From Raw Video to Pedagogical Insights: A Unified Framework for Student Behavior Analysis

Zefang Yu, Mingye Xie, Jingsheng Gao, Ting Liu, Yuzhuo Fu*

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China
{yuzefang, xiemingye, gaojingsheng, louisa_liu, yzfu}@sjtu.edu.cn

Abstract

Understanding student behavior in educational settings is critical in improving both the quality of pedagogy and the level of student engagement. While various AI-based models exist for classroom analysis, they tend to specialize in limited tasks and lack generalizability across diverse educational environments. Additionally, these models often fall short in ensuring student privacy and in providing actionable insights accessible to educators. To bridge this gap, we introduce a unified, end-to-end framework by leveraging temporal action detection techniques and advanced large language models for a more nuanced student behavior analysis. Our proposed framework provides an end-to-end pipeline that starts with raw classroom video footage and culminates in the autonomous generation of pedagogical reports. It offers a comprehensive and scalable solution for student behavior analysis. Experimental validation confirms the capability of our framework to accurately identify student behaviors and to produce pedagogically meaningful insights, thereby setting the stage for future AI-assisted educational assessments.

Introduction

In educational settings, student engagement is a critical determinant of effective learning and overall educational outcomes (Trowler 2010; Finn and Zimmer 2012; Lei, Cui, and Zhou 2018). Extensive research in the field of pedagogy has identified three principal dimensions of student engagement: cognitive engagement, emotional engagement, and behavioral engagement (Fredricks, Blumenfeld, and Paris 2004). Each dimension uniquely contributes to the establishment of effective learning environments and has a positive impact on student outcomes. Notably, behavioral engagement acts as a vital indicator of student participation and attentiveness, directly influencing academic performance(Appleton, Christenson, and Furlong 2008). Traditional methods of assessing this engagement rely heavily on manual classroom observations, although informative, are labor-intensive and subject to observer bias, rendering them unsuitable for large educational settings.

With the development of deep learning and computer vision technologies, there has been a growing interest in automating student behavior analysis in classrooms. However, current approaches present substantial limitations that hinder their broader applicability and effectiveness. Firstly, a considerable portion of the existing work oversimplifies student behavior analysis by reducing it to single or multi-class action detection tasks (Li, Jiang, and Shen 2019; Zhou, Jiang, and Shen 2018; Zheng, Jiang, and Shen 2020), which fails to capture the rich diversity and complexity of student behaviors. Secondly, while there are efforts that have integrated existing machine learning or deep learning techniques to build student behavior analysis systems (Ngoc Anh et al. 2019), they often stop at data analytics, lacking mechanisms to translate the analytical insights into actionable insights for educators, particularly those without a background in pedagogy research. Lastly, both types of approaches frequently rely on proprietary datasets for developing their models. The inability to share these datasets publicly, primarily due to privacy concerns, forms a barrier to their adoption and iterative improvement in different educational settings.

To address the aforementioned challenges and limitations, this paper puts forward a comprehensive, end-to-end framework that facilitates a direct pathway from raw classroom video footage to insightful pedagogical reports. To be specific, our framework initiates with leveraging temporal action detection techniques to capture both the action class and timing of behaviors exhibited in the classroom. This temporal data is subsequently integrated into a state-of-the-art large language model, guided by specific prompts tailored to an educational context, to produce in-depth analysis reports. To navigate the intrinsic privacy concerns associated with video-based analysis utilizing RGB information, we adopt skeleton modality data for our framework. This choice ensures ethical adherence by fundamentally safeguarding student privacy, eliminating the chances of identity disclosure. Additionally, skeleton data is more robust and indifferent to variations in background and environmental settings, thereby enhancing the generalization capability of our framework. In addition to delivering an end-to-end analysis, our framework also produces a diverse array of intermediate data products, forming a valuable reservoir for educational experts and further downstream tasks.

Our main contributions are as follows:

(1) We propose an end-to-end framework that paves a direct and automated pathway from raw classroom videos to the generation of insightful pedagogical reports, serving as

*Corresponding Author.

a readily accessible tool for teachers and a rich data source for educational experts.

(2) Our approach leverages temporal action detection techniques integrated with a large language model, creating a flexible and extensible framework that stands as a rich repository for pedagogical research.

(3) Quantitative and qualitative experimental results demonstrate the validity and robustness of our framework.

Related Work

AI-based Student Behavior Analysis

With the advancement of deep learning and the proliferation of surveillance videos, models that leverage computer vision techniques for analyzing student behavior in classroom videos have begun to emerge. Many works treat student behavior analysis as an object detection task. Some focus on the detection of individual actions, such as detecting student's sleeping gesture (Li, Jiang, and Shen 2019), hand raising (Lin, Jiang, and Shen 2018; Zhou, Jiang, and Shen 2018), and yawn behavior (Wang, Jiang, and Shen 2019). Other studies are dedicated to detecting a variety of actions that may occur within the classroom setting (Zheng, Jiang, and Shen 2020; Tang et al. 2022). Additionally, some works incorporate human pose estimation techniques for student behavior analysis (Lin et al. 2021; Yu et al. 2023). There are also efforts to determine students' classroom behavior with the help of micro-expression recognition (Pei and Shan 2019). Distinct from these studies, we are the first to conceptualize student behavior analysis as a temporal action detection task, integrating additional information along the temporal dimension for a more comprehensive analysis of student behavior.

