Context-Aware Academic Emotion Dataset and Benchmark

Luming Zhao[1*], Jingwen Xuan[1*], Jiamin Lou[2], Yonghui Yu[1], Wenwu Yang[1†]
[1]Zhejiang Gongshang University [2]Zhejiang Yuexiu University

Abstract

Academic emotion analysis plays a crucial role in evaluating students' engagement and cognitive states during the learning process. This paper addresses the challenge of automatically recognizing academic emotions through facial expressions in real-world learning environments. While significant progress has been made in facial expression recognition for basic emotions, academic emotion recognition remains underexplored, largely due to the scarcity of publicly available datasets. To bridge this gap, we introduce RAER, a novel dataset comprising approximately 2,700 video clips collected from around 140 students in diverse, natural learning contexts such as classrooms, libraries, laboratories, and dormitories, covering both classroom sessions and individual study. Each clip was annotated independently by approximately ten annotators using two distinct sets of academic emotion labels with varying granularity, enhancing annotation consistency and reliability. To our knowledge, RAER is the first dataset capturing diverse natural learning scenarios. Observing that annotators naturally consider context cues—such as whether a student is looking at a phone or reading a book—alongside facial expressions, we propose CLIP-CAER (CLIP-based Context-aware Academic Emotion Recognition). Our method utilizes learnable text prompts within the vision-language model CLIP to effectively integrate facial expression and context cues from videos. Experimental results demonstrate that CLIP-CAER substantially outperforms state-of-the-art video-based facial expression recognition methods, which are primarily designed for basic emotions, emphasizing the crucial role of context in accurately recognizing academic emotions. Project page: https://zgsfer.github.io/CAER

1. Introduction

Academic emotions play a crucial role in the learning process, as they are directly linked to factors such as motivation, controllability, and cognition, all of which greatly affect how effectively a student learns [5, 15]. Thus, accurately identifying a learner's emotional state helps in analyzing their engagement levels and understanding their learning process [16], thereby enabling actions that positively impact learning, which not only leads to better academic outcomes but also fosters a more supportive and adaptive learning environment tailored to individual needs.

While academic emotions are internal psychological responses of learners, expert teachers excel at recognizing these emotional states by observing facial expressions and context cues, allowing them to adjust their teaching strategies accordingly. This makes it natural to develop approaches capable of automatically recognizing students' academic emotions or engagement levels from their facial expressions [4, 34]. Most existing facial expression recognition (FER) methods focus on the automatic recognition of basic emotions, such as surprise, fear, disgust, happiness, sadness, anger, and neutral [20, 35–38]. Therefore, many academic emotion recognition studies based on facial expressions [1, 3, 30] have attempted to first use existing FER methods to recognize students' basic emotions and then map them to academic emotions. However, there are significant semantic differences between basic emotions and academic emotions [16, 25], and the correspondence between them can be ambiguous. For instance, a basic expression of happiness might indicate that a student is enjoying the learning experience, or it could mean that the student is distracted by other non-academic, amusing things.

In this work, we investigate the automatic recognition of academic emotions from videos of students in real-world learning settings by leveraging their facial expressions. One major challenge we face is the extreme scarcity of datasets related to academic emotions in natural learning contexts. The number of publicly available academic emotion datasets is limited [8, 18], with the most commonly used ones being DAiSEE [10], EngageWild [14], and OL-SFED [4]. However, as shown in Fig. 1, these datasets have several limitations: (1) They primarily focus on online learning settings, where students interact with a screen, but lack diverse natural learning scenarios, such as classroom instruction and individual study sessions in real-world environments; (2) Affective behaviors are induced by stimulus videos rather than emerging from everyday natural learning

*Equal Contribution
†Corresponding Author (wwyang@zjgsu.edu.cn)

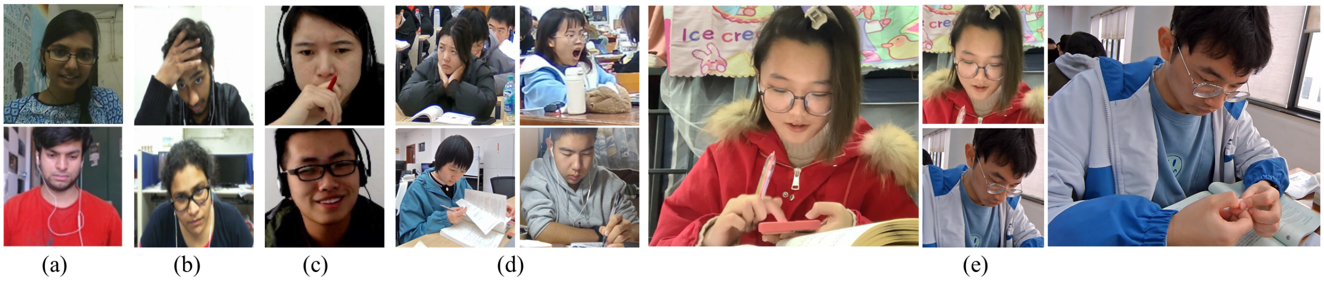

Figure 1. Examples in the (a) DAiSEE [10], (b) EngageWild [14], (c) OL-SFED [4], and (d) RAER. Unlike existing datasets [4, 10, 14] that include emotions induced by watching stimulus videos, RAER captures natural emotions in real learning environments such as classrooms, libraries, labs, and dormitories. Moreover, RAER provides richer context information of the learning environment, which is essential for accurately identifying learners' emotions. As shown in (e), if we only consider the cropped face or upper body video (middle), it may appear that both students are engaged in studying. However, when the full context is included (left or right)—such as the female student using her phone or the male student fiddling with his fingers—it becomes clear that they are actually distracted during the learning process.

processes; (3) The videos capture only the learner's face or upper body, omitting essential learning contexts, which is crucial for a comprehensive representation of emotional responses. These limitations make it challenging for methods developed using these datasets to be effectively applied across diverse real-world learning environments.

