ConfusionBench: An Expert-Validated Benchmark for Confusion Recognition and Localization in Educational Videos

Anonymous CVPR submission

Paper ID 20

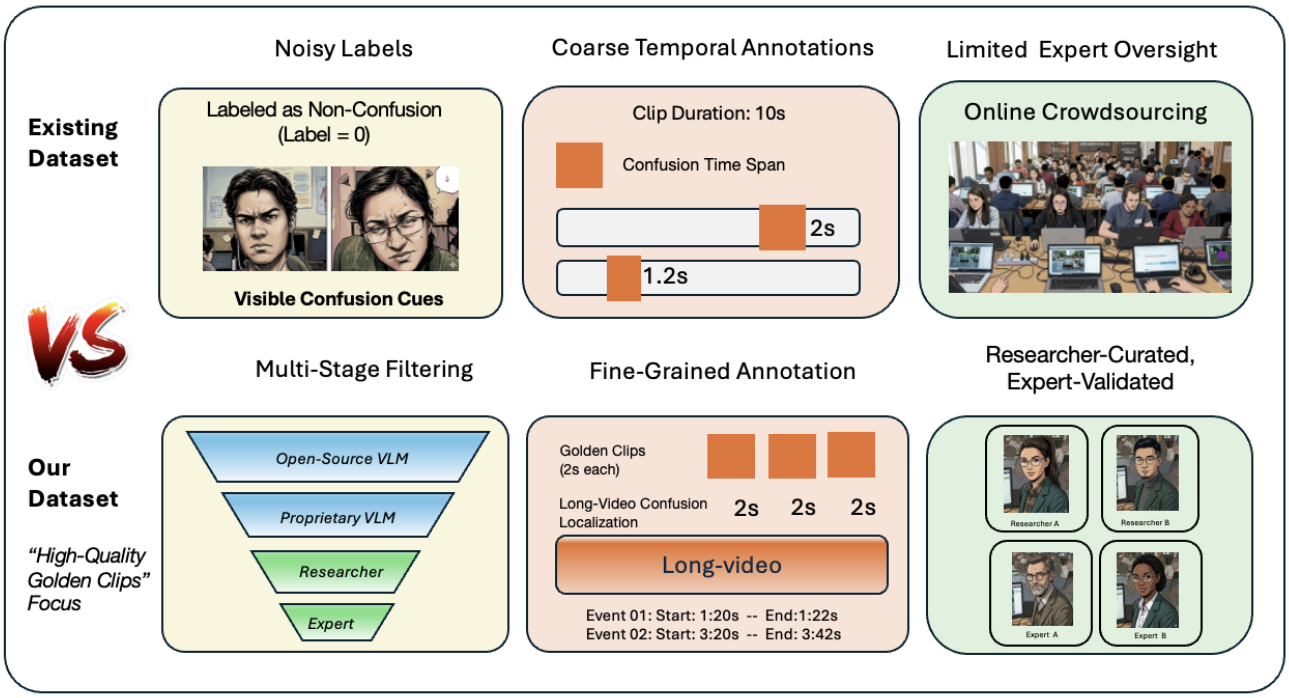

Figure 1. Comparison between existing confusion datasets and our ConfusionBench. Existing datasets often suffer from noisy labels, coarse temporal annotations, and limited expert oversight. In contrast, our ConfusionBench highlights multi-stage filtering, fine-grained annotation, and expert-validated golden clips, together with long-video confusion localization. All human figures shown in this illustration are AI-generated.

Abstract

Recognizing and localizing student confusion from video is an important yet challenging problem in educational AI. Existing confusion datasets suffer from noisy labels, coarse temporal annotations, and limited expert validation, which hinder reliable fine-grained recognition and temporally grounded analysis. To address these limitations, we propose a practical multi-stage filtering pipeline that integrates two stages of model-assisted screening, researcher curation, and expert validation to build a higher-quality benchmark for confusion understanding. Based on this pipeline, we introduce ConfusionBench, a new benchmark for educational videos consisting of a balanced confusion recognition dataset and a video localization dataset. We further provide zero-shot baseline evaluations of a representative open-source model and a proprietary model on clip-level confusion recognition, long-video confusion localization tasks. Experimental results show that the proprietary model performs better overall but tends to over-predict transitional segments, while the open-source model is more conservative and more prone to missed detections. In addition, the proposed student confusion report visualization can support educational experts in making interven-

tion decisions and adapting learning plans accordingly. All datasets and related materials will be made publicly available on our project page.

1. Introduction

Emotions are psycho bio social reactions that motivate actions [21]. These actions are reflected in the nonverbal behavior of people, primarily in the face [8]. Although seven emotions have received the most scrutiny, recently scientists have identified up to 28 different emotions, including those pulling for more cognitive activity such as realization, concentration, and confusion [2]. Of these emotions, confusion turns out to be one of the most relevant to the education process [24]. Identifying the specifics of that particular expression would be imperative to any teacher or automated system designed to assess the efficacy of teaching and learning. Reliable confusion recognition and automatic visualization of student confusion states can further support a broad range of educational AI applications, including intelligent tutoring systems, adaptive feedback, learner monitoring, expert intervention analysis, and human-centered educational interfaces [5, 24, 35]. Meanwhile, recent advances in multimodal foundation models have sparked growing interest in using vision-language models (VLMs) to interpret human affective and cognitive states[34]. However, VLM-driven student confusion analysis remains underexplored.