Temporal Action Detection

Temporal action detection (TAD) aims to localize the start and end timestamps of action instances and recognize their action class at the same time. The vast majority of current TAD efforts are based on RGB video stream, and they can be broadly divided into two-stage and one-stage methods. The two-stage methods (Zhao et al. 2017; Xu, Das, and Saenko 2017; Shou, Wang, and Chang 2016; Lin et al. 2018, 2019; Chao et al. 2018) first generate the proposals, then refine the boundaries and classify actions. The one-stage methods (Zhang, Wu, and Li 2022; Yeung et al. 2016; Lin, Zhao, and Shou 2017; Buch et al. 2019) simultaneously generate action proposals and corresponding action labels in a single model. Compared to RGB video based TAD, skeleton based TAD is still in a very primitive stage. The authors in (Wu et al. 2019) design a YOLOv2 style lightweight detector with a few convolutional layers for skeleton TAD. Recently, the MSGCN (Filtjens, Vanrumste, and Slaets 2022) merges best practices in video-based convolutional networks and skeleton-based action recognition design to provide simple implementation for skeleton-based temporal action recognition.

Large Language Models

Large language models (LLMs) refer to Transformer language models that contain hundreds of billions (or more) of parameters, which are trained on massive text data, such as GPT-3 (Brown et al. 2020), PaLM (Chowdhery et al. 2022), and LLaMA (Touvron et al. 2023; Zhao et al. 2023). LLMs have shown impressive generalization capabilities such as incontext-learning and chain-of-thoughts reasoning (Wei et al. 2022). Compared to other domains (e.g., finance, law, medical), there has been less research on the use of LLM in education domain (Kasneci et al. 2023). Recently, EduQuiz (Dijkstra et al. 2022) used GPT-3 to generate complete multiple-choice questions with the correct and distractor answers, reducing the burden of manual quiz design for teachers. A study by (Moore et al. 2022) employed a fine-tuned GPT-3 model to evaluate student-generated answers in a chemistry education learning environment, which may be powerful tools for assisting teachers in the qualitative and pedagogical evaluation of student answers. Another quantitative analysis (Malinka et al. 2023) shows that students utilizing ChatGPT (either keeping or refining the results from LLMs as their own answers) perform better than average students in some courses from the computer security field (Zhao et al. 2023). To the best of our knowledge, this is the first attempt to use LLM for student behavior analysis.

Methodological Overview

The primary objective of this paper is to develop an end-to-end framework that facilitates the systematic analysis of student behavior in educational settings. The framework leverages advancements in both temporal action detection and large language models to transform raw classroom video footage into pedagogical reports. Figure 1 schematically illustrates the various steps involved in both the training and inference phases of our framework, detailing the flow of data and the integration of modular components.

Training Phase. Given the absence of Temporal Action Detection (TAD) datasets for classroom environments, we constructed a custom dataset by synthesizing long sequences of action skeletons from the publicly available pre-segmented action recognition dataset. We selected those actions that fit the classroom scenario (e.g., writing, reading, playing phone, etc.), and performed randomized splicing while recording the start and end times of each action of interest. To augment the dataset, a multi-view synthesis strategy was employed, converting the long 3D skeleton sequences into diverse 2D perspectives to accommodate different shooting angles encountered in real classroom videos. Additionally, motion interpolation techniques were harnessed at each splice point to foster a more natural transition between actions. These sequences served as inputs for our Temporal Action Detection (TAD) model, which employs a motion encoder followed by a series of Transformer layers and a convolutional decoder. This phase completes the training of our TAD model on our custom dataset.

Inference Phase. For real-world applications, the initial step in our inference phase involves applying a Human Pose Estimation (HPE) model on classroom surveillance videos to extract the 2D skeleton sequences of each student. These student behavior sequences are then processed through our pre-trained TAD model. The resulting temporal action information for each student is extracted in a structured for-

