

Vision Large Language Models Are Good Noise Handlers in Engagement Analysis

Alexander Vedernikov¹, Puneet Kumar¹, Haoyu Chen¹,
Tapio Seppänen¹, Xiaobai Li^{1,2*}

^{1*}Center for Machine Vision and Signal Analysis, University of Oulu,
Oulu, Finland.

²State Key Laboratory of Blockchain and Data Security, Zhejiang
University, Hangzhou, China.

*Corresponding author(s). E-mail(s): xiaobai.li@zju.edu.cn;

Contributing authors: aleksandr.vedernikov@oulu.fi;
puneet.kumar@oulu.fi; chen.haoyu@oulu.fi; tapio.seppanen@oulu.fi;

Abstract

Engagement recognition in video datasets, unlike traditional image classification tasks, is particularly challenged by subjective labels and noise limiting model performance. To overcome the challenges of subjective and noisy engagement labels, we propose a framework leveraging Vision Large Language Models (VLMs) to refine annotations and guide the training process. Our framework uses a questionnaire to extract behavioral cues and split data into high- and low-reliability subsets. We also introduce a training strategy combining curriculum learning with soft label refinement, gradually incorporating ambiguous samples while adjusting supervision to reflect uncertainty. We demonstrate that classical computer vision models trained on refined high-reliability subsets and enhanced with our curriculum strategy show improvements, highlighting benefits of addressing label subjectivity with VLMs. This method surpasses prior state of the art across engagement benchmarks such as *EngageNet* (three of six feature settings, maximum improvement of **+1.21%**), and *DREAMS / PAFE* with F1 gains of **+0.22 / +0.06**.

Keywords: Engagement recognition, Vision-language models, Curriculum learning, Noisy labels, Affective computing

1 Introduction

Affective computing has driven progress in emotion recognition across psychology, education, and human–computer interaction [1]. While Ekman’s basic emotions [2] are well studied, non-typical emotions like engagement remain underexplored [3]. In this context, engagement recognition refers to estimating a learner’s level of involvement from behavioral cues such as gaze, posture, facial affect, and distractions. Video-based engagement recognition seeks to predict this level from short video segments of learners interacting with instructional material. Engagement’s inherent subjectivity and sparse human-generated annotations introduce label disagreements, leading to noise [4, 5], overfitting, reduced generalization [6, 7]. Addressing these challenges and improving dataset quality through annotation refinement [8] is vital for effective engagement recognition [9].

Approaches to Learning with Noisy Labels (LNL), such as module-based [10], loss-based [11], label correction [12], and sample selection [13] methods, attempt to overcome annotation challenges. Yet, in subjective tasks such as emotion recognition [14–16] and engagement recognition, they struggle with annotation variability and data loss [17].

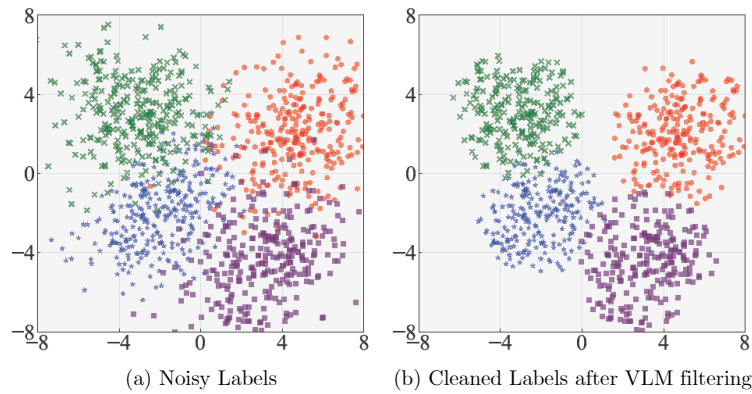


Fig. 1: Dataset refinement through VLM analysis. Visualization of noisy vs. cleaned labels, illustrating VLM filtering on dataset quality. Markers represent dataset’s different classes. In this scheme, “different classes” correspond to the four engagement levels used throughout our experiments (0: Not Engaged, 1: Barely Engaged, 2: Engaged, 3: Highly Engaged), and the plot is illustrative only rather than taken from a specific dataset or VLM configuration.

VLMs integrate computer vision and natural language processing through transformer advances [18], addressing these challenges. By posing structured questions about observable cues, VLMs can uncover subtle annotation inconsistencies and mitigate subjectivity in emotion recognition [19]. Their capacity for automated label correction enhances dataset quality and model generalization. Recent studies have also

begun to explore VLMs for learning engagement and related academic emotion recognition, but typically use VLMs as direct predictors rather than as tools for structuring label reliability and guiding downstream training [20, 21].

We present a framework that uses VLMs to improve analysis of engagement datasets (*EngageNet*, *PAFE*, and *DREAMS*) by addressing label subjectivity and noise. We prompt InternVL2-8B with 15 structured yes/no questions on 16 sampled frames per video to extract behavioral cues (eye gaze, facial expression, body posture, and distractions) to detect annotation inconsistencies. Samples where VLM agrees with the original label form a high-reliability (‘Accepted’) subset, while disagreements form a low-reliability (‘Rejected’) subset. We then train classical computer vision models (e.g., TCCT-Net, ResNet+TCN, EfficientNet+LSTM) using a two-stage curriculum: Stage 1 trains on Accepted samples only; Stage 2 adds Rejected samples and implements curriculum learning with soft label refinement during training, allowing the model to gradually learn from less ambiguous samples while substituting hard one-hot labels with uncertainty-aware soft targets for ambiguous samples. During data augmentation, segment sampling is weighted by VLM-derived reliability. Finally, we evaluate accuracy and F1 on all three datasets and compare against baseline models. Result of dataset refinement is depicted in Figure 1, while detailed refinement process is explained in Figure 2. The Figure 1 is intended as an illustrative schematic of the refinement concept rather than a direct visualization from a particular dataset or VLM run. Concrete datasets (*EngageNet*, *PAFE*, *DREAMS*) and VLM choices are described in detail in Sections 3 and 4. Findings indicate that traditional computer vision models, when trained on selected high-reliability subsets and further enhanced by curriculum strategy, achieve noticeable improvements. The contributions are:

1. We use VLMs to detect and correct label noise in engagement datasets, thereby enhancing annotation consistency without relying on pre-existing noise models.
2. By using a structured questionnaire, we guide VLMs to focus on observable engagement cues, thus reducing ambiguity and improving zero- and few-shot performance.
3. We use VLM-based filtering to partition the dataset into accepted and rejected subsets, which reduces noise and improves the robustness of classical vision models.
4. Our training strategy incorporates curriculum learning and soft label refinement to progressively include samples with varying levels of ambiguity, addressing the challenges posed by label subjectivity.
5. Experiments on multiple datasets demonstrate our approach improves accuracy and robustness while matching human tolerance in engagement annotation.

2 Related Work

2.1 Subjectivity in Engagement Recognition

In contrast to typical emotions with clear indicators, engagement, being non-typical, is subject to ambiguous cues, making its recognition inherently subjective. This subjectivity results in annotations that introduce label noise, weakening model reliability. In *DAiSEE* and similar datasets, inconsistent labeling worsens annotation noise and

impairs model generalization [22]. Moreover, as engagement is inherently nuanced and multidimensional, traditional binary and multi-class classifications often fall short, further impairing model performance [23]. This complexity is underscored by Liao et al. [24], who show that even advanced models like Deep Facial Spatiotemporal Network (DFSTN) achieve less than 60% accuracy on multi-class engagement tasks in *DAiSEE*. As noisy labels degrade generalization and trigger overfitting, Zhong et al. [25] underscore the need for effective noise-handling strategies. Jiang and Deng [26] highlight that without these approaches, emotion recognition performance suffers.

2.2 Learning with Noisy Labels

In the literature, samples whose observed labels match the true labels are called ‘clean’ and those with corrupted or incorrect labels ‘noisy’ [27]. Surveys by Song et al. [28] and Han et al. [29] provide comprehensive taxonomies of these definitions. Noisy labels can cause Deep Neural Networks (DNNs) to overfit, leading to suboptimal performance. While many LNL techniques, including module-based, loss-based, label correction, and sample selection methods, have been proposed to enhance generalization, their application in engagement analysis remains unexplored [30]. Therefore, we examine methods from a broader emotion recognition scope.

Module-based methods modify the neural network modules to be more robust against label noise. In depression detection, Chen et al. [10] leverage soft labels and attention to design a module-based method that enhances neural network generalization despite label noise. Tan et al. [31] present RCL-Net, a module-based approach that uses dual-consistency learning to improve representation quality under noisy label conditions in facial expression datasets. Building on earlier strategies, Liu et al. [32] integrated auxiliary tasks and graph-based semantic reasoning, which significantly improved robustness to noisy labels in facial expression recognition. Subsequently, Mao et al. [33] introduced DMM-CNN, which integrates task-specific features with dynamic weighting and adaptive thresholding to boost generalization in subjective, imbalanced facial attribute classification tasks.

Loss-based methods design the loss function to be more robust against label noise. Khan et al. [11] tackle subject-dependent noise by introducing a Wasserstein Distance-based loss function, thus improving model generalization in a range of emotion recognition applications. Further enhancing loss-based approaches, Jiang and Deng [34] demonstrated that attention mechanisms such as Attention Flipping Consistency (AFC) and label-guided spatial attention dispersing (SAD) losses improve robustness by guiding the model to concentrate on distinct, consistent expression features, mitigating label ambiguity. Chen et al. [35] further advanced loss-based methods by proposing an Emotion-Anchors Self-Distillation Loss, which reduces label ambiguity and enhances dynamic facial expression recognition.