To address these limitations, we introduce RAER, a novel dataset for Real-world Academic Emotion Recognition. The RAER dataset comprises approximately 2,700 video clips from around 140 students, capturing diverse natural learning environments such as classrooms, libraries, laboratories, and dormitories. It includes both classroom sessions and individual study scenarios in real-world settings, as illustrated in Fig. 1d. To enhance the consistency and reliability of annotations, we employ two sets of academic emotion labels with varying levels of granularity: a coarse-grained set (engaged or distracted) and a fine-grained set (enjoyment, neutrality, confusion, fatigue, or distraction). Specifically, our trained annotators label the learning videos using categories from a given label set, with each video independently annotated multiple times, yielding an average of approximately five annotations per label set in our experiment. Using a majority voting strategy and cross-validating annotation consistency between the coarse- and fine-grained label sets, each video is reliably assigned a category from the fine-grained set. To the best of our knowledge, RAER is the first real-world academic emotion dataset covering diverse natural learning contexts.

Compared to existing academic emotion datasets [4, 10, 14], the RAER dataset contains videos with richer context information about the learning environment. Recognizing learners' emotions accurately often requires considering both facial expressions and context cues (see Fig. 1e), a practice commonly employed by expert teachers and our annotators. Building on this insight, we propose a novel framework called CLIP-based Context-aware Academic Emotion Recognition (CLIP-CAER). Our key idea is to leverage learnable text prompts to effectively integrate relevant facial expressions and context information from videos using the vision-language model CLIP [27]. This approach reduces the need for extensive training data while enhancing the accuracy of academic emotion recognition.

Extensive experiments validate the effectiveness of the RAER dataset and the proposed CLIP-CAER framework. Notably, results indicate that incorporating context information significantly improves the model's ability to distinguish between distracted and engaged states during learning activities, achieving, for instance, a **19%** accuracy increase for the "distraction" category in the fine-grained labels.

1.1. Our Contributions

- We introduce the first academic emotion dataset RAER, covering a wide range of natural learning contexts.
- We propose CLIP-CAER, a novel academic emotion recognition framework that effectively integrates facial expressions with context cues in the video sequences.
- We demonstrate that in natural learning contexts, our method significantly outperforms state-of-the-art video-based FER approaches, highlighting the essential role of context in accurately recognizing academic emotions.

2. Academic Emotion Database: RAER

2.1. Collecting Learning Videos in Real World

Unlike existing academic emotion datasets that induce academic emotions by having participants watch given stimulus videos [4, 10, 14] or play cognitive skills training games [34], we collect data on students' emotions in real-world learning settings. To achieve this goal, we contacted seven undergraduate classes at universities and additionally recruited 80 individuals, including both undergraduate and graduate students. The students are all adults, aged between 18 and 25. With the consent of the teachers and students from these classes, as well as the recruited individuals, that the recorded study videos would be made publicly available solely for research purposes, we used industrial cam-

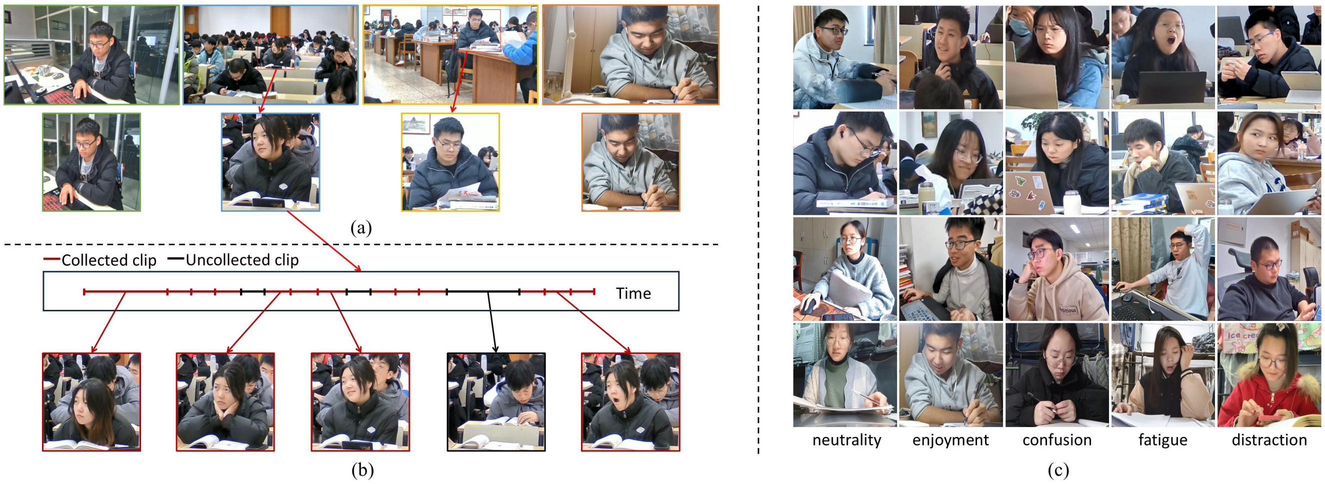

Figure 2. Procedure for building the RAER dataset: (a) extract student clips from the original videos, (b) divide each student's clip into multiple segments, removing redundant parts so that each remaining segmented clip corresponds to an emotion. Finally, the emotion categories are annotated, with examples of the five-class academic emotions shown in (c), corresponding from top to bottom to the learning environments of the classroom, library, laboratory, and dormitory.

eras to record 9 full classroom teaching sessions (covering subjects such as Mathematics, English, and Computer programming) and 177 individual study videos filmed in locations such as libraries, labs, or dormitories. During the recording of individual study sessions, the students were not subject to any special restrictions and could behave as they normally would during self-study. For example, they were free to choose their study materials and could leave or return at their discretion. These videos range from approximately 30 to 150 minutes in length, featuring around 332 students.