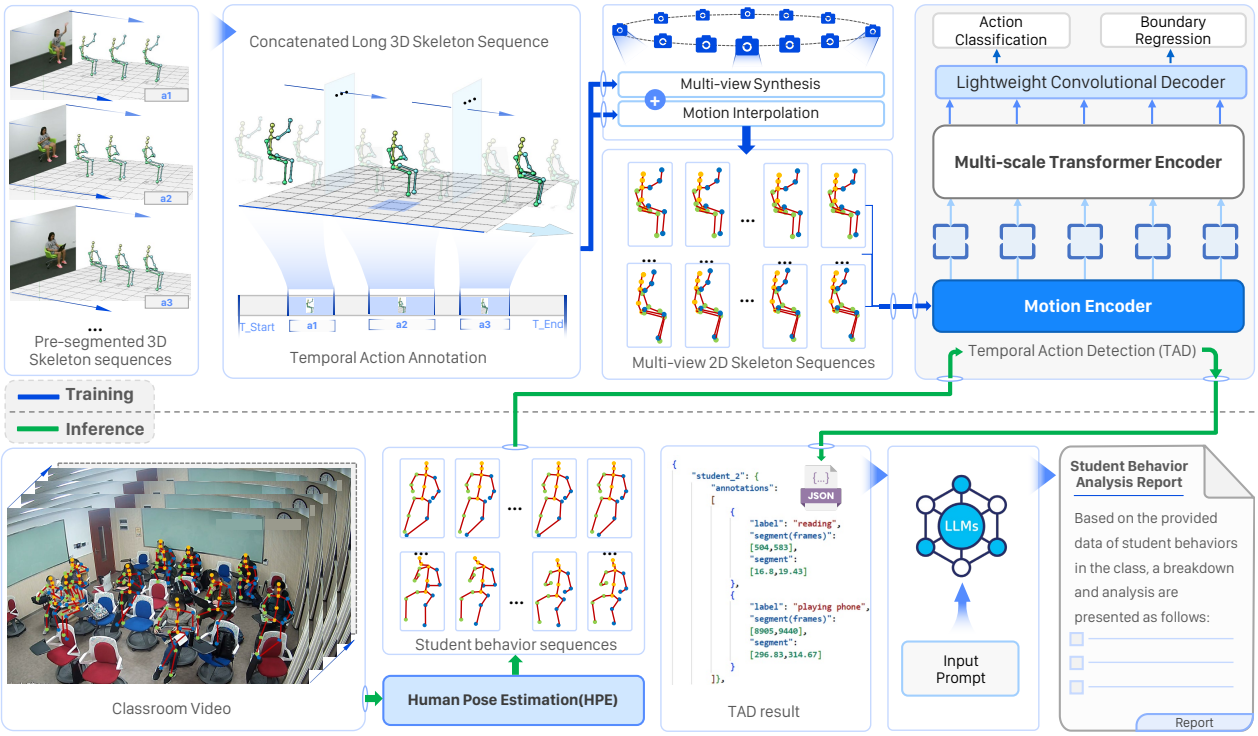

Figure 1: An overview of the entire pipeline of our proposed end-to-end framework, including training and inference phases.

mat that includes timestamps and associated actions. This structured output is then passed into a large language model, guided by specific prompts tailored to an educational context. The resulting pedagogical analysis report serves as the final output of our framework.

Our framework is distinguished by its end-to-end approach, covering the complete pipeline from raw classroom video footage to insightful pedagogical reports. Notably, each component of the framework is designed to be modular and extensible, allowing for continual refinement and adaptation to various educational settings. The detailed implementation of each component will be expanded upon in the subsequent section.

Technical Architecture

Data Construction and Preprocessing

To create long sequences, we leverage the largest available NTU RGB+D dataset (Shahroudy et al. 2016) of segmented actions. The NTU RGB+D dataset was acquired by Kinect at a rate of 30 frames per second, which provided 56,880 actions performed by 40 subjects in 60 categories (here we only use the 49 single-subject categories). Each of the 17 camera settings, varying in heights and distances, captures footage from 3 different viewing angles (-45° , 0°, 45°).

Sequence Concatenation. To mimic real-world classroom scenarios where multiple student behaviors occur in an extended sequence, we design an automated pipeline to concatenate the trimmed action sequences into longer sequences. First, we perform the skeleton-centric normalization where all action sequences are normalized to the position-invariant coordinate system. Subsequently, we establish an 'Action Reference Pool' by filtering actions based on criteria such as the same subject, viewpoint, and settings. This step is crucial to maintain consistency and avoid unnatural scaling and tilting in the skeleton data. We selectively picked classroom-relevant actions (e.g., writing, reading, raising hands) from this reference pool for stitching into longer sequences, while recording their start and end times automatically. To add complexity and distractions akin to a real-world classroom, we also included non-classroom related actions as background actions.

Multi-view Synthesis. Given a 3D skeleton sequence represented as $\mathbf{S} \in \mathbb{R}^{T\times J\times 3}$, where T is the number of frames and J is the number of joints, we strive to map it onto 2D sequences from designated viewpoints. This process leverages a virtual camera model that is characterized by a trio of parameters: d (radial distance), h (height), and α (anticlockwise angle) with respect to the body center (e.g., the spine joint), denoted as $\mathbf{B}$, where B_x, B_y and B_z represent the coordinates of the body center in the 3D space.

To determine the camera position denoted as $\mathbf{C} = (C_x, C_y, C_z)$, we employ the following equations based on the parameters and the coordinates of $\mathbf{B}$:

$$\begin{aligned} C_x &= B_x + d\sin(\alpha), \\ C_y &= B_y + d\cos(\alpha), \\ C_z &= B_z + h. \end{aligned} \tag{1}$$

For each joint j at time t , we define the vector $\mathbf{V}_j^t$ be-

tween the joint position at time t and the camera as:

$$\mathbf{V}_j^t = \mathbf{C} - \mathbf{S}_j^t. \tag{2}$$

Then, the 2D coordinates $P_{x,j}^t$ and $P_{y,j}^t$ of the projected j-th joint at time t are computed using:

$$\begin{aligned} P_{x,j}^t &= C_x + V_{j,x}^t, \\ P_{y,j}^t &= C_y + V_{j,y}^t. \end{aligned} \tag{3}$$

In practice, we randomly choose a set of parameters d, h, and α for each sequence, where $d, h \in \{1, 2, 3\}$,$\alpha \in \{-60°, -50°, \ldots, 50°, 60°\}$. This approach allows us to incorporate a wide variety of perspectives, enhancing the dataset's representational richness.