A key challenge is that existing confusion datasets often do not provide sufficiently reliable annotations for evaluation. DAISEE is a valuable and publicly available dataset with confusion annotations, and has served as an important resource for prior research [10]. However, it has several limitations that hinder its use as a reliable benchmark for confusion recognition. First, the labels can be noisy: some clips annotated as non-confusion still exhibit visible cues of confusion, while some clips labeled as confusion contain little or no observable evidence of confusion [12]. Second, the temporal annotations are overly coarse. Assigning a single confusion label to a 10-second clip may be insufficient, as the actual confusion cue may appear for only 1–2 seconds. As a result, such clip-level annotations fail to accurately capture the video's fine-grained temporal cognitive state [14, 22]. Third, the annotation process often depends on large-scale online crowdsourcing, which can lead to substantial subjective inconsistency in the absence of sufficient expert validation. As a result, the limitations of such datasets make it challenging to reliably evaluate model performance on confusion detection in realistic educational settings. Therefore, high-quality and reliable confusion annotations are essential for meaningful evaluation [10, 12].

To address these issues, expert validation is essential for ensuring annotation quality. However, immersing experts in large volumes of raw data is both inefficient and costly. To this end, we propose an efficient model-assisted benchmark construction pipeline that ensures data quality while making effective use of limited expert resources. The core idea of this pipeline is to construct short, high-confidence golden clips through a multi-stage filtering process that includes coarse screening with open-source VLMs, fine-grained screening with commercial VLMs, researcher curation, and expert validation. These short clips provide a more reliable reference standard for confusion recognition, particularly for defining confusion states and identifying subtle expressions and motion combinations that are often diluted in longer temporal segments.

Moreover, since DAiSEE is originally segmented from longer source videos, the validated golden clips can be traced back to their original temporal positions. Based on these positions and expert annotations as anchors, we can efficiently propagate annotations to neighboring frames with similar visual patterns, thereby enabling timestamp-level labels for long-video confusion event detection. This long-video setting is especially valuable in educational applications, where the frequency and temporal distribution of confusion over time are often more informative than isolated clip-level labels. Based on this pipeline, we present a benchmark consisting of two datasets. The first is a balanced confusion recognition dataset, in which all 2-second confusion samples are expert-validated, while non-confusion samples are selected to exclude any confusion patterns identified in prior literature. The second is a confusion localization dataset, where each long video is annotated with the start and end timestamps of confusion events, enabling evaluation of zero-shot confusion localization performance. In addition, we propose an efficient visualization report to facilitate expert intervention analysis and adaptive learning support.

Beyond benchmark construction, another goal of this work is to study how well modern VLMs can recognize confusion. Although recent open-source and proprietary vision-language models have shown strong zero-shot performance on many visual reasoning tasks, their ability to detect subtle confusion signals in temporally localized educational settings remains unclear. To address this gap, we provide zero-shot baseline evaluations of representative open-source and proprietary models on both clip-level confusion recognition and long-video confusion localization. Our contributions are three-fold:

- We identify several key limitations of existing confusion datasets, including noisy labels, overly coarse temporal annotations, and insufficient expert validation, and propose a practical multi-stage pipeline for constructing higher-quality confusion benchmarks.
- We introduce a high-quality, expert-validated benchmark for confusion understanding in educational videos, consisting of a balanced recognition dataset, long videos with

CVPR #20

Figure 2. Facial Expressions of Confusion [18].

fine-grained temporal annotations, and a confusion report visualization design to support interpretation, intervention analysis, and adaptive learning.
- We benchmark modern VLMs in the zero-shot setting on clip-level confusion recognition and long-video confusion localization, establishing systematic baselines for confusion understanding in educational videos.

2. Related Work

Confusion Research Confusion plays a central role in complex learning activities, such as understanding difficult texts, generating coherent arguments, solving challenging problems, and modeling complex systems, and is often regarded as an inevitable consequence of effortful information processing [6]. Prior research [4, 6, 9, 15] suggests that confusion is associated with characteristic facial, bodily, and temporal cues. In particular, using the Facial Action Coding System (FACS) [25], D'Mello and colleagues [6] identified several facial actions associated with confusion, including brow lowering or frowning (AU4), eyelid tightening or squinting (AU7), upper lip raising (AU10), and lip pressing or tightening (AU24). Among these, AU4, AU7, and especially their combination (AU4+AU7), have been reported as the most reliable facial correlates of confusion, as shown in Fig. 2. Additional contextual cues, such as gaze direction and head pose, may also provide supportive evidence in practice [27].