Label correction methods adjust the loss value or correct the labels to mitigate the impact of label noise. Zhou et al. [12] address label noise in EEG emotion datasets with their Prototypical Representation-based Pairwise Learning (PR-PL) framework, which employs adaptive pseudo-labeling and pairwise similarity learning to boost robustness via domain alignment. Washington et al. [36] revealed that crowdsourced soft-target

labels capture the nuanced subjectivity of emotions, which improves model robustness and mirrors human judgment. Jin et al. [37] adopted Gaussian Mixture Models to generate soft pseudo-labels, effectively reducing label noise and elevating mental state classification accuracy in social media text. Liu et al. [38] introduced ULC-AG, a method that integrates graph-based facial action unit modeling with semantic similarity to effectively correct noisy labels in extensive facial expression datasets. Zhang and Etemad [39] showed that Partial Label Learning with label disambiguation can effectively refine ambiguous EEG emotion labels, mitigating label noise. Complementary to these methods, several recent approaches exploit multimodal large language models (MLLMs) and VLMs as external teachers. Huang et al. [40] formulate “Machine Vision Therapy”, where an MLLM rectifies vision model predictions and improves Out-of-Distribution (OOD) robustness via denoising in-context learning; Wang et al. [41] propose NoiseGPT, which detects and cleanses label noise by analyzing probability curvature and in-context discrepancies; and Zhang et al. [42] introduce AdaNeg, a training-free OOD detector that constructs adaptive negative proxies from test-time feature banks. Our work similarly leverages a VLM as a teacher, but differs in using offline, questionnaire-guided behavioral cues to assign reliability scores and refine subjective engagement labels, which then drive curriculum learning and soft-label training instead of modifying predictions on-the-fly.

Our key methodological contribution is to decouple label reliability from semantic content using a questionnaire and a VLM, and to exploit this reliability in a two-stage curriculum with ambiguity-aware soft labels for adjacent ordinal classes. Unlike prior VLM-based denoising or OOD work, we apply reliability-aware training to subjective engagement and explicitly organize learning by estimated reliability.

Sample selection methods choose clean samples from noisy datasets to improve the model’s generalization performance. Gera and Balasubramanian [13] developed Consensual Collaborative Training, a method that emphasizes clean samples during early training while enforcing consistency loss to improve learning across noisy datasets. Building on previous methods, Chen et al. [43] introduced SSPA-JT, a technique that leverages clean sample filtering and paired augmentation to dynamically balance the dataset and improve generalization in facial expression recognition. Jiang and Deng [26] advanced sample selection by introducing the Progressive Teacher framework, which iteratively discards high-loss samples, improving generalization in noisy facial expression recognition.

2.3 VLMs in Affective Computing

Affective Computing broadly investigates computational approaches for sensing, interpreting, and modelling human emotions and related states in context [3]. In this work, we treat students’ engagement as a non-typical affective state and use VLMs as tools to improve its annotation and recognition under subjective labeling.

VLMs have set new benchmarks in multi-modal tasks, particularly in image captioning and visual question answering [44]. By leveraging lightweight adapters linking visual and language models, they extend their potential to Affective Computing, where the fusion of text, visuals, audio, EEG data helps address challenges like micro-expression recognition, sentiment analysis, brain-computer interactions [45].

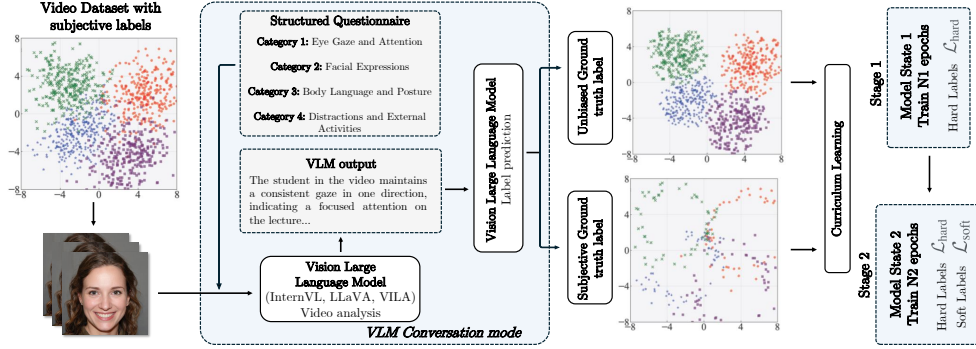


Fig. 2: Overview of our method for engagement analysis with subjective labels. A structured questionnaire guides a VLM to assess label reliability, generating informed outputs. It then uses two-stage curriculum learning, first training on high-confidence data and then introducing ambiguous samples with soft labels. This mitigates label noise and enhances model performance.

Until now, very few works have utilized VLMs in Affective Computing. Recent empirical evaluations of Teotia et al. and Wang et al. consider VLMs for learning engagement detection and academic emotion analysis, treating the models as direct predictors of student states [20, 21]. The results underline significant room for improvement, particularly in handling nuanced and context-dependent engagement cues. Xie et al. [46] presented EmoVIT using GPT-assisted visual tuning for enhanced emotion reasoning, but it does not address subjectivity. Nadeem et al. [47] found VLMs to perform similarly with deep models on typical emotion datasets, restricting their utility for non-typical emotions. While AU-LLaVA [48] and GPT-4V [49] excel in tasks like Action Unit recognition, their limited focus hinders broader applicability. Etesam et al. [50] achieved notable results with image captioning, yet static prompts reduce flexibility in handling subjective non-typical emotions.

Prior applications of VLMs and LNL techniques in Affective Computing cover both typical emotions and initial engagement studies, but mostly rely on simple prediction pipelines with little support for label noise or subjectivity. This highlights the need for VLM-driven strategies that explicitly improve dataset reliability and learning in subjective, non-typical states.

3 Subjective Labels and VLM Filtering

3.1 Datasets

Three engagement datasets are utilized in the course of this study: *EngageNet* [51, 52], *PAFE* (Predicting Attention with Facial Expression) [53], and *DREAMS* (Diverse Reactions of Engagement and Attention Mind States) [54]. Following the original papers, we report classification accuracy and weighted F1, matching their evaluation protocols. Performance is always measured within each dataset’s original engagement

scale and video window. Datasets were licensed, anonymized per GDPR, and bias-evaluated via VLMs.

EngageNet dataset comprises 11,300 video samples from 127 participants aged 18 to 37, interacting with a web-based learning platform. Video samples have been annotated for four levels of engagement: ‘Highly Engaged,’ ‘Engaged,’ ‘Barely Engaged,’ ‘Not Engaged.’ Dataset includes three subject-independent sets: 7,983 training videos from 90 participants, 1,071 validation videos from 11 participants, 2,257 test videos from 26 participants. Each video sample was independently annotated by three experts.

PAFE dataset captures 1,100 video samples from 15 university students attending an online lecture. Each sample was recorded at 640p resolution and 30fps, with probes conducted every 40 seconds to label attentional states as ‘Focused,’ ‘Not-Focused,’ or ‘Skip.’ Probes were annotated by participants’ self-reports and expert review to remove drowsy mismatches.

DREAMS dataset is built from video recordings of 32 participants captured in natural environments. In total, 832 video segments of 40 seconds each were recorded, with 781 segments successfully used for analysis. Participants provided self-ratings of their engagement while watching a range of video clips that were chosen to evoke different emotional reactions. The data is divided into subject-independent sets. Every 40 seconds, participants annotated their engagement levels via self-assessment prompts.

The details of the computational settings used in this work are in Appendix A.

3.2 VLM-based Engagement Classification

We leverage several state-of-the-art open-source VLMs (including InternVL2 (8B and 26B) [55, 56], LLaVA 1.6 (7B) [57–59], and VILA 1.5 (8B) [60]) to address a four-class engagement classification task on the *EngageNet* dataset using zero- and few-shot learning approaches. Due to privacy constraints, proprietary VLMs were not used, as the sensitive video samples cannot be uploaded externally.

Single question mode. We evaluated three input types: single images, multiple images, and video sequences (extracted frames). For multiple images, the model predicts the engagement level for each frame independently, and the final prediction is the mode of these individual predictions.

Conversation mode. To enhance VLM discriminative power, we evaluated InternVL2 in conversation mode with single image and video inputs, using both 8B and 26B model sizes. InternVL2 was chosen for its Conversation mode (absent in LLaVA) and its strong performance in single-question settings. In the ‘few-shot/image’ mode, the model was first provided with a few example images for each engagement class along with corresponding affective descriptions. Subsequently, it was prompted to analyze a new image and finally output a numeric engagement level. In the zero-shot video approach, the model is provided with a video from which it extracts key frames (eight by default) and, through sequential conversational rounds, is instructed first to focus on specific video features and then to classify the engagement level using the same methodology as in the few-shot/image mode. Detailed prompt formulations are provided in Appendix B.

VLM guidance with structured prompts. Without clear, direct instructions defining engagement states, VLM predictions in both Single Question and Conversation modes can suffer from inconsistency and reduced accuracy. To address this, we adopted a structured approach using precise YES/NO prompts that bind the model’s responses to specific, observable behaviors. In our framework, the VLM is guided by a questionnaire (see Appendix C) organized into four key categories: *(1) Eye Gaze and Attention*, *(2) Facial Expressions*, *(3) Body Language and Posture*, *(4) Distractions and External Activities*. This approach allows to investigate whether behavior-based annotation reduces inter-annotator variability and helps the VLM avoid learning annotator-specific biases instead of generalizable engagement features.

