IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 May 2025, accepted 20 June 2025, date of publication 27 June 2025, date of current version 10 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3584025

RESEARCH ARTICLE

Enhancing Engineering and STEM Education With Vision and Multimodal Large Language Models to Predict Student Attention

LUIS MARQUEZ-CARPINTERO, DIEGO VIEJO, AND MIGUEL CAZORLA, (Senior Member, IEEE)
Institute for Computer Research, University of Alicante, 03080 Alicante, Spain

Corresponding author: Luis Marquez-Carpintero (luis.marquez@ua.es)

This work was supported in part by the Generalitat Valenciana (Spain), Department of Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital under Grant CIPROM/2021/17.

ABSTRACT Generative Artificial Intelligence (AI) and Large Language Models (LLMs), including Visual Language Models (VLMs) and Multimodal LLMs (MLLMs), have shown transformative potential in education. These technologies address persistent challenges in fostering classroom engagement and interaction. Our study highlights the efficacy of these models in detecting students' attention levels and emotional states, equipping educators with actionable insights to optimize instructional delivery. However, widespread adoption is hindered by significant barriers such as high computational demands and the limited availability of high-quality datasets. To overcome these challenges, this research proposes the integration of MLLMs with Few-Shot Learning techniques, offering a resource-efficient framework to enable their practical implementation in educational contexts. This study focuses on the application of VLMs and MLLMs to predict student attention in science, technology, engineering and mathematics (STEM) education, evaluating the effectiveness of Few-Shot Training compared to traditional AI methodologies. The research is structured into two phases: the first phase optimizes image frequency and computational costs using MLLMs, while the second phase trains VLMs on classroom data to identify visual cues, including gaze direction and head movement. The results demonstrate that VLMs combined with Few-Shot Learning significantly outperform traditional models in capturing nuanced visual data, allowing for pedagogical adjustments comparable to those made through human labeling. These findings underline the transformative potential of VLMs and MLLMs in education, particularly in resource-constrained environments. Few-Shot Learning emerges as a practical and effective approach for leveraging small datasets to enhance student engagement and instructional quality.

INDEX TERMS Attention prediction, engineering education, few-shot learning, large language models, student engagement.

I. INTRODUCTION

The integration of generative artificial intelligence (AI) and large language models (LLMs) into higher education has the potential to significantly enhance LLMs, such as GPT-4 and LLAMA to generate natural language, provide instant feedback, answer complex questions, and grade assignments, improving personalized learning [1].

Recently, LLMs, VLMs, and MLLMs have garnered significant research interest [2]. However, their application in education remains limited due to challenges in accessing comprehensive multimodal datasets [3]. A recent dataset, collected in classroom settings during STEM and engineering experiments, marks a significant advance, integrating emotional and attentional signals with body sensor data. This integration facilitates the effective training of VLMs and MLLMs, enhancing their potential for educational applications [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

Engineering classes differ from others in their focus on collaboration and the use of advanced technologies. These activities integrate practical elements and teamwork, fostering active learning and problem-solving in authentic engineering contexts. Unlike non-engineering classes, students engage in significant physical movement, highlighting the importance of biometric data for precise monitoring and analysis.

VLMs enhance the understanding of student engagement by analyzing images and videos, integrating visual and textual data with the knowledge base of LLMs. In contrast, MLLMs focus on analyzing multimodal data based on textual information and data from various modalities. Despite their potential, their application in education remains underexplored, primarily due to the current lack of multimodal data.

Among these techniques, Few-Shot Learning (FSL) and Zero-Shot Learning (ZSL) have emerged as innovative approaches [5], offering advantages over traditional machine learning (ML) models as they require fewer data to produce reliable predictive models. FSL, unlike ZSL, leverages labeled images as context. Despite these advances, challenges such as computational limitations, dataset diversity, and ethical considerations (e.g., bias, fairness, and student privacy) persist, warranting further exploration. Future research must prioritize efficient data collection and investigate the synergy between VLMs and MLLMs to address these issues.

Despite the progress made, the potential of multimodal LLMs (MLLMs) and VLMs to improve attention prediction in students through the analysis of visual and textual data in classroom engineering activities has not been sufficiently investigated. The integration of these technologies could offer valuable insights for educators in real-world classroom settings, where real-time feedback on student engagement is crucial for optimizing instructional strategies and fostering a more interactive learning environment. Implementing these models could help tailor educational content to individual needs, even in resource-constrained environments, making it a promising solution for diverse educational institutions.

The use of MLLMs and VLMs in STEM educational environments, where students frequently engage in experiments and hands-on activities, has proven to be particularly promising area. These settings present unique challenges for traditional AI models, as movement, student collaboration, and real-time experiments generate a large amount of visual and behavioral signals. VLMs are particularly effective in capturing and analyzing these signals, enabling more accurate and timely feedback in dynamic educational environments [2]. This article addresses the following key questions:

- What are the predictive capabilities of LLMs, including VLMs and MLLMs, in enhancing student engagement and learning outcomes in STEM and engineering education
- Can conventional Neural Network (CNN) models effectively meet the complex demands of engineering education, and how do they compare to the new LLM architectures that have emerged in recent years

To facilitate comparison and establish a benchmark against traditional methods, we propose using classification methods with CNNs, employing a specific configuration of hyperparameters for optimal performance.

II. LITERATURE REVIEW

In recent years, agent systems, particularly LLMs, have shown great potential in the educational field, especially in detecting student participation and engagement through text analysis. These models use advanced algorithms to process data and generate tailored solutions, such as adaptive learning strategies and personalized feedback, thereby enhancing the educational experience [1]. However, most research has hitherto focused on processing and analyzing textual data, neglecting multimodal analysis, which includes critical visual components for comprehensive assessment of student attention.

One of the most significant limitations of traditional LLMs is their inability to analyze non-verbal signals, including image-based information such as facial expressions, gestures, or postures, which are essential for understanding the degree of engagement in an educational environment. In this context, VLMs and MLLMs have emerged as innovative solutions that integrate visual and other data that can be used to improve the analysis of students' attention and emotions [2], [4].

The application of VLMs and MLLMs in the academic field is still nascent. These models have proven particularly useful in capturing non-verbal cues that text-based models alone cannot detect, such as eye movements, body posture, and gestures [3], [5]. However, the use of these models has been limited by the availability of suitable visual datasets and the high computational demands associated with their large-scale training and use. Despite these challenges, recent studies have highlighted the importance of developing richer and more representative datasets that include not only RGB images but also biometric and depth signals to improve the accuracy of attention prediction [6].

In addition to dataset diversity, the effectiveness of attention and emotion recognition systems critically depends on the strategies employed for feature extraction. Traditional approaches in virtual learning environments often rely on automated facial expression analysis tools that generate probability vectors corresponding to basic emotions (e.g., happiness, sadness, anger, fear, surprise, disgust, and contempt). These features are periodically captured through webcams and statistically normalized to meet assumptions for multivariate analyzes such as MANOVA or ANOVA [7].

Classic facial detection methods, such as Haar Cascades, are commonly used to locate regions of interest, primarily eyes and mouth, before applying edge detection techniques such as the Sobel operator. The extracted features are then fed into neural classifiers trained on standardized facial expression datasets, with a focus on balancing recognition accuracy and computational efficiency in real-time systems [8].

More advanced approaches incorporate temporal dynamics by analyzing transitions in facial landmarks across frames.

By tracking key points—such as eyes, eyebrows, nose, and mouth—these methods compute geometric descriptors (e.g., inter-landmark distances and angles) over sliding temporal windows. Feature selection techniques, including information Gain or Chi-square, are then applied before using classifiers, such as SVMs to detect both discrete emotional states and transitions with pedagogical significance [9]. Together, these methodologies prioritize robust facial detection, temporal modeling, and feature discrimination, forming the foundation for adaptive emotion, and attention, aware learning systems.

Studies have predominantly approached student attention prediction using traditional ML techniques and neural networks with straightforward architectures. However, these methods often present limitations due to the lack of diversity in the datasets, which affects their ability to generalize to different educational contexts. Attention is defined as a cognitive process in which an individual focuses their perception, resources, and capacities on a specific stimulus or set of stimuli, filtering out irrelevant information. This process is essential for learning, memory, and decision-making as it enable the selective processing of relevant information and responses, making it particularly significant in educational contexts [10].

