Context-Aware Academic Emotion Dataset and Benchmark

Supplementary Material

Appendix

In the supplementary material, we provide:

§A Related Work.

§B Additional Implementation Details.

§C Experiments on CAER Dataset.

§D Additional Ablation Study.

§E Ethical Implications.

A. Related Work

Academic Emotion Datasets. Although there are numerous well-known publicly available datasets for basic emotions, such as RAF-DB [21], DFEW [13], and MAFW [23], the availability of academic emotion datasets remains limited, which significantly hinders the progress of research in academic emotion recognition. Existing academic emotion datasets can be broadly classified into two categories: those focused on online learning environments and those focused on real-world classroom settings. For datasets focused on online learning environments, such as HBCU [34], DAiSEE [10], EngageWild [14], and OL-SFED [4], participants typically interact with a computer screen while watching stimulus videos or playing cognitive skill training games to elicit academic emotions. To collect spontaneous emotions in real-world learning environments, [33] introduced the academic emotion dataset BNU-LSVED2.0, which contains 2,117 videos of students engaged in real classroom scenarios. Although the reliability of the BNU-LSVED2.0's annotations was assessed using statistical methods in [33], experimental validation of automatic academic emotion recognition algorithms on this dataset is still lacking. Additionally, [12] introduced a manually annotated facial action unit (AU) database collected from juveniles in real classroom settings. However, the challenge of mapping these AUs to specific academic emotion categories remains unresolved. Overall, existing academic emotion datasets have the following limitations: 1) They lack diversity in natural learning scenarios; 2) They typically include only the learner's face or upper body, missing the context information from the learning environment that is crucial for a comprehensive representation of emotional responses.

In this work, we introduce the first academic emotion dataset that captures a diverse range of natural learning scenarios, including classrooms, libraries, laboratories, and dormitories, encompassing both classroom sessions and individual study in real-world settings, while providing comprehensive context information.

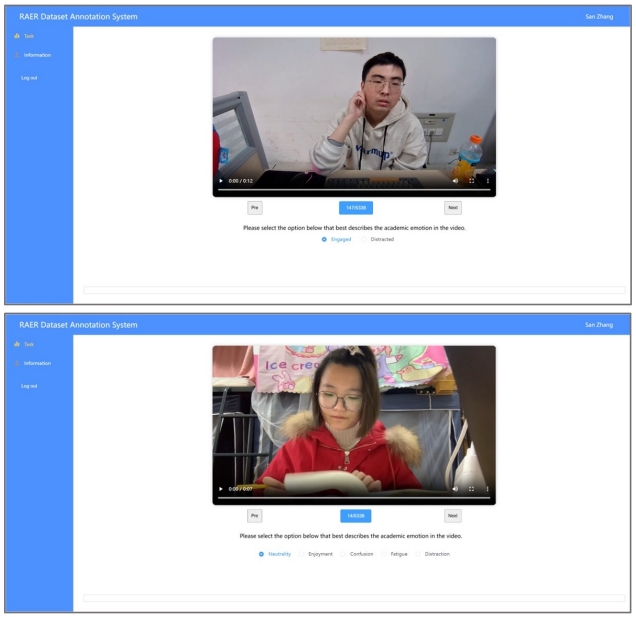

Figure 7. User interface of the annotation website developed.

Video-Based Academic Emotion Recognition. Numerous deep learning-based methods have been developed for video-based emotion recognition: 3D CNN-based [13, 17, 32], RNN-based [2, 7], and Transformer-based [24, 39, 40]. Among these, Transformer-based ones achieve state-of-the-art performance, largely due to the strength of the Transformer's attention mechanism in modeling global dependencies, allowing for more effective long-range feature extraction. However, these methods primarily focus on recognizing basic emotions, and extending them to academic emotion recognition is not straightforward for the following reasons: i) Most existing methods consider only facial expressions, overlooking context information that is crucial for accurately recognizing learners' emotions; ii) Compared to video datasets of basic emotions collected from the internet, such as DFEW [13] or MAFW [23], the academic emotion datasets are typically much smaller, making them less suitable for deep neural models like Transformers, which rely on large-scale training data. Therefore, it is essential to develop a specialized academic emotion recognition framework that can effectively leverage context information from various real-world learning environments without relying on large-scale training data.

Figure 8. Examples from the JuniorRAER dataset. Top row: original video frames; Bottom row: samples of the 5-class academic emotions.

B. Additional Implementation Details

Website for Academic Emotion Annotation. Fig. 7 presents the user interface (UI) of the annotation website developed for labeling academic emotion videos. Through this UI, each annotator can individually review the video clips assigned to them and select an emotion category from the given coarse-grained or fine-grained academic emotion label sets. For challenging videos, annotators can use the progress bar to repeatedly view the video to determine the most appropriate emotion category.

