Context Matters: Peer-Aware Student Behavioral Engagement Measurement via VLM Action Parsing and LLM Sequence Classification*

Ahmed Abdelkawy† Ahmed Elsayed Asem Ali Aly Farag Thomas Tretter Michael McIntyre
University of Louisville
{ahmednady.abdelkawy,ahmed.elsayed,asem.ali,aly.farag, tom.tretter, michael.mcintyre}@louisville.edu

Abstract

Understanding student behavior in the classroom is essential to improve both pedagogical quality and student engagement. Existing methods for predicting student engagement typically require substantial annotated data to model the diversity of student behaviors, yet privacy concerns often restrict researchers to their own proprietary datasets. Moreover, the classroom context, represented in peers' actions, is ignored. To address the aforementioned limitation, we propose a novel three-stage framework for video-based student engagement measurement. First, we explore the few-shot adaptation of the vision-language model for student action recognition, which is fine-tuned to distinguish among action categories with a few training samples. Second, to handle continuous and unpredictable student actions, we utilize the sliding temporal window technique to divide each student's 2-minute-long video into non-overlapping segments. Each segment is assigned an action category via the fine-tuned VLM model, generating a sequence of action predictions. Finally, we leverage the large language model to classify this entire sequence of actions, together with the classroom context, as belonging to an engaged or disengaged student. The experimental results demonstrate the effectiveness of the proposed approach in identifying student engagement. The source code and dataset will be available upon request.

1. Introduction

Students' active participation and engagement in the classroom are essential for successful learning outcomes [14, 30]. Student engagement is a multidimensional construct comprising behavioral, emotional, and cognitive engagement [18]. Behavioral engagement refers to students' self-directed actions to get access to the curriculum, such as studying and doing homework, hand and body movements while attending lectures, as well as engaging cooperatively in classroom activities [4]. Emotional engagement, on the other hand, refers to students' feelings about their learning experience, teachers, and peers [28]. This involves feelings of happiness in learning, disinterest in the subject matter, or frustration and struggle to understand. Lastly, cognitive engagement is the psychological investment in academic achievement, which is usually revealed by a student's ability to comprehend, organize, and analyze in the act of deep learning to fully understand the material [8, 11].

Behavioral engagement is a key indicator of student participation and attentiveness, with a direct impact on academic performance [5]. During a lecture, students' behaviors often reveal whether they are on- or off-task. While an engaged student typically listens, takes notes, and types on a laptop, a disengaged student engages in off-task activities like playing with a mobile and frequently checking the time. Conventional assessments of behavioral engagement typically depend on manual classroom observations [20, 29]. While informative, these observations are labor-intensive and subject to observer bias, which limits their scalability in large educational settings. Measuring students' engagement provides instructors with quantitative feedback, enabling them to adjust lectures and classroom activities to effectively re-engage disengaged students. An important consideration in engagement annotation is temporal resolution, which defines the time scale over which engagement is assessed. Following the standard classroom observation protocols [20, 29], we adopt a 2-minute temporal resolution.

Automating student behavioral engagement in classrooms has received considerable attention due to advances in computer vision and deep learning. Existing studies for student behavior analysis can generally be categorized according to the target outcome into three groups: action recognition, learning analytics, and engagement measurement. Action recognition approaches focus on identifying a particular group of students' behaviors without conducting additional analysis, providing instructors with minimal useful information. Furthermore, a considerable number

*This research was funded by NSF Award Number 2337154.
†Corresponding author

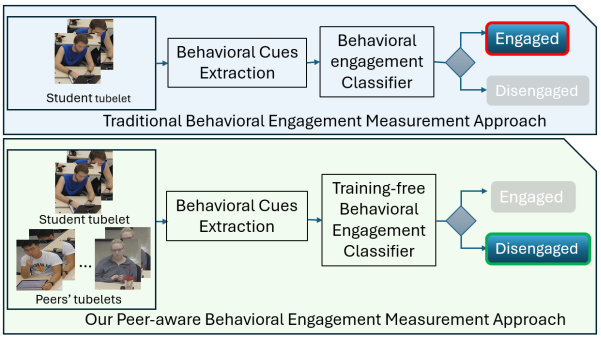

Figure 1. Context matters in engagement classification. Under traditional approaches, a student's action sequence showing continuous typing on a laptop may be labeled as engaged. However, our peer-aware model classifies the same sequence as disengaged when surrounding students are not engaged in similar note-taking behavior.

of studies [21, 38] formulated student behavior analysis as single- or multi-object detection, ignoring the complexity and diversity of student behaviors. Learning analytics approaches [3], which use current computer vision techniques to get a specific student's actions, gaze, and/or facial expression, provide a set of visual features but little useful information for instructors.

Unlike traditional Human Action Recognition (HAR) models that operate on single-action trimmed clips, student behavior analysis demands processing continuous, untrimmed videos in which the execution and order of actions are critical. To address this challenge, research has shifted toward temporal action localization (TAL) [31, 36], which identifies and locates the start and end times of sparse, intermittent actions within long, untrimmed videos. However, TAL tends to ignore "background" segments (e.g., listening, looking to the side/back), which are important in engagement analysis. Furthermore, while Yu et al., [36] leveraged a large language model (LLM) to generate a behavior analysis report, their approach does not explicitly estimate student engagement levels. To measure students' behavioral engagement, Abdelkawy et al. [2] represented a student's actions and their frequencies over an arbitrary time period as a histogram of actions. This histogram, along with the student's gaze, is then used as input to an engagement classifier. However, the student's actions within this time interval are not manually annotated, leaving the evaluation of the behavioral engagement classifier incomplete. Overall, existing approaches face two key limitations:

- Scarcity of publicly available datasets for students actions and engagement, as most studies rely on proprietary datasets due to privacy concerns.
- limited consideration of classroom context, which is defined as actions performed by peers while the student of interest is performing a certain action (see Fig. 1).