In the educational domain, attention is recognized as a prerequisite for effective learning. For example, Bloom's taxonomy of educational objectives includes, within the affective domain, the level of ''receiving'' which involves the willingness to pay attention and be receptive to educational stimuli [11]. In other words, before a student can reach higher levels of learning, they must first attend to the presented material. From the perspective of cognitive psychology, Posner and Cohen [20] proposed influential models on the structure of attention. Posner distinguished attentional components such as alertness (a state of vigilance), orienting (directing focus toward a stimulus), and executive attention (voluntary control of attention) [12]. Posner's tripartite model has served as the basis for the development of cognitive assessments, such as the Attention Network Test, and has influenced our understanding of sustained and selective attention in learning contexts. Other theorists have categorized attention into subtypes such as selective, sustained, or divided attention, emphasizing the importance of maintaining focus on key educational tasks and avoiding distractions [10]. These core concepts underline that a student's ability to select and maintain attention on important information is essential for processing it in memory and constructing knowledge.

When applying these models to a new dataset, such as that used in this study, the results obtained tend to be inferior to those evaluated on the original datasets, possibly due to overfitting to specific educational contexts. For this reason, multimodal approaches, such as VLMs, offer a significant improvement over traditional models by integrating different data sources and enhancing generalization capacity [2].

Advances in FSL have opened up new possibilities for implementing models such as VLMs and MLLMs in educational settings. FSL refers to the ability of a model to generalize from a very small number of examples, which is particularly valuable in contexts where data is limited or costly to obtain. This technique has been shown to reduce computational costs and improve the accuracy of attention prediction, even in contexts with limited datasets [5].

In the present study, Vision–Language Models such as LLaVA v1.6 (built on the Hermes-Yi-34B LLM) were selected over alternatives like Flamingo [13] due to LLaVA's superior generalization capability, which does not require task-specific adjustments. Generative AI methods—including LLaVA [14], DALL·E [15] and CLIP [16]—efficiently integrate visual and textual information, enabling multimodal analysis even under constrained computational budgets.

The incorporation of visual data into these models confers enhanced flexibility and adaptability across a range of educational contexts, including the capacity to predict student attention and to analyze emotions in real-time classroom settings [17]. Their ability to generalize from a limited number of examples, combined with their computational efficiency, makes them particularly suited to dynamic educational environments, such as STEM and engineering classrooms, where diverse visual and behavioral cues are constantly present. In these scenarios, VLMs capture subtle nuances more effectively than traditional approaches, offering more accurate insights. Consequently, the use of these models not only optimizes the allocation of resources but also enhances the quality of personalized instruction and student engagement, without compromising prediction accuracy [18].

Our decision to utilize large-scale language and vision assistants, such as LLaVA, is based on their demonstrated ability to effectively integrate multimodal data and provide precise, context-aware feedback. This capability is particularly critical in educational settings with constrained resources, where maximizing efficiency and accuracy is essential [1].

Moreover, emerging research underscores the potential of integrating physiological and biometric data alongside traditional visual inputs to enhance VLM and MLLM capabilities. For instance, real-time tracking of heart rate variability or other biometric values obtained through smartwatches can offer additional insights into cognitive load and emotional states [19]. This multimodal integration creates opportunities to fine-tune educational interventions by dynamically adjusting content delivery and engagement strategies based on nuanced interpretations of student states.

III. METHODOLOGY

The methodology adopted in this study follows a two-phase approach designed to evaluate the effectiveness of VLMs and MLLMs in predicting student emotion and attention in educational settings. The primary goal was to determine the optimal data frequencies and intervals necessary to maximize model performance while balancing prediction accuracy with computational costs.

First, we analyzed and selected different types of data relevant for evaluating student attention, specifically RGB images and biometric values, all sourced from the DIPSEER dataset [4]. Second, we implemented and assessed VLMs and MLLMs using ZSL and FSL methods. The goal was to determine the least amount of information needed for accurate predictions. These models were chosen for their advanced decision-making and contextual understanding, which allows them to mimic expert-level insights. We then directly compared the results from these deep learning models with manual labeling done by experts to evaluate the different approaches.

The choice of Few-Shot Learning FSL over traditional supervised learning was driven by the limited availability of labeled datasets in educational contexts. Unlike fully supervised models, which require extensive labeled data, FSL enables robust performance with minimal examples per class. Additionally, ZSL facilitates generalization to unseen data without retraining, making it suitable for dynamic classroom environments.

This study highlights the ability of VLMs and MLLMs to integrate multimodal data and provide interpretability in their predictions. Their performance was compared with standard AI models to assess the relative strengths and limitations of both approaches. This comprehensive evaluation provides a detailed understanding of model efficacy across various machine learning methodologies. The following sections outline the dataset used, the procedures for collecting and processing visual and biometric data, and the necessary hyperparameters for the classic neural network.

For all experiments, LLaVA was instantiated from the 'liuhaotian/llava-v1.6-34b' checkpoint, with 8-bit quantization enabled, a temperature of 0 (deterministic, greedy decoding), and a maximum generation length of 512 new tokens.

A. DATASET DESCRIPTION

The DIPSEER dataset (*Dataset for In-Person Student Engagement Recognition*) was specifically designed to analyze student engagement and emotional responses in face-to-face learning environments, contrasting with most prior datasets, which primarily focus on virtual classrooms or controlled simulations. The dataset comprises a combination of RGB images and biometric data collected through smartwatch sensors, alongside a classification of experiments based on classroom activities. The participants in the dataset are pre-service teachers demonstrating their skills across various experiments specifically designed for this purpose. This dataset provides an ideal foundation for multimodal analysis using LLMs within the field of engineering, owing to its Experiments 8 and 9.

The dataset categorizes nine distinct types of experiments, numbered from 1 to 9, according to the activities performed by students.

For this study, only scenarios 8 and 9 were selected for analysis. These two scenarios were chosen because they represent high-engagement, hands-on tasks that reflect real-world engineering and STEM education dynamics.

Other scenarios, such as lectures, brainstorming, or reading activities, involve lower physical activity and limited variability in visual and biometric signals, which reduces their relevance for models designed to detect subtle behavioral cues of attention and emotion.

Scenarios 8 and 9, on the other hand, provide a richer set of multimodal signals, including facial expressions, head and body movements and biometric data from wearable sensors.

These elements make them the most suitable for evaluating the models in dynamic and interactive educational settings, aligning with the research objectives of improving student attention and engagement prediction.

1) RGB DATA

Individual images of each student were collected using a personal camera that recorded their posture and facial expressions. The images had a resolution of 640 × 480 pixels, allowing for detailed analysis of both facial micro-expressions and gestures. The images used for the analysis were one frame every two seconds, which optimizes computational capacity while maintaining fluidity of movement and avoiding significant variations between frames.

One of the challenges encountered in this type of experiment is the students' high level of mobility during manual activities, which often causes them to move out of the camera's field of view. Therefore, it is crucial to ensure continuous capture, minimizing significant variations between images.

2) BIOMETRIC DATA FROM THE SMARTWATCH

Each student in the dataset wore a smartwatch that collected biometric data in real time, which is necessary for analysis with the MLLM models. The following biometric values were used:

- **Heart Rate Sensor**: Measures heart rate in beats per minute (BPM), providing a key indicator of emotional response and physical activity.
- **Accelerometer**: Records linear acceleration along the X, Y, and Z axes at 100 samples per second, allowing for the analysis of body movement.
- **Gyroscope**: Measures angular rotation along the X, Y, and Z axes at 100 samples per second.
- **Wake-Up Sensor**: Provides a binary or threshold-based activity signal, indicating whether the user is actively moving or idle, helping to contextualize body movement and engagement.

In this study, the data selected for analysis included specific visual and biometric inputs. From the RGB data, individual frames captured every two seconds with different resolutions were used, focusing on facial expressions, gaze direction, head posture, and facial landmarks. Regarding biometric data, the selected signals were heart rate (as an indicator of emotional arousal), accelerometer and gyroscope data (to capture body movement and rotation), and wake-up

sensor values (to provide contextual activity information). These data were temporally synchronized with the annotated attention and emotion labels and served as inputs for the VLM and MLLM models evaluated in this work.

3) PROCESSED BASIC DATA

This dataset also includes post-processed data that was been used in the MLLM analysis, described in Section IV. Among the available data are values processed by different models, such as a facial mesh of the student's face, body and head bounding boxes, body landmarks, as well as basic inferred information, including gender, age, and ethnicity. These data facilitate further research and allow other researchers to access already processed results without the need to repeat the initial processing steps.

