A Dual-attention-based Approach for Student Abnormal Behavior Detection in Classroom Videos

Vinothina V [1,*], Augustine George [1], Jasmine Gnanadurai [2], and Vennila J [3]

[1] Kristu Jayanti Institute of Technology, Kristu Jayanti University, Bengaluru, India
[2] Department of Electrical Engineering and Computer Science, George Fox University, Oregon, USA
[3] Manipal College of Health Professions, Manipal Academy of Higher Education, Manipal, Karnataka, India
Email: vinothina.v@kristujayanti.com (V.V.); augustine@kristujauanti.com (A.G.); jgnanadurai@georgefox.edu (J.B.); vennila.j@manipal.edu (V.J.)
*Corresponding author

***Abstract*—As multimedia technology has advanced, surveillance footage has become a vital tool for monitoring student behavior in classrooms, supporting improvements in instructional methods, learning outcomes, and overall discipline. The learning environment might be adversely affected by abnormal activities, including napping during class and using cell phones. Although many researchers have worked on abnormal behavior detection, improvements are still needed to accurately understand how student actions change over time and to detect when the behaviors are subtle or depend on the classroom context. To address this challenge, we propose a Transformer-inspired Attention Model (TAM) designed to classify short classroom videos as normal or abnormal based on student behavior. A novel Context-Aware Attention and Multiscale Detection (CAMD) module was added to extract rich spatial and temporal data at different scales. These frame-level data are subsequently combined into a global temporal representation using a Context-Aware Token (CAT) module, guaranteeing an effective summarization of the full video. To enhance feature integration, the Context-Aware Feature Integration (CAFI) module fuses both local and global contextual insights. Prior to categorization, essential frames were dynamically assigned importance through temporal attention pooling. With a mean Average Precision (mAP) improvement of nearly 4%, the model detection accuracy outperformed current models when tested on classroom videos. These findings show that the modules in the proposed model are effective in detecting classroom conduct from students.**

***Keywords*—dual-attention mechanism, student behavior detection, shot-boundary, keyframes, multi-scale fusion, context-aware feature integration**

I. Introduction

Video surveillance is defined as tracking and reporting the activity or behavior of humans in specific area [1]. Video surveillance can be useful in identifying unusual human behavior or activity in public areas. It also aids in crowd management and the avoidance of unforeseen events at public events with large audiences [2]. Owing to its significant applications in different domains, many researchers have contributed to the detecting the abnormal behavior [3]. Video surveillance in a classroom can be used for a variety of purposes, such as security, discipline, and behavior analysis [4].

From behavior analysis, teachers can measure the focus level of students such as interest in the lecture or any instances of misbehavior. Teachers may assess whether they can improve their teaching methods or take appropriate action if a student is not paying attention to the class. However, manually monitoring the video is time-consuming and requires human intervention. This necessitates the development of an automated system that can identify instances of disruptive student behavior in recordings.

Recent technological advancements have led to the automation of video surveillance [5–7]. The capacity of early methods to capture intricate activities were constrained by their handcrafted features. However, with the rise of deep learning, such as Convolutional Neural Network (CNN), feature extraction has become automated, allowing for more accurate behavior recognition. Recurrent models, such as Long Short-Term Memory (LSTMs), improve temporal analysis, leading to better detection of activities over time. Real-time detection of delicate and complex behavior with high accuracy and minimal human intervention is simplified with the use of 3D CNNs and models based on Vision and Video Transformers [8–10].

Recognition of Student behavior requires an understanding of spatial and temporal information across multiple frames. The goal of spatiotemporal action recognition, for example, is to keep track of each person's activities throughout time and classify them appropriately when there are groups of people in a video. To detect abnormal behavior in actual classroom settings with intricate and dense backgrounds, a more detailed analysis

Manuscript received June 6, 2025; revised July 16, 2025; accepted November 13, 2025; published July 20, 2026.

of how movements vary from frame to frame is required. It is essential to focus on the body parts and objects that directly contribute to disruptive actions while filtering out distractions from the surrounding environment.

Students' disruptive behavior detection encompasses supervised, semi-supervised and unsupervised learning. Supervised Learning offers the best accuracy and performance when sufficient labeled data are available, particularly for well-defined tasks, such as detecting specific disruptive behaviors. Publicly available datasets for classroom behavior detection are scarce, posing challenges in developing robust models for identifying disruptive student behavior. To address this gap and enhance detection accuracy, the contributions of this study are summarized as follows.

- To build a robust dataset for disruptive student behavior detection, the EduNet [11] dataset, along with additional classroom video samples collected from publicly available online repositories, was used to ensure diverse scenarios, including various classroom environments, varying student densities, crowding conditions, and occlusions, enhancing the generalizability of the model.
- To effectively detect subtle and complex anomalies in classroom settings, a Context-Aware Attention and Multiscale Detection Module is incorporated to capture both the global context and localized multiscale patterns. This module dynamically assigns importance to frames or regions based on contextual relevance, thereby improving detection accuracy.
- Context-Aware Feature Integration, a dedicated approach is employed to fuse granular frame-wise details with high-level temporal context, ensuring that both fine-grained and holistic insights contribute to classification.
- The effectiveness of the proposed model was validated through extensive experiments, comparing it with existing deep learning models based on accuracy, precision, recall, and computational efficiency.

The structure of this paper is organized as follows. Section II briefly reviews the related work. The proposed method is presented in Section III. Section IV presents the experimental results. The study is summarized in the conclusion section.

II. Literature Review

Human behavior recognition has garnered significant research interest in recent years, particularly in video surveillance. Recognizing behavior from surveillance videos requires careful attention to specific regions within each frame to accurately identify actions. This section reviews recent literature on video behavior recognition, highlighting both traditional approaches and those that leveraging attention mechanisms.