Motion Interpolation. To tackle abrupt transitions in our multi-view 2D skeleton sequences, we employ cubic spline interpolation as described by (McKinley and Levine 1998). Let $\hat{\mathbf{S}} \in \mathbb{R}^{T \times J \times 2}$ represent the 2D skeleton frames. If t_i marks the end of a previous action and t_{i+1} marks the beginning of the next action in sequence $\hat{\mathbf{S}}$, our aim is to generate smooth intermediate frames. For each joint j, a cubic spline function $F_j(x)$ is fitted between the coordinates $(x_{i,j}, y_{i,j})$ and $(x_{i+1,j}, y_{i+1,j})$ over the interval $[t_i, t_{i+1}]$:

$$F_j(x) = a_j + b_j(x - x_{i,j}) + c_j(x - x_{i,j})^2 + d_j(x - x_{i,j})^3, \tag{4}$$

where a_j, b_j, c_j, and d_j are coefficients optimized for function and derivative continuity. This process generates n intermediate frames, thereby increasing the temporal smoothness of the 2D skeletal movements (see Figure 2).

Skeleton-based Temporal Action Detection

For a 2D skeleton sequence $\hat{\mathbf{S}}$, our skeleton-based temporal action detection model aims to predict a set of action labels $Y = \{y_1, y_2, \ldots, y_M\}$. Each action instance y_i is defined by its onset time s_i, offset time e_i, and action category a_i. Here we employ ActionFormer (Zhang, Wu, and Li 2022) as the foundational architecture of our model. Its proven efficiency and state-of-the-art performance make it an ideal base upon which we build additional functionalities tailored for skeleton sequences.

Motion Encoder. We segment each sequence $\hat{\mathbf{S}}$ into N overlapping clips $\{x_1, x_2, \ldots, x_N\}$ using a sliding window approach. Then, we adapt the MotionBert (Zhu et al. 2023) as a motion encoder E to capture both the spatial configurations and temporal dynamics inherent in the skeleton data. It produces a set of feature vectors as:

$$Z^0 = [E(x_1), E(x_2), \ldots, E(x_N)]^N$$

where $E(x_i) \in \mathbb{R}^D$ is the embedded feature of x_i.

Multi-scale Transformer Encoder. The Transformer encoder further takes Z^0 as input. The encoder has L Transformer layers with each layer consisting of alternating layers of local multi-headed self-attention (MSA) and MLP blocks. Features are calculated as:

$$\begin{aligned} \overline{\mathbf{Z}}^\ell &= \alpha^\ell \, \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{Z}^{\ell-1}\right)\right) + \mathbf{Z}^{\ell-1}, \\ \hat{\mathbf{Z}}^\ell &= \bar{\alpha}^\ell \, \mathrm{MLP}\left(\mathrm{LN}\left(\overline{\mathbf{Z}}^\ell\right)\right) + \overline{\mathbf{Z}}^\ell, \\ \mathbf{Z}^\ell &= \downarrow\left(\hat{\mathbf{Z}}^\ell\right), \ell = 1, 2, \ldots, L, \end{aligned} \tag{5}$$

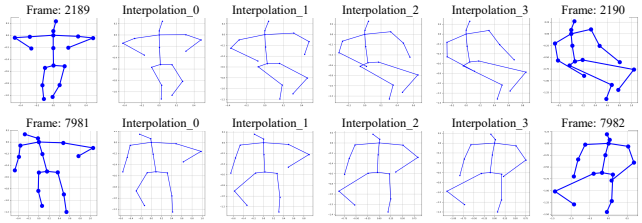

Figure 2: Visualization of Motion Interpolation with n = 4.

where LN stands for LayerNorm and $\downarrow$ is down-sampling operator. More details can be found in (Zhang, Wu, and Li 2022). We further combine several Transformer blocks with down-sampling in between, resulting in a feature pyramid $\mathbf{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^L\}$.

Decoder and Output. The lightweight convolutional decoder takes the feature pyramid $\mathbf{Z}$ and performs the action classification and boundary regression tasks. The action classification task classifies each moment t into one of C pre-defined action categories. The boundary regression task estimates the temporal boundaries of each action, defined by its starting time s_t and ending time e_t. The output $\hat{Y}$ at each time t is $\hat{y}_t = (p(a_t), d_t^s, d_t^e)$, where $p(a_t)$ is the probability of action category a_t, and d_t^s and d_t^e are the distances to the action's onset and offset, respectively.

Large Language Model Integration

During the inference phase, the final stage of our framework leverages a Large Language Model to interpret the structured output $\mathcal{T}$ from the Temporal Action Detection (TAD) model and generate an educational report $\mathcal{R}$. Where $\mathcal{T}$ can be represented as $\mathcal{T} = \{\text{SequenceLength} : \beta, \text{ActionSequence} : \{\text{Student}_i : \{\text{action}_j : [t_{\text{start}}, t_{\text{end}}]\}\}\}$.

Formally, the LLM functions as a generative model $\mathcal{G}$ mapping the TAD output and a carefully designed prompt $\mathcal{P}$ into a readable report $\mathcal{R}$:

$$\mathcal{R} = \mathcal{G}(\mathcal{T}, \mathcal{P})$$

The quality of prompts significantly influences the outcome of text generated by LLMs. Numerous studies (Liu et al. 2023; Gao, Fisch, and Chen 2020; Jiang et al. 2020) have been dedicated to improving the LLM response by optimizing the prompt design. We empirically verify the impact of prompt quality in the Experiments section, offering a comparative analysis of educational reports generated using different prompts.