Student engagement depends not only on their actions but also on the ongoing classroom activity and the behavior of their peers; therefore, classroom context is crucial for accurate engagement assessment. This is particularly important as contemporary pedagogy increasingly moves beyond passive lecture-based learning. The ICAP framework [11] characterizes learning activities along a Passive–Active–Constructive–Interactive hierarchy, where interactive tasks generally reflect deeper cognitive engagement. The same observable behavior can therefore signal opposite engagement states depending on context. For example, a student turning to talk to a peer during a traditional lecture probably shows disengagement, but doing the same during a peer discussion activity that is prompted shows engagement. Similarly, when a student is asked to work independently on a problem, he is engaged by focusing intently on his laptop for problem-solving. In contrast, he is disengaged if he uses his laptop for unrelated activities such as browsing social media while the instructor is lecturing.

To address the aforementioned limitations, we propose a novel three-stage framework for classifying behavioral engagement based on student actions. First, we perform few-shot adaptation for the vision-language model (VLM) [24] to classify human actions in trimmed videos with a limited number of labeled samples per category. Second, because the student actions are continuous, we divide each 2-minute time-resolution student video into non-overlapping segments using the sliding temporal window technique. The VLM model (from stage 1) is then used to assign a predefined label to each segment, resulting in a time-ordered sequence of predicted actions. Third, an LLM uses this sequence of actions and the classroom context as input to determine whether a student is engaged or disengaged.

To the best of our knowledge, prior studies have not investigated whether current vision-language models (VLMs) and large language models (LLMs) can effectively recognize the student's class actions and measure behavioral engagement based on identified actions, respectively. Accordingly, this paper addresses the following research questions:

1. Compared to traditional action classification models, can VLMs accurately classify students' class actions?
2. Which is a more accurate depiction of students' behavior: an action sequence or action histogram?

Our contributions are two-fold:

- We introduce a novel framework for measuring student behavioral engagement, which integrates VLM for action recognition and LLM for engagement classification. To overcome the lack of publicly available

datasets for student actions and engagement, we perform few-shot tuning for VLM to improve student action recognition, and we leverage the LLM in zero-shot prompting for engagement classification.

- Moreover, we construct a novel student behavior dataset with two components: 1) trimmed videos (13 categories, controlled environment) with annotated students' actions and 2) untrimmed videos (3 lectures, 11 students) with annotated students' engagement. The experts' annotations of this dataset include engaged/disengaged labels for each 2-minute video and dense temporal action labels for each segment within them.

2. Related Work

2.1. Student Engagement Measurement

The emerging and promising real-time engagement measurement approaches, enabled by advances in computer vision and machine learning, allows tracking a student's engagement at fine-grained temporal resolutions on the order of seconds or minutes through discrete, objective indicators. Existing research of student behavior analysis in classrooms can be grouped into three categories: student action recognition, learning analytics, and engagement measurement. While student action recognition approaches focus on the student's actions from trimmed videos or a single image, learning analytics approaches rely on facial expressions, student actions, and/or head gaze. For example, Sheng et al. [26] proposed an enhanced YOLOv8-based detector for localizing and classifying student behaviors from single classroom images. Their framework incorporates a Multi-Scale Large Kernel Convolution Module (MLKCM) to extract multiscale spatial features and a Progressive Feature Optimization Module (PFOM) that segments the channel dimension of the input feature map for improved feature refinement. Ahuja et al. [3] presented a classroom sensing system (EduSense) that provides a set of features, including detection of hand raise, hand on face, cross arms, smile, head orientation, and speech duration for both the teacher and the students. These features are extracted from students' skeleton joints and facial landmarks due to the presence of much visual noise, occlusion in the classroom, and privacy preservation. Line et al. [22] recognized student behavior in the classroom based on upper body skeleton joints. They first extract students' poses using bottom-up pose estimation (OpenPose) and remove the incorrect connection of students' joints using YOLOv4 object detection, then generate a feature vector using normalized joint locations, joint distances, and bone angles to represent student posture and use a four-fully-connected-layer neural network to classify the student feature vector into four classes. The limitation of works [3, 22] is that student actions are recognized from single images rather than video segments, thereby ignoring temporal dynamics of the actions. Moreover, Line et al. [22] consider actions with different engagement levels, such as writing and playing with a phone, as one class (bowing).

Student engagement studies, which employ student actions and/or head gaze to measure the student behavioral engagement, provide teachers with actionable insights. In [37], the authors proposed a framework that processes classroom videos by first estimating students' 2D poses per frame and then applying a Temporal Action Detection (TAD) [17] to produce a temporal sequence of actions for each student. These per-student action sequences are subsequently fed into an LLM to generate behavioral reports. The TAD module was trained on the NTU RGB+D dataset [25], using action classes relevant to the classroom environment. However, their approach fails to provide explicit engagement labels and does not account for the peer-aware context when generating a textual behavioral report that includes each student's actions analysis. Abdelkawy el al. [2] utilized a random forest classifier to detect whether the student is on-task or off-task based on a histogram of actions, together with head gaze. Specifically, each student's 2-minute video was divided into overlapping segments using the sliding temporal window technique; the student skeleton sequence of each segment is assigned a predefined action label using 3D CNN model and then the histogram is created from these identified actions. Nevertheless, the student's actions during the given period are not manually annotated. Consequently, the evaluation of the behavioral engagement classifier is still incomplete. Unlike these studies, we analyze the entire 2-minute video to get the student's sequence of actions, which are used for engagement measurement. Moreover, we constructed a dataset that has dual expert annotations: engaged/disengaged for each 2-minute video and its corresponding frame-wise action labels to measure the performance of the engagement classifier. We also investigated the few-shot adaptation of the vision-language model for student action recognition, which learns to differentiate between action classes with a few training samples to overcome the challenge of the scarcity of annotated data required to model the diversity of student behaviors.