Beyond facial cues, researchers have also studied hand-to-face behaviors, body posture, and temporal dynamics as useful signals of confusion. Hand-to-face actions such as touching the chin, pressing the forehead, and covering the mouth may indicate thinking, frustration, or hesitation [18], while body posture and movement patterns can provide complementary evidence [3, 11, 23]. In addition, confusion is often described as a temporally evolving process, ranging from subtle early signs, such as slight frowning, to more developed patterns involving forward leaning and hand-to-face behaviors.

DAiSEE Dataset DAiSEE [10] is one of the most widely used video datasets for affective-state understanding in e-learning, containing 9,068 clips from 112 users with annotations for boredom, confusion, engagement, and frustration. Although DevEmo [20] also provides confusion annotations, it contains only around 50 confusion clips and is therefore much smaller in scale. DAiSEE has been widely used for learner-state analysis, including engagement modeling and confusion-related studies [1, 7, 13, 14, 19, 29, 31, 40]. However, several follow-up studies have noted limitations of DAiSEE, particularly in annotation quality and label distribution [7, 12, 19]. In addition, its clip-level annotations may be too coarse for fine-grained confusion analysis, since brief confusion cues can occupy only a small portion of a longer segment [14, 22]. These limitations highlight the need for high-quality, fine-grained, and balanced benchmarks for confusion analysis.

Vision-language models (VLMs) VLMs have advanced rapidly in recent years and achieved strong performance across a wide range of multimodal understanding tasks. In particular, recent VLMs have shown promising capabilities in video understanding, temporal reasoning, and zero-shot inference, making them attractive tools for evaluating challenging video benchmarks without task-specific training [17, 26, 37]. Beyond direct inference, prior work has also explored using strong multimodal models as teachers to distill knowledge [16, 33] or generate pseudo-labels for downstream datasets [30, 36]. In our setting, we leverage VLMs in two ways. First, we use VLM-based reasoning as an auxiliary tool in the data construction pipeline to help reduce annotation noise during candidate filtering. Second, we evaluate representative open-source and proprietary VLMs on our benchmark to better understand their current ability to recognize subtle confusion cues and perform temporally grounded confusion analysis in educational videos.

3. ConfusionBench Construction

To bridge the gap between large-scale candidate data and limited expert resources, we propose a VLM-assisted multi-stage filtering pipeline that removes irrelevant samples and identifies representative clips for expert validation. See Fig.3. Rather than relying directly on the original clip annotations, we progressively refine candidate samples through multiple stages, including clip segmentation, model-assisted filtering, researcher curation, and expert validation. This pipeline allows us to construct a higher-quality set of golden clips and derive temporally grounded long-video annotations. In the following, we describe each stage of the pipeline in detail.

3.1. Fine-Grained Two-Second Clip Segmentation

Existing 10-second clip annotations are often too coarse for subtle confusion analysis, since brief confusion cues may occupy only a small portion of a longer segment. Prior work on fine-grained emotion recognition suggests using smaller temporal units, and studies of facial expressions indicate that many natural expressions last only a few seconds

CVPR #20

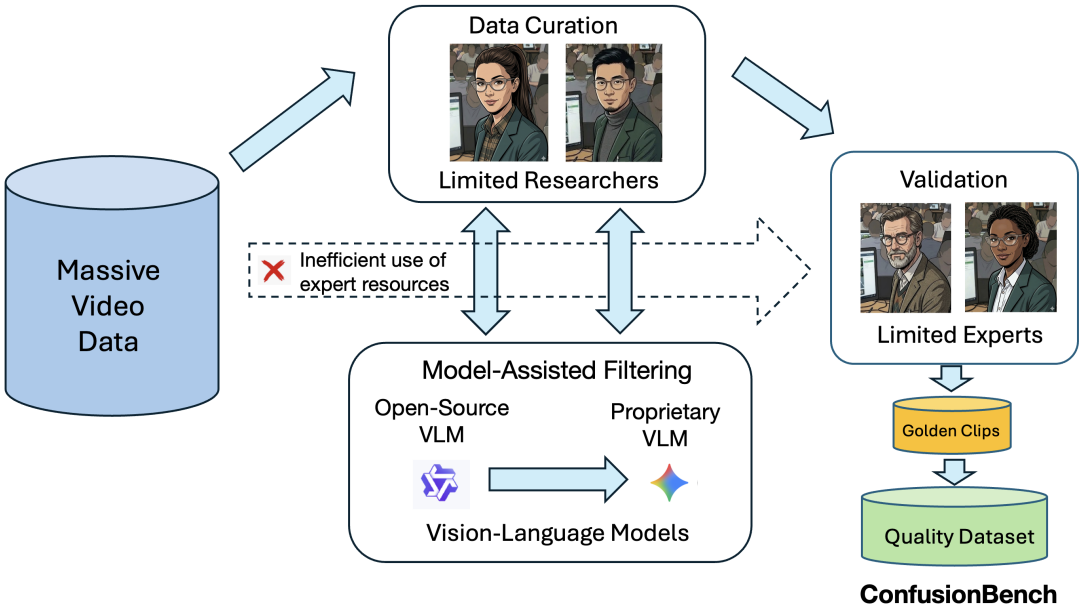

Figure 3. ConfuBench Construction Pipeline

[28, 32, 38, 39]. Based on these observations, we adopt 2-second clips as a practically reasonable and finer-grained unit for confusion analysis. Concretely, each original 10-second DAiSEE clip is divided into five non-overlapping 2-second clips, which better preserve subtle and temporally brief confusion signals for subsequent filtering and validation.

