Learning Emotion Representations from Verbal and Nonverbal Communication

Sitao Zhang* Yimu Pan* James Z. Wang
The Pennsylvania State University, University Park, Pennsylvania, USA
{sbz5211, ymp5078, jwang}@psu.edu

Abstract

Emotion understanding is an essential but highly challenging component of artificial general intelligence. The absence of extensive annotated datasets has significantly impeded advancements in this field. We present EmotionCLIP, the first pre-training paradigm to extract visual emotion representations from verbal and nonverbal communication using only uncurated data. Compared to numerical labels or descriptions used in previous methods, communication naturally contains emotion information. Furthermore, acquiring emotion representations from communication is more congruent with the human learning process. We guide EmotionCLIP to attend to nonverbal emotion cues through subject-aware context encoding and verbal emotion cues using sentiment-guided contrastive learning. Extensive experiments validate the effectiveness and transferability of EmotionCLIP. Using merely linear-probe evaluation protocol, EmotionCLIP outperforms the state-of-the-art supervised visual emotion recognition methods and rivals many multimodal approaches across various benchmarks. We anticipate that the advent of EmotionCLIP will address the prevailing issue of data scarcity in emotion understanding, thereby fostering progress in related domains. The code and pre-trained models are available at https://github.com/Xeaver/EmotionCLIP.

1. Introduction

If artificial intelligence (AI) can be equipped with emotional intelligence (EQ), it will be a significant step toward developing the next generation of artificial general intelligence [46, 92]. The combination of emotion and intelligence distinguishes humans from other animals. The ability to understand, use, and express emotions will significantly facilitate the interaction of AI with humans and the environment [20, 48–50], making it the foundation for a wide variety of HCI [3], robotics [11], and autonomous driving [31] applications.

*equal contribution

Artificial emotional intelligence (AEI) research is still in its nascency [30, 73]. The recent emergence of pre-trained models in CV [10, 24, 62] and NLP [7, 16, 33, 68] domains has ushered in a new era of research in related subjects. By training on large-scale unlabeled data in a self-supervised manner, the model learns nontrivial representations that generalize to downstream tasks [42, 62]. Unfortunately, such a technique remains absent from AEI research. The conventional approaches in visual emotion understanding have no choice but to train models from scratch, or leverage models from less-relevant domains [27, 66], suffering from data scarcity [29, 45]. The lack of pre-trained models greatly limits the development of AEI research.

Research in neuroscience and psychology offers insights for addressing this problem. Extending from the capabilities that have been coded genetically, humans learn emotional expressions through daily interaction and communication as early as when they are infants. It has been shown that both vision [58] and language [41] play crucial roles in this learning process. By absorbing and imitating expressions from others, humans eventually master the necessary
feelings to comprehend emotional states by observing and analyzing facial expressions, body movements, contextual environments, etc.

Inspired by how humans comprehend emotions, we propose a new paradigm for emotion understanding that learn directly from human communication. The core of our idea is to explore the consistency between verbal and nonverbal affective cues in daily communication. Fig. 1 shows how communication reveals emotion. Our method that learns from communication is not only aligned with the human learning process but also has several advantages:

1) Our method bypasses the problems in emotion data collection by leveraging uncurated data from daily communication. Existing emotion understanding datasets are mainly annotated using crowdsourcing [29, 45, 77]. For image classification tasks, it is straightforward for annotators to agree on an image’s label due to the fact that the label is determined by certain low-level visual characteristics. However, crowdsourcing participants usually have lower consensus on producing emotion annotations due to the subjectivity and subtlety of affective labels [90]. This phenomenon makes it extremely difficult to collect accurate emotion annotations on a large scale. Our approach does not rely on human annotations, allowing us to benefit from nearly unlimited web data.

2) Our use of verbal expressions preserves fine-grained semantics to the greatest extent possible. Limited by the data collection strategy, existing datasets usually only contain annotations for a limited number of emotion categories, which is far from covering the space of human emotional expression [86]. Moreover, the categorical labels commonly used in existing datasets fail to precisely represent the magnitude or intensity of a certain emotion.