Dataset of JuniorRAER. We built a small academic emotion dataset, called JuniorRAER, to evaluate the generalization ability of our model. Specifically, video recordings in classrooms are common in first-tier cities, generating large amounts of data daily. We sourced the original video data from 6 open video-recorded courses, with the consent of both teachers and students for research analysis. This dataset captures the emotions of primary school students, approximately 10 years old, in real classroom settings. We processed and annotated these videos using methods similar to those described in Sec. 2.1 and Sec. 2.2 of the main document. As a result, we obtained 468 academic emotion video clips featuring 35 primary school students (17 male and 18 female). Similar to the RAER dataset, the JuniorRAER dataset also exhibits an imbalanced distribution: 351 clips (75%) are labeled as neutrality, 61 clips (13.03%) as distraction, 28 clips (5.98%) as fatigue, 24 clips (5.13%) as enjoyment, and 4 clips (0.85%) as confusion. We split the JuniorRAER dataset into training (60%) and testing (40%) sets, ensuring a nearly identical distribution of academic emotions in both subsets. To protect privacy, this dataset is strictly for non-commercial research purposes. Fig 8 illustrates the original videos alongside specific examples of the 5-class academic emotions.

Descriptors of Academic Emotion Categories. In the context-aware text encoder of the proposed CLIP-CAER, we employ a large language model, such as ChatGPT [26], to generate text descriptions for each academic emotion category, capturing both the associated facial expressions and relevant context behaviors. To generate descriptors, we first provide the LLM with the classification criteria outlined in Sec. 2.2 of the main document, allowing it to form a memory. Then, we use the prompt: "What are the useful visual features for the academic emotion of {classname}, considering both facial characteristics and context information?" The descriptors for each emotion category are as follows:

- **Neutrality:** Relaxed mouth, open eyes, neutral eyebrows, no noticeable emotional changes, engaged with study materials, or natural body posture.
- **Enjoyment:** Upturned mouth corners, sparkling eyes, relaxed eyebrows, focused on course content, or occasionally nodding in agreement.
- **Confusion:** Furrowed eyebrows, slightly open mouth, wandering or puzzled gaze, chin rests on the palm, or eyes lock on learning material.
- **Fatigue:** Mouth opens in a yawn, eyelids droop, head tilts forward, eyes lock on learning material, or hand writing.
- **Distraction:** Shifting eyes, restless or fidgety posture, relaxed but unfocused expression, frequently checking phone, or averted gaze from study materials.

Note that each descriptor consists of two parts: the first describes the corresponding facial expression behaviors, while the second captures relevant contextual learning behaviors, such as body posture, yawning, or using a phone. If only facial expression information from the video sequence is

used, the text descriptions are modified to include only the facial expression behavior component.

Optimization. We train the entire network using the SGD optimizer with a batch size of 8. The base learning rates are set as follows: 1×10^{-5} for the CLIP image encoder, 1×10^{-2} for the temporal visual encoder, 1×10^{-3} for the learnable prompt, and 1×10^{-5} for the fully connected layer. The learning rates are reduced by an order of magnitude at the 10^{th} and 15^{th} epochs. The model is trained for 20 epochs in an end-to-end manner. For each video, we randomly and uniformly select 16 non-overlapping frames and use a face detector to extract the face region from each frame, both of which are used as inputs to the model. Each full frame or face region is resized to 224×224, with the shorter edge padded in black to match the input size required by the CLIP model [27]. During training, we apply data augmentation techniques, including random rotation and random flipping, to enhance robustness.

C. Experiments on CAER Dataset

Table 5 presents the evaluation of our method on the CAER dataset [17], which focuses on basic emotions. The CAER dataset includes not only facial expressions but also rich context information, providing a comprehensive representation of emotional responses. However, unlike the academic emotion dataset RAER, the CAER dataset focuses on basic emotions and is substantially larger, with over 13,200 videos. In our implementation, for the seven basic emotions in CAER, we used descriptors similar to those in [40] to describe facial expressions while also providing additional context descriptions. The details are as follows:

- **Surprise:** Widened eyes, an open mouth, raised eyebrows, and a frozen expression. Sudden stillness, widened eyes on the other person, hands raised or paused mid-motion.
- **Sad:** Tears, a downward-turned mouth, drooping upper eyelids, and a wrinkled forehead. Head down, avoiding eye contact, slow, withdrawn movements.
- **Neutral:** Relaxed facial muscles, a straight mouth, a smooth forehead, and unremarkable eyebrows. Relaxed posture, open stance, steady, calm eye contact.
- **Happy:** A smiling mouth, raised cheeks, wrinkled eyes, and arched eyebrows. Leaning in toward the other person, quick, cheerful movements.
- **Fear:** Raised eyebrows, parted lips, a furrowed brow, and a retracted chin. Hands close to chest or tightly together, small, cautious steps backward.
- **Disgust:** A wrinkled nose, lowered eyebrows, a tightened mouth, and narrow eyes. Slight step back, body angled away, hand raised or shielding face.
- **Anger:** Furrowed eyebrows, narrow eyes, tightened lips, and flared nostrils. Leaning forward, tense stance, fists clenched, or hand pointing.

Table 5. Evaluation of CLIP-CAER compared to 3DCNN [11] and CAER-Net [17] on the CAER benchmark for basic emotions.