3.2. Model-Assisted Filtering

Since our benchmark focuses on confusion analysis and the raw candidate pool is highly imbalanced, we first use a free open-source VLM for coarse filtering. Specifically, Qwen (Qwen3-VL-4B-Instruct) is used to identify 2-second clips likely to contain confusion-related cues. Inference is performed on a local computer with an NVIDIA RTX 3090 GPU. We also employ carefully designed prompts to guide model reasoning and enforce a unified output format. The prompt used in this stage is as shown in Fig. 4. In this process, the outputs includes four confusion levels: *None*, *Low*, *Medium*, and *High*. To focus on more apparent confusion cases, we treat only *Medium* and *High* predictions as positive. Once a 2-second clip is flagged as containing confusion, it is mapped back to its corresponding 10-second source segment, which is then passed to a stronger VLM for further analysis. This stage retains 609 10-second video clips and serves as an efficient first-pass filtering step in our multi-stage benchmark construction pipeline.

In the next stage, we employ a proprietary VLM, Google Gemini (Gemini 3 Flash Preview), to further refine candidate analysis. Gemini is applied to each 2-second clip using the same prompt as in the previous stage. As the 609 selected 10-second source segments are each divided into five 2-second clips, this stage processes 3,045 clips in total. To improve robustness, we adopt a confidence-aware majority voting strategy over the five inference runs. Each prediction is weighted by its confidence level, with High, Medium, and Low assigned scores of 3, 2, and 1, respectively. The scores are then accumulated across the five runs, and the label with

Model-Assisted Filtering Prompt.
Your task is to analyze the facial expressions, head movements, and body movements of the person in the input video to determine whether the person is showing a state of confusion.
Based on your observations, judge whether the person displays confusion and provide the level of confusion.
Requirements:
1. Your judgment must be based only on facial expressions and movements that are actually visible in the video. Do not make unsupported guesses.
2. If the video evidence is insufficient, you must still provide the most reasonable judgment based on the available evidence and explain the source of uncertainty in your analysis.
3. "confusion_level" must be exactly one of the following four values: None, Low, Medium, High.
4. "whether_confusion_is_present" must be exactly one of the following two values: Yes or No.
5. "confidence_in_your_answer" must be exactly one of the following three values: Low, Medium, High.

Figure 4. Prompt Design

the highest total score is selected as the final prediction. Although our final task is binary confusion detection, a more fine-grained prompt encourages better reasoning and more stable predictions. This stage ultimately selects 1,348 2-second clips predicted to contain confusion.

3.3. Researcher Curation

After model-assisted filtering, we further refine the candidate clips through researcher curation based on a confusion-focused behavioral protocol derived from prior literature on confusion [4, 6, 9, 15, 18]. Our protocol considers five criteria: (1) brow lowering or frowning (AU4), eyelid tightening or squinting (AU7), and especially their combination (AU4+AU7) as the most reliable indicators; (2) auxiliary facial cues, including upper lip raising (AU10) and lip pressing or tightening (AU24); (3) smiling as an exclusion cue in most cases; (4) auxiliary hand-to-face behaviors, such as chin touching, forehead pressing, and covering the mouth; and (5) auxiliary body movement and posture cues, such as forward leaning, upright alert posture, and restless movements. The goal of this stage is to remove clearly irrelevant samples, reduce redundant neighboring clips, and retain clips with plausible confusion-related cues for expert validation. Two researchers independently review the clips, and only samples with full agreement are retained. In total, 301 clips are selected for expert annotation.

3.4. Expert Validation

In the next stage, we submit the 301 curated clips with relatively high confidence to experts in facial expression analysis and human cognition for further validation. Each expert is asked to review the clip and assign one of three decisions: *Yes*, *No*, or *Unsure*, depending on whether the clip exhibits confusion-related cues. *Unsure* indicates that additional information is needed for a more confident judgment. This design allows us to separate clear positive samples from clear negative samples. Two experts participated in this process, and the agreed-upon validation results are summarized as follows: 224 clips are labeled as *Yes*, 46 clips as *Unsure*, and 31 clips as *No*. The *Yes* clips form the foundation of our final golden clip benchmark and are retained as the core confusion-positive set.

To construct a balanced dataset, we combine the expert-validated *Yes* and *No* samples with an additional set of non-confusion clips of comparable scale. This step is relatively straightforward, as non-confusion samples are conservatively selected from clips that exhibit none of the five confusion-related criteria described above. As a result, we obtain a balanced confusion recognition dataset of 450 clips, including 224 *Yes* samples and 226 *No* samples.

Figure 5. Sample student confusion visualization reports. These reports provide fast, interpretable cues that help educational experts determine whether intervention is needed and adjust instructional plans accordingly.