2.2. Video Action Recognition

Action recognition aims to assign a predefined action label to a trimmed video, which contains a single action. The majority of existing action recognition approaches are based on RGB video streams. Existing RGB video-based methods can be categorized into four classes based on the network architecture: two-stream 2D convolutional neural network (CNN), 3D CNN, compute-efficient, and transformer-based methods [1]. Two-stream 2D CNN-based methods [27, 34] learn the appearance and motion features separately using two 2D CNN streams. 3D CNN-based methods [10, 16] are introduced to learn spatiotemporal features by tempo-

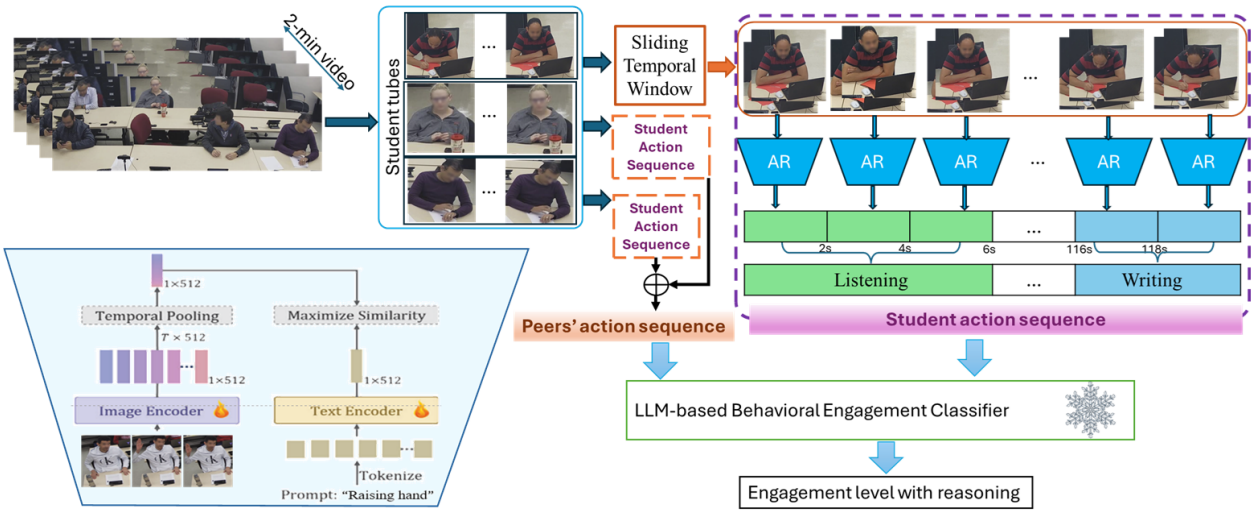

Figure 2. Overview of the proposed framework. Students' tubelets are extracted from a two-minute video, and each student's tube is divided into non-overlapping segments using a sliding temporal window technique. The action sequence is then produced by the fine-tuned VLM model. LLM uses the generated action sequence to classify engagement after integrating it with peers' action sequences.

rally extending the convolution and pooling kernels of a 2D CNN. However, 3D CNNs are computationally intensive. To achieve the computation/accuracy tradeoff, compute-efficient networks [15, 33] are introduced. Transformer-based methods [6,7] adapt the de facto vision transformer to model the long-term temporal dynamic in a video sequence. While most of these methods are unimodal, ignoring the semantic information of action categories, works such as ActionCLIP [35] and XCLIP [23] adapt the CLIP to vidoe action recognition through fusing the vision-language representation of CLIP with introduced temporal modeling components. Rasheed et al. [24] found that these introduced components can negatively impact CLIP's generalization capability, whereas full fine-tuning of CLIP on the Kinetics-400 dataset yields competitive performance. In this work, we adapt the work of [24] in a few-shot setting to boost the performance of student classroom action recognition.

2.3. Large Language Models

Large language models (LLMs) are Transformer-based models with hundreds of billions of parameters, trained on large-scale text corpora (e.g., GPT-3 [9], LLaMA [13], Gemma [32], etc.). They exhibit generalization abilities, enabling them to perform tasks in zero-shot settings without requiring task-specific training examples. While several works have employed LLMs in education for multiple-choice question generation and assignment grading, Yu et al. [36] utilized an LLM to generate a pedagogical report based on the structured temporal action information for each student, including action categories and their timestamps. In contrast, our work uses an LLM to assign an engagement level to each student's sequence of actions.

3. Method

3.1. Overview of the proposed method

The proposed framework consists of three main steps: segment-wise action recognition, VLM action parsing, and engagement classification (Fig. 2). First, a vision-language model is adapted in a few-shot manner using a small set of labeled trimmed student action clips to improve recognition on continuous student videos. Second, the student's untrimmed video is split into non-overlapping segments using a sliding-window technique, and the fine-tuned VLM assigns an action label to each segment. Consecutive identical predictions are merged to form an ordered action sequence that preserves both action identity and duration. Finally, this action sequence, along with the classroom context, is provided to an LLM to determine whether the student is engaged or disengaged.