We processed the collected videos as follows. First, for each original video, we used commercial video editing software to extract clips of each student's study session. These clips included the student's head, visible parts of their body, and the surrounding learning environment, as shown in Fig. 2a. During this process, we excluded students whose heads were heavily obstructed, those with unclear images due to distance, and individuals who inadvertently appeared in the videos without their knowledge. Next, we removed redundant segments from each student's study clip that were unrelated to learning, such as when the student left their seat. Two expert teachers then reviewed each study clip and divided it into multiple segments based on noticeable emotional changes, with each segmented clip representing an academic emotion, as shown in Fig. 2b. For each clip, we retained only the middle 5-15 seconds, trimming any redundant content from the beginning and end.

2.2. Coarse and Fine-grained Annotation

To annotate the video clips with emotion categories, we first need to determine the labels that will represent learners' academic emotions. Unlike basic human emotions, academic emotions are more complex, and there is currently no standardized set of labels. For instance, [14, 34] use four engagement levels to label learning videos, [4, 10, 33] categorize academic emotions into 4, 5, or up to 10 types, while [12] applies 12 action units to capture emotional nuances. In our dataset, we used academic emotion categories to annotate the learning videos. Building on the observation that skilled, attentive tutors and expert teachers typically respond to students by focusing on just a few common emotional cues [16], we defined two sets of academic emotion labels with varying levels of granularity: a coarse-grained set, including engaged and distracted, and a fine-grained set, consisting of enjoyment, neutrality, confusion, fatigue, and distraction. The coarse-grained labels offer a broad assessment of a student's state during learning activities, specifically whether they are engaged or distracted. The fine-grained labels further refine the "engaged" category by introducing specific emotional variations, such as enjoyment, neutrality, confusion, and fatigue. It's worth noting that we treat the "distracted" category as a more homogeneous condition from the standpoint of learning efficiency and, for the sake of simplicity, do not break it down further.

We recruited 29 annotators (students and staff from universities) to assign emotion labels to video clips. For both the coarse-grained and fine-grained label sets, each clip was assigned to the most apparent category from either the two-class or five-class set. We developed a website for RAER annotation that displays each video with category options from a given label set. Clips were randomly and equally distributed among annotators. Each video was independently labeled by approximately 10 annotators, with 5 assigned to the coarse-grained set and 5 to the fine-grained set. The classification for each video in both sets was determined using a majority voting strategy. Based on the relation-

Table 1. Comparison of the RAER with existing academic emotion recognition datasets. RAER provides a broader range of natural learning scenarios and includes more comprehensive learning contexts that are essential for accurately identifying learners' emotions.

Dataset	#Videos	#Subjects	Setting	Annotation Type	Context
DAiSEE [10]	9,068	112	Online	4 Categories	✘
EngageWild [14]	195	78	Online	4 Engagement levels	✘
OL-SFED [4]	1,274	82	Online	5 Categories	✘
HBCU [34]	120	34	Controlled lab	4 Engagement levels	✘
BNU-LSVED2.0 [33]	2,117	81	Real classrooms	10 Categories	✔
RFAU [12]	3,325	1,796	Real classrooms	12 AUs	✘
RAER	2,649	136	Real classrooms, libraries, labs, and dormitories	5 Categories	✔

ship between the coarse- and fine-grained labels—where enjoyment, neutrality, confusion, and fatigue correspond to the "engaged", and distraction is equivalent to the "distracted"—we evaluate the consistency of each clip's classification across both label sets. The result shows that 93.5% of the videos had consistent labels, which supports the overall reliability of the annotations. Only videos with consistent labels were retained in the dataset. Finally, we obtained 2,649 real-world academic emotion videos, featuring 136 students (91 male and 45 female). The Fleiss' κ coefficient [9] for all fine-grained labels assigned by annotators to these videos is 0.832, indicating a high level of consistency and reliability in the annotations. Since the coarse-grained labels for these clips can be inferred from their fine-grained labels, each video was ultimately assigned a single category from the fine-grained set. Fig. 2c presents specific examples of 5-class academic emotions.

2.3. Analysis

Similar to "in the wild" expression datasets that focus on basic emotions [21], the academic emotions captured in RAER's natural learning scenarios also exhibit an imbalanced distribution, with neutrality accounting for 65.23%, followed by distraction at 19.89%, fatigue at 7.40%, enjoyment at 6.08%, and confusion at 1.40%. This imbalance highlights the differences in the distribution of academic emotions between those observed in natural learning environments and those induced by stimulus videos, where the distribution of various academic emotions tends to be more uniform, as seen in existing datasets [4, 10, 14, 34]. This imbalance also aligns with our intuition to some extent, as neutrality often represents a stable emotional state during learning, while emotions such as enjoyment and confusion tend to occur less frequently, especially in structured learning environments like classrooms.