4) ATTENTION AND EMOTION LABELS

DIPSEER provides attention and emotion labels at one-second intervals, which are then propagated to all frames within each second to ensure consistency. Six annotators applied one label per second rather than per individual frame.

The annotators are divided into two groups:

- **Expert evaluators:** Five experts reviewed the videos, assigning attention scores on a 1–5 scale (1 minimal, 5 maximal) and classifying emotions into nine categories (enjoyment, hope, pride, relief, anger, anxiety, shame, despair, boredom).
- **Self-assessment:** Students labeled their own attention and emotions after each class session, using the same scale and the same nine emotion categories.

A simple majority vote determined the final label for each one-second interval. Once a consensus label was reached, it was applied to all subsequent frames until a new majority emerged and, for the initial segment of the video, to preceding frames.

These consensus labels serve as the ground truth for model evaluation. Model predictions are compared directly against them using weighted accuracy, F1 score, MAE and MSE, ensuring a fair and consistent assessment.

Although human annotations directly influence the final labels, inter-annotator agreement was moderate: emotion F1 scores remained below 0.5 (Table 1), and attention agreement was around 0.6(Table 2). This variability highlights the subjective nature of the task and supports the use of majority voting as a robust baseline.

TABLE 1. Comparison of F1 score and accuracy of six human labelers, including self-assessments, for student emotion classification across all experimental frames. Lower scores indicate higher labeling disagreement.

Labeler	Weighted Accuracy	F1 Score
Labeler2	0.176	0.259
Labeler5	0.270	0.330
Labeler3	0.360	0.380
Labeler4	0.400	0.450
Labeler1	0.460	0.480
Self Labeler	0.540	0.610

TABLE 2. Comparison of weighted precision, Mean Squared Error (MSE), Mean Absolute Error (MAE), F1 score, and Cohen's kappa for six human labelers, including self-assessments, in predicting student attention levels.

Labeler	Weighted Precision	Weighted Mean MSE	Weighted Mean MAE	F1 Score
Self Labeler	0.611	0.803	0.530	0.610
Labeler4	0.599	1.629	0.856	0.506
Labeler5	0.539	0.982	0.605	0.581
Labeler1	0.496	0.970	0.614	0.548
Labeler2	0.412	0.731	0.553	0.509
Labeler3	0.405	1.305	0.785	0.414

5) EDUCATIONAL SCENARIOS

The dataset covers nine educational scenarios, each designed to simulate different learning activities. These include reading sessions, robotics testing, educational design experiments, and traditional classes. Each scenarios has a fixed duration of 5 minutes, allowing for the analysis of student interactions and behavior in real teaching contexts.

For this study, only Scenarios 8 and 9 were selected, as they reflect academic dynamics typical of engineering classes:

- **Scenario 1: News Reading** - Students read news articles either projected on a screen or on their personal devices, focusing on content that will be assessed later.
- **Session 2: Brainstorming Session** - During this creative session, students generate ideas for projects or solutions to problems.
- **Scenario 3: Lecture** - A traditional teaching format where the instructor delivers a lecture in front of the class, with minimal to no student interaction.
- **Scenario 4: Information Organization** - Students organize and synthesize information gathered from various sources.
- **Scenario 5: Lecture Test** - A formal assessment concerning the content of Session 3, administered via mobile devices.
- **Scenario 6: Individual Presentation of Work** - Randomly selected students present their projects to the group.
- **Session 7: Knowledge Test** - A formal written assessment on a specific subject area, conducted using Kahoot, to evaluate collective knowledge on the material covered in Session 1.
- **Scenario 8: Robotics Experimentation** - A practical session in which students apply robotics technology to problem-solving, emphasizing computational thinking.
- **Scenario 9: MTINY Activity Design** - Students design and plan an educational activity using the MTINY educational tool, integrating computational thinking principles.

The dataset includes detailed attention and emotion labels, synchronized with sensor signals from smartwatches, together with expert manual labeling. This combination of visual and biometric data is particularly suitable for research

on attention and emotion analysis in educational settings. The images from scenarios 8 and 9 selected for analysis in VLMs and MLLMs include a total of 14,425 analyzed images. Due to computational limitations, frames were sampled every two seconds over five-minute videos, covering 57 subjects, whose demographic characteristics are detailed in Table 3. This table summarizes key information such as gender (male and female distribution), age ranges, and inferred ethnicity categories. Providing this demographic breakdown helps contextualize the dataset, supports the analysis of potential biases, and improves the generalizability of the findings across diverse student populations.

TABLE 3. Demographic distribution of participants in the DIPSEER dataset.

Group	Gender Distribution	Ethnic Composition	Age Distribution (Years Old)
Group 1 (n = 20)	63.6% Female 36.4% Male	86.4% White 13.6% Middle Eastern	19 (20.7%) 20 (13.8%) 21 (10.3%) 22 (13.8%) 23 (10.3%) 24 (6.9%) 25 (10.3%) 27 (6.9%) 31 (6.9%)
Group 2 (n = 21)	69.6% Female 30.4% Male	79.2% White 12.5% Latino/Hispanic 8.3% Asian	16 (13.8%) 18 (10.3%) 19 (20.7%) 20 (13.8%) 21 (17.2%) 23 (6.9%) 24 (10.3%) 26 (6.9%)
Group 3 (n = 16)	66.7% Female 33.3% Male	88.9% White 11.1% Latino/Hispanic	20 (11.5%) 21 (15.4%) 22 (7.7%) 24 (11.5%) 25 (7.7%) 26 (7.7%) 27 (7.7%) 29 (11.5%) 32 (11.5%) 43 (7.7%)

It is worth noting that the data used in this study were collected from a real classroom with no demographic or behavioral balancing. While this decision preserves the ecological validity of the collected data, it may also introduce potential biases related to participant distribution or classroom dynamics. Future research should address this limitation by incorporating more diverse and balanced datasets.

B. PROMPT STRATEGY EMPLOYED

The prompt strategy employed in this study leverages the capabilities of MLLMs and VLMs to efficiently perform complex tasks, such as detecting emotions and attention levels in students. Several prompting approaches were designed and evaluated to enhance the performance and accuracy of predictions within these models. The final prompts that yielded the best results for each specific scenario are detailed in Appendix. Extensive testing was conducted using various prompt configurations [20], [21] and subtle variations applied to one image per second extracted from selected experiments. This included the implementation of ZSL and FSL strategies.

The prompt configurations were assessed based on the quality of the predictions generated, with outputs optimized according to the characteristics of the input data and the visual context.

C. HYPERPARAMETERS OF CLASSICAL NEURAL NETWORKS

For this analysis, a CNN model with hyperparameters that have been thoroughly explored and have performed robustly in similar tasks was utilized [22], [23], [24].

In trials with this CNN model, the hyperparameters yielding the best results were as follows: 48 kernels of size 3×3 with strides of 2×2, a pooling layer with a 2×2 pool size, and a fully connected dense layer containing 50 units, with a batch size of 20 used for training.

The choice of reduced image size is due to the lower complexity of classic models like a CNN, which contrasts with the more complex LLM-based models; thus, larger images may not yield additional precision or performance benefits. This approach allows for greater hyperparameter exploration, optimizing model tuning.

IV. RESULTS

Our analysis classifies the data into three principal categories: ZSL, FSL, and MLLMs. Each approach offers distinct perspectives on model performance, catering to different levels of data exposure and adaptability. This segmentation underscores the varying strengths of each method in handling diverse data contexts, from minimal exposure in ZSL to contextual adaptability in FSL and MLLMs.

Under the ZSL framework, which capitalizes on the model's inherent ability to predict outcomes without exposure to specific examples, experiments adhere to the following configurations:

- **Reduction of image resolution.** The original resolution is tested first, followed by resolutions of 128 and 64 pixels in width, maintaining the aspect ratio.
- **Conversion of the attention label range to a higher scale.** The output format is requested on a scale from 1 to 10, 1 to 50 and 1 to 100, instead of the original 1 to 5 scale. The result is then converted back to one of the initial five classes based on proximity, using the methods of rounding down, rounding up, or rounding to the nearest value.
- **Cropping frames to the face bounding box.** Frames are cropped to the face bounding box with varying margins added to the subject's bounding box: 0, 20, and 40 pixels.
- **Inclusion of historical data from previous moments leading up to the LLM.** Frames from earlier moments are selected to provide context for the processed frame. These prior frames are combined into a single image arranged in either horizontal or vertical format, as shown in Figure 1.