Representing student's behavior during a 2-minute interval as a temporal sequence rather than a histogram, where the duration of each action is accumulated, allows the model to capture not only how long each action occurs but also their ordering, which is shown to be crucial for interpreting engagement. For example, a student may still be deemed engaged if they take notes for the majority of the time period but play on their phone for one continuous 25-second block (e.g., at the beginning or end). However, a student is considered disengaged if they alternate between taking notes and making frequent quick phone checks because this pattern of frequent disruption shows an inability to maintain cognitive focus on the task. Therefore, the temporal sequence and context of actions more accurately reflect engagement than the total amount of time spent in each action (e.g., via a histogram). Furthermore, when representing student ac-

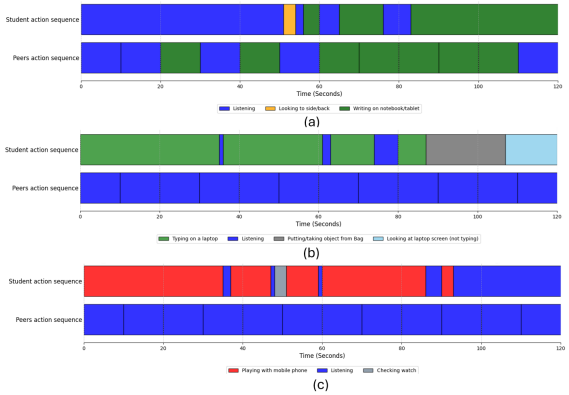

Figure 3. a) for engaged student, b, c) for disengaged student

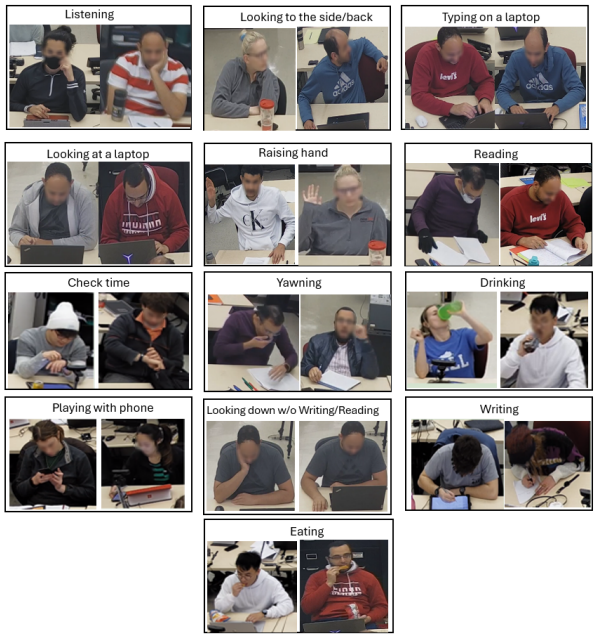

Figure 4. Samples of students' actions during class time.

tion as a sequence, together with peers' action sequences, the LLM predictions are easier to interpret. Fig.3 shows examples of students' actions aligned with classroom actions for engaged (a) and disengaged (b, c) students.

3.2. Actions dictionary

Because this study focuses on college students in introductory engineering STEM courses, the range of classroom actions from which behavioral engagement can be inferred is relatively limited. We adopted the classroom observation protocols in [2, 20], and identified the following 13 actions: Eating meal/snack, Writing on notebook/tablet, Typing on a laptop, Playing with mobile phone, Looking to the side/back, Looking down w/o reading/writing, Looking at laptop screen (not typing), Raising hand, Checking time, Reading, Drinking, Yawning , and Listening. Fig. 4 shows samples of such actions.

3.3. Student action recognition

We investigated few-shot learning for action recognition using vision-language models (VLMs). In particular, we use Video Fine-tuned CLIP (ViFi-CLIP) [24] because of its efficiency and simplicity. For a video segment $S_i \in \mathbb{R}^{TxHxWxC}$, the embeddings of its T frames are obtained using the CLIP image encoder and then average-pooled to get the video-level representation $v_i \in \mathbb{R}^D$. The corresponding class label C_i is embedded in a prompt (e.g., 'a photo of a <category>') and passed through the CLIP text encoder to produce a text embedding $t \in \mathbb{R}^D$. We fine-tuned the ViFi-CLIP model on the student actions dataset in a few-shot setting with 16 video clips per action category by maximizing the cosine similarity between a batch of video-level embeddings v_i and their corresponding text embeddings t_i using cross-entropy loss. Additionally, we incorporate entropy regularization into the loss function to enhance generalization and mitigate overfitting by discouraging overly confident, incorrect predictions.

$$L_{total} = -\frac{1}{N}\sum_{i}^{N} log(p_i) + \frac{1}{N}\sum_{i}^{N}(-\sum_{k=1}^{C} p_{i,k} log(p_{i,k})) \tag{1}$$

where $p_i = \frac{exp(sim(v_i,t_i)/\tau}{\sum_j exp(sim(v_j,t_j/\tau)}$

3.4. Student engagement classification

To determine whether a student is engaged or disengaged, we leverage the LLM as a classifier in zero-shot prompting. Following the prompt-based learning paradigm, the LLM is conditioned on a prompt x_{prompt} to generate a predefined textual label $y \in Y$ (e.g., engaged, disengaged). The prompt x_{prompt} is constructed by concatenating a **task description** x_{desc} and **input** x_{input} as $\{x_{desc};x_{input}\}$. An example of the task description x_{desc}, which describes the engagement classification based on the student's actions and classroom context, is provided in supplementary material (Sec. 1). An example of the input x_{input}, which includes the student action sequence with the duration of each action, along with the classroom context, represented in the majority of peers' actions during the time interval, can be represented as:

- `Student actions: { writing on tablet (00:00-00:20); listening (00:20-01:05; etc.}`
- `Classroom context: {listening (00:00-00:10); writing (00:10-00:50); etc.}`

4. Experiments

4.1. Experimental Setup

Benchmarks. The lack of a public benchmark for classroom engagement analysis makes it necessary to create one.