The resulting checklist is deliberately static and behavior-grounded: items capture gaze, posture, facial and distraction cues commonly used in human engagement protocol [51–54], selected to balance coverage with simplicity. We apply this fixed template across datasets instead of searching for an “optimal” questionnaire, as our goal is to analyse VLM-guided label quality rather than optimize the question set. The full template is given in Appendix C for reuse and future domain-specific refinement.

3.3 Enhancing Data Quality with VLM Filtering

We leverage VLM outputs not only for engagement classification but also to assess label reliability by partitioning the Validation sets of *EngageNet* and *PAFE* datasets into two subsets: Accepted (where VLM predictions agree with the ground truth) and Rejected (where discrepancies exist). Following the clean/noisy nomenclature, we treat our Accepted subset as a proxy for clean data and Rejected as a proxy for noisy data. Unlike standard noise-model approaches (which consider each noisy label as inversion from an assumed ground truth) we use VLM agreement as a model-based measure of sample reliability. To assess this, we evaluate classical computer vision models (TCCT-Net [9], ResNet+TCN [61], EfficientNet+LSTM [62], EfficientNet+Bi-LSTM [62]) across the full Validation set as well as the Accepted and Rejected subsets. This comparative analysis reveals the models’ sensitivity to label noise and highlights their reliance on high-consensus annotations. The filtering criteria and experimental protocol are discussed in Appendix E and F.

From a computational perspective, VLM-based filtering is run once as an offline preprocessing stage (see Appendix A for hardware details), and the resulting Accepted/Rejected indicators and reliability scores are cached for all subsequent analyses and curriculum runs. This step scales roughly linearly with the number of clips, while the per-epoch training throughput of our engagement models stays unchanged. For larger datasets or heavier VLMs, the cost can be further reduced via batched video processing and mixed-precision (e.g., FP16) inference, but a detailed efficiency study is beyond the scope of this work.

3.4 Human Performance in Engagement Analysis

To further assess ground truth label reliability, we conducted a human evaluation comparing annotator consistency with VLM performance, highlighting alignment and

ambiguity in subjective engagement classification. Ten experienced annotators classified 50 video samples from *EngageNet* Validation set into four engagement levels. Annotators received detailed guidelines and were allowed up to two views per video. To account for minor discrepancies, we introduced a tolerance-based accuracy metric (± 1 level). Full experimental details are in Appendix G.

4 Baseline Results: VLM Classification and Data Filtering

4.1 VLM-based Engagement Classification

Table D2 of Appendix D presents the zero- and few-shot performance of VLMs on the *EngageNet* validation set. While LLaVA achieved 40.06% accuracy in Single Question mode using two images, it lacks conversation mode support. In contrast, InternVL2 improved from 28.66% (Single Question mode) to 35.20% in Conversation mode, and further reached 45.28% with structured question guidance and video input using the 8B model (vs. 29.97% for the 26B variant). Temporal and sequential gains motivated our use of InternVL2 (8B) for subsequent experiments.

4.2 Evaluation of VLM Filtering and its Impact on Model Performance

Structured question guidance (Table C1, Appendix G) enabled VLM zero-shot video classification with InternVL2 (8B) to reach 45.28% accuracy on full Validation set of *EngageNet* (thus, resulting in 485 accepted, 586 rejected) and 71.47% on full Validation set of *PAFE* (providing 243 accepted, 97 rejected). This highlights VLM’s capacity to detect annotation inconsistencies, providing refined data for assessing classical models’ overfitting to noisy labels.

Classical computer vision results. Table 1 reveals a strong dependence on label quality. On the *EngageNet* dataset, TCCT-Net achieves 82.06% accuracy on Accepted subset, but this drops by 24.04% to 58.02% on Rejected subset. Similarly, on the *PAFE* dataset, TCCT-Net’s accuracy falls from 95.88% on Accepted samples to 17.53% on Rejected ones with a decline of 78.35%. Such disparities indicate that aggregate Validation scores (e.g., 68.91% for TCCT-Net on *EngageNet* and 74.71% on *PAFE*) may mask the underlying sensitivity of models to label noise. Overall, these results (also observed for ResNet+TCN and EfficientNet-based models) underscore that even minor label inconsistencies impair performance, emphasizing the need for robust data curation and noise mitigation in engagement analysis. Table 1 analyses only expert-labeled *EngageNet* and *PAFE* (expert-confirmed after self-reports), while *DREAMS* dataset (built entirely on participant self-ratings) is discussed in Section 6.

4.3 Human Performance Results

Human evaluation, ground truth annotation and VLM results on the *EngageNet* Validation set revealed subjectivity in engagement classification. Human evaluators achieved 46.25% exact-match accuracy, with 42.00% on Accepted samples and 50.50%

Method	Full Validation	Rejected	Accepted
<i>EngageNet Dataset</i>			
VLM – InternVL2 (8B) [55, 56]	45.28	–	–
TCCT-Net [9]	<u>68.91</u>	58.02 −10.89	82.06 +13.15
ResNet + TCN [61]	<u>54.72</u>	46.65 −8.07	64.47 +9.75
EfficientNet + LSTM [62]	<u>57.57</u>	49.37 −8.20	67.48 +9.91
EfficientNet + Bi-LSTM [62]	<u>58.94</u>	51.17 −7.77	68.33 +9.39
<i>PAFE Dataset</i>			
VLM – InternVL2 (8B) [55, 56]	71.47	–	–
TCCT-Net [9]	<u>74.71</u>	17.53 −57.18	95.88 +21.17
ResNet + TCN [61]	<u>62.37</u>	31.52 −30.85	74.68 +12.31
EfficientNet + LSTM [62]	<u>65.23</u>	35.68 −29.55	77.03 +11.80
EfficientNet + Bi-LSTM [62]	<u>67.12</u>	36.13 −30.99	79.49 +12.37

Table 1: Evaluation of classical computer vision model accuracy across Validation, Accepted, and Rejected subsets in *EngageNet* and *PAFE* (all values represent percentages). **Bold** indicates the best performance and underline shows the second-best.

on Rejected ones, suggesting VLM-filtered “reliable” samples remain ambiguous due to labeling variations. However, when applying ± 1 tolerance metric, human accuracy improved dramatically to 88.75% (with 82.50% on Accepted and 95.00% on Rejected samples), indicating that while exact matches are challenging, annotators generally select engagement levels close to ground truth. In comparison, the VLM achieved an overall accuracy of 50.70% and ± 1 tolerance accuracy of 78.52%, demonstrating that when guided by structured prompts, it can match or even surpass human consistency. Findings show that ground truth labels fail to capture a nuanced nature of engagement: low exact-match accuracies and higher tolerance-based scores. This underscores the need for improved annotation protocols and model designs that better accommodate subjectivity in engagement classification.

4.4 Ablation Study

To assess the impact of input configuration on engagement classification, we did ablation studies on the number of input frames and prompt questions in VLM filtering (Table 2). Complete results for all configurations are provided in Table H3 of Appendix H. Experiments reveal that increasing the number of frames improves performance, with accuracy peaking at 16 frames. Similarly, varying the number of prompt questions, we observed that using all 15 questions yields the best performance, as it provides richer set of engagement cues for VLM filtering. Results indicate optimal balance between temporal context and prompt detail, critical for mitigating effects of subjective labeling.

Frames / Questions	Full Validation Accuracy (%)	Rejected Accuracy (%) / Samples		Accepted Accuracy (%) / Samples	
F2	43.04	62.13	-6.78	610	77.87 +8.96
F16	50.70	50.19	-18.72	528	87.11 +18.20
F20	47.99	52.78	-16.13	557	86.38 +17.47
Q3	36.23	69.84	+0.93	683	67.27 -1.64
Q9	43.79	55.48	-13.43	602	86.14 +17.23
Q15	50.70	50.19	-18.72	528	87.11 +18.20

Table 2: Performance of VLM over different frame counts (F) and prompt question (Q) counts. Red/green values indicate accuracy differences from the TCCT-Net [9] baseline of 68.91%. Total number of samples in the Validation set of *EngageNet* is 1071. **Bold** indicates the best performance. Detailed results are provided in Table H3 of Appendix H.

5 Curriculum Learning with Soft Label Refinement

Due to label inconsistencies in Sections 3 and 4, we propose a training strategy integrating curriculum learning with soft label refinement, guided by labels derived from VLM analysis. Gradually incorporating challenging samples and using soft labels, our approach mitigates label noise and subjectivity. Validation set evaluations confirm its effectiveness. This framework is implemented to improve TCCT-Net, a two-stream architecture for fast engagement analysis by Vedernikov et al. [9], and is adaptable to other models. We select TCCT-Net as our main backbone because it is a recent, efficient engagement estimator that already demonstrates strong performance on the benchmarks considered in this work.