3.5. Long-Video Confusion Localization

DAiSEE is originally derived from longer recordings for each participant and segmented into 10-second clips, with clip identifiers following a consistent naming rule that allows the original temporal order to be recovered. In other words, the long videos can be reconstructed from clip IDs. Leveraging this property, we map each expert-validated golden clip back to its corresponding participant and source segment, and reconstruct long videos by concatenating consecutive clips from the same recording. We then use the validated golden clips as reference anchors and further refine the start and end boundaries of each confusion event according to the confusion cues and curation principles described in the previous stages. In this way, we obtain higher-quality timestamp-level confusion localization annotations for long-video evaluation. In our benchmark, we construct a collection of ten 5-minute long videos to create a more realistic setting for long-video confusion localization.

4. Experiments

Given the rapid progress of vision-language models (VLMs) and their strong multimodal reasoning ability, their capability for confusion understanding remains underexplored. We therefore report zero-shot baselines on the proposed benchmark. Specifically, we evaluate Qwen3-VL-4B-Instruct (*Qwen*) and Gemini 3 Flash Preview (*Gemini*)

Table 1. VLM Zero-shot clip-level confusion recognition results on the proposed balanced dataset.

VLM	Acc	Prec	Rec	F1
Qwen	0.6911	0.7815	0.5268	0.6293
Gemini	0.7978	0.7509	0.8884	0.8139

Table 2. Zero-shot long-video confusion localization results using a proprietary VLM (Gemini 3 Flash Preview).

Source	Acc	Prec	Rec	F1	GT_ev	Pr_ev	tIoU	P@.1	R@.1	F1@.1	P@.3	R@.3	F1@.3
001	0.9067	0.8684	0.7857	0.8250	4	10	0.7021	0.4000	1.0000	0.5714	0.2000	0.5000	0.2857
002	0.7467	0.9111	0.5467	0.6833	11	19	0.5190	0.5263	0.9091	0.6667	0.3684	0.6364	0.4667
003	0.7200	0.7273	0.9072	0.8073	17	19	0.6769	0.6842	0.7647	0.7222	0.5263	0.5882	0.5556
004	0.6267	0.6116	0.8916	0.7255	17	23	0.5692	0.6087	0.8235	0.7000	0.4783	0.6471	0.5500
005	0.6800	0.5926	0.9412	0.7273	16	22	0.5714	0.3636	0.5000	0.4211	0.3182	0.4375	0.3684
006	0.9000	0.7903	0.9608	0.8673	11	20	0.7656	0.5000	0.9091	0.6452	0.5000	0.9091	0.6452
007	0.6867	0.6327	0.8493	0.7251	11	22	0.5688	0.4091	0.8182	0.5455	0.3636	0.7273	0.4848
008	0.7133	0.7407	0.3571	0.4819	7	19	0.3175	0.2632	0.7143	0.3846	0.1053	0.2857	0.1538
009	0.5933	0.8727	0.4706	0.6115	6	25	0.4404	0.1600	0.6667	0.2581	0.1200	0.5000	0.1935
010	0.3200	1.0000	0.3200	0.4848	1	22	0.3200	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[Micro-Avg]	**0.6893**	**0.7289**	**0.6612**	**0.6934**	**101**	**201**	**0.5307**	**0.3831**	**0.7624**	**0.5099**	**0.2985**	**0.5941**	**0.3974**
[Macro-Avg]	**0.6893**	**0.7747**	**0.7030**	**0.6939**	**10.1**	**20.1**	**0.5451**	**0.3915**	**0.7106**	**0.4915**	**0.2980**	**0.5231**	**0.3704**

Table 3. Zero-shot long-video confusion localization results using a open source VLM (Qwen3-VL-4B-Instruct).

Source	Acc	Prec	Rec	F1	GT_ev	Pr_ev	tIoU	P@.1	R@.1	F1@.1	P@.3	R@.3	F1@.3
001	0.7267	1.0000	0.0238	0.0465	4	1	0.0238	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
002	0.4933	0.0000	0.0000	0.0000	11	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
003	0.6067	0.7317	0.6186	0.6704	17	22	0.5042	0.5455	0.7059	0.6154	0.4091	0.5294	0.4615
004	0.6333	0.6346	0.7952	0.7059	17	19	0.5455	0.6842	0.7647	0.7222	0.5263	0.5882	0.5556
005	0.6733	0.6462	0.6176	0.6316	16	23	0.4615	0.5217	0.7500	0.6154	0.4348	0.6250	0.5128
006	0.8067	0.7750	0.6078	0.6813	11	15	0.5167	0.6000	0.8182	0.6923	0.4667	0.6364	0.5385
007	0.5933	0.6034	0.4795	0.5344	11	15	0.3646	0.4000	0.5455	0.4615	0.2000	0.2727	0.2308
008	0.6400	0.7500	0.0536	0.1000	7	4	0.0526	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
009	0.3667	0.6522	0.1471	0.2400	6	9	0.1364	0.2222	0.3333	0.2667	0.0000	0.0000	0.0000
010	0.1133	1.0000	0.1133	0.2036	1	12	0.1133	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
[Micro-Avg]	**0.5653**	**0.6835**	**0.3388**	**0.4530**	**101**	**121**	**0.2928**	**0.4463**	**0.5347**	**0.4865**	**0.3223**	**0.3861**	**0.3514**
[Macro-Avg]	**0.5653**	**0.6793**	**0.3456**	**0.3814**	**10.1**	**12.1**	**0.2719**	**0.2974**	**0.3918**	**0.3374**	**0.2037**	**0.2652**	**0.2299**

without any task-specific training or fine-tuning, to assess how well current VLMs can directly recognize confusion-related cues.