So, we created a behavioral engagement dataset with two subsets of annotated videos—one for student actions and the other for student engagement—in order to independently assess each step of the proposed framework. All videos were recorded from a wall-mounted 4K camera at 15 fps in classrooms of 6–11 students. For the student action dataset, we recorded 7 subjects performing 13 classroom-related actions in a controlled environment. This resulted in 208 trimmed clips for training (16 per action) and 46 trimmed video clips for testing across the 13 action categories. This trimmed video dataset is designed to evaluate the proposed action recognition stage (Stage 1). The student engagement dataset comprises videos from three half-hour lectures involving eleven students. Each video is segmented into 2-minute clips per student, yielding a total of 455 video clips. Experts in psychology and education labeled each two-minute clip as either engaged or disengaged, resulting in 400 engaged and 55 disengaged samples. Each two-minute video segment is also temporally partitioned manually, and every resulting segment is assigned a predefined action label. The entire student engagement dataset is used as a test set to evaluate the remaining stages—proposed temporal action segmentation (Stage 2) and engagement classification (Stage 3)—since both stages are training-free: Stage 2 employs a temporal sliding window, while Stage 3 uses an LLM with zero-shot prompting.

Implementation details. We implemented the proposed framework in PyTorch on a workstation with two NVIDIA TITAN RTX GPUs (24 GB each). The VIFI-CLIP model (ViT-B/16 backbone) is fine-tuned in a few-shot setting with **K**=16 samples per class, using a batch size of 32 and a learning rate 2×10^{-6} for 50 epochs. We use the AdamW optimizer with a weight decay of 0.001 and apply cosine annealing for learning rate scheduling. Each video clip is represented by 16 uniformly sampled frames, which are resized to 224×224. For VLM-based action parsing, both the sliding window size and stride are set to 3 seconds (45 frames), resulting in 40 non-overlapping segments for each 2-minute video. For an LLM-based engagement classifier, we set the temperature to 0.1 to ensure a deterministic and non-creative output.

Our proposed framework consists of distinct tasks: action recognition, action parsing, and engagement classification. Each stage is evaluated with task-appropriate metrics. For action recognition, we report the top-1 accuracy. For action parsing, we adopt standard temporal action segmentation metrics, which fall into two categories: frame-based and segment-based metrics [12]. The frame-based metric, Mean over Frames (MoF), measures the ratio of correctly predicted frames to the total number of frames. However, the MoF metric suffers from two major problems. First, it is unreliable for imbalanced datasets, where long-duration action classes can dominate the sequence. Second, it fails to capture segment quality. A high MoF can still be achieved even if the action segments are fragmented (over-segmented). To account for the over-segmnetation phenomenon, we also report two segment-based metrics: Edit Score and $F1@\tau$. The Edit Score evaluates the similarity between the predicted and ground-truth action sequences without requiring exact frame alignment, capturing the correctness of action order. The $F1@\tau$ score measures the Intersection-over-Union (IoU) between predicted and ground-truth segments under threshold τ, counting a segment as a true positive only when its IoU exceeds τ (with $\tau \in \{10, 25, 50\}$). Multiple predictions overlapping a single ground-truth segment result in one true positive, and the rest are counted as false positives. Finally, we evaluate the engagement classification stage using precision, recall, and F1 score. These metrics are selected over simple accuracy to effectively assess performance in the presence of class imbalance between the engaged and disengaged labels.

Table 1. Comparison of action recognition models.

Model	Student Action Dataset			
	K=2	K=4	K=8	K=16
TC-CLIP [19]	52.2%	52.2%	80.4%	93.5%
XCLIP [23]	43.2%	79.6%	88.6%	97.9%
ViFi-CLIP [24]	47.7%	**81.8.0%**	**93.2%**	**97.9%**

4.2. Evaluation of the proposed framework stages

Stage #1: Action Recognition. We evaluated state-of-the-art video-language models, including XCLIP [23], TC-CLIP [19], and ViFi-CLIP [24] on our student classroom action test set (46 instances) under zero-shot and few-shot protocols. For the zero-shot setting, the ViFi-CLIP achieved 52.1% top-1 accuracy, which is far from satisfactory. We attribute this poor performance to the model's fine-tuning on general web videos (Kinetics-400). While this bridges the image-video gap for general tasks, it fails to capture the subtle, fine-grained differences in our domain, such as differences between reading and writing. For the few-shot setting, as shown in Tab. 1, we note that the performance improves for all methods as the number of shots increases. Across most values of K, ViFi-CLIP achieves the highest accuracy, with particularly noticeable gains in the low- and mid-shot regimes, demonstrating that it can be effectively adapted to our specialized domain with minimal data. For example, ViFi-CLIP achieves gains of 4.5%,2.2%, and 4.6%, compared to XCLIP at k=2, K=4, and K=8, respectively.

Stage #2: Temporal action segmentation (TAS). Tab. 2 shows the comparison of Gemini 2.5 Pro and Gemini 2.5 Flash, which are multimodal large language models (MLLMs), with the proposed VLM action parsing on the student engagement dataset, which provides dense action

Table 2. Comparison of temporal action segmentation using VLM-based, Gemini-2.5-Flash, and Gemini-2.5-Pro-based.

TAS	Accuracy	Edit	F1@10	F1@25	F1@50
Gemini-2.5-flash	57.2	37.4	39.0	34.5	25.3
Gemini-2.5-pro	69.8	51.8	58.2	55.2	45.2
VLM-based	67.0	45.7	48.3	43.9	31.4

labels for each 2-minute video. For MLLMs, we curated a prompt describing the task to be performed on the given 2-min video (see supplementary material Sec. 2 for details).

Dense student action labeling is challenging due to fine-grained, visually similar actions (e.g., 'looking at laptop' vs. 'looking down') and cluttered environments. The task is further complicated by inter-student occlusion, where objects from one student's space appear in another's, creating ambiguous visual data. It is noticeable from Tab. 2 that Gemini-2.5-pro achieves higher MoF, edit score, and F1 score compared to both Gemini-2.5-flash and the proposed VLM action parsing, while the proposed VLM action parsing outperforms Gemini-2.5-flash across all three metrics. Moreover, the proposed VLM action parsing is more consistent in identifying the on- and off-task actions than both Gemini MLLMs. This leads to accurate engagement classification, as shown next.