Early machine learning models, such as decision trees and rule-based systems, were used to recognize basic activities; however, these systems required handcrafted features and were limited in scalability. Classical machine learning models, such as Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and Hidden Markov Models (HMMs), have become more prevalent, allowing for improved classification of human actions in surveillance footage. Because the model learns slowly with more data, human feature extraction forces the model to have low accuracy [12, 13].

Unlike manual feature extraction, deep CNN uses hierarchical learning techniques to automatically extract features from massive amounts of data without the need for human participation [14]. This improved the accuracy of the model at the cost of network complexity. To detect subtle and complex patterns, Recurrent Neural Networks (RNN) [15] and Long Short-Term Memory (LSTM) [16] have been proposed. Subsequently, hybrid approaches that combines RNN, LSTM and deep learning have improved accuracy by capturing both spatial and temporal features [17].

To detect individuals with spatio-temporal abnormalities in small and large-scale crowds, a hybrid classifier using deep learning (CNN) and a machine learning algorithm of Random Forest has been proposed [18]. However, the model does not achieve better performance for large-scale crowds owing to low-resolution, low illumination and occlusion. Paulraj and Vairavasundaram [19] proposed Swin Transformer- Convolutional-Long Short-Term Memory (C-LSTM) networks to capture global and local temporal dependencies in order to improve the accuracy of abnormal event detection. However, the model's inability to capture hierarchical patterns at different scales restricts its ability to identify complex behaviors. Ren *et al.* [20] proposed transformer-based model Slow-Fast Swin, to detect abnormal behavior by focusing on temporal dependencies between adjacent frames while prioritizing spatial information. However, the model's performance is constrained by small datasets.

Dense Visual Attention Augmented Spatio-Temporal (DAST-Net) employs a dense attention-aware autoencoder architecture with residual connections to enhance spatiotemporal feature representation [21]. Visual attention mechanisms are used for spatial dimensions, whereas ConvLSTM autoencoders handle temporal patterns. The unsupervised approach, validated on the UCSD dataset, addresses the shortcomings of traditional frameworks, such as Autoencoders, ResNet, U-Net, and GANs, which inadequately differentiate the foreground from background in surveillance video analysis. DAST-Net demonstrated improved performance by incorporating dense attention mechanisms.

TransCNN integrates Convolutional Neural Networks (CNNs) for spatial feature extraction with Transformers to learn temporal dependencies and long-term relationships [22]. The temporal self-attention mechanism enhances the fusion of spatial and temporal information. The model was evaluated on datasets like ShanghaiTech, UCSD Ped2, and CUHK Avenue, excelling in anomaly detection. However, its computational overhead and intricate design remain challenge in real-time applications.

Conv-ViT proposes a hybrid extraction framework that combines Inception V3 and ResNet to capture texture

features, whereas Transformers are employed for shape-based feature learning [23]. This innovative hybrid model facilitates a comprehensive understanding of spatio-temporal patterns for retinal disease detection. Despite their success on datasets such as Mendeley and OCTID (Optical Coherence Tomography Image Database), their computational complexity poses a significant hurdle. Future work will focus on optimizing the framework to reduce time complexity. To overcome the above-mentioned limitations and challenges, the proposed model employs a context-aware attention-multiscale detection module to record spatiotemporal features and, for improved performance, record the features at several scales.

III. Materials and Methods

In this section, the Transformer-Inspired Attention Model (TAM) for abnormal behavior detection in classrooms is described in detail. To classify the video as normal or abnormal, the proposed model considered the brief actions that students displayed during lectures. The actions considered for classification are listed in Table I.

A. Overall Architecture

A Transformer-inspired Attention Model Architecture for identifying aberrant behavior in classroom video is discussed in this section. A feature extractor processes the input, which is a series of video frames ($T \times H \times W \times C$), and then adds positional embeddings to enhance the input. The Context-Aware Attention Multiscale Detection Module receives these features and uses a Dual-Attention Module that combines temporal and spatial attention to capture temporal dependencies and spatial details. Features from various resolutions were integrated using a multi-scale fusion technique. To improve the contextual representation, enriched features are delivered to the Context-Aware Feature Integration Module and Context-Aware Token Generator. Finally, a classification head predicts the normality or abnormality of the conduct observed in the classroom.

B. Dataset Preparation

Video clips from the EduNet dataset [11] and additional classroom videos sourced from online repositories were used to train and evaluate the model. The videos with different students' behaviors are categorized in to abnormal and normal, as mentioned in Table I. Each video is labelled using the format actionclass_vX.mp4, for example, Sleeping_v15.mp4. Each directory included a corresponding CSV file that provides metadata such as video names, file locations, and labels of the clips within that directory.

C. Preprocessing

To detect abnormal student behaviors in a classroom, such as throwing objects, sleeping, walking, using mobile phones, eating at desks, sitting on desks, falling, listening to lectures, and writing in textbooks, a systematic video preprocessing pipeline is essential. First, raw videos with average size of 10 s are acquired and standardized in terms of resolution 720 p or 480 p and frame rate (30 fps) for consistency. Subsequently, all videos were resized to a uniform dimension of 224 × 224 pixels with 3 color channels (RGB) to ensure compatibility with the proposed model. This preprocessing ensured that the videos were standardized and optimized for the efficient classification of student behaviors.

TABLE I. Dataset Classes

Student Behaviour	No. of Videos	Class
Using Mobile Phone [11]	90	Abnormal
Sitting on Bench [11]	72	Normal
Eating during Lecture [11]	60	Abnormal
Writing in Textbook [11]	95	Normal
Throwing Object	40	Abnormal
Rising Hand	52	Normal
Sleeping during Lecture	61	Abnormal
Walking during Lecture	55	Abnormal
Listening to Lecture	70	Normal

D. TAM Architecture

The model processes video inputs with an average duration of 10 s, N frames of spatial dimensions $W \times H$ and three color channels. Dividing videos into frames allows the model to extract and analyze temporal patterns, which are crucial for detecting short-period abnormal behaviors in the videos. The architecture is shown in Fig. 1 and its modules are described in the following subsections.