The results from the four experiments conducted in the MLLM are shown in Table 4 for the attention cases and in

FIGURE 1. Example of a horizontally concatenated image sequence using two consecutive frames to provide temporal context for attention prediction using Zero-Shot Learning. The black frame is added for visualization purposes to highlight the margin of the image that is fed to the language model.

TABLE 4. Performance of MLLMs using different input combinations: smartwatch data, basic visual information, and RGB images. Evaluated by weighted accuracy and F1 score.

Labeler	Weighted Accuracy	F1 Score
MLLM: Watch, Basic and RGB Info	0.249	0.216
MLLM: Basic Info	0.252	0.216
MLLM: RGB Info	0.247	0.220
MLLM: Watch Info	0.253	0.225

TABLE 5. Performance of MLLMs in student emotion classification using smartwatch, visual, and RGB data, reported by F1 score and accuracy.

Modelo Accuracy	Weighted F1 Score	
MLLM: Watch Info	0.256	0.266
MLLM: Basic Info	0.248	0.259
MLLM: Watch, Basic and RGB Info	0.238	0.244
MLLM: RGB Info	0.230	0.245

Table 5 for the emotion results. The first experiment analyzes all available information, including demographic data from the profile (age, gender, and emotion, inferred by another model), biometric data from the smartwatch (wake-up sensor and rotation sector), and inferred information from the RGB camera (head rotation, eye-opening area, mouth-opening area). All of this is combined with the visual information already available from the RGB images.

The FSL approach employs an inter-subject testing strategy, where the same subject is consistently used across the entire dataset. This enables a more controlled analysis of model performance. The FSL Results for attention, shown in Table 6, provide insights into the model's generalization capacity, while Table 8 presents its effectiveness in capturing emotional nuances.

TABLE 6. Weighted accuracy, Mean Squared Error (MSE), and Mean Absolute Error (MAE) of Few-Shot Learning (FSL) models using intra- and inter-subject configurations with 8 and 16 example shots.

Configuration for Attention Labels	Weighted Accuracy	Weighted MSE	Weighted MAE
FSL: 16 Shots. Inter-Subject. (All experiments)	0.116	9.852	2.823
FSL: 16 Shots. Intra-Subject. (scenarios 08 and 09)	0.386	5.760	1.753
FSL: 08 Shots. Intra-Subject. (scenarios 08 and 09)	0.413	4.983	1.635

TABLE 7. Summary of best-performing models and configurations for student attention and emotion prediction, compared with human labeler' performance.

Configuration for Attention Labels	Weighted Accuracy	Configuration for Emotion Labels	Weighted Accuracy
Self-Labeler	0.611	Self-Labeler	0.540
Labeler 4	0.599	Labeler 1	0.460
Labeler 5	0.539	Labeler 4	0.400
Labeler 1	0.496	Labeler 3	0.360
Labeler 2	0.420	ZSL: Horizontal History (2 Frames)	0.321
FSL: 08 Shots, Intra-Subject (scenarios 08 and 09)	0.413	FSL: 16 Shots, Intra-Subject (scenarios 08 and 09)	0.317
Labeler 3	0.4053	ZSL: Horizontal History (5 Frames)	0.314
FSL: 16 Shots, Intra-Subject (scenarios 08 and 09)	0.386	ZSL: Face Crop (0px Margin)	0.287
CNN: Inter-Sample (scenarios 08 and 09)	0.325	ZSL: Vertical History (5 Frames)	0.287

For the ZSL approach, the attention an emotion results are shown in Table 9 and in Table 10, respectively. Only configurations yielding outcomes superior to chance are included, highlighting reliable and robust results.

The best-performing model configuration was selected for further testing consistently achieving reliable results, even with only eight shots. This level of efficiency demonstrates the model's applicability in educational settings, underscoring the value of FSL in balancing predictive accuracy and computational efficiency.

TABLE 8. F1 score and weighted accuracy of FSL models using 16 example shots in intra-subject configuration for student emotion classification.

Configuration for Emotion Labels	Weighted Accuracy	F1 Score
FSL: 16 Shots. Intra-Subject. (scenarios 08 and 09)	0.317	0.374

Metrics used to evaluate the performance of VLMs include measures of weighted accuracy and MSE. The selection of these metrics is critical because they offer complementary insights into different aspects of model evaluation. Weighted accuracy metrics provide insight into the overall correctness of the model by taking into account the varying importance of different classes or outcomes. This provides a nuanced picture of performance, especially in cases where certain tasks or classes are more critical than others. On the other hand, the weighted MSE is crucial for tackling the problem as a regression problem with continuous predictions, as it quantifies the root mean square difference between predicted and actual values, making it especially useful for identifying model accuracy and bias in regression-based tasks. Taking together, these metrics provide a comprehensive evaluation framework, facilitating a more robust assessment of the model's ability to generalize across a variety of tasks and conditions, while gauging its computational efficiency.

The most commonly used metric for these types of problems in state-of-the-art models is the accuracy, which evaluates the model's classification performance by weighting across each class. Another approach some researchers

TABLE 9. Weighted accuracy, Mean Squared Error (MSE), Mean Absolute Error (MAE), and F1 Score for ZSL models under various configurations including image scaling, face cropping, and historical context inclusion.

Method	Weighted Accuracy	Weighted MSE	Weighted MAE	F1 Score
ZSL: Over 10 Ceil	0.289	2.056	1.115	0.270
ZSL: Over 10 Round	0.230	2.719	1.261	0.250
ZSL: Over 50 Ceil	0.279	1.689	1.008	0.232
ZSL: Over 50 Round	0.289	1.694	1.010	0.231
ZSL: Over 50 Floor	0.240	1.694	1.010	0.231
ZSL: Over 10 Ceil (Width 128px)	0.301	2.293	1.206	0.230
ZSL: Over 100 Ceil	0.230	1.519	0.938	0.220
ZSL: Over 100 Round	0.223	1.519	0.938	0.220
ZSL: Over 100 Floor	0.222	1.519	0.938	0.220
ZSL: Over 10 Round (Width 128px)	0.202	3.167	1.390	0.220
ZSL: Over 10 Ceil (Width 64px)	0.230	2.340	1.221	0.220
ZSL: Face Crop (40px Margin)	0.253	0.984	0.769	0.207
ZSL: Face Crop (20px Margin)	0.250	0.997	0.774	0.202
ZSL: Width 128px	0.262	0.860	0.725	0.200
ZSL: Width 64px	0.262	0.860	0.725	0.200
ZSL: No modifications	0.260	0.849	0.724	0.200
ZSL: Face Crop (0px Margin)	0.245	0.999	0.777	0.189
ZSL: Vertical History (2 Frames)	0.250	0.731	0.697	0.156
ZSL: Over 10 Floor	0.226	4.130	1.704	0.140
ZSL: Over 10 Floor (Width 64px)	0.196	5.042	1.946	0.110
ZSL: Over 10 Floor (Width 128px)	0.198	4.856	1.904	0.110

TABLE 10. F1 score and weighted accuracy of ZSL models using historical image sequences and face cropping techniques to classify student emotions.

Configuration for Emotion Labels	Weighted Accuracy	F1 Score
ZSL: Horizontal History (2 Frames)	0.321	0.331
ZSL: Horizontal History (5 Frames)	0.314	0.331
ZSL: Face Crop. Margin 0px	0.287	0.291
ZSL: Vertical History (5 Frames)	0.287	0.300
ZSL: Vertical History (2 Frames)	0.276	0.272
ZSL: No modifications	0.275	0.270
ZSL: Width 128px	0.253	0.236
ZSL: Width 64px	0.243	0.222
ZSL: Width 32px	0.194	0.203
ZSL: Horizontal History (2 Frames and 1 Frame in the Future)	0.148	0.051

adopt is the use of MAE and MSE, treating the attention problem as a regression task that penalizes the difference between decisions. These metrics need to be adjusted for class imbalance in the dataset, as Class 3 is the majority class, meaning a model that classifies all frames as belonging to Class 3 would outperform other approaches.

A particular emphasis was placed on the analysis of the Mean Squared Error (Weighted MSE), a metric that quantifies the magnitude of the errors produced by each method, with greater weight assigned to larger deviations. For the purposes of this analysis, solely the model configurations that demonstrated a performance level that exceeded chance in the emotion-labelling task, as determined by Cohen's Kappa [25], were considered.

In the field of ZSL, configurations that utilize the full image size without prior context or rescaling yield better results than Labelers 1, 3, 4, and 5, even when the image size is reduced, without significantly affecting the MSE (see Table 7).

