusion strategy is adopted, as vision serves as the primary vehicle for behavioural expression. Specifically, visual features are employed as the Query, with audio and text features functioning as Key and Value respectively, to conduct cross-attention computations.

Visual and audio fusion enables the model to recalibrate the importance of visual features using audio cues , such as sudden applause or loud questions , as illustrated in Equation 4. When audio features indicate ‘applause,’ the model prioritises segments showing ‘students standing’ or ‘teachers gesturing’ in the visual data.

$$F^{v2a} = CrossAttention(Q = \tilde{F}^v, K = \tilde{F}^a, V = \tilde{F}^a) \tag{4}$$

The fusion of visual and textual information enables semantic cues to guide the allocation of visual attention, as illustrated in Equation 5. For instance, when the text prompts 'Let's discuss this in groups,' the model will purposefully seek visual patterns such as 'students turning their heads' or 'forming groups' within the corresponding visual segments.

$$F^{v2t} = CrossAttention(Q = \overset{\succ}{F}^{v}, K = \overset{\succ}{F}^{t}, V = \overset{\succ}{F}^{t}) \tag{5}$$

Finally, we fuse the original enhanced visual features with the two cross-modal attention outputs, as shown in Equation 6.

$$F^{fusion} = MLP(\overset{\succ}{F}^{v} \oplus F^{v2a} \oplus F^{v2t}) \tag{6}$$

Here, $\oplus$ denotes the vector concatenation operation, while MLP represents a multi-layer perceptron used to project the concatenated high-dimensional features onto a unified fusion feature space $F^{fusion} \in R^{L \times d_f}$.

3.3 Adaptive Proposal Search Module

Traditional dense generation-evaluation strategies prove inefficient for long-form videos. Inspired by human viewing behaviour—specifically the cycle of "skimming through—identifying points of interest—pausing to examine in detail"—we have designed an adaptive proposal search module. This module operates iteratively, with its core principle being the dynamic determination of the "direction" and "step size" for the next search step based on the context surrounding the current search position. As illustrated in Figure 5, the adaptive proposal search process emulates intelligent human browsing behaviour during long-form video consumption. It abandons the traditional sliding window strategy with fixed strides, instead dynamically adjusting search granularity and direction based on contextual semantic information from the current video segment and predicted behavioural confidence. When the search pointer resides in behaviourally sparse regions, the algorithm employs larger strides to rapidly skip irrelevant segments, enhancing efficiency. Conversely, upon detecting regions of high behavioural confidence, it automatically switches to a fine-grained small-step search mode. Within these zones, it densely generates proposals and performs boundary fine-tuning. This achieves an optimal balance between efficiency and precision within lengthy lecture videos, avoiding computational waste on irrelevant background content while ensuring the capture of fleeting or marginally defined behavioural patterns.

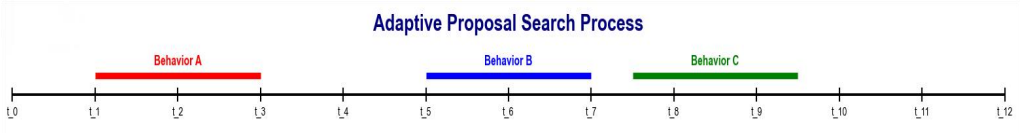

Figure 5 Adaptive Proposal Search Process

In each iteration i, the model maintains a current search pointer p_i and observes a local context window C_i centred on a feature p_i. The window's features F^{fusion} are encoded by a small neural network g_{ctx} applied to the corresponding fragments from the fused features, as shown in Equation 7.

$$s_i = g_{ctx}(C_i) \tag{7}$$

Here, s_i denotes the current state representation.

4 EXPERIMENTS AND DISCUSSION

4.1 Experimental Setup

Regarding dataset design, as existing public datasets (such as ActivityNet and THUMOS) are unsuitable for classroom scenarios, we have developed our own 'Edu-Action' dataset. This dataset comprises 500 hours of authentic classroom videos spanning different educational stages (primary, secondary, and sixth form) and subjects (Chinese, mathematics, English, etc.). A team of educational experts was engaged to annotate over 20,000 time intervals according to rigorous standards. These annotations encompass ten core teaching behaviours: 'teacher instruction', 'board writing', "questioning", 'individual student responses', 'collective student responses', 'group discussions', 'individual practice', 'teacher-student interaction', and 'student-student interaction'.