1) Feature extraction

To reduce computational complexity while preserving essential video information, Large Model-based Sequential Keyframe Extraction (LMSKE) [24] is employed to extract a reduced set of keyframes that best represent the video content while maintaining temporal consistency. The steps are illustrated in Fig. 2. Instead of processing all frames the in a video, LMSKE selects an optimal subset K keyframes where $K \ll N$, reducing the video representation to $K \times H \times W \times 3$. The working of LMSKE is as follows:

LMSKE first segments the video into shots based on scene changes using TransNetV2 [25], a deep learning model trained for shot boundary detection.

Each frame within a shot was encoded using Contrastive Language-Image Pretraining (CLIP) [26], a pre-trained vision-language model. Instead of full feature extraction, CLIP's vision encoder maps frames into a high-dimensional semantic space, allowing similarity comparison based on learned contextual relationships. Every frame was encoded into a 512-dimensional feature vector using CLIP's vision encoder. If there are K frames in the video, the output is $F = [F_1, F_2, ..., F_K] \in \mathbb{R}^{K \times 512}$ where each F_i is a 512-dimensional vector representing the semantic meaning of the ith frame. A similarity matrix S is computed using cosine similarity between frame embeddings. It is calculated as follows:

$$S_{ij} = \frac{F_i . F_j}{\|F_i\| \, \|F_j\|} \quad (1)$$

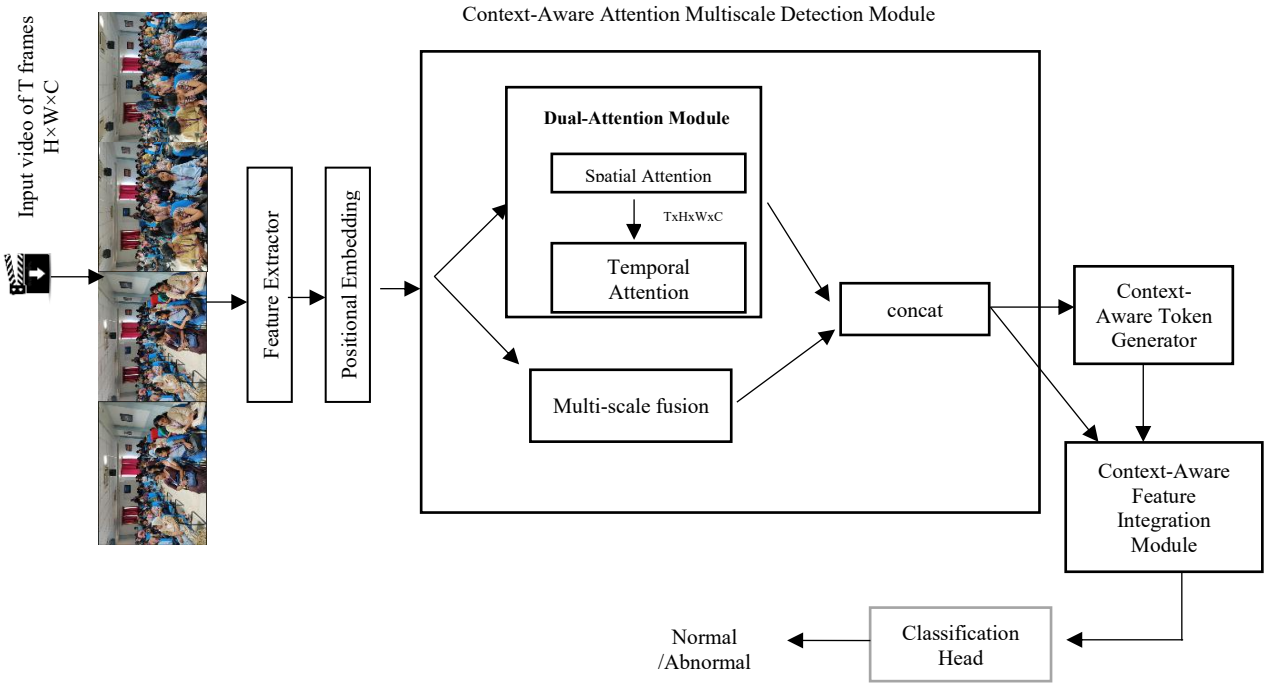

Fig. 1. Transformer-inspired attention model architecture.

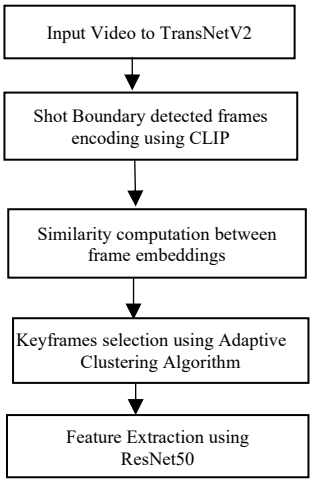

Fig. 2. Keyframe selection and feature extraction pipeline.

An adaptive cluster algorithm is applied within Ref. [24] is applied within each shot to automatically determine optimal K clusters representing distinct visual patterns. Each cluster denoted by Cluster ID represents a group of frames that share common spatiotemporal features. For instance, frames clustered together may correspond to students exhibiting similar activities, such as listening to lectures, writing on notebooks or abnormal behaviors such as sleeping, talking, or using mobile phones. Thus, cluster IDs serve as an intermediate representation that maps low-level frame features to high-level behavior categories, enabling more interpretable and structured behavior analysis.

From each cluster, the most representative frame closest to the cluster center is selected as a candidate keyframe, while redundant frames are eliminated. The final set of selected keyframes was concatenated in the original temporal sequence, ensuring the preserve video flow of content.