For the confusion recognition task, we evaluate model performance using standard classification metrics, including Accuracy (Acc) , Precision (Prec), Recall(Rec), and F1 score. These metrics provide complementary views of model behavior, allowing us to measure overall correctness as well as the model's ability to reliably identify confusion samples that are relatively salient to human observers.

For long-video confusion localization, we evaluate both event detection and temporal boundary quality. Consecutive confusion-positive segments are grouped into predicted events (Pr_ev) and compared with ground-truth events (GT_ev). Temporal overlap is measured by (tIoU), defined as the ratio between the overlap duration and the union duration of a predicted event and a ground-truth event. We further report event-level Precision, Recall, and F1 under tIoU thresholds of 0.1 and 0.3, corresponding to relatively loose and moderately strict matching criteria. (P@tIoU) measures the proportion of predicted events that correctly match a ground-truth confusion event under the specified overlap threshold, while R@tIoU measures the proportion of ground-truth confusion events that are successfully detected by the model. (F1@tIoU) provides the harmonic mean of Precision and Recall. Finally, we report both micro-average (Micro-Avg) and macro-average (Macro-Avg) metrics for overall aggregation.

The confusion recognition results in Table 1 show that *Gemini* outperforms *Qwen* by a large margin in terms of F1 score (0.8139 vs. 0.6293). This gap is mainly attributable to Qwen's relatively low recall (0.5268), which suggests that many clips containing confusion are misclassified as non-confusion. These results indicate that *Qwen* is more conservative in recognizing confusion and lacks sufficient sensitivity to subtle confusion-related cues.

The confusion video localization results in Table 2 suggest that *Gemini* achieves reasonably strong segment-level confusion recognition, with a Micro-Avg F1 of 0.6934. At the event level, the model attains a Micro-Avg tIoU of 0.5307, indicating moderate temporal alignment for matched confusion events. However, the gap between F1@0.1 (0.5099) and F1@0.3 (0.3974) shows that, although Gemini can often roughly identify the presence

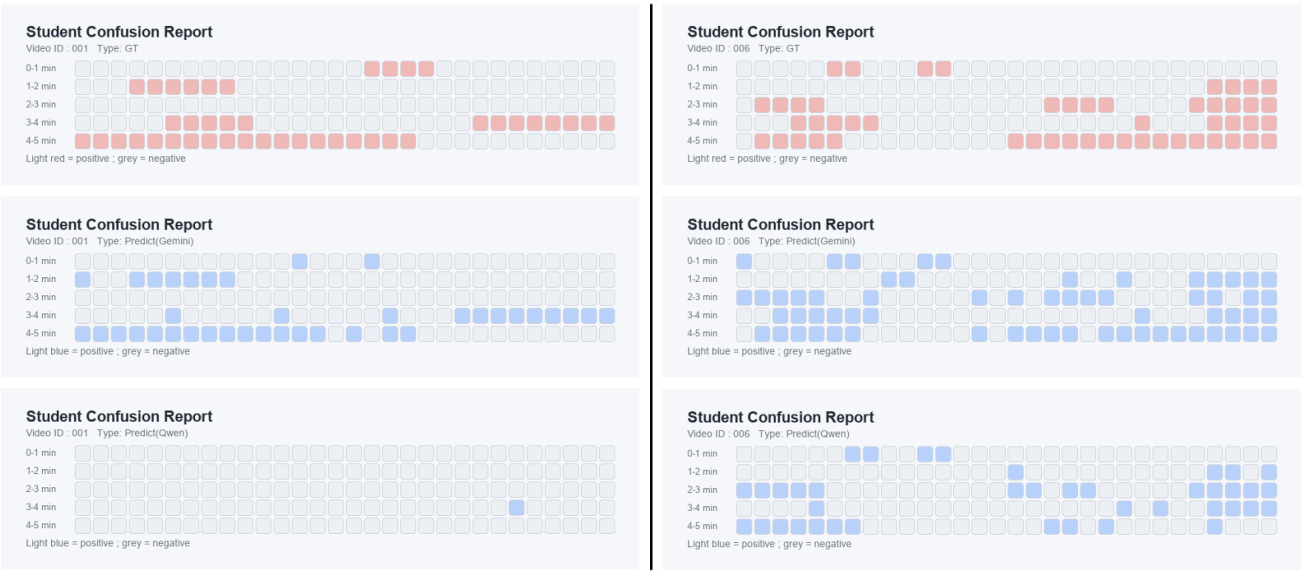

Figure 6. **Comparison of Student Confusion Reports.** The first row in red presents the ground-truth annotations. The second row in blue shows the predictions from Gemini (Gemini 3 Flash Preview), a proprietary multimodal model optimized for high-throughput inference. The third row in blue shows the predictions from Qwen (Qwen3-VL-4B-Instruct), a lightweight local multimodal model with strong video understanding capabilities.

of confusion events, accurately localizing their temporal boundaries remains challenging. In addition, the relatively high Recall@0.1 (0.7624) compared with Precision@0.1 (0.3831) suggests that *Gemini* is more sensitive than precise in long-video confusion detection, tending to cover more true events at the cost of additional false positives.