When this is treated as a classification problem and weighted accuracy is used, as is common in the literature, it is observed that even a slight reduction in resolution can improve the results, with a significant enhancement when scaling by 10 with a ceil rounding. However, ZSL alone is not enough to match human labelers' performance 9.

Despite their widespread use, the performance of CNNs applied to this task falls considerably short compared to the results achieved by VLMs, obtaining a balanced accuracy of 0.259 for emotion and 0.320 for attention. These results are marginal in comparison to those of models derived from LLMs.

When addressing the problem using FSL, performance reaches that of human labelers, provided that example images of the same subject from the same experiment type are available. In the attention results, the best performance when using up to eight samples per class is likely because using more samples leads to poorer generalization across subjects. This approach outperforms Labeler 3. However, the MLLM approach makes no contribute significant improvements in regression or classification tasks in the attention field.

In terms of emotion detection, using ZSL the F1 Score and Accuracy metrics, as shown in Table 10, reveals that providing a horizontal history of previously processed images and applying facial cropping significantly improves the baseline performance of ZSL, even surpassing Human Labelers 5 and 2 7. Meanwhile, contrary to expectations, incorporating biometric data into the model provides no substantial benefits and, in some cases, results in poorer performance compared to other approaches. Furthermore, in specific instances, the LLM-based model failed to generate a response or clarify the solution, with these cases being classified as direct errors in the analysis.

V. DISCUSSION

One of the most significant findings from the data presented are that models based on LLMs outperform traditional AI models in ZSL experiments. These models achieve performance levels comparable to human labelers in terms of weighted MSE and weighted accuracy. This outcome is particularly striking given that the final labels were determined by human labelers through a simple majority voting system.

Notably, Labelers 1, 2, 4, and 5 demonstrate superior performance in attention tasks compared to certain ZSL model configurations. A key observation is that reducing image resolution to 128 or 64 pixels does not significantly impact model performance and, in some cases, even improves it. This robustness is evident in metrics such as weighted

accuracy and weighted MSE, suggesting the models are robust to variations in input image resolution.

In terms of precision, FSL models outperform human labelers, achieving their best performance when utilizing up to eight samples per class in the same subject and experiment. This approach enables the model to surpass Labeler 3 in weighted accuracy, highlighting the importance of incorporating subject-specific examples. These results were observed in inter-subject contexts, using images from engineering experiments as examples.

The classic CNN model yielded suboptimal results, likely due to insufficient refinement of input image data (e.g., cropping and aligning facial features) and limited experimentation with alternative model configurations. This limitation highlights an advantage of vision-language models (VLMs), which can achieve effective performance without extensive preprocessing of input data.

In the domain of emotion detection, ZSL models equipped with a historical context of prior images and facial crops outperformed human labelers, including Labelers 5 and 2. These findings suggest that incorporating a historical context significantly enhances the model's ability to identify emotional patterns more accurately compared to processing individual images. The inclusion of prior images proved critical for improving performance, particularly in metrics such as the F1 score and precision. Furthermore, transitioning from ZSL approaches to FSL ones, such as the ''ZSL: History Horizontally 2 Frame'' configuration, demonstrated enhanced results.

For attention-related tasks, the FSL approach closely approximated the performance of Labeler 2, substantially outperforming Labeler 3 and demonstrating significant improvements over the traditional CNN. These outcomes are detailed in Table 8.

Although the dataset comprised only 57 students, which may introduce selection bias, further studies should investigate larger and more diverse samples to enhance the generalizability of the findings. Expanding the dataset would provide a more comprehensive understanding of model performance across different demographics and learning environments.

Overall, FSL methods without image resolution reduction or cropping consistently outperform human labelers with lower accuracy levels. While ZSL does not yet achieve human-level accuracy in attention classification, it surpasses the performance of traditional learning methods. In cases where the VLM failed to generate a result for specific images, these instances were attributed to errors in capturing the image's specific value rather than systemic flaws in the model's design, indicating room for further optimization.

VI. LIMITATIONS

Notwithstanding the encouraging results obtained in this study, several limitations were identified that may have ramifications for the practical deployment and generalizability of the proposed models.

Firstly, the models demonstrated performance degradation in uncontrolled classroom scenarios, particularly under suboptimal lighting conditions, partial occlusions, or when students moved outside the camera's field of view. These limitations indicate the reliance of the models on high-quality visual inputs, giving rise to concerns regarding their robustness in real-world educational environments.

Secondly, the integration of biometric data from wearable devices, such as smartwatches, did not result in consistent performance enhancements. This underscores the necessity for more rigorous validation and contextualisation of biometric signals prior to their incorporation into multimodal learning models.

Thirdly, the models demonstrated a propensity to favour the majority class, thereby constraining their sensitivity to detect less frequent yet pedagogically salient states, such as extreme engagement or disengagement. This imbalance in class composition poses a risk of minority behaviours being overlooked, which may be critical for instructional adaptation.

Furthermore, the dataset utilised in this study exhibited a high degree of demographic homogeneity, characterised by a paucity of representation with respect to age, cultural background, and ethnicity. This restricts the external validity of the findings and may impact the model's effectiveness when applied to more diverse educational settings.

Another limitation that was identified was the occurrence of silent failures in certain Zero-Shot Learning (ZSL) configurations, where the models failed to produce outputs for specific instances. Although these failures are not common, they do pose reliability concerns for practical deployment in live classroom settings.

With regard to the evaluation framework, whilst weighted accuracy, F1-score, MAE, and MSE were utilised to assess model performance, the study did not incorporate confidence intervals or statistical significance testing. This determination was informed by the deterministic nature of the evaluated models, which consistently generate outputs devoid of stochastic variability. However, future work involving randomised prompt ordering, cross-subject generalisation, or large-scale testing would benefit from the inclusion of statistical significance and uncertainty quantification to strengthen the robustness of the evaluation.

Finally, although Cohen's Kappa was calculated in order to assess inter-annotator agreement on emotion labels, confirming moderate agreement above chance, the specific values were not reported in order to avoid redundancy. This finding signifies a methodological limitation, and future research should incorporate more detailed inter-rater reliability analyses, such as Krippendorff's Alpha, to provide a more comprehensive assessment of annotation quality.

A. RECOMMENDATIONS FOR FUTURE WORK

To address these limitations, future research should prioritize:

- Expanding the dataset with a more diverse participant population to improve model generalizability.

- Enhancing model robustness to low-quality visual inputs through advanced preprocessing techniques or data augmentation.
- Developing adaptive multimodal fusion strategies that contextualize biometric data according to the learning environment.
- Implementing and carefully following the IEEE Ethically Aligned Design framework[1] to guide ethical considerations and ensure responsible, context-aware deployment in real classroom environments.

Furthermore, it is essential to examine how such systems might be implemented ethically and effectively in real classroom environments, paying close attention to student privacy, equitable access, instructor training, and adaptation to diverse pedagogical contexts, so that these solutions are not only technically robust but also socially responsible and practically beneficial.

VII. CONCLUSION

The findings of this study have significant potential for application in real-world educational settings, particularly in STEM education. The proposed use of Vision-Language Models combined with Few-Shot Learning presents a feasible solution for educators and institutions seeking to leverage advanced AI models without the need for large datasets. Since FSL only requires a small number of examples per class, institutions with limited data can quickly calibrate these models using their own student data. This approach allows for real-time analysis of student engagement and learning patterns, making it highly relevant for classrooms that face constraints in resources, such as access to large datasets or high computational power.

Beyond their predictive performance, the true value of these models lies in their ability to actively support instructional decision-making. By integrating explainability features, the models allow educators not only to receive engagement predictions but also to query the reasons behind them. This transparency helps teachers understand the underlying factors, such as body language patterns or emotional cues, that influence student engagement. When deployed in interactive dashboards, these models can provide real-time recommendations, alternative teaching strategies, or explanations tailored to specific classroom situations. This shifts the role of the system from a passive analytics tool to an active pedagogical assistant, enabling educators to refine their instruction, personalize learning, and respond more effectively to the dynamic needs of their students. Such capabilities have the potential to foster more reflective, data-informed teaching practices and improve learning outcomes across diverse educational environments.

Furthermore, the implementation of VLMs and MLLMs in a typical classroom environment appears viable. The model's ability to process and analyze both biometric and visual data offers valuable insights into student engagement without the need for intensive data preprocessing, a common barrier in many educational institutions. By minimizing the data and computational requirements, this system could be adapted to existing classroom technology, such as standard desktop computers or affordable wearable devices. While there may be challenges in terms of the hardware needed to process these models in real-time, such as the demand for advanced GPUs or other computational resources, these limitations could be addressed with future advancements in hardware accessibility and optimization algorithms.