Experiments

Dataset and Implementation Details

We conduct experiments utilizing our newly constructed dataset, as detailed previously in the Technical Architecture section. The dataset comprises of 4710 sequences, with a total length over 76 million frames. Each sequence spans from 5 to 25 minutes. We divide 60% of the dataset for training and the remaining 40% for testing purposes. The temporal

information in our constructed dataset automatically annotates 8 distinct classroom behaviors: *listening, reading, writing, playing phone, typing on keyboard, raising hand, showcasing signs of tiredness,* and *standing*. The training phase followed the experimental setups from prior works (Zhang, Wu, and Li 2022; Zhu et al. 2023), while the inference phase incorporated the Alphapose (Fang et al. 2022) as the human pose estimator to ensure accurate results. In the upcoming subsections, we present a series of experiments conducted to validate the efficacy of our framework. The performance in the first two experiments is evaluated using mean average precision (mAP) at various temporal intersection over union (tIoU) thresholds set at [0.3 : 0.7 : 0.2], with an additional report on the aggregate mAP computed as the mean value across all tIoU thresholds.

Ablation Studies on Dataset Construction

To systematically examine the individual contributions of various factors incorporated during the dataset construction, we conduct a series of ablation studies. This involved omitting certain factors to assess their respective impacts on the performance of our temporal action detection model.

The primary elements under investigation in this study include *Splicing Consistency*, *Multi-View Synthesis*, and *Motion Interpolation*.

The term *"consistency"* refers to maintaining a coherent narrative when merging various trimmed action clips into longer sequences, ensuring uniformity in subject, viewpoint, and settings. *Multi-View Synthesis*, on the other hand, pertains to the mapping of 3D skeleton sequences to multiple 2D viewpoints using a virtual camera model, thus enhancing the richness of the perspectives captured. *Motion Interpolation* focuses on the seamless transitioning between different action segments through appropriate frame interpolations.

Table 1 shows the performance that reflect the significance of each factor in dataset construction. Analyzing the initial five rows of the table allows us to pinpoint the individual and cumulative effects of each incorporated technique on the model's performance. Notably, the integration of the Multi-View Synthesis (M-View) emerges as a substantial contributor to enhancing the accuracy, suggesting that diverse viewpoints harbor a rich vein of information that is vital for the accurate detection of actions. This finding underscores the necessity for future researchers to prioritize diversity, especially in the context of viewpoint variation while constructing similar datasets.

Moreover, we investigated the influence of varying training data sizes on the model's performance. Analyzing the latter portion of Table 1, a clear trend of increasing model precision parallel to the augmentation of the training sample size is noticeable. However, this growth starts to decelerate as the sample size approaches 2000, hinting at a potential saturation point where further increase in data size may offer diminishing returns in performance enhancement.

Ablation Experiments for TAD Architectures

The Temporal Action Detection (TAD) component in our framework is vital in identifying and delineating various student actions within a classroom setting. In this subsection, we scrutinize the choices of backbone architectures employed during the design phase of our model, and assess their impacts on the overall performance. The initial phase in our TAD model involves feature extraction from the input motion data. In this regard, we explore two motion encoder backbones: ST-GCN (Yan, Xiong, and Lin 2018) and MotionBert (Zhu et al. 2023), tasked with extracting critical skeleton features. Subsequently, we consider a convolutional network (Qiu, Yao, and Mei 2017) as an alternative backbone for the TAD encoder. The results derived from experimenting with various encoder backbones are presented in Table 2. Our empirical observations demonstrate that the MotionBert coupled with Transformer blocks deliver optimal performance under our experimentation setup. Going forward, we believe that incorporating newer models into our framework could potentially enhance the performance.

Factors				Performance on test set			
Cons.	Interp.	M-View	D-size	0.3	0.5	0.7	Avg.
			all	23.63	21.16	15.47	20.35
✓			all	31.11	29.89	25.16	28.99
✓	✓		all	46.25	45.06	42.14	44.62
✓		✓	all	70.11	68.91	66.71	68.77
✓	✓	✓	all	**78.80**	**78.17**	**76.36**	**77.89**
✓	✓	✓	500	33.40	25.53	15.53	25.03
✓	✓	✓	1000	53.35	50.35	42.76	49.26
✓	✓	✓	1500	65.64	64.34	61.59	64.08
✓	✓	✓	2000	76.14	75.38	74.29	75.36

Table 1: Ablation studies on various factors (Cons.: Consistency, Interp.: Motion Interpolation, M-View: Multi-View Synthesis, D-size: training data size, with 'all' indicating the use of the complete training set.) affecting performance on the test set.

Backbones		Performance on test set			
Motion Enc.	TAD Enc.	0.3	0.5	0.7	Avg.
ST-GCN	Conv	32.30	27.61	21.46	27.22
ST-GCN	Transformer	48.78	44.75	38.24	44.06
MotionBert	Conv	54.45	51.69	42.63	50.10
MotionBert	Transformer	**78.80**	**78.17**	**76.36**	**77.89**

Table 2: Ablation studies on different combinations of motion and TAD encoders on the test set.