The long-video confusion localization results reported in Table 3 show that *Qwen* performs clearly worse than *Gemini*. Its Micro-Avg recall is low (0.3388) despite a moderate precision (0.6835), suggesting that the model is conservative in predicting confusion and misses many true confusion segments. This is also reflected in the event-level results, where *Qwen* achieves only 0.4865 F1@0.1 and 0.3514 F1@0.3, indicating limited ability in both event discovery and temporal localization. In addition, the relatively low Micro-Avg tIoU (0.2928) suggests weak temporal alignment between predicted and ground-truth confusion events. This may partly reflect the limited capacity of the 4B model. Overall, these results suggest that this version of *Qwen* remains limited in both sensitivity and temporal grounding for long-video confusion localization.

The student confusion visualization report in Figure 6 further validates the above discussion. *Gemini* generally produces more accurate confusion predictions than the open-source *Qwen* for both clearly confused and clearly non-confused states. In particular, *Gemini* recovers obvious confusion segments more effectively while maintaining more reasonable coverage of negative regions, whereas *Qwen* misses many true confusion intervals and appears overly conservative. This difference is especially evident in transitional or ambiguous segments, where *Gemini* more often labels subtle indicators, such as hand-on-chin or other emerging confusion-related behaviors, as confusion, while *Qwen* tends to refrain from assigning confusion labels to these intermediate states, leading to substantial missed detections, as illustrated in the left example of Figure 6. Overall, these qualitative results are consistent with the quantitative findings: *Gemini* shows stronger sensitivity and more complete temporal coverage, whereas *Qwen* remains limited by conservative predictions and weaker temporal grounding.

5. Conclusion

In this work, we identify key limitations of existing confusion datasets, including noisy labels, coarse temporal annotations, and insufficient expert validation. To address these issues, we propose a VLM-assisted multi-stage benchmark construction pipeline that balances expert effort, model cost, and data quality. Based on this pipeline, we construct an expert-validated balanced confusion recognition dataset, a collection of long videos with fine-grained temporal annotations for confusion localization, and a confusion report visualization design. We further benchmark representative VLMs in the zero-shot setting on both confusion recognition and localization tasks, providing a clearer picture of current model capabilities in recognizing confusion. This marks an important step toward more effective automated

interactive systems.

In future work, we plan to expand the video localization benchmark, evaluate a wider range of models, and develop improved confusion localization frameworks for more reliable student confusion reports.

6. Limitations

Our current recognition dataset contains 450 video clips, including 224 expert-validated golden confusion clips and 226 non-confusion clips, while the video localization benchmark includes 10 long videos of approximately 5 minutes each. Although these datasets provide an initial testbed for confusion recognition and localization, their scale remains limited and could be improved by expanding both the golden clip set and the long-video benchmark.

In addition, our current formulation focuses on whether confusion is present, without distinguishing among different stages or intensities of confusion. In practice, confusion may manifest as early, moderate, or strong confusion. In this work, all such cases are treated as a single category. Developing a more fine-grained evaluation of confusion, in collaboration with facial expression and educational experts, remains an important direction for future work.

7. Ethics Consideration

This work does not involve new human-subject data collection. We use data curated from the publicly available DAiSEE dataset, which has been released since 2016. According to the dataset documentation, participants provided signed consent for their videos to be shared with the research community. We follow the dataset usage agreement and use the data only for academic research. For vision-language model analysis, we use both local and cloud-based models. We run Qwen locally on our computer for data analysis. We access Gemini through Google Cloud Vertex AI. According to Google, customer data is not used to train its AI models, and API communications are encrypted in transit using TLS.

Overall, we consider the ethical risk of this work to be limited, since it relies on previously released research data and does not involve new participant recruitment. Nevertheless, possible misinterpretation of students' affective states, especially under limited or variable video quality, remains an important consideration. Accordingly, the proposed benchmark is intended for research use only, and model predictions should be treated as supportive signals rather than as a basis for high-stakes educational decisions.