Furthermore, we compare RAER with other academic emotion datasets, such as DAiSEE [10], EngageWild [14], OL-SFED [4], HBCU [34], BNU-LSVED2.0 [33], and RFAU [12], as shown in Table 1. DAiSEE [10], EngageWild [14], and OL-SFED [4] are relatively common publicly available datasets. However, these datasets primarily focus on online learning environments, where academic emotions are induced by watching stimulus videos. Additionally, they consist solely of face-cropped or upper-body-cropped videos, lacking the surrounding learning context. Furthermore, compared to BNU-LSVED2.0 [33] and RFAU [12], which also collect videos from real classrooms, RAER includes a broader spectrum of natural learning environments, including classrooms, libraries, laboratories, and dormitories, while covering both classroom sessions and individual study. Moreover, different from RFAU [12], which only includes facial information, RAER provides more comprehensive context information of the learning environment. Additionally, while RFAU [12] uses 12 action units (AUs) and 6-level intensities for each action unit in its annotations, it remains unclear how these AUs are mapped to specific academic emotion categories.

3. Context-Aware Academic Emotion Analysis

Numerous deep learning-based methods have been developed for video-based emotion recognition [2, 7, 13, 17, 22, 24, 29, 32, 39, 40]. However, these methods primarily focus on recognizing basic emotions, making their adaptation to the RAER dataset for academic emotion recognition challenging for the following reasons: i) Most existing methods consider only facial expressions, neglecting 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 RAER dataset is significantly smaller, making it less suitable for deep neural models like Transformers, which require large-scale training data. To overcome these limitations, we propose CLIP-CAER, a novel context-aware academic emotion recognition framework that harnesses the power of the vision-language pretraining model CLIP [27]. Our approach leverages learnable text prompts within CLIP to extract and integrate relevant facial expressions and context information from videos, reducing the need for extensive training data while improving the accuracy of academic emotion recognition.

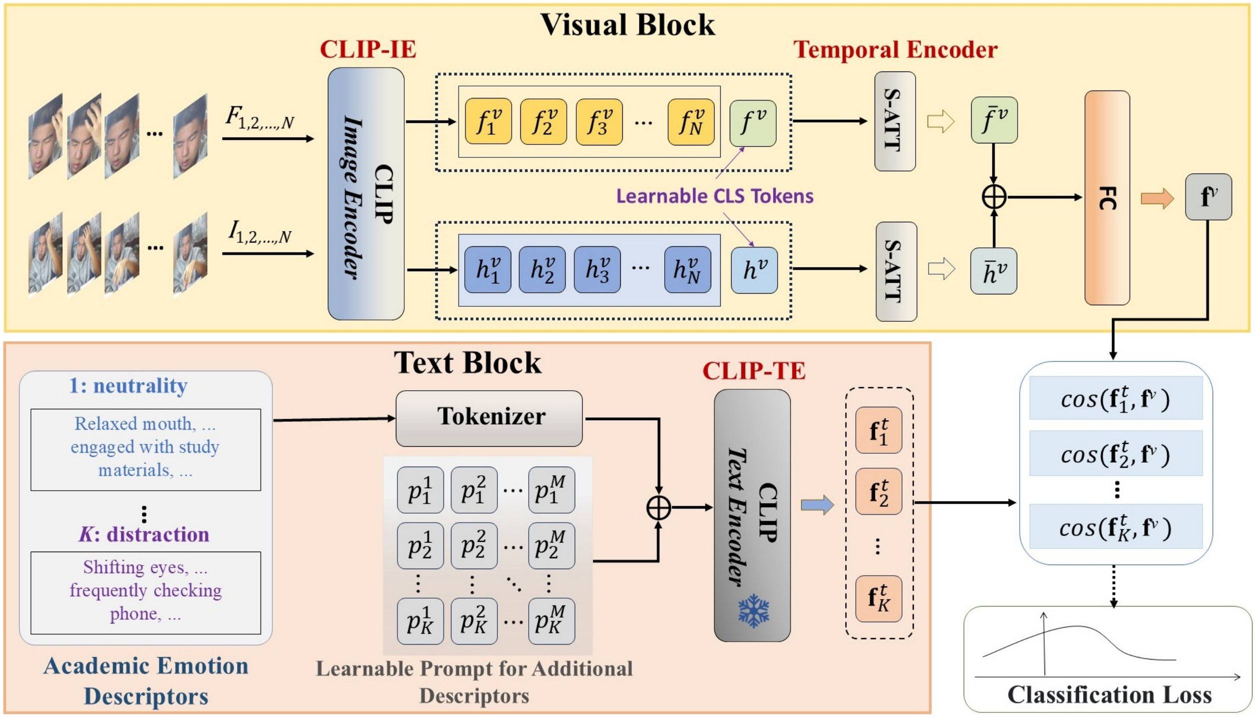

Figure 3. The framework of the proposed CLIP-CAER. Given a video sequence, we first crop the face region from each frame, forming a sequence of face images. Both the face and full-frame sequences are fed into the **Visual Block**, which outputs a visual feature token ($\mathbf{f}^v$) that captures both facial expression and context information from the video. In the **Text Block**, a description along with a learnable text prompt is used to generate a text feature token for each academic emotion category ($\{\mathbf{f}_k^t\}_{k=1}^K$). The input video is classified by computing the cosine similarity between the visual feature token and the text feature tokens corresponding to each academic emotion category.

3.1. Overview

As shown in Fig. 3, our framework, similar to the vision-language model CLIP [27], primarily consists of two components: a text block and a visual block. In the text block, for each academic emotion category, a fixed text is pre-generated to describe the associated facial expressions and learning contexts, complemented by a learnable text prompt to capture additional relevant details during training. Subsequently, by inputting the fixed text and the learnable text prompt together into the *CLIP text encoder*, we obtain a text feature token for each emotion category. Given an input video, the visual block uses the *CLIP image encoder* to separately extract facial expression features and context features from each video frame. These visual features are then processed through a *temporal encoder module* to capture their sequential relationships, resulting in a visual feature token that effectively represents both the facial expression and context information within the video. Given the aligned visual and text feature spaces in the pre-trained CLIP model, we classify the input video by calculating the similarity between its visual feature token and the text feature tokens for each academic emotion category.