The K keyframes are selected while maintaining the original sequence of the video. The final keyframe set is $K_f = [F_{k1},F_{k2},...,F_{kk}] \in \mathbb{R}^{K\times 512}$, where K $\ll$ N, reducing computational cost while preserving the important information of the video. This approach significantly reduces the computational load while retaining meaningful video information, making it well-suited for abnormal student behavior detection.

As ResNet performs well in low-level extraction, ResNet50 [27] was employed for spatial feature extraction. The K selected keyframes were resized to 224 × 224 × 3. The video has dimensions K × 224 × 224 × 3, where K represents the number of frames. ResNet50 extracts spatial features from only these K keyframes, instead of all frames. For each frame, ResNet50 outputs a feature map of size 7 × 7 × 2048 which captures rich spatial information. After processing all frames, the output becomes K × 7 × 7 × 2048, which represents the spatial feature maps for all k frames. To simplify further processing, these feature maps are typically flattened using average pooling across the spatial dimensions of 7 × 7, resulting in a single feature vector of size 2048 for each frame. For further processing, the spatial and temporal features of all K keyframes were retained.

2) Positional embedding

Positional embedding is applied to preserve the order of temporal and spatial context in the extracted feature maps. The input is a 4D tensor of dimensions K × 7 × 7 × 2048, positional embeddings are added to these feature maps as follows:

- A tensor of size K × 1 × 1 × 2048 was created, with one embedding vector for each frame. This tensor is broadcast across the spatial grid 7 × 7 and added

element-wise to the feature maps, enriching the feature representation of each frame with temporal context. A learnable temporal embedding matrix E_T is defined as $E_T \in \mathbb{R}^{K \times 1 \times 1 \times 2048}$ where each frame k has an embedding vector $E_T(k) \in \mathbb{R}^{2048}$. This tensor is broadcast across the spatial grid 7 × 7 and added element-wise to the feature maps as follows:

$$\tilde{F}(k,x,y,c) = F(k,x,y,c) + E_T(k,1,1,c) \quad (2)$$

where (x,y) are the spatial indices, c is the feature channel index, k is the frame index, E_T (k,1,1,c) is broadcast to all spatial positions.

- A tensor of size 1 × 7 × 7 × 2048 was created, to encode positional information for each spatial location. This tensor is broadcast across the temporal dimension K and added element-wise to the feature maps, to provide spatial context. A learnable spatial embedding matrix E_S is defined as $E_S \in \mathbb{R}^{K \times 7 \times 7 \times 2048}$ where spatial location (x,y) has an embedding vector $E_S(x,y) \in \mathbb{R}^{2048}$. This tensor is broadcast across the temporal dimension K and added element-wise to the feature maps as follows:

$$\tilde{F}(k,x,y,c) = F(k,x,y,c) + E_S(1,x,y,c) \quad (3)$$

where E_s (1,x,y,c) is broadcast across all K frames.

The output of positional embedding retains the original dimensions (K × 7 × 7 × 2048) but is now enriched with temporal and spatial contexts, which aids in refining feature representation in subsequent steps. The final positionally enriched feature map F_{pos} is computed as:

$$F_{(pos)}(k,x,y,c) = F(k,x,y,c) + E_T(k,1,1,c) + E_S(1,x,y,c) \quad (4)$$

3) Context-aware Attention and Multiscale Detection (CAMD)

Context-aware Attention and Multiscale Detection (CAMD) integrates dual attention and multiscale fusion mechanisms to refine feature extraction for detecting anomalous student behavior in classrooms. The key idea is to simultaneously capture context-aware global representations via dual attention and localized multiscale patterns via multiscale fusion, enabling the model to detect subtle anomalies effectively. The input to the CAMD step comprised features of dimension K × 7 × 7 × 2048, where K corresponds to the number of frames, 7 × 7 represents the spatial grid size from ResNet50 feature maps, and 2048 is the channel depth. These features are processed through two parallel pipelines as Dual Attention and Multi-scale fusion.

4) Dual attention mechanism

It focuses on spatial and temporal aspects to identify global dependencies. Spatial attention selectively emphasizes critical regions in each frame, whereas temporal attention identifies important frames across the video sequence. The output from this pipeline is aggregated into a global context representation with dimensions of 1 × 2048, summarizing the entire video's contextual features. Temporal attention helps the model to focus on the most relevant frames in the video. Let $F \in \mathbb{R}^{K \times 7 \times 7 \times 2028}$. The temporal attention score is computed as:

$$A_T = Softmax\left(\frac{Q_T K_T^T}{\sqrt{d_k}}\right) \quad (5)$$

Q_T, K_T, V_T are the query, key, and value matrices for temporal attention respectively, where Q_T, K_T, $V_T \in \mathbb{R}^{K \times d_k}$. d_k is the feature dimension. The temporal attention output is $F_T = A_T \cdot V_T$ where F_T is the temporal attended feature map of size $K \times d_k$. Similarly spatial attention helps the model focus on important regions within each frame. The spatial attention score was computed as follows:

$$A_s = Softmax\left(\frac{Q_S K_S^T}{\sqrt{d_k}}\right) \quad (6)$$

Finally, the spatial attention output is $F_s = A_s \times V_s$, where F_s is the spatially attended feature map of size (7 × 7) × d_k. The global context representation is formed by combining both temporal and spatial attention as follows:

$$F_{Global} = W_T F_T + W_S F_S \quad (7)$$

where W_T and W_S are learnable weights that balance temporal and spatial attention contributions. The final feature vector F_{global} has a shape of 1 × 2048 and serves as the global context representation for video classification.