5.1 Curriculum Learning Strategy

Our curriculum learning strategy (Figure 3) is built on the idea that not all training samples are equal, some are more reliable than others. To that end, we rank and split the training data based on the difference between the ground truth (GT) label and the VLM-predicted label, which serves as a proxy for sample reliability. We treat samples with zero or with differences of 2 or 3 (typically VLM errors) as high-confidence, using them in the first training stage, and reserve samples with a one-level difference for a later stage due to their subjectivity. This design reflects the ordinal nature of engagement: adjacent classes ($|y - \hat{y}| = 1$) differ only slightly in intensity and are therefore more subjective. In contrast, larger gaps ($|y - \hat{y}| \geq 2$) indicate clearly distinct engagement regimes; we interpret them mainly as VLM failures and trust the human label, treating these samples as high-confidence in the first stage. By training first on reliable samples and only later on ambiguous ones, the model learns more robust features and generalizes better. We explicitly do not treat VLM agreement as absolute

truth; it is merely a proxy for ambiguity that guides the optimizer from the least ambiguous to the most difficult samples.

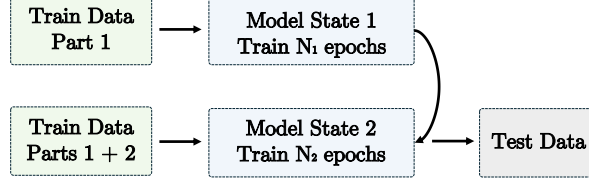


Fig. 3: Two-stage curriculum learning scheme for engagement estimation: Stage 1 uses high-confidence samples; Stage 2 incorporates ambiguous samples.

5.2 Soft Label Generation and Integration

To further mitigate the impact of noisy or subjective labels, we generate soft labels for samples with minor discrepancies (one-level differences), converting hard one-hot vectors into probability distributions that express uncertainty. This integration into the loss function mitigates the negative effects of mislabeling.

Let $y \in \{0, 1, 2, 3\}$ denote the ground truth label and \hat{y} the VLM-predicted label. We define the soft label vector $\mathbf{s} \in \mathbb{R}^4$ as:

$$s_i = \begin{cases} \alpha, & \text{if } |y - \hat{y}| = 1 \text{ and } i = y, \\ 1 - \alpha, & \text{if } |y - \hat{y}| = 1 \text{ and } i = \hat{y}, \\ 1, & \text{if } |y - \hat{y}| \neq 1 \text{ and } i = y, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $\alpha \in (0, 1)$ controls confidence in the GT versus the VLM prediction. In practice, we set the smoothing coefficient α through a grid search. This formulation applies soft labels (i.e., probability distributions) to samples with a one-level discrepancy, while reverting to standard one-hot encoding for unambiguous cases. The integration of soft labels into training is achieved by modifying the loss function. For ambiguous samples, i.e. a one-level discrepancy ($|y - \hat{y}| = 1$), we use a soft loss defined via Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\text{soft}} = \sum_{i=0}^3 s_i \log \frac{s_i}{p_i}, \quad (2)$$

where $p_i = \text{softmax}(z_i)$ is the predicted probability for class i . For unambiguous samples, we apply the standard cross-entropy loss:

$$\mathcal{L}_{\text{hard}} = -\log(p_y). \quad (3)$$

The per-sample loss is therefore:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{soft}}(\mathbf{s}, p), & \text{if } |y - \hat{y}| = 1, \\ \mathcal{L}_{\text{hard}}(p_y), & \text{otherwise,} \end{cases} \quad (4)$$

with total training loss computed as:

$$\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{j=1}^N w_j \mathcal{L}_j, \quad (5)$$

where w_j reflects sample reliability and N is the batch size.

This loss function enables the network to learn from ambiguous samples, reducing overfitting to hard labels by incorporating insights from both GT and VLM. Although our soft targets come from VLM-GT discrepancies rather than multi-annotator distributions, they nevertheless inject a useful uncertainty signal that stabilizes training.

5.3 Weighted Data Augmentation via Segmentation and Recombination

To incorporate sample reliability into training, we modify the standard Segmentation and Recombination (S&R) augmentation, originally used in TCCT-Net, by including sample weights. Each sample is assigned a weight w reflecting its label reliability, defined by the discrepancy between the GT and VLM prediction. Weights are used to compute a sampling probability, ensuring that segments from more reliable samples are more likely to be selected, and they determine the average weight of each augmented example.

Let the dataset be $\{(x_i, y_i, w_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^L$ is a signal of length L , y_i its label, and w_i its weight. Each signal is segmented into n equal parts of length $l = L/n$. For class c , denote its sample indices by $\mathcal{I}_c = \{i \mid y_i = c\}$. Sampling probability for each sample is defined as:

$$p_i = \frac{w_i}{\sum_{j \in \mathcal{I}_c} w_j}. \quad (6)$$

To form an augmented sample $x_{\text{aug}} \in \mathbb{R}^L$ for class c , we:

1. For each segment j ($1 \leq j \leq n$), randomly select a sample index $i_j \in \mathcal{I}_c$ with probability p_{i_j} .
2. Extract the j th segment as $x_{i_j}^{(j)} = x_{i_j}[(j-1)l : jl]$.
3. Concatenate segments $x_{\text{aug}} = [x_{i_1}^{(1)}, x_{i_2}^{(2)}, \dots, x_{i_n}^{(n)}]$.
4. Compute the augmented weight as $w_{\text{aug}} = \frac{1}{n} \sum_{j=1}^n w_{i_j}$.

This strategy ensures that segments from more reliable samples (with higher w_i) are preferentially selected, preserving signal integrity and reducing noise. The augmented data $\{(x_{\text{aug}}, c, w_{\text{aug}})\}$ is then merged with the original dataset, increasing effective dataset size while integrating label confidence into training. Modified S&R

augmentation provides data diversification along with a refined weighting mechanism reflecting label reliability.

5.4 VLM-guided Label Refinement

VLM output serves as a critical guide for quantifying label noise and informing training stages. The discrepancy between the GT and VLM-predicted labels provides an objective measure of reliability informing sample ranking in curriculum learning strategy and the generation of soft labels. By integrating VLM guidance into sample ordering and label adjustment, our method selectively weights training examples and mitigates the effects of noisy annotations.

Curriculum learning, by starting with high-confidence, low-noise samples [63], establishes a solid foundation for handling ambiguous data. Soft label refinement smooths the supervision signal, decreasing sensitivity to errors and enhancing generalization [64]. Leveraging VLM insights motivated by ensemble and co-training paradigms [65] further mitigates overfitting and improves accuracy.

Together, these components enable the network to adapt effectively to ambiguous samples, ultimately leading to improved generalization and robustness on unseen test data. However, the agreement between VLM and human labels is only used as a proxy for sample reliability and does not imply that the VLM predictions are always correct. In our design, (i) VLM-human agreement and large discrepancies are used only to stage the curriculum and define sample weights, (ii) one-level disagreements are encoded as α -weighted soft targets that still preserve the original ground-truth label, and (iii) samples never have their human labels replaced by VLM predictions. As shown in Section 4, VLM accuracy is prompt- and setting-dependent, and can vary substantially, so relying on it blindly would be inappropriate. A formal convergence analysis under VLM-induced noise is outside the present scope, and we position the method as an empirically validated effective scheme whose formal analysis is left to future work.

6 Experiment Results

6.1 Overall Performance

EngageNet dataset: Table 3a shows that conventional methods (LSTM, TCN, CNN-LSTM) achieve accuracies around 67%–68%, while TCCT-Net reaches 68.91% with Head Pose features. With VLM guidance, Head Pose accuracy slightly drops to 68.16%, whereas for Eye Gaze a modest improvement is observed (64.61% vs. 64.33%). In addition, the results with combinations of features further demonstrate the strength of VLM approach: when using Head Pose with Action Units, TCCT-Net with VLM outperforms its baseline (66.29% vs. 65.08%), and similarly, for the Action Units and Eye Gaze combination, VLM guidance brings the accuracy to 65.73% compared to 65.17%. It suggests that while TCCT-Net is robust on its own, incorporating VLM guidance is very useful as it consistently leverages complementary information from

Feature	Transformer	LSTM	TCN	CNN-LSTM	TCCT-Net	TCCT-Net + VLM
EG	55.45	61.34	63.03	60.78	64.33	64.61
HP	–	67.41	67.51	67.88	68.91	<u>68.16</u>
AU	–	62.75	64.24	62.00	66.29	66.01
EG + HP	64.45	65.17	–	–	65.64	<u>65.55</u>
HP + AU	–	–	–	–	<u>65.08</u>	66.29
AU + EG	–	–	–	–	<u>65.17</u>	65.73

(a) **EngageNet**: Classification accuracies (%) for Transformer, LSTM, TCN, and CNN-LSTM from [51]; TCCT-Net from [9].

Feature	ST	MT	TL	TCCT-Net	TCCT-Net + VLM
EG	0.185	0.185	0.185	<u>0.357</u>	0.423
HP	–	–	–	<u>0.185</u>	0.405
AU	–	–	–	<u>0.316</u>	0.343
EG + HP	0.181	<u>0.218</u>	0.165	0.185	0.434
HP + AU	–	–	–	<u>0.185</u>	0.316
EG + AU	–	–	–	<u>0.339</u>	0.386
EG + HP + AU	0.256	<u>0.306</u>	0.216	0.185	0.384
EG + HP + AU + LMK	<u>0.226</u>	0.208	0.293	–	–
EG + HP + AU + LMK + PDM	0.305	<u>0.299</u>	0.284	–	–

(b) **DREAMS**: Weighted F1 scores for various feature combinations. Results for Single-task (ST), Multi-task (MT), and Transfer Learning (TL) from [54]; TCCT-Net from [9].