3) Our approach provides a way to directly model expressed emotion. Ideally, AEI should identify the individual’s emotional state, i.e., the emotion the person desires to express. Unfortunately, it is nearly impossible to collect data on this type of “expressed emotion” on a large scale. Instead, the current practice is to collect data on “perceived emotion” to approximate the person’s actual emotional state, which inevitably introduces noise and bias to labels.

In general, learning directly from how humans express themselves is a promising alternative that gives a far broader source of supervision and a more comprehensive representation. This strategy is closely analogous to the human learning process and provides an efficient solution for extracting emotion representations from uncurated data.

We summarize our main contributions as follows:

- We introduce EmotionCLIP, the first vision-language pre-training paradigm using uncurated data to the visual emotion understanding domain.
- We propose two techniques to guide the model to capture salient emotional expressions from human verbal and nonverbal communication.

Extensive experiments and analysis demonstrate the superiority and transferability of our method on various downstream datasets in emotion understanding.

2. Related Work

Emotion Recognition from Visual Clues. Facial expression recognition [19, 38, 69] has been well studied in the field of emotion recognition, mainly because faces are not only expressive but also easy to model. Handcrafted features have been developed to describe different facial expressions [13, 43, 74]. Recently, deep learning-based approaches have begun to emerge [38]. Principal research focuses on the design of novel modules for conventional network architectures [88], distinct loss functions for facial tasks [15, 71, 85], and addressing label uncertainties [8, 80].

Recently, with the growing interest in recognizing emotion in the wild, the focus of research has gradually shifted to modeling body language [21, 45, 59] and context [29, 53, 54]. Several datasets for understanding human emotional states in unconstrained environments have been proposed [5, 60, 77]; Kosti et al. [29] and Yu et al. [45] established the first benchmark for image and video data, respectively. Follow-up work mainly focuses on context-aware emotion recognition, which usually adopts a multi-branch structure where one branch focuses on the face or body and the other focuses on capturing context [12, 34, 53, 59]. Moreover, some approaches take into account temporal causality [54] or represent context information via graphs [96]. To the best of our knowledge, there are no pre-trained models or effective methods for leveraging unlabeled data in the domain of visual emotion recognition.

Vision-Language Pre-training. Visual-language pre-training has achieved remarkable progress recently. CLIP [62] demonstrated the feasibility of using contrastive learning [10, 56] to learn transferable and powerful visual representations from large-scale image-text pairs [72]. Many follow-up approaches have been proposed to transfer the pre-trained model to downstream tasks [22, 97, 100] or leverage the scheme for different domains [37, 57, 64, 83, 98]. A line of research endeavors to expand CLIP for general video understanding [26, 35, 40, 44, 55, 82, 87]. The majority of the effort focuses on fine-tuning datasets with textual annotations [27, 32, 59]. However, not only are these curated annotations challenging to obtain, but they also limit the model’s potential in various applications. Another line of work [36, 87] extends the image-level pre-training by utilizing unlabeled narrated videos [2, 52], similar to ours. However, we aim to learn abstract emotion representations rather than low-level visual patterns, which are beyond the reach of current models. EmotionNet [84] and its sequel [99], which likewise seeks to learn visual emotion repre-
sentiment scores by a text encoder and sentiment analysis model, respectively. The model learns emotion representations under sentiment guidance in a contrastive manner, by exploring the consistency of verbal and nonverbal communication.

3. Methodology

Our core idea is to learn directly from human communication how they express their emotions, by exploring the consistency between their verbal and nonverbal expressions [41]. We tackle this learning task under the vision-language contrastive learning framework [62], that is, the model is expected to learn consistent emotion representations from the verbal expressions (e.g., utterance and dialogue) and nonverbal expressions (e.g., facial expression, body language, and contextual environment) of the same individuals. We give a brief introduction of our data collection procedure in Sec. 3.1 before presenting the overview of EmotionCLIP in Sec. 3.2. We further discuss how the model is guided to learn emotion-relevant representations from the nonverbal perspective in Sec. 3.3, and from the verbal perspective in Sec. 3.4. Please see Appendix for details of the dataset and implementations.