IV. Multiscale Fusion

The features are analyzed at multiple spatial resolutions i.e., full resolution $F_7 \in \mathbb{R}^{K \times 7 \times 7 \times 2048}$, half resolution $F_3 \in \mathbb{R}^{K \times 3 \times 3 \times 2048}$, and quarter resolution $F_2 \in \mathbb{R}^{K \times 2 \times 2 \times 2048}$. These are obtained using adaptive pooling: $F_3 = \text{Pool}_{3 \times 3}(F_7)$, $F_2 = \text{Pool}_{2 \times 2}(F_7)$, where Pool denotes an adaptive average pooling operation that reduces spatial resolution. Each feature map is flattened and concatenated as $F_{ms} = \text{Concat}(\text{Flatten}(F_7), \text{Flatten}(F_3), \text{Flatten}(F_2))$. This captures fine-grained and coarse-grained patterns simultaneously. After processing, these multiscale features are concatenated into a combined representation, yielding an output of dimension K × 2048.

Finally, the outputs from both pipelines are broadcasted and concatenated to form a unified feature representation with dimension K × 4096, combining global context and multiscale detail effectively as $F_{final} = \text{Concat}(F_{ms}, F_{global})$, where $F_{final} \in \mathbb{R}^{K \times 4096}$. This module ensures the robustness in detecting abnormal behaviour.

A. Context-Aware Token (CAT)

To get the global representation of the entire video without losing the temporal dependencies, temporal attention is applied to assign varying levels of importance to individual frames based on their contribution to the detection task. For instance, frames with higher anomaly relevance are given greater weight. After this aggregation process, CAT produces a global context token, typically represented by a compact dimension such as 1 × 512. This

token encapsulates the temporal essence of the video, serving as a summary of the most critical temporal and contextual information. The full transformation process can be computed as follows.

$$T_{CAT} = W_c \sum_{i=1}^{K} W_i \left[SoftMax\left(\frac{(W_q F_{CAMD})(W_k F_{CAMD})^T}{\sqrt{d}}\right)(W_v F_{CAMD})\right] \quad (8)$$

where $F_{CAMD} \in R^{K \times 4096}$ is the input frame-level feature representation, W_q, W_k, $W_v \in R^{4096 \times d}$ are attention weight matrices, $W_c \in R^{4096 \times 512}$ is a learnable projection and $T_{CAT} \in R^{1 \times 512}$ is the final context-aware token, summarizing the most important temporal and contextual features. By condensing the frame-level details into a global token, CAT ensures that the subsequent classification step operates on a feature that retains the core temporal and contextual insights, making it both efficient and effective for anomaly detection.

B. Context-Aware Feature Integration (CAFI)

CAMD produces detailed multiscale frame-level features for all the Keyframes K, capturing both spatial and temporal variations across the video. These features are particularly valuable for identifying subtle behaviors such as throwing objects or using mobile phones. Meanwhile, CAT processes the CAMD output to generate a global context token with a dimension of 1 × 512, summarizing the overall temporal context of the video while assigning dynamic importance to frames that carry the most significant information about potential anomalies. CAFI module combines the multiscale frame-level details from CAMD and the global context insights from CAT. This unified representation ensures that the model retains both granular information about individual frames and an overarching perspective of the entire video.

Frame-Level Features from CAMD is $F_{CAMD} \in R^{K \times 4096}$ means that K frames, and each frame has a 4096-dimensional feature vector. The dimension of Global Context Token from CAT is $T_{CAT} \in R^{1 \times 512}$. This is a single vector summarizing the whole video with a 512-dimensional feature vector. Before we can combine these two, they need to have the same dimensionality. Learnable projection matrices W_f and W_c are used to transform both feature spaces into a common dimensional space, say d = 1024 or any chosen feature size. The projected CAMD features into d dimension is as $F_{CAMD}' = F_{CAMD} W_f$, $W_f \in R^{4096 \times d}$ and projected CAT features into d dimension is as $T_{CAT}' = T_{CAT} W_c$, $W_c \in R^{512 \times d}$.

To effectively integrate these representations, a weighted fusion mechanism is employed as $F_{CAFI} = \alpha F_{CAMD}' + \beta T_{CAT}'$, Where α and β are learnable fusion weights that control the contribution of frame-level and global features. Since T_{CAT}' is 1 × d, we broadcast it across K frames for element-wise fusion as T_{CAT}' = repeat(T_{CAT}',K,1). Thus, the fused representation is $F_{CAFI} \in R^{K \times d}$. This ensures that each frame-level feature incorporates both its local multiscale details and the global temporal context.

This combined feature set is then passed to the classification head, which uses it to make accurate predictions about whether the behavior in the video is normal or abnormal. CAFI thus enhances the model's ability to detect complex and subtle student behaviors, providing a robust and comprehensive solution for anomaly detection in classroom settings.

C. Classification Head

The transformed vector from the CAFI module is passed through fully connected layers that learn higher-level representations of the contextual features. The output of the CAFI module is a set of frame-wise feature vectors, denoted as $F_{CAFI} \in R^{K \times d}$, where K is the number of frames and d is the feature dimension. To ensure that frames containing salient information like anomalies like throwing objects or using mobile phones receive higher importance, temporal attention weight is computed for each frame as:

$$\alpha_k = \frac{\exp(w^T F_{CAFI,k})}{\sum_{i=1}^{k} \exp(w^T F_{CAFI,i})} \quad (9)$$

where $w \in R^d$ is a learnable weight vector, $F_{CAFI,k}$ is the feature vector for frame k, α_k is the attention score for frame k, ensuring that higher scores correspond to more relevant frames. The denominator ensures that all scores sum to 1. The final global feature representation for the entire video is obtained by computing a weighted sum over all frames.

$$F_{video} = \sum_{k=1}^{K} \alpha_k F_{CAFI,k} \quad (10)$$

where $F_{video} \in R^{1 \times d}$ represents the temporal-pooled video-level feature. F_{video} is passed through a Fully Connected (FC) layer as $z = WF_{video} + b$, where $W \in R^{d \times 1}$ is the weight matrix, $b \in R^1$ is the bias term and $z \in R^1$ is the raw score before activation. Since the model performs binary classification such as normal vs. abnormal behavior, sigmoid activation function is employed to convert the score into a probability.