Method	5 sec	10 sec	20 sec
SVM	0.570	0.610	0.600
XGBoost	0.620	0.640	0.650
DNN	0.620	0.640	0.680
TCCT-Net Methods (no windowing needed)			
TCCT-Net (HP)		<u>0.72</u>	
TCCT-Net + VLM (HP)		0.74	

(c) **PAFE**: Weighted F1 scores for different window sizes (SVM, XGBoost, DNN) and TCCT-Net methods (no windowing). Results for SVM, XGBoost, and DNN from [53]; TCCT-Net from [9].

Table 3: Combined comparison of results across the *EngageNet*, *DREAMS*, and *PAFE* datasets. Abbreviations: **EG** = Eye Gaze, **HP** = Head Pose, **AU** = Action Units, **LMK** = Landmarks, **PDM** = Point Distribution Model, **ST** = Single-Task, **MT** = Multi-Task, **TL** = Transfer Learning. **Bold** indicates the best performance and underline the second-best.

different feature combinations to boost performance. To give more insight into class-wise performance, per-class confusion matrices for the vanilla TCCT-Net and our TCCT-Net + VLM are provided in Appendix I.

DREAMS dataset: Table 3b reports weighted F1 scores for various behavioral feature combinations. Notably, TCCT-Net with VLM guidance outperforms both the vanilla TCCT-Net and baseline methods even with fewer features. Using only Eye Gaze features, TCCT-Net + VLM achieves a weighted F1 of 0.423, and combining Eye Gaze with Head Pose reaches 0.434. These results show that integrating VLM guidance improves performance and allows superior results with a reduced feature set, highlighting the effectiveness of our noise-robust training strategy. It is worth noting that TCCT-Net achieves a score of 0.185 when using Head Pose alone or in combinations (with Action Units, Eye Gaze or both), indicating that the Head Pose feature by itself does not contribute to an improvement in performance. Moreover, the experiments reveal that fusing additional features does not always enhance the weighted F1 score, regardless of the method employed, which emphasizes the challenge of identifying complementary feature combinations.

PAFE dataset: Traditional models (SVM, XGBoost, DNN) yield weighted F1 scores between 0.57 and 0.68, depending on the window size. In contrast, TCCT-Net with Head Pose features reaches 0.72 (Table 3c), and with VLM-guided strategy, it improves to 0.74, demonstrating the power of such approaches in addressing subjective labeling. Importantly, this improvement comes without any hand-crafted fixed-length time windows, simplifying the pipeline.

Across three datasets, the enhanced TCCT-Net, integrating curriculum learning, label refinement, weighted augmentation, and VLM guidance, consistently improves robustness and generalization. Although performance gains vary by dataset, our approach mitigates noisy, subjective labels, highlighting its potential for engagement recognition. In absolute terms, the improvements are modest (maximum of +1.21% on *EngageNet* accuracy and +0.22/ +0.06 weighted F1 on *DREAMS* / *PAFE*), but they are obtained repeatedly across three independently collected datasets and several feature sets. With the exception of a single configuration (Head Pose alone on *EngageNet*, -0.75%), we do not observe notable drops, suggesting that the VLM-guided curriculum adds robustness to label subjectivity without destabilizing a strong baseline. These results therefore support the use of our framework as a reliability- and robustness-oriented enhancement rather than a mechanism for large accuracy jumps on any single benchmark.

6.2 Ablation Study

Table 4 shows the impact of VLM-guided, curriculum learning-based training on TCCT-Net across multiple datasets. On the *EngageNet* dataset, using only Eye Gaze features gives a baseline accuracy of 64.33%. With a one-stage VLM-guided approach, accuracy drops to 59.29%, but the two-stage curriculum raises it to 64.61%. Similarly, for the combination of Head Pose and Action Units, performance improves from 65.08% to 66.29%, and for Eye Gaze with Action Units from 65.17% to 65.73%. In the *DREAMS* dataset, the weighted F1 score for Eye Gaze goes from 0.357 to 0.423, and

for Head Pose it jumps from 0.185 to 0.405. Combining features, such as Eye Gaze with Head Pose, increases the score from 0.185 to 0.434, with similar improvements seen for the other combinations. On the *PAFE* dataset, using Head Pose features, the weighted F1 score improves from a baseline of 0.72 to 0.74 with the two-stage method, compared to 0.65 for the one-stage approach. Overall, these results show that our two-stage VLM-guided curriculum not only recovers but also improves the baseline performance, enhancing the model’s ability to handle ambiguous and noisy labels and leading to a more robust and effective system across all datasets.

Feature	TCCT-Net	TCCT-Net + VLM (CL: one stage)	TCCT-Net + VLM (CL: two stage)
<i>EngageNet</i> (Accuracy %)			
EG	64.33	59.29 -5.04	64.61 $+0.28$
HP + AU	<u>65.08</u>	61.62 -3.46	66.29 $+1.21$
EG + AU	<u>65.17</u>	61.44 -3.73	65.73 $+0.56$
<i>DREAMS</i> (Weighted F1)			
EG	0.357	0.317 -0.040	0.423 $+0.066$
HP	0.185	<u>0.301</u> $+0.116$	0.405 $+0.220$
AU	0.316	<u>0.334</u> $+0.018$	0.343 $+0.027$
EG + HP	0.185	<u>0.288</u> $+0.103$	0.434 $+0.249$
EG + AU	0.339	0.292 -0.047	0.386 $+0.047$
HP + AU	0.185	<u>0.284</u> $+0.099$	0.316 $+0.131$
EG + HP + AU	0.185	<u>0.354</u> $+0.169$	0.384 $+0.199$
<i>PAFE</i> (Weighted F1)			
HP	<u>0.72</u>	0.65 -0.07	0.74 $+0.02$

Table 4: Ablation study comparing baseline TCCT-Net [9] with our curriculum learning-based, VLM-guided training, showing incremental performance gains on *EngageNet*, *DREAMS*, and *PAFE* datasets. Abbreviations: **EG** = Eye Gaze, **HP** = Head Pose, **AU** = Action Units, **CL** = Curriculum Learning. **Bold** indicates the best performance and underline the second-best.

7 Conclusion

This study has addressed the challenge of subjective annotations in engagement recognition by integrating VLM guidance into both dataset refinement and training strategies. By employing a structured questionnaire, our method effectively distinguishes high- from low-reliability samples, thereby enabling the selective training of classical models on cleaner data. Furthermore, introduction of a curriculum learning framework paired with soft label refinement allows models to progressively adapt to ambiguous cases, reducing overfitting to annotator-specific biases. Experimental results across multiple datasets demonstrate that our approach not only enhances

overall classification accuracy but also improves robustness against label noise. Our ablation studies reveal that careful design of prompt detail and temporal context is crucial for maximizing VLM performance and downstream model benefits. Moving forward, we envision further integration of interactive VLM-assisted annotation protocols and development of evaluation metrics that better capture the inherent subjectivity of engagement. We will also validate our framework on the gaming benchmarks (including *AGAIN* [66] and others) to assess its generalizability in dynamic, interactive environments. Ultimately, our work sets the stage for more reliable affective computing systems by advancing data quality and noise-robust learning in subjective emotion recognition tasks.

Declarations

Acknowledgments: The authors would like to acknowledge the support of the Center for Machine Vision and Signal Analysis, University of Oulu.

Funding: This research received no external funding.

Competing interests: The authors have no competing interests to declare.

Availability of data and material: The datasets used in this study were obtained from the original authors upon request and are not publicly available due to data sharing restrictions. The authors do not have permission to redistribute these datasets. Data can be requested directly from the respective dataset owners.

Authors’ contributions: *Alexander Vedernikov:* Conceptualization, Data Curation, Methodology, Software, Investigation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing. *Puneet Kumar:* Investigation, Formal Analysis, Writing – Review & Editing. *Haoyu Chen:* Investigation, Formal Analysis, Supervision, Writing – Review & Editing. *Tapio Seppänen:* Supervision, Writing – Review & Editing. *Xiaobai Li:* Supervision, Writing – Review & Editing.

Ethics approval: Not applicable.

Consent for publication: Not applicable.

Appendix A Computational Settings

Computational tasks for VLM-based engagement classification (Section 3.2) were conducted on a supercomputer, utilizing between one and four Nvidia V100 GPUs, depending on model size. Classical computer vision methods (Section 5) were processed

on a supercomputer, utilizing 32 SMT-enabled CPU cores and 32 GB RAM, complemented by an AMD MI250 GPU with 128GB of memory. We ensured consistency of the experiments by utilizing the identical computational environment.

Appendix B Conversation Mode Prompt Details

In our few-shot/image Conversation mode experiments, InternVL2 was provided with a sequence of prompts to guide its analysis and classification of engagement levels. The prompting protocol consisted of the following stages:

1. Example provision stage: model was provided with a few example images for each engagement class, along with corresponding affective descriptions. These examples served to establish a reference for the further analysis.

2. Analysis stage: next, model received a prompt instructing it to analyze a new image based on the previously provided examples: *“Using the previous examples as a reference, analyze the person’s affective state in the new image.”*

3. Classification stage: finally, the model was prompted to output a numeric engagement level, where mapping is defined as: *“Based on your analysis and observations, classify the engagement level in the new image. Output only the engagement level as a single number: 0 - Not engaged, 1 - Low engagement, 2 - Moderate engagement, 3 - High engagement. Provide no additional text or explanation in the output.”*

Appendix C Full Questionnaire Details

Table C1 presents the complete set of YES/NO questions used to guide the VLM analysis for engagement assessment. The questions are organized into four key categories: (1) Eye Gaze and Attention, (2) Facial Expressions, (3) Body Language and Posture, and (4) Distractions and External Activities. These targeted prompts were designed to capture specific, observable behaviors in order to reduce ambiguity and inter-annotator variability.