A key distinction between previous methods that utilize the pre-trained CLIP model for recognizing basic emotions from facial expressions (*e.g.*, DFER-CLIP [40] and CLIPER [19]) and our approach is that we integrate context cues from learning activities. This significantly enhances the model's ability to accurately identify academic emotions. To achieve this, we design a novel context-aware temporal encoder within the visual block and incorporate context prompt into the text block, as detailed below.

3.2. Context-Aware Visual-Text Encoding

Given a frame sequence $\mathcal{S} = \langle \mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N \rangle$ containing N sample frames from an academic emotion video, the goal of the visual block is to generate a visual feature token that effectively represents both the facial expression and context information within the video, denoted as

$$\mathbf{f}^v = \text{VIS-B}(\mathcal{S}), \tag{1}$$

where $\text{VIS-B}(\cdot)$ denotes the visual block, and $\mathbf{f}^v$ is the visual feature token. In our implementation, $\mathbf{f}^v \in \mathbb{R}^{512}$, matching the class token output of the CLIP image encoder we used.

Context-Aware Temporal Visual Encoder. In $\text{VIS-B}(\cdot)$, we first use a shared CLIP image encoder to extract visual features of both facial expressions and context information from each frame $\mathcal{I}_i$, $i \in \{1, 2, \cdots, N\}$. A straightforward approach would be to directly feed each image frame into the CLIP image encoder. However, this method has two limitations: (1) Because the face region occupies a relatively small portion of the frame compared to the surrounding learning environment, the model tends to focus more on the context, potentially overlooking facial expressions; (2) The facial expression and context information for each frame are represented within the same feature vector, making it impossible to separately model their temporal dynamics. To address these issues, we detect and crop the face region from each frame $\mathcal{I}_i$ to obtain a face image $\mathcal{F}_i$. Then, we separately input $\mathcal{I}_i$ and $\mathcal{F}_i$ into the CLIP image encoder to extract context features from $\mathcal{I}_i$ and facial expression features from $\mathcal{F}_i$, denoted as

$$\begin{cases} f_i^v = \text{CLIP-IE}(\mathcal{F}_i) \\ h_i^v = \text{CLIP-IE}(\mathcal{I}_i) \end{cases}, \tag{2}$$

where CLIP-IE$(\cdot)$ denotes the CLIP image encoder, and f_i^v and h_i^v represent the feature vectors of facial expression and context information within each frame $\mathcal{I}_i$, respectively.

To effectively combine the facial expression and context features extracted from all frames in the sequence $\mathcal{S}$, we employ the Transformer's self-attention mechanism, which has proven effective in capturing long-range dependencies within sequences [31]. Specifically, we apply the self-attention module separately to the facial expression feature sequence $\bar{\mathcal{S}} = \langle f^v, f_1^v, f_2^v, \ldots, f_N^v \rangle$ and the context feature sequence $\tilde{\mathcal{S}} = \langle h^v, h_1^v, h_2^v, \ldots, h_N^v \rangle$, to capture the temporal dynamics of the facial expression and context information across video frames, respectively. Note that f^v in $\bar{\mathcal{S}}$ and h^v in $\tilde{\mathcal{S}}$ are special learnable vectors for the class token. After being updated, these class tokens encode the temporal information of the facial expression and context features learned by the self-attention module, respectively, *i.e.*,

$$\begin{cases} \bar{f}^v = \text{S-ATT}(\bar{\mathcal{S}}) \\ \bar{h}^v = \text{S-ATT}(\tilde{\mathcal{S}}) \end{cases}, \tag{3}$$

where S-ATT$(\cdot)$ denotes the self-attention module [31], and $\bar{f}^v$ and $\bar{h}^v$ represent the updated class tokens for f^v and h^v, respectively. In our implementation, the S-ATT module adheres to the conventional Transformer encoder architecture [31]. Finally, $\bar{f}^v$ and $\bar{h}^v$ are concatenated and fed into a fully connected layer, producing the visual feature token $\mathbf{f}^v$, which represents both the facial expression and context information within the academic emotion video:

$$\bar{f}^v \oplus \bar{h}^v \xrightarrow[\text{network}]{\text{fully connected}} \mathbf{f}^v, \tag{4}$$

where $\oplus$ denotes the concatenation operation.

Context-Aware Text Encoder. Following the approach in [40], we use text descriptions instead of class names to represent each emotion category. For each academic emotion, we describe not only the associated facial expressions but also the relevant context learning behaviors. In our implementation, rather than manually designing these descriptions, we employ a large language model, such as ChatGPT [26], to generate them automatically. Further details are provided in the supplementary material. Additionally, a learnable text prompt [41] is included to capture additional relevant details for each category during training. The structure of the prompt embeddings for each category is

$$\mathcal{P}_k = [p]_k^1 [p]_k^2 \cdots [p]_k^M [\text{Tokenizer}(\mathcal{T}_k)], \tag{5}$$

where $\mathcal{P}_k$ and $\mathcal{T}_k$ denote the prompt embeddings and the fixed text description for the k-th category, respectively, where $k \in \{1, 2, ..., K\}$, K is the number of academic emotion classes, and M is the number of tokens for the learnable text prompt, set to $M = 8$ in our implementation. Each learnable token $[p]_k^j \in \mathbb{R}^{512}$, with $j \in \{1, 2, ...M\}$, matches the word embeddings input to the CLIP text encoder. Finally, each $\mathcal{P}_k$ is fed into the CLIP text encoder, producing the text feature token for the corresponding category:

$$\mathbf{f}_k^t = \text{CLIP-TE}(\mathcal{P}_k), \tag{6}$$

where CLIP-TE$(\cdot)$ is the CLIP text encoder, and $\mathbf{f}_k^t$ is the text feature token for the k-th academic emotion category.