3.1. Data Collection

Publicly available large-scale vision-and-language datasets do not provide desired verbal and nonverbal information because they either comprise only captions of low-level visual elements [2, 72] or instructions of actions [52]. The captions mostly contain a brief description of the scene or activity, which is insufficient to reveal the underlying emotions; the instruction videos rarely include humans in the scene or express neutral emotions, which fail to provide supervision signals for emotion understanding. To overcome such problems, we gather a large-scale video-and-text paired dataset. More specifically, the videos are TV series, while the texts are the corresponding closed captions. We collected 3,613 TV series from YouTube, which is equivalent to around a million raw video clips. We processed them using the off-the-shelf models to group the words in closed caption into complete sentences [23], tag each sentence with a sentiment score [68], and extract human bounding boxes [79].

3.2. Overview of EmotionCLIP

Fig. 2 presents an overview of our approach. We follow the wildly adopted vision-language contrastive learning paradigm [62] where two separate branches are used to encode visual inputs (i.e., nonverbal expressions) and textual inputs (i.e., verbal expressions), respectively.

**Video Encoding.** The visual branch of EmotionCLIP takes two inputs, including a sequence of RGB frames $X_v$ and a sequence of binary masks $X_m$. The binary mask has the same shape as the frame and corresponds to the frame one-to-one, indicating the location of the subject within the frame. The backbone of the subject-aware frame encoder $f_s$ is a Vision Transformer [17]. In particular, it extracts $m$ non-overlapping image patches from the frame and projects them into 1D tokens $z_i \in \mathbb{R}^d$. The sequence of tokens passed to the following Transformer encoder [76] is $z = [z_1, \ldots, z_m, z_{cls}, z_{hmn}]$, where $z_{cls}, z_{hmn}$ are two additional learnable tokens. The mask $X_m$ is converted to an array of indices $P$ indicating the image patches containing the subject. The frame encoder further encodes $z, P$ into a frame-level representation. All frame representa-
We intend to guide the model to focus on the interaction between the subject of interest and context. As shown in Fig. 3, the cropped character and the whole image are usually encoded by two separate networks and fused at the ends [34, 53, 59]. This approach is inflexible and inefficient since it overlooks the dependency between subject and context and encodes redundant image portions. Following this line of thought, we propose two potential subject-aware context encoding strategies, i.e., subject-aware attention masking (SAAM) and subject-aware prompting (SAP). The former can be regarded as an efficient implementation of the traditional two-stream approach but avoids the problem of redundant encoding. The latter is a novel encoding strategy that enables adaptive modeling of the interaction between the context and the subject by providing necessary prompts.

3.3. Subject-Aware Context Encoding

Context encoding is an important part of emotion understanding, especially in unconstrained environments, as it has been widely shown in psychology that emotional processes cannot be interpreted without context [47, 51, 65]. We intend to guide the model to focus on the interaction between the subject of interest and context. As shown in Fig. 3, the cropped character and the whole image are usually encoded by two separate networks and fused at the ends [34, 53, 59]. This approach is inflexible and inefficient since it overlooks the dependency between subject and context and encodes redundant image portions. Following this line of thought, we propose two potential subject-aware context encoding strategies, i.e., subject-aware attention masking (SAAM) and subject-aware prompting (SAP). The former can be regarded as an efficient implementation of the traditional two-stream approach but avoids the problem of redundant encoding. The latter is a novel encoding strategy that enables adaptive modeling of the interaction between the context and the subject by providing necessary prompts.