Regarding evaluation metrics, we employ the most prevalent assessment protocol within the behavioural proposal domain: average recall across different IoU thresholds. We report average recall at the tIoU threshold set {0.5, 0.55, ..., 0.95}, calculating AR@50 and AR@100 for average proposal counts of 50 and 100 respectively. Additionally, we output the AUC, representing the area under the recall curve across the tIoU threshold range [0.5:0.05:0.95].

In terms of implementation details, the video is downsampled to 5 frames per second. I3D, VGGish, and BERT all utilise pre-trained weights which are fixed, with only the subsequent fusion network being fine-tuned. The Adam optimiser is employed, with an initial learning rate of 1e-4. Training of the APS module adopts curriculum learning, commencing with simpler videos.

4.2 Experimental Comparison

In terms of experimental comparisons, we contrast our approach with several state-of-the-art generalised action proposal methods, including SSN, BMN, MGG, and RTD-Net. To ensure fair evaluation, all baseline methods were retrained on the Edu-Action dataset using the identical multimodal features provided by us, as detailed in Table 1.

Table 1 Performance Comparison of Behavioural Proposals on the Edu-Action Test Set

Mehtod	Modal	AUC	@0.5	@0.7	@0.9
SSN[11]	RGB	28.1	42.5	28.9	8.1
BMN[12]	RGB	32.5	52.1	36.8	12.5
MGG[29]	RGB	33.8	54.3	38.1	13.2
RTD-Net[16]	RGB	35.2	56.7	40.5	14.8
BMN[12]	RGB+Audio	34.9	55.8	39.4	13.7
BMN[12]	RGB+Audio+Text	36.1	57.5	41.0	14.5
Ours	RGB+Audio+Text	41.7	68.7	53.4	21.2

The experimental results demonstrate that our approach achieves significant and consistent superiority over all baseline models across all evaluation metrics. Particularly under stringent metrics measuring boundary localisation accuracy, it achieves absolute performance gains of 12.0% and 12.9% respectively compared to the strongest baseline, RTD-Net. This conclusively demonstrates our model's distinct advantage in generating precise, high-quality boundary proposals. It is noteworthy that while incorporating multimodal information into baseline methods yields some performance gains, these improvements remain limited. This indicates that simple feature concatenation strategies struggle to fully exploit the deep correlations between multimodal information. In contrast, the hierarchical attention mechanism proposed herein achieves more effective fusion through dynamic weight allocation. Furthermore, our approach maintains a recall rate exceeding 21%, whereas all baseline methods fall below 15%. This outcome robustly validates our model's exceptional precision in behavioural boundary localisation, demonstrating superior alignment with actual behavioural intervals.

4.3 Ablation Experiment

The ablation experiments aim to systematically validate the effectiveness of each core component within the model, following the implementation process outlined below. First, building upon the complete model, we sequentially removed or substituted specific modules. This included: - Isolating audio and text modalities to validate multimodal necessity. Replacing hierarchical attention with simple feature concatenation to assess fusion efficacy. Ablating intra-modal and cross-modal attention submodules to analyse their respective contributions. Substituting adaptive search strategies with traditional fixed-step sliding windows to evaluate efficiency advantages. All comparative experiments were conducted under identical training/validation/test dataset partitions, employing consistent hyperparameter settings and evaluation metrics to ensure comparability. This approach precisely quantifies each component's contribution to final performance. Specific results are presented in Table 2.