Classification Loss. The probability of the input video belonging to each academic emotion category is computed:

$$P_\theta(y = k|\mathcal{S}) = \frac{e^{\cos(\mathbf{f}_k^t, \mathbf{f}^v)}}{\sum_{j=1}^{K} e^{\cos(\mathbf{f}_j^t, \mathbf{f}^v)}}, \tag{7}$$

where $P_\theta(y = k|\mathcal{S})$ represents the probability that the input video $\mathcal{S}$ belongs to the k-th category, $\cos(\cdot, \cdot)$ denotes cosine similarity, and θ represents the parameters of our model, which primarily consist of a shared CLIP model and a temporal encoder module. We adopt the cross-entropy loss as the classification loss function. During training, we keep the CLIP text encoder fixed, fine-tune its image encoder, and optimize the temporal encoder module in the visual block, with the entire model trained end-to-end.

4. Experiments

4.1. Experimental Settings

We have conducted evaluations on RAER for the automatic recognition of academic emotions in real-world learning scenarios. To objectively evaluate the models, we split the RAER dataset into training (80%) and testing (20%) sets, striving to exclude overlapping individuals while maintaining a nearly identical distribution of academic emotions across both sets. Given the class imbalance in our dataset, we use unweighted average recall (UAR), defined as the average per-class accuracy, as the evaluation metric, instead of the accuracy metric, which is sensitive to bias and ineffective for imbalanced data. We implemented CLIP-CAER

Table 2. Comparison with SOTA methods on the academic emotion benchmark RAER. 'TF' denotes Transformer.

Method	Temporal Model			Context	UAR(%)
	3DCNN	LSTM	TF		
3DResNets18 [11]	✓			✘	43.98
I3D [6]	✓			✘	44.63
CAER-Net [17]	✓			✔	45.92
M3DFEL [32]	✓	✓	✓	✘	50.82
Former-DFER [39]			✓	✘	51.26
CLIPER [19]			✓	✘	53.18
DFER-CLIP [40]			✓	✘	61.19
CLIP-CAER (Ours)			✓	✔	**68.00**

using PyTorch, utilizing the CLIP model [27] with a ViT-B/32 architecture and its pre-trained weights. For more details, please refer to the supplementary material.

4.2. Comparison with State-of-the-art Methods

We first compare the performance of our proposed CLIP-CAER against current state-of-the-art video-based approaches on RAER for academic emotion recognition in natural learning contexts. Additionally, we further demonstrate the effectiveness of our method on the basic emotion dataset CAER [17], which contains context information.

Results on our RAER Dataset. To ensure a fair comparison, we re-implemented several state-of-the-art methods using their publicly available code. For methods like CLIPER [19], DFER-CLIP [40], and our proposed CLIP-CAER, which are built upon the CLIP pretraining model, no additional pretraining was required. In contrast, other methods were first pre-trained on the basic emotion dataset DFEW [13] before undergoing fine-tuning on the RAER dataset. Table 2 presents the quantitative results of various methods on the RAER test set. The results indicate that methods using Transformers to capture temporal dynamics across frames outperform those relying on 3D CNNs for extracting video temporal information. Additionally, due to the typically smaller size of academic emotion datasets, leveraging the vision-language pre-training model CLIP has proven effective in boosting performance, as seen in the results of CLIPER [19], DFER-CLIP [40], and our CLIP-CAER. Furthermore, by integrating context information from the learning environment, our CLIP-CAER achieves the highest classification performance, outperforming the runner-up by as much as **6.81%**. As shown in Fig. 4, CLIP-CAER consistently outperforms across nearly all categories. Notably, it achieves a significant performance boost of nearly **20%** in the "distraction" category, underscoring the critical role of context information in distinguishing between distraction and engagement. Note that although CAER-Net [17] also considers context information, it struggles to capture relevant learning contexts due to the relatively small size of RAER, as demonstrated in Fig. 5.

Results on the CAER Dataset. We further evalu-

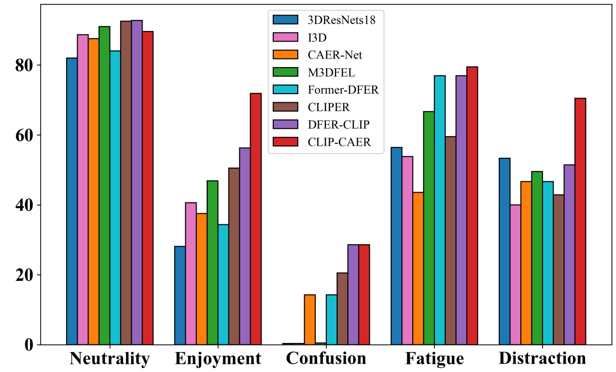

Figure 4. Comparison (UAR %) with SOTA methods on each category in the academic emotion benchmark RAER.