Moreover, it is important to recognize potential privacy concerns when collecting and analyzing biometric data from students. Ensuring that these systems comply with data protection regulations and ethical standards is crucial for their widespread adoption in educational environments. Addressing these concerns through secure data handling practices and transparency with stakeholders can help mitigate these challenges.

This study contributes to a growing body of research on how AI-driven tools can be integrated into educational contexts to enhance learning outcomes. Future research should focus on the practical deployment of these systems in various educational settings, considering not only technical feasibility but also the social and ethical implications of their use. With continued advances in AI technology and data availability, the proposed systems have the potential to revolutionize how we understand and support student learning in real-time.

The integration of VLMs and MLLMs with FSL into educational environments offers an actionable framework for enhancing student engagement. In classrooms where resources may be limited, the use of FSL can enable rapid model training with minimal data, allowing educational institutions to make the most of their available data. This is particularly valuable in settings where large-scale datasets are not available, such as smaller schools or universities with limited access to specialized educational data. Furthermore, these models can be calibrated to meet the specific needs of educators, making them adaptable to a variety of learning environments.

However, beyond the technical and practical feasibility of such systems, it is crucial to address their ethical implications. AI systems that process student biometric and emotional data raise significant privacy concerns. Even when such systems do not directly identify individuals, there is a growing—albeit weak—consensus among stakeholders on the necessity of safeguarding emotional data, driven by various factors, ranging from ethical responsibility to institutional liability [26]. The urgency of protecting student data privacy is particularly pronounced in educational environments, where AI technologies introduce novel risks to the security of personal information [27]. Among the most critical issues are the need for informed consent, establishing robust regulatory frameworks, and preventing surveillance or misuse of sensitive information [28], [29].

[1] https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Equally important is the principle of fairness in the deployment of AI tools that analyze students' emotional and biometric data. Studies reveal that fairness is often overlooked in current AI/ML research on student mental health, with few works reporting demographic breakdowns or addressing biases in their models [30]. These omissions risk reinforcing existing inequalities, particularly when intersectional and structural disparities are not considered [31]. Moreover, both students and future AI developers require explicit training and guidance to understand the implications of fairness and bias, especially when such technologies are applied to vulnerable or marginalized groups.

Beyond privacy and fairness, the ethical deployment of AI in education must also consider transparency, accountability, and inclusive stakeholder engagement [32]. The concept of *algorithmovigilance*—continuous oversight and evaluation of AI systems—has been proposed to mitigate unintended negative outcomes and to foster trust in educational settings [33]. Ethical frameworks should thus prioritize transparent decision-making, iterative evaluation, and the active involvement of educators, students, and policymakers. According to Drira et al. [30], many AI education tools neglect key ethical dimensions such as demographic bias and explainability. Their review found that fewer than 10% of works considered fairness frameworks or included stakeholder feedback.

Not only does the present study demonstrate potential of AI-driven educational tools but it also provides a path forward for practical implementation. Future work should explore the scalability of these models in diverse educational settings, addressing both technical challenges and ethical considerations, to ensure their broad and responsible application.

DATA AVAILABILITY STATEMENT

The data supporting this study are available upon reasonable request to the corresponding author. Currently, the data are under review as part of a manuscript submitted for publication.

COMPLYING WITH ETHICS OF EXPERIMENTATION

This study did not require new ethical approval, as it was based on previously collected data that had already received ethical approval in a prior project. Participant data was anonymized in accordance with GDPR guidelines, ensuring that no personally identifiable information (PII) was stored or shared.

User: <image> On the image provided, indicate the detected emotion, according to your face. If you are unsure, indicate the emotion you think is most likely. Emotion is categorized into nine categories: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair' or 'Boredom'. Respond with the category only

PROMPT 1. Template used in Zero-Shot Learning for emotion detection using VLMs. The model is asked to select one of nine predefined emotion categories based solely on a provided image.

User: I want you to complete the task below by providing concise answers.

Based on the provided image, identify the emotion displayed. If uncertain, select the emotion you believe is most likely.

Example {{ COUNTER SAMPLE GIVEN }}.:

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'

Respond with the selected emotion. <image>

Response: {{ PREVIOUS RESULT }}.

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', "'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'. Respond with the selected emotion <image>. Response:

PROMPT 2. Template for Few-Shot Learning in emotion detection using VLMs. Multiple context examples are provided before asking the model to classify a new target image.

User: I want you to complete the task below by providing concise answers.

Based on the provided image, identify the emotion displayed. If uncertain, select the emotion you believe is most likely.

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'

Respond with the selected emotion. <image1>

Response: Boredom.

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'

Respond with the selected emotion. <image2>

Assistant: Boredom.

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'

Respond with the selected emotion. <image3>

Assistant: Enjoyment.

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'

Respond with the selected emotion. <image4>

Assistant: Enjoyment.

Question: What is the emotion in the image between the following: 'Enjoyment', 'Hope', 'Pride', 'Relief', "'Anger', 'Anxiety', 'Shame', 'Despair', or 'Boredom'. Respond with the selected emotion <image>. Response:

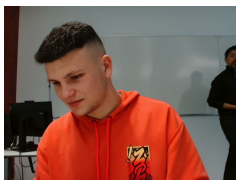
PROMPT 3. Example of a fully structured Few-Shot Learning prompt using four context images to guide the model in classifying the emotion displayed in a fifth image.

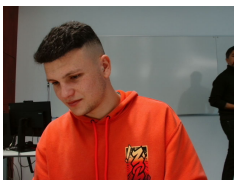
User: <image>
The age of the person is: {{ AGE }}
The reflected emotion of the person is: {{ EMOTION }}
The gender of this person is: {{ GENRE }}
The new assigned head rotation values are Pitch: {{ Value PITCH }}, Yaw: {{ Value Yaw }}, Roll: {{ Value ROLL }}, while the averages are Pitch: {{ AVG PITCH }}, Yaw: {{ AVG YAW }}, Roll: {{ AVG ROLL }}
The left eye is open {{ VALUE EYE LEFT }}%. The average is {{ AVG EYE LEFT }}%
The right eye is open {{ VALUE EYE RIGHT }}% while the average is {{ AVG EYE RIGHT }}%
The mouth area is {{ VALUE MOUTH }}% while the average is {{ AVG MOUTH }}%
The hand sensor rotation values, recorded {{ RECORD NUMBER }} data entries ago, are {{ ROTATION X VALUE }}, {{ ROTATION Y VALUE }}, {{ ROTATION Z VALUE }} The wake-up sensor values is {{ WAKE-UP VALUE }}
On the image provided, indicate the detected emotion, according to your face. If you are unsure, indicate the emotion you think is most likely. Emotion is categorized into nine categories: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair' or 'Boredom'. Respond with the category only"

PROMPT 4. **Template for Zero-Shot Learning in emotion detection using MLLMs, incorporating biometric and demographic data alongside visual input.**


(a) Ref. image 1


(b) Ref. image 2


(c) Ref. image 3


(d) Ref. image 4

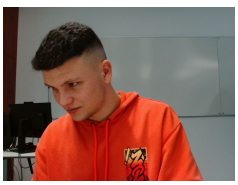

(e) Target image

FIGURE 2. **Visual layout of historical context images used in FSL to guide the model in emotion classification tasks. (a)-(d) Reference images, (e) Target image.**

The existing consents permit the analysis and publication of the data for academic purposes while fully adhering to the conditions set by the participants and relevant ethical standards.