References

[1] Ali Abedi and Shehroz S Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *2021 18th Conference on Robots and Vision (CRV)*, pages 151–157. IEEE, 2021. 3

[2] Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90, 2019. 2

[3] Sidney S D'Mello, Patrick Chipman, and Art Graesser. Posture as a predictor of learner's affective engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2007. 3

[4] Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012. 3, 5

[5] Sidney D'Mello, Scotty Craig, Karl Fike, and Arthur Graesser. Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In *International Conference on Human-Computer Interaction*, pages 595–604. Springer, 2009. 2

[6] Sidney K D'Mello and Arthur C Graesser. Confusion. In *International handbook of emotions in education*, pages 289–310. Routledge, 2014. 3, 5

[7] Yu Fang, Shihong Huang, and Amy Ogan. A cross-cultural confusion model for detecting and evaluating students' confusion in a large classroom. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 473–483, 2025. 3

[8] Mark G Frank and Allison Z Shaw. Evolution and nonverbal communication. 2016. 2

[9] Joseph F Grafsgaard, Kristy Elizabeth Boyer, and James C Lester. Predicting facial indicators of confusion with hidden markov models. In *International Conference on Affective computing and intelligent interaction*, pages 97–106. Springer, 2011. 3, 5

[10] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016. 2, 3

[11] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4):341–359, 2013. 3

[12] Shehroz Khan and Sadaf Safa. Revisiting annotations in online student engagement. In *Proceedings of the 2024 10th International Conference on Computing and Data Engineering*, pages 111–117, 2024. 2, 3

[13] Shoroog Ghazee Khenkar, Salma Kammoun Jarraya, Arwa Allinjawi, Samar Alkhuraiji, Nihal Abuzinadah, and Faris A Kateb. Deep analysis of student body activities to detect engagement state in e-learning sessions. *Applied Sciences*, 13 (4):2591, 2023. 3

[14] Purushottama Rao Komaravalli and B Janet. Detecting academic affective states of learners in online learning environ-

ments using deep transfer learning. *Scalable Computing: Practice and Experience*, 24(4):957–970, 2023. 2, 3

[15] Blair Lehman, Sidney D'Mello, and Art Graesser. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, 15(3):184–194, 2012. 3, 5

[16] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 3

[17] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024. 3

[18] Marwa Mahmoud and Peter Robinson. Interpreting hand-over-face gestures. In *International Conference on Affective Computing and Intelligent Interaction*, pages 248–255. Springer, 2011. 3, 5

[19] Somayeh Malekshahi, Javad M Kheyridoost, and Omid Fatemi. A general model for detecting learner engagement: implementation and evaluation. *arXiv preprint arXiv:2405.04251*, 2024. 3

[20] Michalina Manikowska, Damian Sadowski, Adam Sowinski, and Michal R Wrobel. Devemo—software developers' facial expression dataset. *Applied Sciences*, 13(6):3839, 2023. 3

[21] David Matsumoto, Mark G Frank, and Hyi Sung Hwang. *Nonverbal communication: Science and applications*. Sage Publications, 2012. 2

[22] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. Automatic recognition of student engagement using deep learning and facial expression. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 273–289. Springer, 2019. 2, 3

[23] Yondu Mori and Marc D Pell. The look of (un) confidence: visual markers for inferring speaker confidence in speech. *Frontiers in Communication*, 4:487004, 2019. 3

[24] Mark H Myers. Automatic detection of a student's affective states for intelligent teaching systems. *Brain Sciences*, 11 (3):331, 2021. 2

[25] Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. Facial action coding system. *Environmental psychology & nonverbal behavior*, 1, 2015. 3

[26] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 3

[27] Paul Rozin and Adam B Cohen. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans. *Emotion*, 3(1):68, 2003. 3

[28] Karen L Schmidt, Sharika Bhattacharya, and Rachel Denlinger. Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. *Journal of nonverbal behavior*, 33(1):35–45, 2009. 4

[29] Rui Su, Lang He, and Mengnan Luo. Leveraging part-and-sensitive attention network and transformer for learner engagement detection. *Alexandria Engineering Journal*, 107: 198–204, 2024. 3

[30] Xin Xing, Zhexiao Xiong, Abby Stylianou, Srikumar Sastry, Liyu Gong, and Nathan Jacobs. Vision-language pseudo-labels for single-positive multi-label learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7799–7808, 2024. 3

[31] Yaping Xu, Yaqian Zheng, Keru Li, and Yanyan Li. Automatic recognition and analysis of academic emotions based on facial expressions during online learning environments. In *Proceedings of the 15th International Conference on Education Technology and Computers*, pages 328–334, 2023. 3

[32] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 4

[33] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 3

[34] Shuzhen Yu, Alexey Androsov, and Hanbing Yan. Exploring the prospects of multimodal large language models for automated emotion recognition in education: Insights from gemini. *Computers & Education*, 232:105307, 2025. 2

[35] Ziheng Zeng, Snigdha Chaturvedi, and Suma Bhat. Learner affect through the looking glass: Characterization and detection of confusion in online courses. *International Educational Data Mining Society*, 2017. 2

[36] Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. *arXiv preprint arXiv:2406.10502*, 2024. 3

[37] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3

[38] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors. *Sensors*, 21(1):52, 2020. 4

[39] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. Weakly-supervised learning for fine-grained emotion recognition using physiological signals. *IEEE Transactions on Affective Computing*, 14(3): 2304–2322, 2022. 4

[40] Xianwen Zheng, Shinobu Hasegawa, Wen Gu, and Koichi Ota. Addressing class imbalances in video time-series data for estimation of learner engagement:"over sampling with skipped moving average". *Education Sciences*, 14(6):556, 2024. 3