Method	UAR(%)
3DResNets18 [11] w/o Context	68.22
CAER-Net [17] w/o Context	74.13
CAER-Net [17] w/ Context	77.04
CLIP-CAER w/o Context	75.36
CLIP-CAER w/ Context	**81.77**

Table 6. Ablation study on the number of layers in the temporal encoder and learnable prompt tokens. '# Layer' and '# Tokens' denote the number of layers in the temporal encoder and learnable prompt tokens, respectively.

# Layer	# Tokens	UAR(%)
1	8	**68.00%**
2	8	64.78%
3	8	64.29%
1	4	65.54%
1	8	**68.00%**
1	12	64.15%
1	16	64.27%

It can be observed from Table 5 that incorporating context information in addition to facial expressions significantly improves recognition performance. Furthermore, compared to the state-of-the-art method CAER-Net [17], which also leverages context, the proposed CLIP-CAER achieves an improvement of up to **4.73** points, reaching an accuracy of **81.77%**. These results demonstrate that the proposed CLIP-CAER method is highly effective for both academic and basic emotion recognition, significantly surpassing current state-of-the-art methods.

D. Additional Ablation Study

Number of Layers in the Temporal Encoder and Learnable Prompt Tokens. Table 6 examines the impact of varying the number of layers in the temporal encoder and the effect of different numbers of learnable prompt tokens for each category. The self-attention module S-ATT in Eq. 3, used in the context-aware temporal visual encoder, consists of several identical self-attention layers sequentially stacked together [31]. In general, increasing the number of layers tends to improve model performance; however, in the proposed CLIP-CAER, the best performance is achieved with a single-layer temporal encoder. This finding aligns with the conclusion in [40]. The primary reason is that the academic emotion dataset RAER is relatively small, and the temporal encoder is trained from scratch, which makes it prone to overfitting if the model is overly complex, thereby degrading generalization performance on the test data. This consideration also applies to the learnable

Figure 9. Visualization of attention on full-frame images using Grad-CAM [28]: a) Full-frame input; b) Model utilizing full-frame video sequences to jointly capture facial expressions and context information; c) Model integrating facial image sequences with full-frame sequences to separately capture facial expressions and context information.

prompt tokens, where using 8 learnable tokens achieves the best performance, while increasing the number of tokens does not lead to further improvements.

Attention Visualization. Fig. 9 visualizes the attention regions of our model on input images using Grad-CAM [28]. In this visualization, we compare two input strategies for the model: (a) using full-frame video sequences to jointly capture facial expression features and context information, and (b) combining facial image sequences with full-frame sequences to separately model facial expression features and context information. As shown, both strategies effectively capture context information relevant to academic emotion recognition within the full-frame images, such as using a phone or reading a book, thanks to the robust alignment between text and visual feature spaces provided by the pre-trained CLIP model. However, both tend to overlook facial expression information in the full frame due to the relatively small size of the face region compared to the surrounding context, which may cause the model to disregard it. To address this issue, our model adopts strategy (b), which incorporates an additional facial image sequence to specifically capture facial expression features. These results highlight the effectiveness and robustness of our model's design, as well as the importance of incorporating both facial image sequences and full-frame sequences as inputs.

E. Ethical Implications

The RAER dataset, which consists of real-world academic emotion videos, involves the collection and processing of student data. Ensuring privacy protection is paramount, given that facial expressions and context cues are sensitive personal data. To address this, we adhere to strict data anonymization protocols, ensuring that personally identifiable information is removed. Storage and access control mechanisms are also implemented to prevent unauthorized use of the dataset. Data sharing is regulated to ensure compliance with relevant legal and ethical standards, such as local data protection laws. Researchers using RAER must agree to ethical data usage policies to minimize the risk of privacy breaches.

To mitigate potential bias across different cultural backgrounds, we introduce JuniorRAER as an indirect validation of the model's generalization ability. However, this does not fully resolve cross-cultural bias, as learning environments and emotional expressions can vary significantly across cultures. Future work should focus on diversifying the dataset by including students from different ethnicities, socioeconomic backgrounds, and educational settings to enhance fairness and robustness. Additionally, the subjective nature of emotion annotation introduces another layer of bias. Human annotators may interpret emotions differently based on their own experiences and cultural backgrounds, leading to inconsistencies in labeling. To reduce this bias,

we employed a majority voting strategy across multiple annotators, ensuring greater reliability in emotion classification. Further studies could explore leveraging self-reported emotions or multimodal signals to enhance label accuracy.

The real-world application of academic emotion recognition systems must be approached with caution. While such models have the potential to enhance personalized learning and provide insights into student engagement, they also carry risks of misuse. For instance, over-reliance on AI-based emotion recognition in educational settings could lead to unintended consequences, such as automated decision-making that lacks human oversight. To prevent ethical misuse, AI-based academic emotion recognition systems should be used as assistive tools rather than absolute evaluators of student emotions or performance. Educators and stakeholders must be trained to interpret model outputs critically, using them to complement rather than replace human judgment.