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \quad (11)$$

where $\hat{y}$ represents the probability of the behavior being abnormal. During training, the predicted probability $\hat{y}$ is compared with the ground truth label $y \in \{0,1\}$ using binary cross-entropy loss. This loss is minimized via backpropagation, updating model parameters to improve classification performance. In this way, the classification head makes the final decision on classifying student behavior as either normal or abnormal.

V. Results and Discussion

To evaluate the impact of the modules within the proposed transformer inspired model, EduNet dataset and real-time classroom videos were used. The performance of these variations is compared against state-of-the-art models with the same objective such as SBD model [28], Deep Learning using ResNet [29], OpenPose [30], YOLOv3_D47 [31] and MSTA-SlowFast [32]. The evaluation is conducted using standard classification

metrics, including accuracy, precision, recall, F1-Score, and AUC-ROC, to analyze the effectiveness of the proposed approach in detecting abnormal student behavior in classroom settings.

A. Setup

The study is conducted using a deep learning pipeline implemented in PyTorch, leveraging high-performance GPU hardware to accelerate computations. The model is trained with a batch size of 16, utilizing the AdamW optimizer, while Binary Cross-Entropy (BCE) is employed for classification as follows.

$$L_{BCE} = \frac{-1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (12)$$

The training pipeline follows an 80-10-10 split, where 80% of the dataset is used for training, 10% for validation, and 10% for testing. For classification, fully connected layers with Sigmoid activation are used to determine the final prediction as follows.

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (13)$$

When validation performance reaches a plateau, the learning rate is lowered dynamically to ensure optimal convergence.

B. Fine Tuning

Initially only higher layers in ResNet50 are trained to adapt to dataset-specific features. Then as training progresses, all layers are gradually fine-tuned with a lower learning rate to retain the general representation learned from the dataset. For keyframe selection, the ViT-L/14 variant of CLIP is utilized to extract embeddings that effectively capture the most informative frames from classroom videos. A dropout rate between 0.2 and 0.4 is applied to mitigate overfitting and improve generalization.

The proposed model's detection accuracy is evaluated by comparing it with existing models SBD [24], DL-ResNet [25], OpenPose [26], DarkNet-47 [27], and MSTA-SlowFast [28]. Precision, Recall and mean Average Precision (mAP) metrics are considered for evaluation and it is computed as follows:

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

$$mAP = \frac{1}{N}\sum_{n=1}^{N}(AP)_n \quad (16)$$

where *FP* is the number of predicted positive samples that turned out to be negative, *FN* is the number of predicted negative samples that turned out to be positive, and *TP* is the number of correctly identified positive samples. The area under the Precision-Recall curve for a single class is called Average Precision (AP). *N* represents the total number of classes, which in this case are normal and abnormal.

The average of all classes' AP scores is known as mAP. Fig. 3 shows the output following Shot Boundary detection, clustering of frames, and keyframe selection for a sample video. Identical visual patterns are grouped into clusters, each denoted by a Cluster ID. After shot boundary detection and clustering, the proposed pipeline is selecting only few keyframes i.e, the final representative frames. Fig. 4 represents the Temporal visualization of video frames. Normal frames are shown in white, shot boundary regions are highlighted in grey, and selected keyframes are marked in red. Out of the 145 frames in a sample video, the proposed model selectively utilized only a subset of frames for behavior detection. This representation indicates which parts of the video contribute most to behavior detection. For short videos, these few frames may be enough to capture student behaviors. Table II contains a tabulation of the experimental outcomes and the same is shown in Fig. 5.

(a)

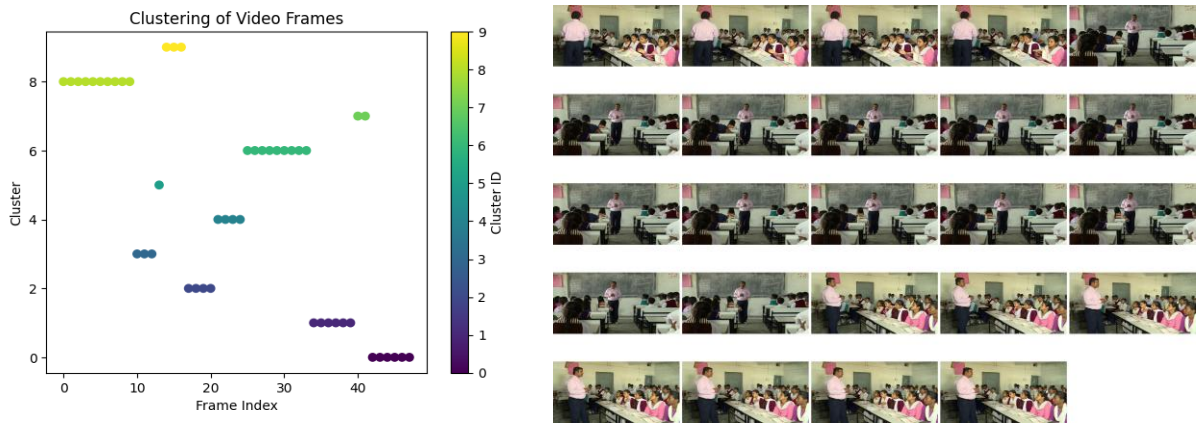

Fig. 3. Illustrates the shot boundary detection and keyframe selection process. (a) Frames after applying TransNetV2 for shot boundary detection; (b) Clustered fames using adaptive clustering algorithm; (c) Selected keyframes.