Table 2 Ablation Experiment Results

Model Configuration	AUC	@0.5	@0.7
Complete Model	41.7	68.7	53.4
- w/o Audio modal	38.9	64.1	48.5
- w/o Text modal	39.5	65.0	49.3
- w/o HMAF	37.2	61.5	45.7
- w/o Modal attention	40.1	66.3	51.0
- w/o Cross-modal attention	40.5	66.8	51.5
- w/o APS	35.9	59.8	43.6

Regarding the necessity of multimodality, removing either the audio or text modality resulted in a significant decline in performance, decreasing the AUC by 2.8 and 2.2 points respectively. This demonstrates the indispensable role of multimodal information in classroom behaviour analysis. Regarding the efficacy of the HMAF module [30], performance plummeted when this module was replaced with simple feature concatenation, with AUC dropping from 41.7 to 37.2. This demonstrates the critical role of our proposed attention mechanism in information fusion. Furthermore, removing either the intra-modal or cross-modal attention submodules separately also resulted in performance degradation, indicating both are effective, with intra-modal attention playing a slightly greater role than cross-modal attention. Regarding the effectiveness of the APS module, replacing it with a traditional fixed-step sliding window yielded the most pronounced performance decline, with AUC dropping to 35.9, demonstrating the substantial advantages of adaptive search strategies in enhancing both accuracy and efficiency. We also measured inference time, where the full model achieved approximately 3.5 times faster processing compared to the sliding-window variant.

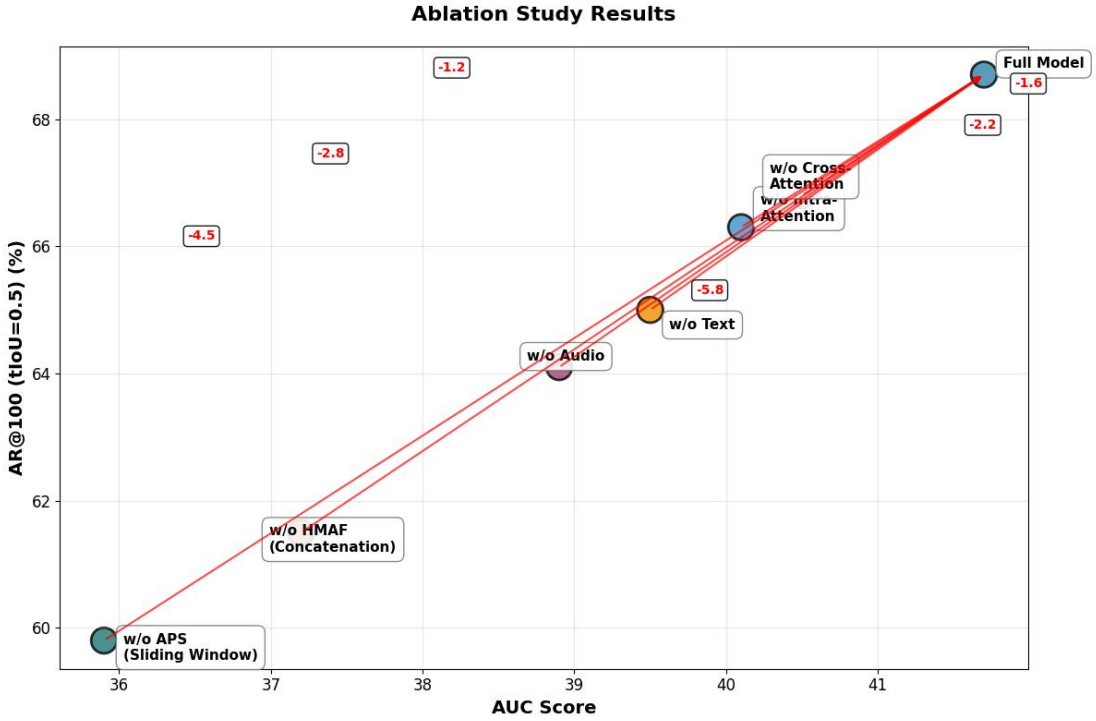

Figure 6 Scatter Plot of Performance Metrics for Each Model Configuration in Ablation Experiments