Stage #3: Engagement Classification.

To quantify the performance of the context-aware engagement classifier, we first evaluate LLMs using the manual dense student action annotation, as shown in Tab. 3. We can notice that both Llama-3-8B [13] and Gemma-2-9B [32] achieve similar performance in terms of F1 score. The moderate performance of the LLMs on the disengaged class can be attributed to the inherent ambiguity of student engagement. In our dataset, some students were doodling on tablets rather than taking notes; experts labeled these cases as disengaged, but our action labels do not distinguish fine-grained behaviors such as writing versus drawing. Additionally, students who were daydreaming were sometimes annotated as listening because their posture suggested attention despite their lack of focus. Finally, student seated at the classroom periphery may appear to look forward without actually paying attention to the instructor. Since the context-aware engagement classifier (Gemma) achieves the highest F1-score of 92% using manually annotated actions, we subsequently employ Gemma with automated action parsing. As shown in Tab. 4, it achieves an F1-score of 86% with Gemini-based parsing and 88% with VLM-based parsing. This performance gap is attributable to the difficulty of the student action recognition task, as even the advanced Gemini-2.5-Pro model struggles to accurately parse the 2-minute-long student videos.

4.3. Comparison with Existing Works

Direct comparison with prior work is difficult due to differences in datasets, action classes, and annotation protocols. Therefore, we implemented the approach in [2] by leveraging the predicted actions from the fine-tuned VLM model and applying its reported best-performing classifier. For each student tublet, we obtain the sequence of predicted actions and construct the histogram of actions that represents the frequency of actions within a 2-minute interval. This histogram is then used as an input for the Random Forest classifier for engagement measurement. Since the method in [2] requires training, we evaluate it using Leave-One-Out Cross-Validation (LOOCV) on the entire dataset, which contains 3 30-minute lectures. As shown in Tab. 5, the proposed framework surpasses the compared method by 3% in F1-score. It also achieves a higher F1-score for the disengaged class (52% vs. 47%). This improvement highlights the effectiveness of our approach, which uses a student's action sequence augmented with the actions of their peers.

Table 3. Comparison of LLMs for context-aware engagement classification based on manual action annotation of student tubelets.

LLM		Recall	Precision	F1-score
Llama-3-8B [13]	disengaged	0.76	0.56	0.65
	engaged	0.92	0.97	0.94
	weighted avg	0.90	**0.92**	0.91
Gemma-2-9B [32]	disengaged	0.69	0.67	0.68
	engaged	0.95	0.96	0.95
	weighted avg	**0.92**	**0.92**	**0.92**
GPT-3.5-turbo [9]	disengaged	0.71	0.52	0.60
	engaged	0.91	0.96	0.93
	weighted avg	0.89	0.90	0.89

Table 4. Comparison of context-aware engagement classifier based on VLM action parsing and Gemini-2.5-pro-based temporal segmentation of student tubelets using Gemma-2-9B.

Action parsing		Recall	Precision	F1-score
Gemini-2.5-Pro-based	disengaged	0.58	0.43	0.49
	engaged	0.89	0.94	0.92
	weighted avg	0.85	**0.88**	0.86
Our VLM-based	disengaged	0.53	0.52	0.52
	engaged	0.93	0.93	0.93
	weighted avg	**0.88**	**0.88**	**0.88**

Table 5. Comparison with existing work based on the proposed VLM-based action parsing of student tubelets using Gamma-2.9B.

Method		Recall	Precision	F1-score
[2]	disengaged	0.70	0.36	0.47
	engaged	0.84	0.97	0.90
	weighted avg	0.83	0.91	0.85
Ours	disengaged	0.53	0.52	**0.52**
	engaged	0.93	0.93	**0.93**
	weighted avg	0.88	0.88	**0.88**

4.4. Cross-Course Evaluation

Our student engagement dataset, consisting of 3 30-minute lectures, was collected from a single course. This raises a question: Can the proposed framework generalize to different classes with new students, instructors, and course content? To address this, we collected an addi-

Table 6. Evaluation of a context-aware engagement classifier on a 30-minute lecture from different class using Gemma-2-9B.

Student tubelet annotation		Recall	Precision	F1-score
Manual annotation	disengaged	0.85	0.81	0.83
	engaged	0.89	0.92	0.90
	weighted avg	**0.88**	**0.88**	**0.88**
Gemini-2.5-Pro-based	disengaged	0.90	0.56	0.69
	engaged	0.62	0.92	0.74
	weighted avg	0.72	0.79	0.72
Our VLM-based	disengaged	0.65	0.84	0.73
	engaged	0.93	0.83	0.88
	weighted avg	0.83	0.83	0.83

tional 30-minute lecture from a separate class involving eight different students. This yielded 113 two-minute segments, consisting of 73 engaged and 40 disengaged samples. Tab. 6 presents a comparison of the context-aware engagement classifier based on ground truth manual annotation, Gemini-2.5-Pro-based temporal action segmentation (TAS), and our VLM-based TAS on a 30-minute lecture from a different class. The results show that the proposed framework with VLM-based TAS surpasses the Gemini-2.5-Pro-based variant by a substantial margin of 9% in weighted F1-score. This improvement highlights the generalization capability of the proposed VLM–LLM framework, demonstrating its ability to accurately parse and classify student action sequences despite variations in student cohorts and student–teacher dynamics.

