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.