As illustrated in Figure 6, the performance scatter plot from the ablation experiments clearly reveals the contribution of each component to model performance and their intrinsic relationships. The complete model occupies the optimal position with 68.7% AR@0.5 and 41.7% AUC, demonstrating the best overall performance. When the adaptive search strategy was removed in favour of a traditional sliding window approach, performance declined most markedly, AR@0.5 decreased to 59.8% and AUC fell to 35.9%, confirming the critical role of adaptive search in enhancing detection efficiency and boundary accuracy. Replacing the hierarchical attention mechanism with simple feature concatenation caused AR@0.5 drop in a to 61.5% and AUC to 37.2%, highlighting the necessity of refined multimodal fusion. Removing either the audio or text modality individually caused varying degrees of performance degradation, confirming the complementary value of multimodal information. The greater impact observed when the audio modality was absent indicates that audio cues are particularly crucial for behaviour recognition in classroom settings. Notably, performance degradation from removing intra-modal attention slightly exceeded that from cross-modal attention removal, indicating that capturing intra-modal temporal dependencies contributes more significantly to final performance than cross-modal alignment. These results collectively demonstrate that the model's components synergistically enhance behaviour detection performance, with adaptive search strategies making the greatest contribution, followed by hierarchical attention mechanisms, while multimodal information provides indispensable complementary cues.

4.4 Discussion of Results

We conducted a case study to visualise the temporal attention weights of successful examples, as illustrated in Figure 7. The figure depicts the visualisation of attention weights during a “teacher-question-student-answer” process. The top section displays video frames, the middle shows audio waveforms and transcribed text, while the bottom presents tri-modal attention weights. It is evident that at the start of the question, textual attention focuses on the word ‘why’, while audio attention concentrates on the rising intonation. During the student's response, visual attention centres on the student's area, audio attention shifts to the student's voice, and textual attention aligns with the content of the student's answer. Our model successfully localises the entire interaction process.

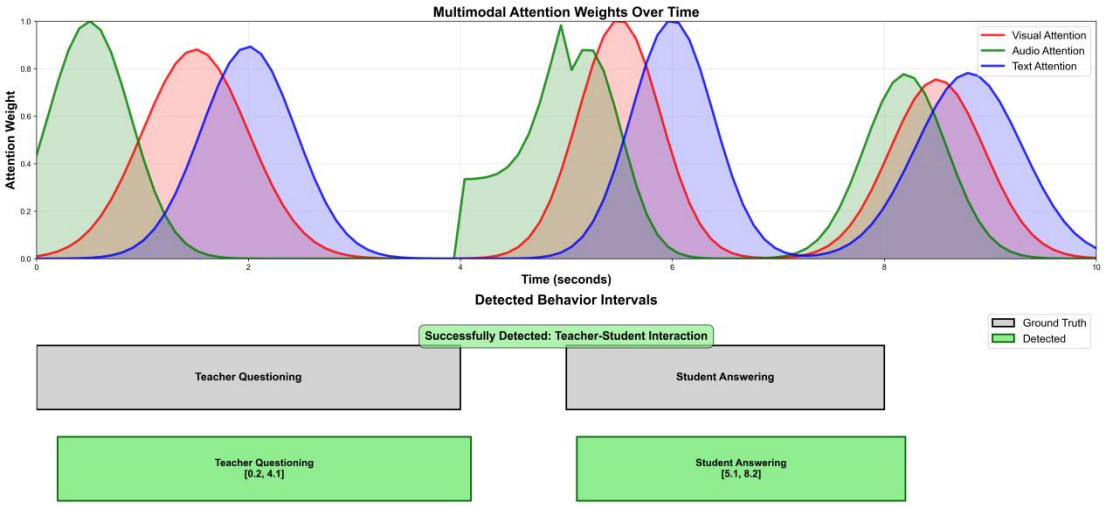

Figure 7 Visualisation of Temporal Attention Weights

Moreover, we have also observed instances of failure. Firstly, extremely ambiguous boundaries, such as a group discussion that commences slowly without clear linguistic markers. Additionally, multimodal signals of extremely poor quality, such as severe camera shake, audio containing significant noise, or entirely erroneous ASR transcriptions. These too represent challenges that require continued attention in future work.