Subject-Aware Attention Masking. The canonical attention module [76] in a Transformer is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^\top}{\sqrt{d}} \right) V .$$  (1)

We model the context and subject in a synchronous way by modifying the attention module to

$$\text{Attention}^*(Q, K, V, U) = \text{softmax} \left( \frac{QK^\top}{\sqrt{d}} \right) (J - A)V + \text{softmax} \left( \frac{QK^\top}{\sqrt{d}} \right) AUV ,$$  (2)

where $J$ is a matrix with all ones, $A$ is a learnable parameter containing values in range $[0, 1]$, and $U$ is a weight matrix constructed using $P$. Intuitively, we shift $A$ amount of attention from a total of $J$ amount of attention from the context to the subject. To partition the $A$ amount of attention to all image patches containing subject, we compute $U$ as the following:

$$U = \text{softmax} \left( \frac{QK^\top + M}{\sqrt{d}} \right) .$$  (3)

The masking matrix $M$ is defined as:

$$M = \begin{bmatrix} M^{(1)} & M^{(2)} \\ M^{(3)} & M^{(4)} \end{bmatrix} ,$$  (4)

where $M^{(1)} = 0_{(m+1) \times (m+1)}$, $M^{(4)} = 0_{1 \times 1}$, $M^{(2)}_{i\notin P} = -\infty$, $M^{(3)}_{(m+1) \notin P} = -\infty$, and all other entries are zero. Intuitively, $M^{(2)}$ and $M^{(3)}$ represent the attention between all image patches $z_i$ and the human token $z_{hmn}$ to model the subject stream; we mask out all attention between non-human patches and the human token. Moreover, we mask out attention from $z_{hmn}$ to $z_{cls}$, $M^{(3)}_{(m+1)}$, to ensure $z_{hmn}$ only encodes the subject.

Subject-Aware Prompting. Prompting is a parameter-free method that restricts the output space of the model by shaping inputs. In our case, we hope to prompt the model to distinguish between the context and the subject. A recent visual prompting method, CPT [89], provides such a prompt by altering the original image, i.e., imposing colored boxes on objects of interest. It shows that a Transformer is able to locate objects with the help of positional hints. However, introducing artifacts on pixel space may not be optimal as it causes large domain shifts. To address this issue, we propose to construct prompts in the latent space based on positional embeddings, considering that they are inherently designed as indicative information. Formally, let $e_i$ be
the positional embedding corresponding to the patch token $z_i$, and $P$ are the indicator set of the subject location. The prompting token is designed as $z_{bmn} = \sum_{i \in P} e_i$.

We argue the sum of positional embeddings is enough to provide hints about the subject location. The previous study [78] demonstrates a Transformer treats all tokens without positional embedding uniformly but with positional embedding differently. This result shows positional embeddings play a vital role in guiding model attention.

### 3.4. Sentiment-Guided Contrastive Learning

We train the model to learn emotion representations from verbal and nonverbal expressions in a contrastive manner. In the traditional contrastive setting, the model is forced to repel all negative pairs except the only positive one. However, many expressions in daily communication indeed have the same semantics from an emotional perspective. Contrasting these undesirable negative pairs encourages the model to learn spurious relations. This problem comes from false negatives, i.e., the affectively similar samples are treated as negatives. We address this issue by introducing a trained sentiment analysis model [68] from the NLP domain for the suppression of false negatives, thereby guiding our model to capture emotion-related concepts from verbal expressions. Specifically, we propose a sentiment-guided contrastive loss:

$$
SNCE(v, t, s) = -\sum_{i \in B} \log \frac{\exp (v_i \cdot t_i / \tau)}{\sum_{j \in B} \exp (v_i \cdot t_j / \tau - w_{i,j})},
$$

where $B$ is a batch. The reweighting term $w_{i,j}$ is defined as

$$
w_{i,j} = \begin{cases} 
\beta \cdot \text{KL} (s_i || s_j) ^{-1} & i \neq j \\
0 & i = j 
\end{cases},
$$

where $\beta$ is a hyper-parameter for controlling the reweighting strength. The total loss is defined as:

$$
L = \frac{1}{2|B|} \left( SNCE(v, t, s) + SNCE(t, v, s) \right).
$$

As shown in Fig. 4, the false negative sample with similar emotion to the positive sample is greatly suppressed, while other negatives are not affected. Note that when $i \neq j$ but $s_i = s_j$, we have $w_{i,j} = \infty$, which is equivalent to removing $j$th sample from the negative pairs; when $s_i$ and $s_j$ are very different, $w_{i,j}$ is negligible; we set $w_{i,i} = 0$ to not affect the true positive pair. Since $s$ is the sentiment score, the sentiment-related differences are emphasized and weighted more during training. Therefore, the proposed contrastive loss is expected to provide cleaner supervision signals for learning emotion-related representations.