4.5. Ablation Studies

To ablate our model, we conducted two experiments: a) removing the classroom context to assess its contribution, and b) comparing our sequence-based method with a histogram-based counterpart. In this section, all engagement classification experiments were performed using Gemma-2.9B, based on the action sequences obtained from our VLM-based action parsing for each 2-minute student video.

a) Engagement classification without classroom context: Tab. 7 shows a comparison between the proposed peer-aware approach and its peer-free counterpart, respectively. It is noticeable that the proposed peer-aware approach outperforms its peer-free counterpart, achieving a 1% higher F1-score for the disengaged class. Moreover, a context-free approach performs badly ($< 50\%$precision) in identifying disengaged students, which is critical. The prompt of the context-free engagement classification is provided in the supplementary material (Sec. 1).

b) Input representation: The sequence-based and histogram-based input representations of a student's time-stamped actions are compared in Tab. 8. We note that the proposed sequence-based method achieved an 88% F1-score compared to 86% for its histogram-based counterpart. Specifically, the sequence-based approach clearly outperforms the latter in terms of F1-score for the disengaged class (52% vs. 39%). This substantial difference shows the student engagement is better captured by the temporal action sequence rather than by aggregated action durations represented as a histogram. Furthermore, temporal action sequences preserve the order and timing of events, enabling relational analysis of a student's behavior in the context of peers' concurrent actions. In contrast, action histograms lack temporal structure and therefore fail to capture such contextual dynamics.

Table 7. Comparison of engagement classifier with/out classroom context based on the introduced VLM-based action parsing of student tubelets using Gemma-2-9B.

w/o		Recall	Precision	F1-score
Context-free	disengaged	0.62	0.44	0.51
	engaged	0.89	0.94	0.92
	weighted avg	0.86	0.88	0.87
context-aware	disengaged	0.53	0.52	0.52
	engaged	0.93	0.93	0.93
	weighted avg	**0.88**	**0.88**	**0.88**

Table 8. Comparison of input representation for engagement classification based on the introduced VLM action parsing of student tubelets Using Gemma-2-9B.

Representation		Recall	Precision	F1-score
Histogram	disengaged	0.35	0.44	0.39
	engaged	0.94	0.91	0.93
	weighted avg	0.87	0.86	0.86
Sequence	disengaged	0.53	0.52	0.52
	engaged	0.93	0.93	0.93
	weighted avg	**0.88**	**0.88**	**0.88**

5. Conclusion

We presented a novel context-aware framework to measure student behavioral engagement in a classroom setting. The proposed framework employs a Vision-Language Model (VLM)-based action parser to get a sequence of student actions during an arbitrary time interval and then feeds this student sequence, along with the classroom context (i.e., peers' activities), to a Large Language Model (LLM) for engagement measurement. The experimental results demonstrated that context awareness is crucial to take into account to detect disengaged students because, depending on the context, the same action may indicate engagement or disengagement. On the constructed dataset, we evaluated the proposed framework and obtained engagement classification F1 scores of 92%, 86%, and 88% using manual, Gemini-based, and modified VLM-based temporal action segmentation, respectively. Furthermore, the proposed framework overcomes the challenge of the scarcity of annotated data required to model the diversity of student behaviors.

6. Ethical Statement

This study was approved by IRB # 19.0513 and informed consent was obtained from all participants.

References

[1] Ahmed Abdelkawy, Asem Ali, and Aly Farag. Epam-net: An efficient pose-driven attention-guided multimodal network for video action recognition. *Neurocomputing*, 633:129781, 2025. 3

[2] Ahmed Abdelkawy, Aly Farag, Islam Alkabbany, Asem Ali, Chris Foreman, Thomas Tretter, and Nicholas Hindy. Measuring student behavioral engagement using histogram of actions. *Pattern Recognition Letters*, 186:337–344, 2024. 2, 3, 5, 7

[3] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. Edusense: Practical classroom sensing at scale. *Proc. on Inter., Mobile, Wearable and Ubiquitous Tech.*, 3, 2019. 2, 3

[4] Valentina Rita Andolfi, Chiara Di Nuzzo, and Alessandro Antonietti. Opening the mind through the body: The effects of posture on creative processes. *Thinking Skills and Creativity*, 24:20–28, 2017. 1

[5] James J Appleton, Sandra L Christenson, and Michael J Furlong. Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, 45(5):369–386, 2008. 1

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 4

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, volume 2, page 4, 2021. 4

[8] Brandon M Booth, Nigel Bosch, and Sidney K D'Mello. Engagement detection and its applications in learning: a tutorial and selective review. *Proceedings of the IEEE*, 111(10):1398–1422, 2023. 1

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4, 7

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[11] Michelene TH Chi and Ruth Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4), 2014. 1, 2

[12] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1011–1030, 2023. 6

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 4, 7

[14] Aly A Farag, Asem Ali, Islam Alkabbany, James Christopher Foreman, Tom Tretter, Marci S DeCaro, and Nicholas Carl Hindy. Toward a quantitative engagement monitor for stem education. In *2021 ASEE Annual Conference Content Access*, 2021. 1

[15] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 4

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[17] Benjamin Filtjens, Bart Vanrumste, and Peter Slaets. Skeleton-based action segmentation with multi-stage spatial-temporal graph convolutional neural networks. *IEEE Transactions on Emerging Topics in Computing*, 12:202–212, 1 2024. 3

[18] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109, 2004. 1

[19] Minji Kim, Dongyoon Han, Taekyung Kim, and Bohyung Han. Leveraging temporal contextualization for video action recognition. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 6

[20] Erin S Lane and Sara E Harris. A new tool for measuring student behavioral engagement in large university classes. *Journal of College Science Teaching*, 44(6):83–91, 2015. 1, 5

[21] Wen Li, Fei Jiang, and Ruimin Shen. Sleep gesture detection in classroom monitor system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7640–7644. IEEE, 2019. 2

[22] Feng-Cheng Lin, Huu-Huy Ngo, Chyi-Ren Dow, Ka-Hou Lam, and Hung Linh Le. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. *Sensors*, 21(16):5314, 2021. 3

[23] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pages 1–18. Springer, 2022. 4, 6

[24] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6545–6554, 2023. 2, 4, 5, 6