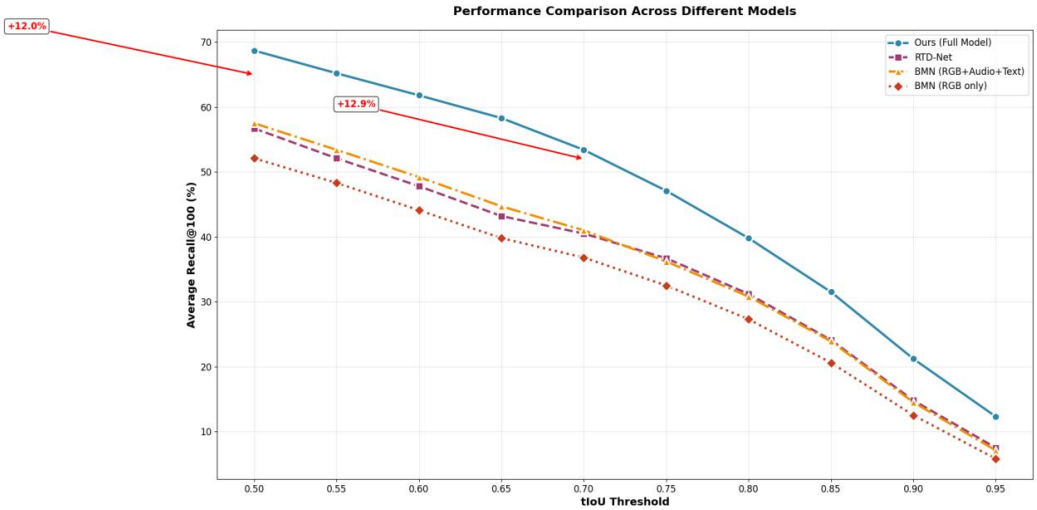

Figure 8 Performance Curves of Different Models at Various tIoU Thresholds

For the comparative experiments conducted in this paper, we performed a performance analysis, as detailed in Figure 8. The figure illustrates the average recall trends across different models as the tIoU threshold varies from 0.5 to 0.9. Our proposed model maintains a leading position across all thresholds, exhibiting the most gradual decline in performance curves. This indicates that the generated behavioural proposals demonstrate superior boundary accuracy and robustness. Specifically, under stringent thresholds tIoU@0.5 and @0.7, our model achieves absolute performance improvements of 12.0% and 12.9% respectively compared to the strongest baseline RTD-Net, highlighting its significant advantage in precisely locating behavioural boundaries. Even as the threshold increases to @0.9, the proposed model maintains a recall rate of 21.2%, substantially exceeding the baseline model's rate below 15%, further validating its capability to capture extreme precision boundaries. This outcome stems from the hierarchical attention mechanism's dynamic fusion of multimodal information, effectively leveraging complementary visual, auditory, and textual cues. Concurrently, the adaptive search strategy intelligently adjusts granularity within lengthy videos, mitigating background redundancy while enhancing detection of critical behavioural regions. This holistic approach elevates the model's performance across varying levels of strictness.

5 CONCLUSION

This paper addressed the core challenges in temporal action proposal generation for classroom videos, namely the extreme variation in action durations, complex semantic hierarchies, and the heterogeneous nature of multimodal information. We proposed a novel proposal generation model centered on a Hierarchical Multimodal Attention Fusion

(HMAF) module and an Adaptive Proposal Search (APS) strategy. Comprehensive experiments and in-depth analysis on the collected Edu-Action dataset lead to the following principal conclusions.

First, the proposed HMAF module and APS algorithm are conclusively identified as the key drivers for the performance superiority of our framework. The significant performance gains, evidenced by absolute improvements of 12.0% and 12.9% in AR@100 at tIoU thresholds of 0.5 and 0.7, respectively, over the strongest baseline RTD-Net, demonstrate a substantial advancement in generating high-quality proposals with precise temporal boundaries. The model's robustness is further highlighted by its maintained recall of over 21% at the highly stringent tIoU threshold of 0.9, significantly surpassing all baseline methods and underscoring its exceptional capability in localizing actions with ambiguous boundaries or short durations.