Appendix D VLM-based Engagement Classification

Table D2 summarizes the zero- and few-shot performance of VLMs on the *EngageNet* validation set in both Single Question and Conversation modes. In Single-Question mode, the LLaVA model achieved the highest accuracy of 40.06% with two images, while InternVL2 reached a peak of 28.66%. InternVL2’s conversation mode (absent in LLaVA) shows a modest improvement, with accuracy increasing from 23.44% to 27.00% in few-shot, single-image settings. This suggests that a single frame does not suffice for accurate engagement detection. Sequential questioning, by utilizing temporal context, shifts the analysis from isolated snapshots to a more dynamic and integrated evaluation of engagement. We then evaluated InternVL2 in conversation mode using video input, comparing the 8B and 26B models. The 8B model achieved 35.20% accuracy, while the 26B model reached only 29.97%, suggesting that larger models may introduce additional noise. It further reached 45.28% with structured question guidance. Thus, we adopt InternVL2 (8B) for future experiments. In summary, our experiments demonstrate that without clear, targeted prompts, VLMs can

Specific Area and Questions
Category 1: Eye Gaze and Attention Q1: Does the student frequently shift their gaze significantly away from the screen or primary point of focus? Q2: Does the student’s eye gaze remain concentrated in one consistent direction while watching the video? Q3: Does the student keep their attention directed at a single point during most of the video? Q4: Does the student’s eye movement stay within a limited area, suggesting focused attention on the content? Q5: Does the student frequently glance away from the screen?
Category 2: Facial Expressions Q6: Does the student appear uninterested or bored based on their facial expressions? Q7: Does the student show signs of fatigue or sleepiness, such as yawning or nodding off? Q8: Does the student barely open their eyes, appearing tired? Q9: Does the student appear to like or enjoy the content being presented?
Category 3: Body Language and Posture Q10: Is the student fidgeting restlessly in their chair? Q11: Does the student exhibit a passive posture with minimal movement? Q12: Does the student lean forward, appearing highly engaged with the content?
Category 4: Distractions and External Activities Q13: Is the student using a phone or other device during the lecture? Q14: Does the student talk to others or engage in activities unrelated to the lecture? Q15: Are there any signs of the student being distracted from the lecture?
Table C1: Categorized areas and their associated questions for engagement assessment.

vary in accuracy by as much as 15%, highlighting the inherent challenges in engagement estimation due to both model limitations and human label ambiguity. These results emphasize the critical role of prompt clarity in achieving robust performance in nuanced tasks.

Appendix E Enhancing Data Quality with VLM Filtering

Although the structured questionnaire design from Appendix C helps target observable engagement cues, further refinement is required to reduce subjective label inconsistencies. Therefore, we used VLMs to divide the Validation set of the used datasets into so-called Accepted and Rejected subsets, enabling an in-depth evaluation of classical computer vision models. This approach (1) assesses how label quality impacts classical computer vision model performance, particularly in tasks as subjective as engagement classification, (2) highlights the impact of label quality on model robustness, and (3) provides insights into dataset reliability.

We employed the InternVL2 (8B) model to analyze the *EngageNet* and *PAFE* datasets, then split each into Accepted and Rejected subsets. The process was structured around a two-round conversation approach using video inputs. Video input involved eight frames extracted evenly across the entire video for the *EngageNet* dataset and 20 frames for the *PAFE* dataset. In the first round, the model responded

Model	Input	Validation Accuracy (%)
Single Question Mode		
InternVL2 (8B)	0-shot/1 image	23.44
	0-shot/2 images	28.66
	0-shot/3 images	27.45
	0-shot/4 images	24.65
	0-shot/5 images	24.28
LLaVA 1.6 (7B)	0-shot/1 image	33.05
	0-shot/2 images	40.06
	0-shot/3 images	35.48
	0-shot/4 images	35.11
	0-shot/5 images	31.84
VILA 1.5 (8B)	0-shot/1 image	19.89
	0-shot/2 images	23.72
	0-shot/3 images	21.85
	0-shot/4 images	23.34
	0-shot/5 images	21.94
Conversation Mode		
InternVL2 (8B)	few-shot/1 image	27.00
	0-shot/video	35.20
	structured question guidance	<u>45.28</u>
InternVL2 (26B)	0-shot/video	29.97
LLaVA1.6 (7B)	X	X
VILA1.5 (8B)	X	X

Table D2: Zero-shot and few-shot performance of VLMs on *EngageNet* validation set: four-class classification accuracy using single-question and conversation modes. **Bold** denotes best result per model in each mode, while **bold and underline** denotes the best result across all experiments. X - Conversation mode isn't supported.

to a comprehensive set of YES/NO questions (see Table C1) designed to detect key engagement indicators, such as eye movement consistency, signs of distraction, body posture, and expressions of interest or disinterest. After gathering this information, the model proceeded to a second round where it assigned each sample an engagement level on a scale from 0 to 3 for *EngageNet* dataset case, and 0 to 1 for *PAFE* dataset case. Based on the model's predictions, we divided the Validation set of each dataset into an Accepted subset, where model classifications matched the ground truth labels, and Rejected subset, where discrepancies occurred.

Appendix F Comparative Analysis of Classical Computer Vision Models

In engagement recognition, classical computer vision models may overfit to annotator biases, particularly when faced with subjective and inconsistent labels. VLM-based

filtering aims to assess these classical computer vision models on subsets with varying label clarity, providing an opportunity to explore the effects of annotation reliability on model accuracy and to reveal dependencies on high-consensus labels. This part of experiment explores whether certain portions of the dataset better align with classical computer vision model expectations, offering an indirect assessment of annotation reliability. By examining performance across these subsets and using classical computer vision models (TCCT-Net [9], ResNet + TCN [61], EfficientNet + LSTM [62], EfficientNet + Bi-LSTM [62]; each trained on the *EngageNet* and *PAFE* datasets) we assess robustness across the entire validation set, as well as its Accepted and Rejected subsets. This method provides an approach, where VLM models help filter ambiguous labels and highlight areas for data quality improvement, systematically identifying and potentially filtering out unreliable annotations.

Appendix G Human Performance in Engagement Analysis

To further examine the reliability of ground truth engagement labels, we conducted a human evaluation experiment, measuring annotator consistency with predefined labels and comparing it against VLM performance, highlighting areas of alignment and ambiguity. It provides a benchmark for assessing model performance on subjective tasks and highlights inherent challenges in engagement classification.

Participant selection: We selected 10 participants experienced in affective computing and familiar with emotion analysis, aiming to reduce variability and establish a realistic baseline for annotating subjective engagement data. Each participant received a sweet reward for participation.

Engagement classification task: Participants were presented with 50 video samples from the Validation set of the *EngageNet* dataset and tasked with categorizing each one into predefined engagement levels (0 to 3). To improve consistency in this subjective task, annotators followed detailed guidelines provided by the dataset authors. Each video sample could be viewed up to two times, with the goal of assigning engagement levels that matched or closely approximated the ground truth labels.

Tolerance-based accuracy metric: To account for slight discrepancies in engagement levels, we introduced a ‘tolerance-based accuracy’ metric, allowing a margin of error of ± 1 level in classification, in addition to the standard *EngageNet* accuracy metric. This approach provided insight into how closely human annotators could approximate the correct label without an exact match, offering a realistic perspective on performance in this subjective task.

Appendix H Extended Ablation Studies on the Validation Set

To investigate the impact of subjective annotations and potential model biases, we conducted an ablation study analyzing both InternVL2 (8B) using a video-based zero-shot approach and the classical computer vision model TCCT-Net. It examined how

variations in input configurations (specifically, the number of frames per video and the selection of specific questions from our structured guidance list (see Table C1) affect the performance of both models in engagement classification. By assessing these factors, we aimed to identify optimal configurations that could mitigate the influence of subjective labels, thereby enhancing model robustness and improving classification accuracy for both VLMs and classical computer vision approaches in the context of subjective emotion recognition tasks.

Frames / Questions	Full Validation Accuracy [%]	Rejected Accuracy [%] / Samples	Accepted Accuracy [%] / Samples
F2	43.04	62.13 -6.78 610	77.87 +8.96 461
F4	42.11	61.45 -7.46 620	79.16 +10.25 451
F8	45.28	58.02 -10.89 586	82.06 +13.15 485
F10	44.44	56.13 -12.78 595	84.87 +15.96 476
F16	50.70	50.19 -18.72 528	87.11 +18.20 543
F20	47.99	52.78 -16.13 557	86.38 +17.47 514
Q3	36.23	69.84 +0.93 683	67.27 -1.64 388
Q6	33.99	71.57 +2.66 707	63.74 -5.17 364
Q9	43.79	55.48 -13.43 602	86.14 +17.23 469
Q12	45.56	57.63 -11.28 583	82.38 +13.47 488
Q15	50.70	50.19 -18.72 528	87.11 +18.20 543

Table H3: Performance of VLM over different frame counts (F) and prompt question (Q) counts. Red/green values indicate accuracy differences from the TCCT-Net [9] baseline of 68.91%. Total number of samples in the Validation set of *EngageNet* is 1071. **Bold** indicates the best performance.