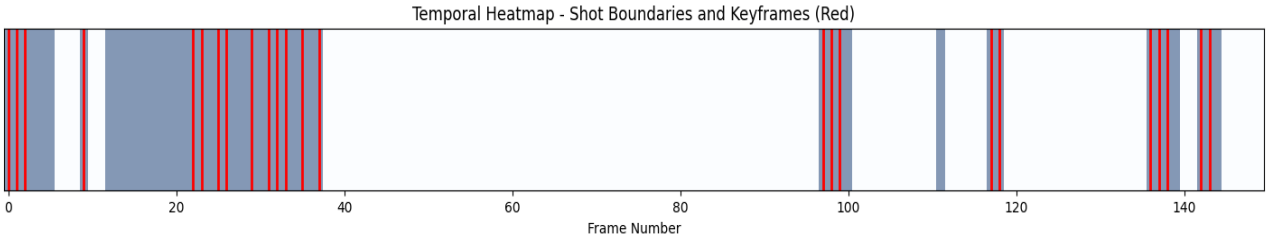

Fig. 4 Temporal visualization of video frames. Normal frames are shown in white, shot boundary regions are highlighted in grey, and selected keyframes are marked in red.

TABLE II. EXPERIMENTAL RESULTS

Model	AP (%)		mAP (%)
	Normal	Abnormal	
SBD Model [24]	78.5	80.1	79.3
DL-ResNet [25]	81.2	82.5	81.9
OpenPose [26]-Self-Built dataset	84.5	85.1	84.8
YOLOv3_D47 [27]	86.3	87	86.7
MSTA-SlowFast [28]	88.9	89.4	89.2
TAM*	91.8	92.6	92.2

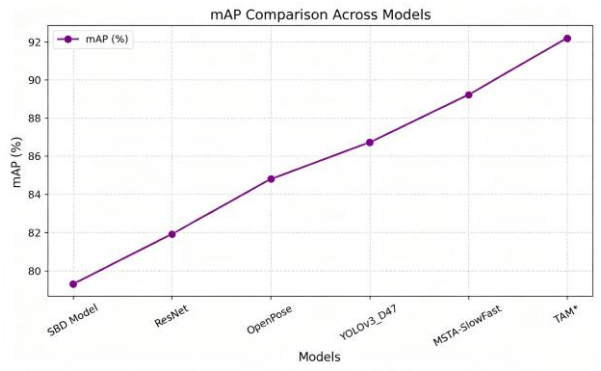

Fig. 5. Performance comparison across models.

The experimental results demonstrate that TAM outperforms other models owing to its transformer-inspired architecture, which integrates multiscale fusion with dual spatial-temporal attention, enabling more precise detection of normal and abnormal activities by effectively capturing complex motion patterns and contextual interactions. Through efficient keyframe selection, the proposed TAM model outperforms existing models while significantly reducing computational time. For anomalous behaviors, the model shows superior Average Precision (AP), demonstrating its capacity to extract fine-grained characteristics for the detection of small anomalies in classroom settings.

C. *Ablation Study*

To evaluate the impact of keyframes selection in detection accuracy, the model is experimented with and without keyframe selection prior to keyframe selection using the same dataset used for evaluation. By comparing these variations against the full model, the ablation study provides a detailed understanding of how the keyframes selection contributes to the overall performance of the abnormal behavior detection framework. Table III presents a performance comparison between the full model and the model without keyframe selection.

TABLE III. EXPERIMENTAL RESULTS OF ABLATION STUDY

Model Variant	AP (%)		mAP(%)
	Normal	Abnormal	
TAM—Full model	91.8	92.6	92.2
TAM—No Keyframe Selection	91.5	90.9	92.0

Although the performance of model variants is nearly the same, TAM without keyframe selection is computationally expensive as it processes a larger number of frames compared to the full model, leading to increased computation time and resource usage. The proposed model enhances feature representation through keyframe selection, positional encoding, and dual attention with

multiscale fusion. It integrates context-aware tokens and feature fusion to improve abnormal behavior detection. However, challenges remain in handling rapid motion changes, occlusions, and diverse classroom environments. A key limitation is its sensitivity to keyframe selection, where ineffective clustering may lead to missing critical abnormal behaviors, especially fast actions like sudden movements or quick gestures.

VI. Conclusion

To detect unusual student behavior in classroom settings, a Transformer-inspired model has been designed and evaluated. After identifying the shot boundary and clustering the frames, keyframes are chosen in order to minimize the number of frames needed for computation. By leveraging the attention mechanism, the model efficiently focuses on critical spatial-temporal regions within video frames, enhancing the accuracy of abnormal behavior classification. The multiscale fusion enables the model to capture the features at different scales. The CAFI module helps the model preserve the global representation of the full video as well as frame information. The experimental results demonstrated that the proposed model detection accuracy is better than the existing models. However, the model still needs to be improved in order to handle occlusion and camera view point.

Future research will focus on improving temporal consistency, integrating multi-modal inputs such as physiological signals and audio cues, and optimizing the model for real-time performance. By advancing automated classroom behavior monitoring, this research contributes to smart education systems, assisting educators in maintaining an engaging and disciplined learning environment while ensuring students' well-being.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Vinothina contributed to the conceptualization of the study, methodology design, data analysis, and manuscript writing; Augustine George was involved in supervision, validation of results, and critical revision of the manuscript; Jasmine contributed to data collection, implementation, and drafting of initial sections of the manuscript; Vennila assisted in literature review, methodological design and manuscript writing. All authors reviewed and approved the final version of the manuscript.