Analysis Report Generation: The Role of Prompts and Model Variants

This section evaluates how different prompts and large language model variants influence the pedagogical analysis report generation.

Prompt Influence on Report Quality. Prompts serve as an anchor point, guiding language models towards a specific context or perspective. In this section, we seek to explore the strategy for improving the prompts fed into the LLM to generate more nuanced and rich pedagogical analysis. To

investigate this, we utilize output action sequences derived from our TAD model, providing them to the LLM alongside varied prompts. We initiate our study with a basic prompt: *"Please generate an educational analysis report based on the data."* Subsequently, we enhance this prompt, emphasizing different dimensions to guide the LLM towards deeper insights. The prompt evolved through iterative refinements across the following dimensions:

- **Identity Implication**: Introducing the role of a pedagogist to lend an expert's perspective in the analysis (e.g., *"You are an expert in education and classroom behavior analysis,..."*).
- **Theoretical Guidance**: Incorporating specific educational theories to guide the analysis (e.g., *"Considering the theories of [specific pedagogical perspective],..."*).
- **Recommendations**: Instructing the LLM to provide actionable recommendations to improve classroom engagement (e.g., *"Provide recommendations to enhance classroom engagement..."*).
- **Contextual Analysis**: Associating students' actions to possible classroom triggers based on traditional classroom practices (e.g., *"Relate the student's actions to potential classroom stimuli..."*).
- **Classroom Dynamics**: Encouraging an analysis of the overall classroom dynamics to provide a holistic view (e.g., *"Analyze the overall classroom dynamics, teacher strategies, and classroom environment on the observed behaviors...."*).

To validate the effectiveness of our refined prompt strategies in eliciting deeper analyses from LLM, we conducted a user study with a panel of 15 graduate and doctoral students from educational backgrounds. Prior to evaluating the reports generated by various prompt enhancement techniques, each evaluator was required to watch a specific video relevant to the reports and review a sequence of student actions generated by the TAD model. The study adopted a blind review methodology where the evaluators were unaware of the variations in the prompt strategies. We report the result in Table 3. It can be observed that carefully crafted prompt strategies can indeed nurture more detailed and insightful pedagogical reports. Strategies that combine expert perspectives and theoretical grounding emerge as the most potent in guiding the LLM to generate deep analyses. Also, a nuanced approach to context elucidation paired with strategic guidance and insights into classroom dynamics creates a more grounded and insightful analysis, which is appreciated by the evaluators. Future work might further explore the intricate dynamics of prompt strategies to facilitate even more nuanced and rich analyses.

Performance Across Model Variants. In this experiment, we aim to explore the impact of utilizing different LLM variants in generating analysis reports, maintaining a constant enhanced prompt with Identity Implication and Theoretical Guidance. We employed four notable LLMs for this experiment: LLaMA-7B, LLaMA-13B, GPT-3.5, and GPT-4. Each model was used to generate five reports per sequence to mitigate bias. To evaluate the reports, we utilized the GPT-4-based automatic evaluation system recently proposed by (Chiang et al. 2023). This system assesses the depth of analysis, relevance of insights, actionable recommendations, theoretical grounding, and overall quality while allocating ratings. It assigns ratings based on predefined criteria that align with established educational assessment principles. The depth of analysis is evaluated in terms of the report's thoroughness in exploring classroom dynamics(Merriam and Tisdell 2015), while the theoretical grounding is assessed based on the report's alignment with established pedagogical theories(Lodico, Spaulding, and Voegtle 2010). The actionable recommendations are gauged for their practical applicability in educational settings.

Prompt Enhancement	Score (out of 10)
Basic Prompt	4.25
+ Identity Implication	5.75
+ Theoretical Guidance	6.17
+ Identity Implication + Theoretical Guidance	**7.91**
+ Contextual Analysis	5.08
+ Recommendations	7.08
+ Comprehensive Classroom Dynamics	**7.33**

Table 3: Average scores awarded by evaluators for reports generated with different enhanced prompts.

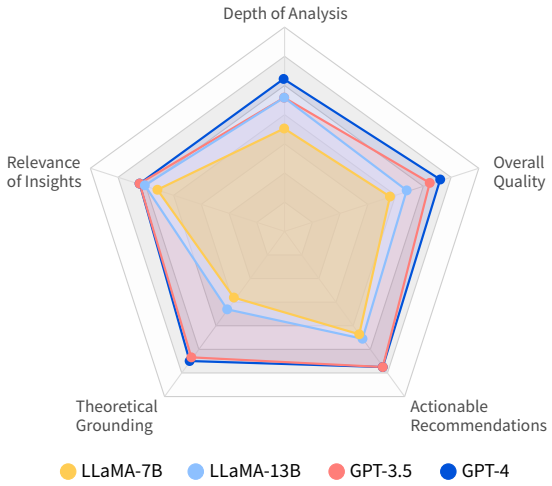

Figure 3: Comparative performance of various Large Language Models (LLM) variants.