Second, the results of the ablative studies, clearly visualized via a performance scatter plot, quantitatively dissect the contribution of each component. The Adaptive Proposal Search mechanism is confirmed to be the most critical innovation, as its replacement with a sliding window approach resulted in the most severe performance degradation. This underscores its indispensable role in achieving an optimal balance between efficiency and accuracy in long, untrimmed videos. The Hierarchical Multimodal Attention Fusion module is the second most significant contributor. Its performance gain far exceeded that of a simple feature concatenation baseline, validating the effectiveness of its dynamic, fine-grained fusion of visual, acoustic, and linguistic cues through intra- and cross-modal attention for deep semantic alignment and enhancement. Furthermore, the performance drop observed from removing either the audio or text modality confirms the necessity of multimodal information, with the slightly larger impact from ablating audio suggesting the particularly strong discriminative power of vocal and acoustic events in the classroom context.

In summary, this work not only delivers a model that significantly outperforms the state-of-the-art for classroom behavior analysis but also, through meticulous experimentation, elucidates the critical roles and underlying mechanisms of sophisticated multimodal fusion and intelligent search strategies for fine-grained action localization in long videos. It provides a reliable tool for automated classroom behavior analysis in the domain of smart education and offers a valuable framework and insights for the broader field of long-form, multimodal video understanding.

Future work will focus on model lightweighting for practical deployment, enabling online real-time processing, and enhancing cross-scenario generalization to foster application in real-world intelligent classroom environments and provide more powerful support for teaching analytics and assessment.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

1. Research on Government Coordination for Building Shenzhen into a Leading Hub of the Low-Altitude Economy Industry (6025210051S).
2. Research on the Paths of In-depth Industry-University-Research Integration in the Guangdong-Hong Kong-Macao Greater Bay Area (6025260077S).
3. Research on Classroom Scene Understanding and Behavior Analysis Method Based on Multimodal Attention Mechanisms (7024310268).
4. A Process Tracing Study of the Construction of Vocational Undergraduate Curriculum System Based on Multiple Causal Mechanism Framework (23YJA880048).
5. A Sociological Analysis of the Motivations Behind Shenzhen Enterprises' Participation in Industry-Education Integration from the Perspective of Rational Choice Theory (6025310024S).

REFERENCES

[1] Jain A, Dubey K A, Khan S, et al. A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems. Scientific Reports, 2025, 15(1): 17463-17463.
[2] Chen J, Qian L, Ni H. The Smart Classroom Practices in Science Courses. Higher Education and Practice, 2024, 1(8).
[3] Yanqiu Z ,Yinghua S, Rongxia H, et al. The Development and Application of the Metaverse in Smart Education in Chinese Universities. Frontiers in Educational Research, 2024, 7(6).
[4] Xu Z, Zhou Q, Li Z, et al. Adaptive Multi-Function Radar Temporal Behavior Analysis. Remote Sensing, 2024, 16(22): 4131-4131.
[5] Andriani R, Disman, Ahman E, et al. Polychronic Behaviors: The Role of Job Residency and Education Level. International Journal of Entrepreneurship, 2019, 23(3).
[6] Petrov M. Analyzing and classifying a range of incorrect actions made by students during an educational process using an interval temporal behavior observation. Educational Alternatives, 2019, 17(1): 117-126.
[7] Sarra A, Leila J M, Fahima H, et al. Fuzzy Vikor Application for Learning Management Systems Evaluation in Higher Education. International Journal of Information and Communication Technology Education (IJICTE), 2021, 17(2): 17-35.
[8] Ukpong E D, George N I. Length of Study-Time Behaviour and Academic Achievement of Social Studies Education Students in the University of Uyo. International Education Studies, 2013, 6(3): 172.