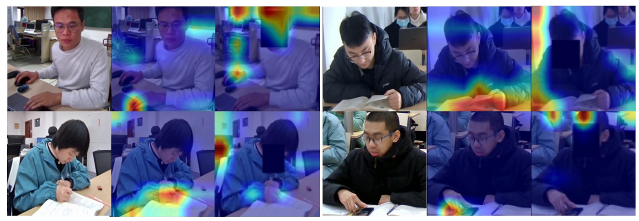

Figure 5. Visualization of context attention using Grad-CAM [28]. For each example: left, input image; middle, our model; right, CAER-Net [17] with face area masked as input.

ate our method on the basic emotion dataset CAER [17]. Due to space limitations, the results are presented in the supplementary material. These results show that the proposed CLIP-CAER, by incorporating context information, achieves an **81.77%** accuracy, significantly surpassing facial expression-based methods and outperforming the state-of-the-art context-aware CAER-Net [17] by **4.73** points.

4.3. Ablation Study

To assess the contribution of each component of CLIP-CAER, and evaluate the impact of different prompt strategies on performance, we conduct extensive ablation experiments on the real-world academic emotion dataset RAER.

The Design of CLIP-CAER. Fig. 6 evaluates the impact of different visual block designs: (a) using facial image sequences to model only facial expression features; (b) using full-frame video sequences to jointly capture facial expression features and context information; (c) combining facial image sequences and full-frame sequences to separately model facial expression features and context information. The results indicate that using only facial expressions yields an accuracy of 61.19%, whereas using full-frame video sequences to jointly capture facial expressions and context information not only fails to enhance accuracy but reduces it by **3.16%**. By combining facial image sequences with full-frame sequences, we effectively integrate facial expressions and context cues, resulting in a **6.81%** improvement in performance. The confusion matrix in Fig. 6 reveals that for the "distraction" category, both model (b) and model

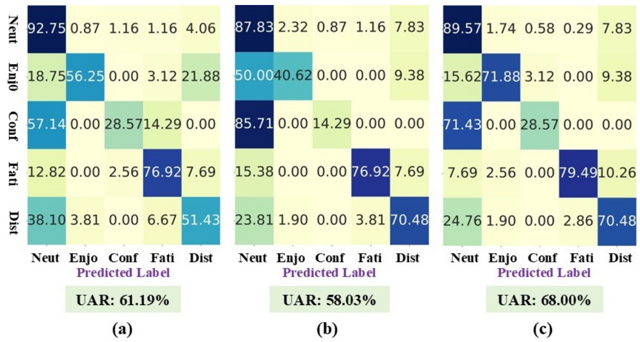

Figure 6. Ablation of different input designs in CLIP-CAER.

(c), which incorporate context information, achieve an accuracy of **70.48%**, while model (a), which considers only facial expressions and ignores context information, reaches just **51.43%**. This aligns with our observations and feedback from annotators, who often rely on context information to determine whether a learner is distracted or engaged in studying. However, with model (b), accuracy drops significantly for emotions like "enjoyment" and "confusion," which are typically identified based on facial expressions. This supports our observation that the face region, being relatively small compared to the surrounding context, may be overlooked by this model.

Impact of Different Prompt Strategies. The proposed CLIP-CAER differs from conventional classification models in that it utilizes prompts to create classifier-free predictions, making prompt engineering a crucial component of the approach. We compare different prompt strategies in Table 3. We see that the strategy incorporating the learnable prompt consistently outperforms its counterpart without it. Moreover, our method surpasses the approach of using class names as prompts with a learnable prompt, highlighting the effectiveness of using descriptive prompts. We believe this is primarily because, compared to class names, descriptive prompts offer a more detailed and accurate representation of behaviors, including specific expressions and actions associated with academic emotions, such as yawning in fatigue, fiddling with fingers, or using a phone in distraction.

Cross-Cultural Generalization via Indirect Validation. Our RAER dataset primarily includes Asian students. Unlike basic emotions, authentic academic emotions occurring in natural educational settings are difficult to obtain directly from online sources. Moreover, gathering real-world educational videos featuring students from diverse cultural backgrounds poses significant challenges, especially in a non-immigrant context. This raises a critical question: Can a model trained on RAER effectively generalize across different cultural contexts? To address this question, we conducted an indirect validation experiment using a newly collected dataset named JuniorRAER. Compared to RAER ($\sim$2700 videos), JuniorRAER is smaller ($<$ 470 videos). Furthermore, unlike RAER, which focuses on

Table 3. Evaluation of different prompts used in CLIP-CAER.

Prompts	UAR(%)
an emotion of [Class] during studying	62.14
[Learnable Prompt][Class] during studying	64.26
[Descriptors]	65.43
[Learnable Prompt] [Descriptors]	**68.00**

Table 4. Evaluation of CLIP-CAER on JuniorRAER test set: a) trained on RAER; b) trained on JuniorRAER; c) fine-tuned on a).

Models	Neut.	Enjo.	Conf.	Fati.	Dist.	UAR (%)
a)	58.16	10.00	0.00	75.00	64.00	41.43
b)	100.00	0.00	0.00	0.00	0.00	20.00
c)	95.04	80.00	0.00	66.67	64.30	61.20

adult university students, JuniorRAER captures academic emotions from primary school students around 10 years old during classroom activities. Details about JuniorRAER are provided in the supplementary material. This experiment explores the model's generalization capability beyond its original training domain, given notable differences in facial appearances, study habits, and cognitive developmental stages between primary school and university students. Additionally, variations in educational environments and cognitive engagement levels further enhance insights into the model's potential for cross-cultural applicability.

As presented in Table 4, the model trained on RAER demonstrates robust generalization to the JuniorRAER test set. Directly training on JuniorRAER alone leads to overfitting to the dominant neutrality class due to the dataset's limited size. Conversely, fine-tuning a model pre-trained on RAER substantially enhances performance, indicating that the RAER-trained model possesses strong generalization capabilities. We attribute the model's generalization ability to two key factors. First, the pre-trained CLIP model utilized in our approach was initially trained on a large-scale dataset, capturing a broad spectrum of facial features across various age groups and ethnicities. Second, fundamental academic emotions (engagement, enjoyment, confusion, fatigue, and distraction) are universally experienced across diverse educational contexts. Although specific teaching methods and cultural nuances may differ, the core emotional context associated with learning remains broadly consistent.