The results in Figure 3 suggest that increasing the parameter count from LLaMA-7B to LLaMA-13B enhances all aspects of the generated reports. However, there is a significant lack in the theoretical depth of these two models, presumably due to a lack in the pre-training corpus. Comparatively, GPT-3.5 and GPT-4 exhibit a closer performance level, with GPT-4 having a slight edge in quality. From our experiments, it becomes evident that larger models tend to provide more detailed and nuanced reports. However, this benefit comes with higher computational demands. Balancing computational efficiency with depth of analysis will be a consideration for real-world implementations.

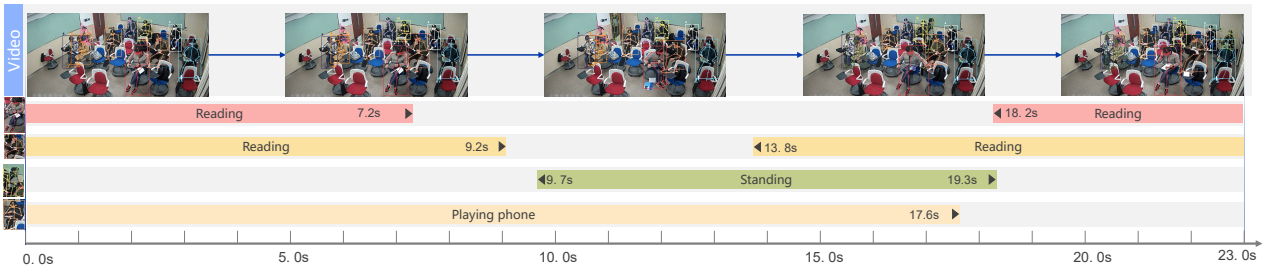

Figure 4: Temporal Action Detection (TAD) output for a segment of a classroom surveillance video.

Case Study: Deriving Insights from Authentic Classroom Footage

In this section, we showcase the qualitative results and discussions that emerged from applying our framework to real-world classroom scenarios.

Visualization of TAD and Report. Figure 4 illustrates the TAD output results of a classroom surveillance video, consisting of students' bounding boxes, human poses, and temporal action labels. It was observed that our model could accurately recognize actions such as "reading" and "writing", while pinpointing the start and end times for each action. These results, paired with an enhanced prompt, were fed into GPT-4 to generate the corresponding report. As demonstrated in Figure 5, we offer an excerpt from a pedagogical analysis report generated through GPT-4. The excerpt delineates a nuanced analysis drawn from the motion data processed through our TAD component. Utilizing the enhanced prompt strategy guided by Identity Implication and Theoretical Guidance dimensions, the LLM succeeds in crafting a report that not only identifies and interprets student actions but also relates them to established pedagogical theories and potential strategies for enhancing classroom engagement. Furthermore, the report goes a step further to offer actionable recommendations based on the observations, thus proving to be a powerful tool in the hands of educators seeking to understand and improve the learning landscape in their classrooms. This illustrative example underlines the potential of our framework in translating raw motion data into insightful analyses and practical solutions, affirming its utility in real-world educational settings. In our supplement, we furnish the complete video and report coupled with the application of available prompts.

Limitation and Future work. Despite the encouraging results garnered from our initial foray into real-world settings, we acknowledge the nascent stage of this research. One notable limitation is that our training data does not originate from real-world educational settings, potentially causing performance discrepancies when applied to actual classroom environments. Additionally, the current framework primarily focuses on individual actions commonly observed in classroom settings, such as reading, writing, and raising hands. However, it does not capture interactive or group activities, such as discussions among classmates. Future work will aim to integrate group action recognition algorithms to provide a more comprehensive analysis of student engagement. Furthermore, we aim to extend our work by integrating it with classroom lesson plans and students' self-assessments. This multi-dimensional approach will combine behavioral engagement with cognitive engagement, offering a more holistic understanding of student engagement.

Student Behavior Analysis Report

Based on the provided data of student behaviors in the class, a breakdown and analysis are presented as follows:

1. General Overview:

- Almost all students have segments of time where they are listening. It is considered a fundamental behavior expected in the class.
- .....

2. Individual analysis:

Student 1:

- Listening: Student 1 engages in listening for most of the observed time, which aligns with the class norm.
- .....

3. Standout Patterns:

- **Student7**: Spends a large portion of time (0s-1143s) on their phone. This is noticeably higher than any other student.
- .....

4. Analysis Based on Pedagogical Theories:

1. Behaviorism: Behaviorism emphasizes the role of external stimuli and reinforcement in shaping behavior. In this context, teachers can employ strategies such as positive reinforcement, praise, and rewards to encourage desired behaviors .....

5. Recommendations:

1. Collaborative learning: Encourage peer collaboration and cooperative learning activities to foster a positive peer influence and create a sense of community within the classroom. This can increase student engagement and motivation .....

References:

- Woolfolk, A. E. (2014). Educational psychology: Active learning edition (12th ed.). Pearson.
- Bandura, A. (1977). Social learning theory. General Learning Press.

Figure 5: An excerpt from a pedagogical analysis report generated by LLM.

Conclusion

In this paper, we introduce a comprehensive, end-to-end framework that analyzes student behavior from classroom videos and automatically derives pedagogically meaningful reports.Through the creation of a custom dataset and the employment of skeleton-based temporal action detection, we achieved nuanced insights into student behavior analysis. Integrating student behavioral sequence data and large language models to automatically generate reports provides a unique layer of analytical depth. Our work stands as a comprehensive approach to classroom behavior analysis, balancing technical rigor with ethical considerations, thereby contributing to a richer understanding of student engagement and pedagogical effectiveness.