[25] Amir Shahroudy, Jun Liu, Tian Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:1010–1019, 12 2016. 3

[26] Xiaojing Sheng, Suqiang Li, and Sixian Chan. Real-time classroom student behavior detection based on improved yolov8s. *Scientific Reports*, 15:1–11, 12 2025. 3

[27] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 3

[28] Ellen A Skinner and Michael J Belmont. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of educational psychology*, 85(4):571, 1993. 1

[29] Michelle K Smith, Francis HM Jones, Sarah L Gilbert, and Carl E Wieman. The classroom observation protocol for undergraduate stem (copus): A new instrument to characterize university stem classroom practices. *CBE—Life Sciences Education*, 12(4):618–627, 2013. 1

[30] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027, 2021. 1

[31] Bo Sun, Yong Wu, Kaijie Zhao, Jun He, Lejun Yu, Huanqing Yan, and Ao Luo. Student class behavior dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Comp. & App.*, 33, 2021. 2

[32] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 4, 7

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 4

[34] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3

[35] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 4

[36] Zefang Yu, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. From raw video to pedagogical insights: A unified framework for student behavior analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23241–23249, 2024. 2, 4

[37] Zefang Yu, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. From raw video to pedagogical insights: A unified framework for student behavior analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:23241–23249, 3 2024. 3

[38] Rui Zheng, Fei Jiang, and Ruimin Shen. Intelligent student behavior analysis system for real classrooms. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9244–9248. IEEE, 2020. 2

Context Matters: Peer-Aware Student Behavioral Engagement Measurement via VLM Action Parsing and LLM Sequence Classification

Supplementary Material

1. LLM Prompts

In this section, we present the LLM prompts used for:

- Context-free engagement classification (see Fig. 2).
- Context-aware engagement classification (see Fig. 4).

The context-free engagement classification uses only a student's action sequence as input and outputs the student's engagement level along with the reasoning behind that assessment. In contrast, context-aware engagement classification takes both a student's action sequence and the majority action sequence of their peers as input and outputs the student's engagement level together with the reasoning for that assessment.

Context-free Prompt

You are an educational expert. You are given a student's sequence of time-stamped actions. Your task is to classify the student behavioral engagement into one of the following categories: High engagement: the student is Focused and performs active learning behaviors (e.g., writing in a notebook/taplet, typing on a laptop, listening, etc.). Low engagement: the student is frequently distracted or disengaged, with only brief moments of on-task behavior. Instructions: 1-if multiple actions are present, weigh them by duration. 2-The action 'looking at laptop screen (not typing)' can be treated as a reading action if it is interspersed between engaged actions (e.g., writing, typing, listening attentively). 3- The longer-lasting behaviors should influence the classification more. Provide both the classification and a brief reasoning. Format the generated response in the form of a Python dictionary string with key 'Category' for engagement classification and 'reasoning' for the reasoning of your choice.

Figure 2. The prompt for context-free engagement classification.

2. MLLM Prompt

To assess the capability of advanced multimodal large language models, e.g., Gemini-2.5 Pro, in parsing a student's actions within a 2-minute video, we designed a prompt that instructs the MLLM to perform temporal action segmentation as shown in Fig. 6.

Context-aware Prompt

You are an expert educational behavior analyst specializing in classroom observation and learning analytics. You are given: Student_Sequence: a sequence of the student's time-stamped actions over a fixed time interval. Peers_Aggregate: the aggregate (majority) sequence of peers' actions over the same interval, representing classroom context. Your task is to classify the student's behavioral engagement level into one of the following categories:
High engagement: Predominantly learning-oriented behaviors, either active or passive (e.g., Writing on notebook/tablet, Typing for class-related work, Reading, Listening attentively, Asking or answering questions, Pointing to instructional material).
Low engagement: Predominantly distracted or disengaged behaviors, with only brief or sporadic moments of on-task activity (e.g., Playing with a mobile phone, off-task laptop use, repeated checking of personal items, extended inattentive behavior).
Decision Guidelines
1-Interpretation of ambiguous actions
-Treat "Looking at laptop screen (not typing)" as a reading action only if it is interspersed between clearly engaged actions (e.g., Writing, Typing, Listening attentively). Otherwise, consider it potentially off-task.
2- Primary evidence: student behavior
-Prioritize the student's own actions and their durations when determining engagement.
-Sustained on-task behavior outweighs brief off-task interruptions.
3-Secondary evidence: peer context
-Use Peers_Aggregate as supporting evidence, not the primary determinant.
-If the student's actions slightly align with the majority of the class (not necessarily at every time segment), this reinforces engagement.
-If the student's actions consistently diverge from peers' instructional activities, this may slightly lower engagement, but only insofar as it helps interpret ambiguous student actions.
4-Overall judgment
-Base the final label on the dominant pattern across the interval, not isolated moments.
-Consider whether the student's behavior plausibly supports learning given the classroom context.
Provide your response as a Python dictionary string with 'Category' and 'reasoning' keys. Do not repeat or return the input.

Figure 4. An example of the task description x_{desc}

Action Segmentation Prompt

You are an expert in video understanding. You are given a 2-minute-long video of a student in-class (or a sequence of frames). Your task is to perform temporal action segmentation as follows: 1- partition the video temporally into segments, 2- assign an action label to each segment, where the action label is one of the following Writing on notebook/tablet, Typing on a laptop, Playing with mobile phone, Reading, Raising hand, Drinking, Eating meal/snack, Yawning, Listening, Looking at laptop screen (not typing), Checking time, Looking to side/back, Looking down without reading/writing], and 3- ensure the segmentation is consistent over time. Format the output as a structured JSON list, where each entry includes: start time (in 'MM:SS' format), end time (in 'MM:SS' format), action label. Avoid overlapping intervals, and cover the full video duration

Figure 6. The prompt for temporal action segmentation.