Number of input frames. The results, presented in Table H3, reveal that increasing the number of frames enhances the performance of both the VLM and TCCT-Net models. For the VLM, accuracy increases slightly with more frames, peaking at 50.70% with 16 frames before declining at 20 frames (47.99%). This indicates that additional temporal information from more frames allows the VLM to capture dynamic engagement cues more effectively, up to a point where adding more frames may introduce redundancy or noise. The slight decline beyond 16 frames suggests a saturation point where the model’s capacity to utilize additional information is exceeded. Similarly, the accuracy on Accepted subset for TCCT-Net improves as the frame count increases, reaching 87.11% with 16 frames. This improvement underscores the importance of temporal dynamics in engagement recognition, as more frames provide a richer temporal context for the model to learn from. However, for Rejected subset, TCCT-Net’s accuracy decreases with more frames, suggesting that increased frame counts may amplify the impact of label noise on these samples. The reduction in the number of Rejected samples as frames increase (from 610 with 2 frames to 528 with 16 frames) supports

the idea that increased temporal data helps reduce ambiguity in engagement classification. Overall, these findings indicate that there is an optimal number of frames (around 16 in this case) that balances the benefits of additional temporal information with the potential drawbacks of redundancy and increased computational load.

Number of questions in the prompt. We show process of forming a logical approach to selecting which questions to include when reducing list from 15 questions to fewer numbers for ablation study (Table H4). It ensures that the selection process is systematic and grounded in the relevance of each question to engagement prediction. To systematically reduce the number of questions, we: (1) Group the questions into categories based on the aspects of engagement they assess. (2) Identify possible redundancy among questions. (3) Rank the questions based on their importance and uniqueness. (4) Select questions that cover the broadest range of engagement indicators with minimal overlap. The results of ablation study are presented in Table H3. The analysis reveals that question count significantly impacts the performance of both VLM and TCCT-Net. For VLM, accuracy steadily increases from 36.23% with 3 questions to 50.70% with the full 15 questions, indicating that a broader set of questions aids in capturing engagement cues. Notably, TCCT-Net achieves 87.11% accuracy on accepted samples with all 15 questions, while accuracy drops as question count reduces, reaching 67.27% with only 3 questions. This indicates that a more detailed and diverse prompt enhances the effectiveness of the VLM filtering, resulting in a higher-quality Accepted subset that improves the downstream performance of TCCT-Net. The number of Rejected samples also decreases as the question count increases, highlighting the role of comprehensive prompts in identifying and filtering out samples with inconsistent or noisy annotations. These findings emphasize the importance of carefully designing prompts to include sufficient detail for capturing the multifaceted nature of engagement, thereby improving model robustness and classification accuracy.

Appendix I Per-Class Confusion Matrices

Across both feature-set configurations, integrating VLM guidance improves class discrimination, increasing true positives (diagonal entries) and cutting adjacent-class confusions. In the Action Units + Head Pose case (Fig. I1), class 2 ('Engaged') true positives rise from 66 to 108 (+64%), while misclassifications into classes 1 ('Barely Engaged') and 3 ('Highly Engaged') each fall by 17 and 19 points, respectively, indicating clearer mid-level engagement boundaries. Similarly, in the Action Units + Eye

Number of Questions	Questions (by Number) Included
15	All questions (1 to 15)
12	1, 2, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
9	1, 2, 6, 7, 9, 10, 12, 13, 14
6	1, 2, 9, 10, 12, 13
3	1, 12, 13

Table H4: Ablation study question configurations: subsets of questions selected for each experiment.

True Class	Not Engaged	Barely Engaged	Engaged	Highly Engaged
	93	13	10	16
	14	16	27	40
	28	16	66	163
Highly Engaged	10	6	31	522
Predicted Class				

(a) TCCT-Net

True Class	Not Engaged	Barely Engaged	Engaged	Highly Engaged
	83	14	23	12
	9	10	46	32
	11	10	108	144
Highly Engaged	4	1	55	509
Predicted Class				

(b) TCCT-Net + VLM

Fig. I1: Confusion matrices on the *EngageNet* validation set using Action Units + Head Pose features. (a) Baseline TCCT-Net; (b) TCCT-Net augmented with VLM-guided curriculum learning and soft-label refinement.

True Class	Not Engaged	Barely Engaged	Engaged	Highly Engaged
	95	20	6	11
	12	26	19	40
	22	26	69	156
Highly Engaged	22	8	31	508
Predicted Class				

(a) TCCT-Net

True Class	Not Engaged	Barely Engaged	Engaged	Highly Engaged
	99	9	16	8
	20	11	31	35
	25	8	101	139
Highly Engaged	20	1	55	493
Predicted Class				

(b) TCCT-Net + VLM

Fig. I2: Confusion matrices on the *EngageNet* validation set using Action Units + Eye Gaze features. (a) Baseline TCCT-Net; (b) TCCT-Net augmented with VLM-guided curriculum learning and soft-label refinement.

Gaze case (Fig. 12), class 2 (‘Engaged’) true positives jump from 69 to 101 (+46%), and class 3 (‘Highly Engaged’) from 156 to 139 (−17 false positives), showing both stronger recall and reduced misclassifications into neighboring labels. Overall off-diagonal errors decrease by roughly 10–15% across most classes, underscoring that VLM-filtered training both tightens core predictions and mitigates ambiguous label noise. Complementary qualitative analysis of these confusion matrices and representative clips indicates that the remaining errors are dominated by borderline cases. Many segments in which participants quietly follow the video or read questions with stable gaze and minimal facial movement are annotated as “Highly Engaged”, yet the VLM-guided model tends to assign them to “Engaged”, as the visible cues of enthusiasm (frequent expressions, stronger nods, pronounced forward lean) are weak. We also observe subject-specific differences in expressiveness: for some individuals, highly engaged behaviour remains visually subtle and is systematically under-estimated, whereas more animated participants are more reliably recognised as “Highly Engaged”. Taken together, these observations suggest that VLM guidance sharpens mid-scale boundaries while leaving residual ambiguity almost exclusively between neighbouring labels.

References

- [1] Noroozi, F., Corneanu, C.A., Kaminska, D., Sapinski, T., Escalera, S., *et al.*: Survey on Emotional Body Gesture Recognition. *IEEE Transactions on Affective Computing* **12**(2), 505–523 (2021)
- [2] Ekman, P., *et al.*: An Argument for Basic Emotions. *Cognition & Emotion* **6**(3-4), 169–200 (1992)
- [3] Kumar, P., *et al.*: Computational Analysis of Stress, Depression and Engagement in Mental Health. *arXiv:2403.08824* (2024)
- [4] Plank, B.: The ‘Problem’ of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *arXiv:2211.02570* (2022) [2211.02570](#)
- [5] Kumar, P., *et al.*: Multimodal Interpretable Depression Analysis using Visual Physiological Audio and Textual Data. In: *WACV* (2025)
- [6] Tu, Y., *et al.*: Learning from noisy labels with decoupled meta label purifier. In: *CVPR*, pp. 19934–19943 (2023)
- [7] Wen, K., *et al.*: Benign Overfitting in Classification: Provably Counter Label Noise with Larger Models. In: *ICLR* (2023)
- [8] Kumar, P., *et al.*: VISTANet: VISual Spoken Textual Additive Net for interpretable multimodal emotion recognition. *arXiv:2208.11450* (2024)
- [9] Vedernikov, A., *et al.*: TCCT-Net: Two-Stream Network Architecture for Fast and Efficient Engagement Estimation via Behavioral Feature Signals. In: *CVPR*

Workshops, pp. 4723–4732 (2024)

- [10] Chen, T., Guo, Y., Hao, S., Hong, R.: Exploring Self-Attention Graph Pooling With EEG-Based Topological Structure and Soft Label for Depression Detection. *IEEE Transactions on Affective Computing* **13**(4), 2106–2118 (2022)
- [11] Khan, N., et al.: A Novel Loss Function Utilizing Wasserstein Distance to Reduce Subject-Dependent Noise for Generalizable Models in Affective Computing. *arXiv:2308.10869* (2023)
- [12] Zhou, R., Zhang, Z., Fu, H., Zhang, L., *et al.*: PR-PL: A Novel Prototypical Representation Based Pairwise Learning Framework for Emotion Recognition Using EEG. *IEEE Transactions on Affective Computing* **15**(2), 657–670 (2024)
- [13] Gera, D., *et al.*: Noisy Annotations Robust Consensual Collaborative Affect Exp. Recognition. In: *ICCV Workshops*, pp. 3578–3585 (2021)
- [14] Wu, J., Dang, T., Sethu, V., Ambikairajah, E.: A novel sequential monte carlo framework for predicting ambiguous emotion states. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8567–8571 (2022). <https://doi.org/10.1109/ICASSP43922.2022.9746350>
- [15] Chou, H.-C., Lee, C.-C.: Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5886–5890 (2019). <https://doi.org/10.1109/ICASSP.2019.8682170>
- [16] Wu, Y.-T., Chang, C.-F., Lee, M.-J., Kumar, A., Ruan, J., Smith, J.: Can modelling inter-rater ambiguity lead to noise-robust continuous emotion predictions? In: *Proceedings of Interspeech 2024* (2024)
- [17] Karim, N., *et al.*: UniCon: Combating Label Noise Through Uniform Selection and Contrastive Learning. In: *CVPR*, pp. 9676–9686 (2022)
- [18] McKinzie, B., Gan, Z., et al.: MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. *arXiv:2403.09611* (2024) [arXiv:2403.09611](https://arxiv.org/abs/2403.09611) [cs.CV]
- [19] Yang, Q., Ye, M., Du, B.: EmoLLM: Multimodal Emotional Understanding Meets Large Language Models (2024)
- [20] Teotia, J., Zhang, X., Mao, R., Cambria, E.: Evaluating vision language models in detecting learning engagement. In: *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 496–502 (2024). <https://doi.org/10.1109/ICDMW65004.2024.00069>