5. Conclusion

In this paper, we introduce RAER, a dataset for academic emotion recognition in real-world learning scenarios. Additionally, we propose CLIP-CAER, which uses CLIP with learnable text prompts to combine facial expressions and context cues. Our method significantly outperforms state-of-the-art methods, emphasizing the crucial role of context in accurately recognizing academic emotions. Our work provides a foundational dataset and approach for advancing research in academic emotion analysis.

Acknowledgment

This work was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (2024C01167) and the Fundamental Research Funds for the Provincial Universities of Zhejiang (FR24005Z).

References

[1] Mohammed Aly. Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. *Multimedia Tools and Applications*, 2024. 1

[2] Wissam J. Baddar and Yong Man Ro. Mode variational lstm robust to unseen modes of variation: application to facial expression recognition. In *AAAI*, 2019. 4

[3] Mohamed Ben Ammar, Mahmoud Neji, Adel. M. Alimi, and Guy Gouardères. The affective tutoring system. *Expert Systems with Applications*, 37(4):3013–3023, 2010. 1

[4] Cunling Bian, Ya Zhang, Fei Yang, Wei Bi, and Weigang Lu. Spontaneous facial expression database for academic emotion inference in online learning. *IET Computer Vision*, 13(3):329–337, 2019. 1, 2, 3, 4

[5] Monique Boekaerts. Understanding students' affective processes in the classroom. In *Emotion in Education*, pages 37–56. 2007. 1

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 7

[7] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, page 445–450, 2016. 4

[8] Bei Fang, Xian Li, Guangxin Han, and Juhou He. Facial expression recognition in educational research from the perspective of machine learning: A systematic review. *IEEE Access*, 11:112060–112074, 2023. 1

[9] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. 4

[10] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild, 2022. 1, 2, 3, 4

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 7

[12] Qiaoping Hu, Chuanneng Mei, Fei Jiang, Ruimin Shen, Yitian Zhang, Ce Wang, and Junpeng Zhang. Rfau: A database for facial action unit analysis in real classrooms. *IEEE Transactions on Affective Computing*, 13(3):1452–1465, 2022. 3, 4

[13] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *ACM MM*, page 2881–2889, 2020. 4, 7

[14] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2018. 1, 2, 3, 4

[15] ChanMin Kim and Charles B. Hodges. Effects of an emotion control treatment on academic emotions, motivation and achievement in an online mathematics course. *Instructional Science*, 40(1):173–192, 2012. 1

[16] B. Kort, R. Reilly, and R.W. Picard. An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In *Proceedings IEEE International Conference on Advanced Learning Technologies*, pages 43–46, 2001. 1, 3

[17] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-Aware Emotion Recognition Networks . In *ICCV*, pages 10142–10151, 2019. 4, 7

[18] Jeniffer Xin-Ying Lek and Jason Teo. Academic emotion classification using fer: A systematic review. *Human Behavior and Emerging Technologies*, 2023(1):9790005, 2023. 1

[19] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. 5, 7

[20] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022. 1

[21] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2584–2593, 2017. 4

[22] Baolong Liu, Tianyi Zheng, Peng Zheng, Daizong Liu, Xiaoye Qu, Junyu Gao, Jianfeng Dong, and Xun Wang. Litemkd: A multi-modal knowledge distillation framework for lightweight few-shot action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7283–7294, 2023. 4

[23] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *ACM MM*, page 24–32, 2022. 4

[24] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023. 4

[25] Reinhard Pekrun. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educ Psychol Rev*, 18: 315–341, 2006. 1

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 6

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, page 8748–8763, 2021. 2, 4, 5, 7

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 7

[29] Shengkai Sun, Daizong Liu, Jianfeng Dong, Xiaoye Qu, Junyu Gao, Xun Yang, Xun Wang, and Meng Wang. Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2973–2984, 2023. 4

[30] Xiao-Yu Tang, Wang-Yue Peng, Si-Rui Liu, and Jian-Wen Xiong. Classroom teaching evaluation based on facial expression recognition. In *Proceedings of the 2020 9th International Conference on Educational and Information Technology*, page 62–67, 2020. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, page 6000–6010, 2017. 6

[32] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *CVPR*, pages 17958–17968, 2023. 4, 7

[33] Qinglan Wei, Bo Sun, Jun He, and Lejun Yu. Bnu-lsved 2.0: Spontaneous multimodal student affect database with multi-dimensional labels. *Signal Processing: Image Communication*, 59:168–181, 2017. 3, 4

[34] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1): 86–98, 2014. 1, 2, 3, 4

[35] Zhi-Fan Wu and Jinshi Cui. La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In *ICCV*, pages 20698–20707, 2023. 1

[36] Jing Yang, Yordan Hristov, Jie Shen, Yiming Lin, and Maja Pantic. Toward robust facial action units' detection. *Proceedings of the IEEE*, 111(10):1198–1214, 2023.

[37] Wenwu Yang, Jinyi Yu, Tuo Chen, Zhenguang Liu, Xun Wang, and Jianbing Shen. Multi-threshold deep metric learning for facial expression recognition. *Pattern Recognition*, 156:110711, 2024.

[38] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *ECCV*, pages 418–434, 2022. 1

[39] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *ACM MM*, page 1553–1561, 2021. 4, 7

[40] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. In *BMVC*, pages 1–14, 2023. 4, 5, 6, 7

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022. 6