[9] Derya S, Felix H, Magdalena B, et al. Early and middle latency auditory event-related potentials do not explain differences in neuropsychological performance between schizophrenia spectrum patients and matched healthy controls. Psychiatry Research, 2021, 304(prepublish): 114162-114169.
[10] Tucholka I, Gold B. Analysing classroom videos in teacher education— How different instructional settings promote student teachers' professional vision of classroom management. Learning and Instruction, 2025: 97102084-102084.
[11] Stapleton N J, Richardson R M. Social Support Network and Sedentary Behavior Among US Adults With and Without Mobility Impairment. American journal of health promotion: AJHP, 2024, 38(7): 8901171241252526-8901171241252526.
[12] Peng Z, Jitong W, Mengshu L, et al.Structure, electrical properties and energy storage performance of BNKT-BMN ceramics. Journal of Materials Science: Materials in Electronics, 2022(prepublish): 1-12.
[13] Ghaderi A, Athitsos V. Selective Unsupervised Feature Learning with Convolutional Neural Network (S-CNN). CoRR, 2016.
[14] Moss H E, Tantry K E, Le E, et al. Distinct patterns of PV and SST GABAergic neuronal activity in the basal forebrain during olfactory-guided behavior in mice. The Journal of neuroscience: the official journal of the Society for Neuroscience, 2025.
[15] Zeinab H Z, Shekoufeh R K, Esmaeil F, et al. Distributed RMI-DBG model: Scalable iterative de Bruijn graph algorithm for short read genome assembly problem. Expert Systems With Applications, 2023, 233.
[16] Changyou D, Hong L, Han Z, et al. New Understanding on Relationship Between RTD Curve and Inclusion Behavior in the Tundish. Metallurgical and Materials Transactions, 2024, 55(4): 2224-2239.
[17] Tu L, Hong H. Multimodal Learning Data Analysis and Algorithmic Teaching Effectiveness Evaluation Model Construction. International Journal of High Speed Electronics and Systems, 2024(prepublish).
[18] Cai Q, Bajuri R M, Leong E K, et al. Multimodal Learning Interactions Using MATLAB Technology in a Multinational Statistical Classroom. Multimodal Technologies and Interaction, 2025, 9(10): 106-106.
[19] Sun C, Huang S, Sun B, et al. Personalized learning path planning for higher education based on deep generative models and quantum machine learning: a multimodal learning analysis method integrating transformer, adversarial training and quantum state classification. Discover Artificial Intelligence, 2025, 5(1): 29-29.
[20] Gajghate S S , Noor M M, Kumar S, et al. A transformer guided multi modal learning framework for predictive and causal assessment of thermal runaway in high energy batteries. Scientific Reports, 2025, 15(1): 37054-37054.
[21] Zhang X, Bahri A, Desrosiers C, et al. SegMamba: Mamba-based Incomplete Multimodal Learning for Brain Tumor Segmentation with Few Samples. IEEE journal of biomedical and health informatics, 2025. DOI: 10.1109/JBHI.2025.3600652.
[22] Huafeng W, Hanlin L, Wanquan L, et al. Temporal information oriented motion accumulation and selection network for RGB-based action recognition.I mage and Vision Computing, 2023, 137.
[23] Ng L H D, Chia T R T, Young E B, et al. Study protocol: infectious diseases consortium (I3D) for study on integrated and innovative approaches for management of respiratory infections: respiratory infections research and outcome study (RESPIRO). BMC infectious diseases, 2024, 24(1): 123-123.
[24] Abd I E L E, Emad N S E, K K M, et al. VGGish transfer learning model for the efficient detection of payload weight of drones using Mel-spectrogram analysis. Neural Computing and Applications, 2024, 36(21): 12883-12899.
[25] Iancu B. Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources. Informatica Economica, 2019, 23(1): 17-25.
[26] CallMiner Combines Conversation Analytics Platform With Microsoft Azure Speech to Text. Telecomworldwire, 2021.
[27] Huang P, Zhu H, Wang Y, et al. Enhanced Semantic BERT for Named Entity Recognition in Education. Electronics, 2025, 14(19): 3951-3951.
[28] Shylaja R, Kumari R V L. ArrhythTransform: Multi-Head Attention-Based Transformer Encoder for Arrhythmia Classification. Engineering Letters, 2025, 33(11).
[29] Dixit A, Kumar S, Kumar N, et al. Assessing the impact of process and design variations on reliability of complementary FET. Solid State Electronics, 2025: 230109226-109226.
[30] Ying L, Wujie Z. Hierarchical Multimodal Adaptive Fusion (HMAF) Network for Prediction of RGB-D Saliency. Computational Intelligence and Neuroscience, 2020: 20208841681-8841681.