- [21] Wang, D., Yang, C., Chen, G.: Using Vision Language Models to Detect Students' Academic Emotion through Facial Expressions (2025). <https://arxiv.org/abs/2506.10334>
- [22] Khan, S., Safa, S.: Revisiting Annotations in Online Student Engagement. In: International Conference on Computing and Data Engineering, pp. 111–117 (2024)
- [23] Abedi, A., *et al.*: Detecting Disengagement in Virtual Learning as an Anomaly using Temporal Convolutional Net. Autoencoder. Signal, Image and Video Processing **17**(7), 3535–3543 (2023) <https://doi.org/10.1007/s11760-023-02578-z>
- [24] Liao, J., *et al.*: Deep Facial Spatiotemporal Net For Engagement Prediction in Online Learning. Applied Intelligence **51**(10) (2021) <https://doi.org/10.1007/s10489-020-02139-8>
- [25] Zhong, P., Wang, D., Miao, C.: EEG-Based Emotion Recognition Using Regularized Graph Neural Networks. IEEE Transactions on Affective Computing **13**(3), 1290–1301 (2022) <https://doi.org/10.1109/TAFFC.2020.2994159>
- [26] Jiang, J., *et al.*: Boosting Facial Exp. Recognition by A Semi-Supervised Progressive Teacher. IEEE Transactions on Affective Computing (2023) <https://doi.org/10.1109/TAFFC.2021.3131621>
- [27] Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International Conference on Machine Learning, pp. 312–321 (2019). PMLR
- [28] Song, H., Kim, M., Park, D., Shin, Y., Lee, J.-G.: Learning from Noisy Labels with Deep Neural Networks: A Survey (2022). <https://arxiv.org/abs/2007.08199>
- [29] Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I.W., Kwok, J.T., Sugiyama, M.: A Survey of Label-noise Representation Learning: Past, Present and Future (2021). <https://arxiv.org/abs/2011.04406>
- [30] Northcutt, C.G., *et al.*: Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv:2103.14749 (2021) [arXiv:2103.14749](https://arxiv.org/abs/2103.14749) [stat.ML]
- [31] Tan, Y., Xia, H., Song, S.: Robust Consistency Learning for Facial Expression Recognition Under Label Noise. Visual Computer (2024) <https://doi.org/10.1007/s00371-024-03558-1>
- [32] Liu, Y., *et al.*: Uncertain Facial Expression Recognition via Multi-Task Assisted Correction. IEEE Transactions on Multimedia **26** (2024) <https://doi.org/10.1109/TMM.2023.3301209>
- [33] Mao, L., Yan, Y., Xue, J.-H., Wang, H.: Deep Multi-Task Multi-Label CNN for Effective Facial Attribute Classification. IEEE Transactions on Affective

- Computing (2022) <https://doi.org/10.1109/TAFFC.2020.2969189>
- [34] Jiang, J., Deng, W.: Improving Multi-Label Facial Expression Recognition With Consistent and Distinct Attentions. *IEEE Transactions on Affective Computing* **15**(3), 1279–1288 (2024) <https://doi.org/10.1109/TAFFC.2023.3333874>
 - [35] Chen, Y., Li, J., Shan, S., Wang, M., *et al.*: From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos. *IEEE Transactions on Affective Computing* (2024) <https://doi.org/10.1109/TAFFC.2024.3453443>
 - [36] Washington, P., Kalantarian, H., Kent, J., *et al.*: Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels. *Cognitive Computation* **13**(5), 1363–1373 (2021) <https://doi.org/10.1007/s12559-021-09936-4>
 - [37] Jin, Z., *et al.*: A Psychological Evaluation Method Incorporating Noisy Label Correction Mechanism. *Soft Computing* **28**(11-12) (2024) <https://doi.org/10.1007/s00500-023-09479-w>
 - [38] Liu, Y., *et al.*: Uncertain Label Correction via Auxiliary Action Unit Graphs for Facial Expression Recognition. In: *ICPR*, pp. 777–783 (2022). <https://doi.org/10.1109/icpr56361.2022.9956650>
 - [39] Zhang, G., Etemad, A.: Partial Label Learning for Emotion Recognition from EEG. *arXiv:2302.13170* (2023)
 - [40] Huang, Z., Liu, C., Dong, Y., Su, H., Zheng, S., Liu, T.: Machine Vision Therapy: Multimodal Large Language Models Can Enhance Visual Robustness via Denoising In-Context Learning (2024). <https://arxiv.org/abs/2312.02546>
 - [41] Wang, H., Huang, Z., Lin, Z., Liu, T.: NoiseGPT: Label noise detection and rectification through probability curvature. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*, vol. 37, pp. 120159–120183. Curran Associates, Inc., ??? (2024). <https://doi.org/10.52202/079017-3819> . https://proceedings.neurips.cc/paper_files/paper/2024/file/d95cb79a3421e6d9b6c9a9008c4d07c5-Paper-Conference.pdf
 - [42] Zhang, Y., Zhang, L.: AdaNeg: Adaptive Negative Proxy Guided OOD Detection with Vision-Language Models (2024). <https://arxiv.org/abs/2410.20149>
 - [43] Chen, T., Zhang, D., Lee, D.-J.: A New Joint Training Method for Facial Expression Recognition with Inconsistently Annotated and Imbalanced Data. *Electronics* **13**(19) (2024)
 - [44] Shvetsova, N., Kukleva, A., *et al.*: HowToCaption: Prompting LLMs to Transform Video Annotations at Scale. *arXiv: 2310.04900* (2024)

- [45] Zhang, Y., , et al.: Affective Computing in the Era of Large Language Models: A Survey from the NLP Perspective. arXiv:2408.04638 (2024)
- [46] Xie, H., Peng, C., et al.: EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning. arXiv:2404.16670 (2024)
- [47] Nadeem, M., Sohail, S.S., *et al.*: Vision-Enabled Large Language and Deep Learning Models for Image-Based Emotion Recognition. *Cognitive Computation* **16**(5), 2566–2579 (2024)
- [48] Hu, G., Lan, X., et al.: Towards Unified Facial Action Unit Recognition Framework by Large Language Models. arXiv:2409.08444 (2024)
- [49] Lu, H., Niu, X., et al.: GPT as Psychologist? Preliminary Evaluations for GPT-4V on Visual Affective Computing. arXiv:2403.05916 (2024)
- [50] Etesam, Y., Nilay Yalçın, et al.: Contextual Emotion Recognition using Large Vision Language Models. arXiv:2405.08992 (2024)
- [51] Singh, M., *et al.*: Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines. In: ICMI, pp. 174–182 (2023). <https://doi.org/10.1145/3577190.3614164>
- [52] Dhall, A., Singh, M., Goecke, R., Gedeon, T., Zeng, D., *et al.*: EmotiW 2023: Emotion Recognition in the Wild Challenge, pp. 746–749 (2023)
- [53] Lee, T., Kim, D., Park, S., *et al.*: Predicting Mind-Wandering with Facial Videos in Online Lectures. In: CVPR Workshops, pp. 2103–2112 (2022)
- [54] Singh, M., Sharma, G., *et al.*: DREAMS: Diverse Reactions of Engagement and Attention Mind States Dataset. In: ICPR, pp. 163–179 (2024)
- [55] Chen, Z., et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. arXiv:2312.14238 (2023)
- [56] Chen, Z., Wang, W., Tian, H., Ye, S., et al.: How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. arXiv:2404.16821 (2024)
- [57] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., et al.: LLaVA-NeXT: Improved Reasoning, OCR And World Knowledge (2024). <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- [58] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 (2023)
- [59] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. NeurIPS (2023)

- [60] Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., et al.: VILA: On Pre-training for Visual Language Models. arXiv:2312.07533 (2023)
- [61] Abedi, A., Khan, S.S.: Improving State-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. In: Conference on Robots and Vision, pp. 151–157 (2021)
- [62] Selim, T., *et al.*: Students Engagement Level Detection in Online e-Learning Using Hybrid EfficientNetB7 Together With TCN, LSTM, and Bi-LSTM. IEEE Access **10**, 99573–99583 (2022)
- [63] Bengio, Y., *et al.*: Curriculum Learning. In: International Conference on Machine Learning (ICML), pp. 41–48 (2009). <https://doi.org/10.1145/1553374.1553380>
- [64] Szegedy, C., *et al.*: Rethinking the Inception Architecture for Computer Vision. In: CVPR (2016)
- [65] Zhou, Z.-H.: Ensemble methods: Foundations and algorithms. CRC Press, Taylor & Francis Group, 1–18 (2012) <https://doi.org/10.1201/b12207>
- [66] Melhart, D., Liapis, A., Yannakakis, G.N.: The arousal video game annotation (again) dataset. IEEE Transactions on Affective Computing **13**(4), 2171–2184 (2022) <https://doi.org/10.1109/taffc.2022.3188851>