Ethical Statement

In compliance with institutional guidelines, we have obtained authorization to collect and use classroom surveillance videos for research purposes. All identifying features in the videos, such as faces or identifiable markers, are blurred out to maintain anonymity. We are committed to ensuring that the footage will only be used for the aims specified in this research and not for any other purposes. We take extensive precautions to protect the database from unauthorized access, thus ensuring that the data is used responsibly. We advocate for continuous ethical reviews and community consultations to iteratively refine the framework's ethical considerations and implementation guidelines.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No.61977045).

References

Appleton, J. J.; Christenson, S. L.; and Furlong, M. J. 2008. Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, 45(5): 369–386.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Buch, S.; Escorcia, V.; Ghanem, B.; Fei-Fei, L.; and Niebles, J. C. 2019. End-to-end, single-stream temporal action detection in untrimmed videos. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association.

Chao, Y.-W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D. A.; Deng, J.; and Sukthankar, R. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1130–1139.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Dijkstra, R.; Genç, Z.; Kayal, S.; Kamps, J.; et al. 2022. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers.

Fang, H.-S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; and Lu, C. 2022. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Filtjens, B.; Vanrumste, B.; and Slaets, P. 2022. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*.

Finn, J. D.; and Zimmer, K. S. 2012. Student engagement: What is it? Why does it matter? In *Handbook of research on student engagement*, 97–131. Springer.

Fredricks, J. A.; Blumenfeld, P. C.; and Paris, A. H. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1): 59–109.

Gao, T.; Fisch, A.; and Chen, D. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8: 423–438.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

Lei, H.; Cui, Y.; and Zhou, W. 2018. Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46(3): 517–528.

Li, W.; Jiang, F.; and Shen, R. 2019. Sleep gesture detection in classroom monitor system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7640–7644. IEEE.

Lin, F.-C.; Ngo, H.-H.; Dow, C.-R.; Lam, K.-H.; and Le, H. L. 2021. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. *Sensors*, 21(16): 5314.

Lin, J.; Jiang, F.; and Shen, R. 2018. Hand-raising gesture detection in real classroom. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6453–6457. IEEE.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3889–3898.

Lin, T.; Zhao, X.; and Shou, Z. 2017. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, 988–996.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Lodico, M. G.; Spaulding, D. T.; and Voegtle, K. H. 2010. *Methods in educational research: From theory to practice*. John Wiley & Sons.

Malinka, K.; Peresíni, M.; Firc, A.; Hujnak, O.; and Janus, F. 2023. On the educational impact of ChatGPT: Is Artificial Intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 47–53.

McKinley, S.; and Levine, M. 1998. Cubic spline interpolation. *College of the Redwoods*, 45(1): 1049–1060.

Merriam, S. B.; and Tisdell, E. J. 2015. *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Moore, S.; Nguyen, H. A.; Bier, N.; Domadia, T.; and Stamper, J. 2022. Assessing the quality of student-generated short answer questions using GPT-3. In *European conference on technology enhanced learning*, 243–257. Springer.

Ngoc Anh, B.; Tung Son, N.; Truong Lam, P.; Phuong Chi, L.; Huu Tuan, N.; Cong Dat, N.; Huu Trung, N.; Umar Aftab, M.; and Van Dinh, T. 2019. A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences*, 9(22): 4729.

Pei, J.; and Shan, P. 2019. A Micro-expression Recognition Algorithm for Students in Classroom Learning Based on Convolutional Neural Network. *Traitement du Signal*, 36(6).

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, 5533–5541.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1049–1058.

Tang, L.; Xie, T.; Yang, Y.; and Wang, H. 2022. Classroom Behavior Detection Based on Improved YOLOv5 Algorithm Combining Multi-Scale Feature Fusion and Attention Mechanism. *Applied Sciences*, 12(13): 6790.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trowler, V. 2010. Student engagement literature review. *The higher education academy*, 11(1): 1–15.

Wang, Z.; Jiang, F.; and Shen, R. 2019. An effective yawn behavior detection method in classroom. In *International conference on neural information processing*, 430–441. Springer.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, J.; Li, Y.; Wang, L.; Wang, K.; Li, R.; and Zhou, T. 2019. Skeleton based temporal action detection with yolo. In *Journal of Physics: Conference Series*, volume 1237, 022087. IOP Publishing.

Xu, H.; Das, A.; and Saenko, K. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, 5783–5792.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yeung, S.; Russakovsky, O.; Mori, G.; and Fei-Fei, L. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2678–2687.

Yu, Z.; Hu, Y.; Xiang, S.; Liu, T.; and Fu, Y. 2023. CC-PoseNet: Towards Human Pose Estimation in Crowded Classrooms. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*, 2914–2923.

Zheng, R.; Jiang, F.; and Shen, R. 2020. Intelligent student behavior analysis system for real classrooms. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9244–9248. IEEE.

Zhou, H.; Jiang, F.; and Shen, R. 2018. Who are raising their hands? Hand-raiser seeking based on object detection and pose estimation. In *Asian Conference on Machine Learning*, 470–485. PMLR.

Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; and Wang, Y. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.