User: <image>
The age of the person is: 16.44
The reflected emotion of the person is: sad
The gender of this person is: female
The new assigned head rotation values are Pitch: -33.788, Yaw: -28.092, Roll: 3.869, while the averages are Pitch: -46.225, Yaw: -14.389, Roll: -0.443
The left eye is open 3.89% while the average is 25.07%
The right eye is open 2.883% while the average is 19.36%
The mouth area is 0.11% while the average is 3.43%
The hand sensor rotation values, recorded 0 data entries ago, are 0.082, 0.229, -0.027
The wake-up sensor values is 84
On the image provided, indicate the detected emotion, according to your face. If you are unsure, indicate the emotion you think is most likely. Emotion is categorized into nine categories: 'Enjoyment', 'Hope', 'Pride', 'Relief', 'Anger', 'Anxiety', 'Shame', 'Despair' or 'Boredom'. Respond with the category only"

PROMPT 5. **Example of a complete MLLM prompt showing the fusion of demographic, biometric, and visual data for emotion classification.**

FIGURE 3. **Example of a multimodal prompt combining visual, demographic, and biometric data as input to a MLLM for emotion detection.**

User: <image> On the image provided, rate the individual's level of attention focused on what is happening in the classroom. They are doing manual testing at the table and interacting with classmates so this is considered attention in the context of the lesson. If you are unsure, indicate the rating you think is most likely. Rate attention on a scale of 1 to 5, where 5 indicates the highest level of attention and 1 indicates no attention. Respond only with the digit. For example, '5', '4', '3', '2' or '1'

PROMPT 6. **Template for Zero-Shot Learning prompts to predict attention levels in classroom settings using a single frame and predefined engagement scale.**

APPENDIX
DETAILED PROMPT STRUCTURES AND EXAMPLES

This appendix provides a detailed analysis of the prompt structures used in this study, focusing on their role in guiding multimodal and language models. These prompts form the

User: <image> Based on the previous image, assess the individual's attention in the classroom. They are interacting with peers and performing a task at the table, which shows engagement. If you're uncertain, provide your best guess. Rate attention from 1 to 5. Respond only with the digit.

Assistant: {{ VALUE }}.

User: <image> Look at the image provided and evaluate the individual's attention in the classroom. They are performing manual testing and interacting with classmates, showing relevant attention. Rate attention on a scale of 1 to 5, where 5 indicates the highest level of attention and 1 indicates no attention. Respond only with the digit: '5', '4', '3', '2', or '1'.

PROMPT 7. One-shot learning template providing a single labeled example before requesting attention prediction on a new image.

User: <image> Based on the previous image, assess the individual's attention in the classroom. They are interacting with peers and performing a task at the table, which shows engagement. If you're uncertain, provide your best guess. Rate attention from 1 to 5. Respond only with the digit.

Assistant: 4.

User: <image> Based on the previous image, assess the individual's attention in the classroom. They are interacting with peers and performing a task at the table, which shows engagement. If you're uncertain, provide your best guess. Rate attention from 1 to 5. Respond only with the digit.

Assistant: 4.

User: <image> Based on the previous image, assess the individual's attention in the classroom. They are interacting with peers and performing a task at the table, which shows engagement. If you're uncertain, provide your best guess. Rate attention from 1 to 5. Respond only with the digit.

Assistant: 3.

User: <image> On the image provided, rate the individual's level of attention focused on what is happening in the classroom. They are doing manual testing at the table and interacting with classmates so this is considered attention in the context of the lesson. If you are unsure, indicate the rating you think is most likely. Rate attention on a scale of 1 to 5, where 5 indicates the highest level of attention and 1 indicates no attention. Respond only with the digit. For example, '5', '4', '3', '2' or '1'

PROMPT 8. Example of a complete Few-Shot Learning prompt using four labeled examples to improve model performance in attention classification.

foundation for consistent and accurate outputs, ensuring alignment with study objectives.

The appendix is organized into sections covering:

1) **Emotion prompt structures**, including ZSL and FSL configurations.

(a) Ref. image 1 (b) Ref. image 2

(c) Ref. image 3 (d) Ref. image 4

(e) Target image

FIGURE 4. Visual layout of historical context images used in FSL to guide the model in attention classification tasks. (a)–(d) Reference images, (e) Target image.

2) **Attention prompt structures**, with examples demonstrating different learning scenarios.

Each section includes template designs and practical examples, highlighting their importance in optimizing model performance and minimizing errors in complex tasks.

A. EMOTION PROMPT STRUCTURE

1) ZSL PROMPTS

ZSL prompts leverage the model's pre-trained knowledge to generate responses without requiring prior task-specific examples. A detailed task description and context guide the model's predictions.

2) FSL PROMPTS

Unlike the case of attention prediction, in FSL prompts for emotion adding a counter of previous examples helps to obtain more consistent results, reducing noise in the system output.

3) EXAMPLE FSL PROMPTS (4 SHOTS)

The following example illustrates a typical FSL prompt configuration for emotion detection using four shots, giving the model an initial context to improve predictive accuracy in educational settings. By presenting a sequence of relevant examples, the model adapts to the task at hand, enhancing its ability to generalize from limited data. An example of the images used is shown in Figure 3

In multimodal linguistic models (MLLMs), instructions are designed to handle various types of data, such as images and text. In this context, the prompt retains the prompt structure used for attention.

User: <image>
The age of the person is: {{ AGE }}
The reflected emotion of the person is: {{ EMOTION }}
The gender of this person is: {{ GENRE }}
The new assigned head rotation values are Pitch: {{ Value PITCH }}, Yaw: {{ Value Yaw }}, Roll: {{ Value ROLL }}, while the averages are Pitch: {{ AVG PITCH }}, Yaw: {{ AVG YAW }}, Roll: {{ AVG ROLL }}
The left eye is open {{ VALUE EYE LEFT }}%. The average is {{ AVG EYE LEFT }}%
The right eye is open {{ VALUE EYE RIGHT }}% while the average is {{ AVG EYE RIGHT }}%
The mouth area is {{ VALUE MOUTH }}% while the average is {{ AVG MOUTH }}%
The hand sensor rotation values, recorded {{ RECORD NUMBER }} data entries ago, are {{ ROTATION X VALUE }}, {{ ROTATION Y VALUE }}, {{ ROTATION Z VALUE }} The wake-up sensor values is {{{ WAKE-UP VALUE }}
On the image provided, rate the individual's level of attention focused on what is happening in the classroom. They are doing manual testing at the table and interacting with classmates so this is considered attention in the context of the lesson. If you are unsure, indicate the rating you think is most likely. Rate attention on a scale of 1 to 5, where 5 indicates the highest level of attention and 1 indicates no attention. Respond only with the digit. For example, '5', '4', '3', '2' or '1'

PROMPT 9. Template for Zero-Shot Learning in attention detection using Multimodal Large Language Models (MLLMs), integrating visual, biometric, and contextual data.

4) EXAMPLE PROMPTS IN MULTIMODAL LANGUAGE MODELS

This section provides an example of prompts used in Multimodal Language Models (MLLMs) for emotion detection. An illustration of a typical MLLM prompt is shown in Figure 2.

B. ATTENTION PROMPT STRUCTURE

The final prompts used in this study can be categorized into three main sections based on the type of learning and the model's capabilities:

1) ZSL PROMPTS

The following example shows the ZSL prompt used in attention prediction.

2) FSL PROMPTS (1 SHOT)

In FSL, prompts are crafted by giving the model a limited set of input-output examples before prediction. This approach can substantially improve model performance on complex tasks (see Table 7), as it provides context about the task's nature. Here, labeled classroom examples are used in the prompt to assist the model in accurately detecting attention and emotions, thereby enhancing its adaptability to diverse data.

3) EXAMPLE FSL PROMPTS (4 SHOTS)

The following example illustrates a typical FSL prompt configuration for attention detection using four shots, giving the model an initial context to improve predictive accuracy in educational settings. By presenting a sequence of relevant examples, the model adapts to the task at hand, enhancing its ability to generalize from limited data 3.

4) PROMPTS IN MULTIMODAL LANGUAGE MODELS

This section provides an example of prompts used in MLLMs for emotion detection. Image used is shown in Figure 2.

ACKNOWLEDGMENT

This study employed the generative artificial intelligence model ChatGPT-4o Mini to enhance grammar in the translation to the target language.