References

[1] Y. Myagmar-Ochir and W. Kim, "A survey of video surveillance systems in smart city," *Electron.*, vol. 12, no. 17, 2023. doi: 10.3390/electronics12173567

[2] A. Jan and G. M. Khan, "Real-world malicious event recognition in CCTV recording using Quasi-3D network," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 8, pp. 10457–10472, 2023. doi: 10.1007/s12652-022-03702-6

[3] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Futur. Gener. Comput. Syst.*, vol. 96, pp. 386–397, 2019. doi: 10.1016/j.future.2019.01.029

[4] Y. Li, X. Qi, A. K. J. Saudagar, A. M. Badshah, K. Muhammad, and S. Liu, "Student behavior recognition for interaction detection in the classroom environment," *Image Vis. Comput.*, vol. 136, 104726, 2023. doi: 10.1016/j.imavis.2023.104726

[5] K. Nithesh, N. Tabassum, D. D. Geetha, and R. D. A. Kumari, "Anomaly detection in surveillance videos using deep learning," in *Proc. IEEE Int. Conf. Knowl. Eng. Commun. Syst. ICKES 2022*, 2022. doi: 10.1109/ICKECS56523.2022.10059844

[6] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis. Comput.*, vol. 106, 104078, Feb. 2021. doi: 10.1016/J.IMAVIS.2020.104078

[7] V. Vinothina, A. George, G. Prathap, and J. Beulah, *A Study on Surveillance System Using Deep Learning Methods BT-Ubiquitous Intelligent Systems*, P. Karuppusamy, F. P. García Márquez, and T. N. Nguyen, Eds. Singapore: Springer Nature Singapore, 2022, pp. 147–162.

[8] N. Combining, C. Block, A. Module, and U. M. Frames, "Anomaly detection based on a 3D convolutional neural network combining convolutional block attention module using merged frames," *Sensors*, vol. 23, no. 23, 9616, 2023.

[9] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL : A vision transformer network for image anomaly detection and localization," arXiv Preprint, arXiv:2104.10036, 2021.

[10] W. Sun, Y. Ma, and R. Wang, "k-NN attention-based video vision transformer for action recognition," *Neurocomputing*, vol. 574, 127256, 2024.

[11] V. Sharma, M. Gupta, A. Kumar, D. Mishra, M. Martínez-Zarzuela, and D. G. Ortega, "EduNet: A new video dataset for understanding human activity in the classroom environment," *Sensors*, vol. 21, no. 17, 2021. doi: 10.3390/s21175699

[12] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2259–2322, 2021. doi: 10.1007/s10462-020-09904-8

[13] N. C. Tay, T. Connie, T. S. Ong, A. B. J. Teoh, and P. S. Teh, "A review of abnormal behavior detection in activities of daily living," *IEEE Access*, vol. 11, pp. 5069–5088, 2023. doi: 10.1109/ACCESS.2023.3234974

[14] D. Weimer, B. Scholz-Reiter, and M. Shpitalni, "Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection," *CIRP Ann.*, vol. 65, no. 1, pp. 417–420, 2016.

[15] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc. the 25th ACM International Conference on Multimedia*, 2017, pp. 863–871. doi: 10.1145/3123266.3123328

[16] R. Vohra, K. Goel, and J. K. Sahoo, "Modeling temporal dependencies in data using a DBN-LSTM," in *Proc. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–4. doi: 10.1109/DSAA.2015.7344820

[17] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, 102946, 2024.

[18] T. Alafif *et al.*, "Hybrid classifiers for spatio-temporal abnormal behavior detection, tracking, and recognition in massive Hajj crowds," *Electron.*, vol. 12, no. 5, pp. 1–19, 2023. doi: 10.3390/electronics12051165

[19] S. Paulraj and S. Vairavasundaram, "Transformer-enabled weakly supervised abnormal event detection in intelligent video surveillance systems," *Eng. Appl. Artif. Intell.*, vol. 139, 109496, 2025.

[20] Y. Ren, C. Li, W. Bao, X. Chen, and Y. Jing, "A study of student action recognition in smart classrooms based on improved slowfast swin transformer," in *Proc. 2023 8th Int. Conf. Signal Image Process. ICSIP 2023*, 2023, pp. 59–63. doi: 10.1109/ICSIP57908.2023.10270896

[21] R. Kommanduri and M. Ghorai, "DAST-Net: Dense visual attention augmented spatio-temporal network for unsupervised video anomaly detection," *Neurocomputing*, vol. 579, 127444, 2024.

[22] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for

surveillance anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 123, 106173, 2023.
[23] P. Dutta, K. A. Sathi, and A. Hossain, "Conv-ViT: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection," *Journal of Imaging*, vol. 9, no. 7, 140, 2023.
[24] K. Tan, Y. Zhou, Q. Xia, R. Liu, and Y. Chen, "Large model based sequential keyframe extraction for video summarization," in *Proc. of the International Conference on Computing, Machine Learning and Data Science*, 2024, pp. 1–5.
[25] T. Souček and J. Lokoč, "TransNet V2: An effective deep network architecture for fast shot transition detection," in *Proc. the 32nd ACM International Conference on Multimedia*, 2024, pp. 11218–11221. doi: 10.1145/3664647.3685517
[26] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Mach. Learn. Res.*, vol. 139, 2021, pp. 8748–8763.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90
[28] Q. Jia and J. He, "Student behavior recognition in classroom based on deep learning," *Appl. Sci.*, vol. 14, no. 17, 2024. doi: 10.3390/app14177981
[29] M. Lu, D. Li, and F. Xu, "Recognition of students' abnormal behaviors in English learning and analysis of psychological stress based on deep learning," *Front. Psychol.*, vol. 13, no. 11, pp. 1–13, 2022. doi: 10.3389/fpsyg.2022.1025304
[30] Y. Ding, K. Bao, and J. Zhang, "An intelligent system for detecting abnormal behavior in students based on the human skeleton and deep learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022. doi: 10.1155/2022/3819409=0
[31] W. Yajun, "A deep learning recognition method for students' abnormal behaviors in smart classroom teaching scenarios," *Int. J. High Speed Electron. Syst.*, 2540291, Jan. 2025. doi: 10.1142/S0129156425402918
[32] S. Zhang *et al.*, "MSTA-SlowFast: A student behavior detector for classroom environments," *Sensors*, vol. 23, no. 11, pp. 1–13, 2023. doi: 10.3390/s23115205