REFERENCES

[1] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 23716–23736.
[2] M. Ashok, R. Madan, A. Joha, and U. Sivarajah, "Ethical framework for artificial intelligence and digital technologies," *Int. J. Inf. Manage.*, vol. 62, Feb. 2022, Art. no. 102433.
[3] M. Bustos-López, N. Cruz-Ramírez, A. Guerra-Hernández, L. N. Sánchez-Morales, N. A. Cruz-Ramos, and G. Alor-Hernández, "Wearables for engagement detection in learning environments: A review," *Biosensors*, vol. 12, no. 7, p. 509, Jul. 2022.
[4] A. B. Cantor, "Sample-size calculations for Cohen's Kappa," *Psychol. Methods*, vol. 1, no. 2, pp. 150–153, 1996.
[5] H.-C. Chu, W. W.-J. Tsai, M.-J. Liao, and Y.-M. Chen, "Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning," *Soft Comput.*, vol. 22, no. 9, pp. 2973–2999, May 2018.
[6] M. Drira, S. B. Hassine, M. Zhang, and S. Smith, "Machine learning methods in Student mental health research: An ethics-centered systematic literature review," *Appl. Sci.*, vol. 14, no. 24, p. 11738, Dec. 2024.
[7] G. Tonguç and B. O. Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture" *Comput. Educ.*, vol. 148, 2020, Art. no. 103797, doi: 10.1016/j.compedu.2019.103797.
[8] H. Hardjadinata, R. S. Oetama, and I. Prasetiawan, "Facial expression recognition using Xception and DenseNet architecture," in *Proc. 6th Int. Conf. New Media Stud. (CONMEDIA)*, Oct. 2021, pp. 60–65.
[9] L. Huang, "Ethics of artificial intelligence in education: Student privacy and data protection," *Sci. Insights Educ. Frontiers*, vol. 16, no. 2, pp. 2577–2587, Jun. 2023.
[10] M. Jovanovic and P. Voss, "Towards incremental learning in large language models: A critical review," 2024, *arXiv:2404.18311*.
[11] S. N. Karimah and S. Hasegawa, "Automatic engagement recognition for distance learning systems: A literature study of engagement datasets and methods," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, Jan. 2021, pp. 264–276.
[12] B. Klimova, M. Pikhart, and J. Kacetl, "Ethical issues of the use of AI-driven mobile apps for education," *Frontiers Public Health*, vol. 10, no. 1, pp. 1–10, Jan. 2023.
[13] Z. Li, H. Miao, V. Pascucci, and S. Liu, "Visualization literacy of multimodal large language models: A comparative study," 2024, *arXiv:2407.10996*.
[14] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023, *arXiv:2310.03744*.
[15] L. Marquez-Carpintero, S. Suescun-Ferrandiz, C. L. Álvarez, J. Fernandez-Herrero, D. Viejo, R. Roig-Vila, and M. Cazorla, "DIPSER: A dataset for in-person student engagement recognition in the wild," 2025, *arXiv:2502.20209*.

[16] N. McDonald and S. Pan, "Intersectional AI," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. 2, pp. 1–19, Oct. 2020.
[17] A. McStay, "Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy," *Big Data Soc.*, vol. 7, no. 1, Jan. 2020, Art. no. 205395172090438.
[18] A. North-Samardzic, "Biometric technology and ethics: Beyond security applications," *J. Bus. Ethics*, vol. 167, no. 3, pp. 433–450, Dec. 2020.
[19] M. Organt, "Community-engaged programming: Applying lessons learned from the VLM green teens program," *Proc. 21st Annu. Undergraduate Graduate Student Res. Conf. (Paideia)*. Newport News, VA, USA: Christopher Newport University, Apr. 2023.
[20] M. I. Posner and Y. Cohen, "Components of visual orienting," *Attention Perform., Control Lang. Processes*, vol. 32, pp. 531–556, Jul. 1984.
[21] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.
[22] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021, *arXiv:2102.12092*.
[23] H. R. Saeidnia, S. G. H. Fotami, B. Lund, and N. Ghiasi, "Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact," *Social Sci.*, vol. 13, no. 7, p. 381, Jul. 2024.
[24] M. M. Santoni, T. Basaruddin, and K. Junus, "Convolutional neural network model based students' engagement detection in imbalanced DAiSEE dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, pp. 1–10, 2023.
[25] J. Stamper, R. Xiao, and X. Hou, "Enhancing LLM-based feedback: Insights from intelligent tutoring systems and the learning sciences," in *Proc. Int. Conf. Artif. Intell. Educ.* Cham, Switzerland: Springer, 2024, pp. 32–43.
[26] G. Tonguç and B. O. Ozkara, "Automatic recognition of Student emotions from facial expressions during a lecture," *Comput. Educ.*, vol. 148, Apr. 2020, Art. no. 103797.
[27] F. Villena, L. Miranda, and C. Aracena, "llmNER: (Zero|Few)-shot named entity recognition, exploiting the power of large language models," 2024, *arXiv:2406.04528*.
[28] A. O. R. Vistorte, A. Deroncele-Acosta, J. L. M. Ayala, A. Barrasa, C. López-Granero, and M. Martí-González, "Integrating artificial intelligence to assess emotions in learning environments: A systematic literature review," *Frontiers Psychol.*, vol. 15, Jun. 2024, Art. no. 1387089.
[29] J. Wu, Z. Zhang, Y. Xia, X. Li, Z. Xia, A. Chang, T. Yu, S. Kim, R. A. Rossi, R. Zhang, S. Mitra, D. N. Metaxas, L. Yao, J. Shang, and J. McAuley, "Visual prompting in multimodal large language models: A survey," 2024, *arXiv:2409.15310*.
[30] D. Yang, A. Alsadoon, P. W. C. Prasad, A. K. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Proc. Comput. Sci.*, vol. 125, pp. 2–10, Jul. 2018.
[31] W.-C. Yeh, Y.-P. Lin, Y.-C. Liang, and C.-M. Lai, "Convolution neural network hyperparameter optimization using simplified swarm optimization," 2021, *arXiv:2103.03995*.
[32] J. Zhang, "Cognitive functions of the brain: Perception, attention and memory," 2019, *arXiv:1907.02863*.
[33] M. Zhao, J. Wang, Z. Li, J. Zhang, Z. Sun, and S. Zhou, "Effectively enhancing vision language large models by prompt augmentation and caption utilization," 2024, *arXiv:2409.14484*.

LUIS MARQUEZ-CARPINTERO received the degree in computer science from the University of Alicante in 2020, a specialized master's degree from the Complutense University of Madrid in 2021, and the Ph.D. degree in computer science from the University of Alicante.

He initially conducted research in the private sector before joining the Institute for Computer Research at the University of Alicante in 2023 as a Predoctoral Researcher. His research focuses on the application of artificial intelligence in educational contexts. He has authored more than 15 publications in conferences and workshops and has actively participated in several national and regional research projects. In 2024, he completed an international research stay at the University of Tokyo, where he collaborated on projects related to AI-driven educational technologies.

DIEGO VIEJO received the degree and Ph.D. degree in computer science from the University of Alicante, in 2002 and 2008, respectively.

In 2004, he began working with the University of Alicante as an Assistant Professor. Since then, he has held various positions (a collaborator and a contracted professor). Since 2019, he has been a Tenured University Professor. He has been the Principal Investigator in one regional-level project and one local-level project, and he has also participated as a research team member in numerous national-level projects. Additionally, he has completed a research stay at the Australian Center for Field Robotics, University of Sydney, under the supervision of Dr. Eduardo Nebot. His primary research focus has always been artificial vision, specifically methods that enable mobile robots to acquire three-dimensional vision for performing tasks in real-world environments. His work has contributed to improving robot localization and navigation in three-dimensional spaces, object identification, and object manipulation. In recent years, his research has shifted towards social robotics, aiming to use robots to assist dependent individuals.

MIGUEL CAZORLA (Senior Member, IEEE) received the degree and Ph.D. degree in computer science from the University of Alicante, in 1995 and 2000, respectively.

In 1995, he joined the University of Alicante as a Professor, progressing through various academic positions (an associate, an assistant, a school professor, and a university professor). Since 2017, he has been a Full Professor. He has served as the Deputy Director, the Secretary, and the Director of the Institute for Computer Research. Additionally, he has been the Coordinator of the Robotics Engineering degree and the Director of the Master's Program in Artificial Intelligence, Polytechnic School. He was the proposer and later the Coordinator of the Ph.D. Program in Computer Science, University of Alicante. He has undertaken several research stays at institutions abroad, including Carnegie Mellon University, the University of Sydney, and the University of Edinburgh. He has published more than 70 JCR-indexed articles (more than 30 in Q1) and more than 100 conference papers at both national and international levels. He has supervised 22 Ph.D. theses and is the Principal Investigator in several national projects (CICYT), Retos, and has carried out multiple technology transfer contracts with companies. He is a member of various program committees for both national and international conferences. His research has always focused on computer vision. From the beginning, he applied these techniques to solving robotic tasks. Since almost the start of his research career, he has worked on 3D data processing. In recent years, he has diversified his research to apply deep learning techniques to various fields, including medical imaging, object recognition, depth estimation, and traffic object identification. He is also focusing his research on using large language models (LLMs) to address different problems. In recent years, all his research has been directed towards social robotics, applying these techniques to assist dependent individuals.