### 4. Experiments and Results

We first introduce the datasets for evaluation in Sec. 4.1 before analyzing various components of EmotionCLIP in Sec. 4.2. Then, we compare EmotionCLIP with the state-of-the-art methods on various datasets in Sec. 4.3. Please see Appendix for more experimental results.

#### 4.1. Datasets and Evaluation Metrics

We evaluate the performance of EmotionCLIP on a variety of recently published challenging benchmarks, including four video datasets and an image dataset. The annotations of these datasets are mainly based on three physiological models: Ekman’s basic emotion theory [18] (7 discrete categories), the fine-grained emotion model [14] (26 discrete categories), and the Valence-Arousal-Dominance emotion model [67] (3 continuous dimensions). The evaluation metrics are consistent with previous methods.

**BoLD** [45] is a dataset for understanding human body language in the wild, consisting of 9,827 video clips and 13,239 instances, in which each instance is annotated with 26 discrete categories and VAD dimensions.

**MovieGraphs** [77] is a dataset for understanding human-centric situations consisting of graph-based annotations on social events that appeared in 51 popular movies. Each graph comprises multiple types of nodes to represent actors’ emotional and physical attributes, as well as their relationships and interactions. Following the preprocessing and evaluation protocol proposed in previous work [29, 54], we extract relevant emotion attributes from the graphs and group them into 26 discrete emotion categories.

**MELD** [60] is an extension to the EmotionLines [9], which is an emotion corpus of multi-party conversations initially proposed in the NLP domain. It offers the same dialogue examples as EmotionLines and includes audio and visual modalities along with the text. It contains around 1,400 dialogues and 13,000 utterances from the Friends tv show, where each example is annotated with 7 discrete categories.

**Liris-Accede** [5] is a dataset that contains videos from a set of 160 professionally made and amateur movies covering a
Table 1. Component-wise analysis of our method on BoLD.

<table>
<thead>
<tr>
<th></th>
<th>mAP</th>
<th>AUC</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmotionCLIP (vanilla)</td>
<td>21.97</td>
<td>68.85</td>
<td>0.130</td>
</tr>
<tr>
<td>+ SAAM</td>
<td>21.53</td>
<td>68.56</td>
<td>0.137 ±0.007</td>
</tr>
<tr>
<td>+ SAP</td>
<td>22.28</td>
<td>69.06</td>
<td>0.131 ±0.001</td>
</tr>
<tr>
<td>+ SAP &amp; SNCE</td>
<td>22.51</td>
<td>69.30</td>
<td>0.133 ±0.003</td>
</tr>
</tbody>
</table>

4.2. Ablation Study

4.2.1 Analysis of Subject-Aware Context Encoding

In this series of experiments, we start with a vanilla model and analyze it by adding various subject-aware approaches. As shown in Table 1, decent results can be achieved in downstream tasks using the vanilla EmotionCLIP. This result supports our argument that models can learn non-trivial emotion representations from human verbal and nonverbal expressions by matching them together.

The SAP achieves better results and improves over the baseline by a reasonable margin. This improvement demonstrates the design of SAP can incorporate location-specific information to guide the model in acquiring target-related content without impacting global information modeling.

Additionally, we note that the model with SAAM yields mediocre performance. As discussed earlier, SAAM can be regarded as an efficient implementation of the multi-stream strategy in the Transformer. This outcome suggests that the multi-stream strategy, commonly used in previous methods, may not be optimal. To rule out the possibility of fusion at inappropriate layers, we explore the impact of different fusion positions by applying SAAM up to a certain layer in the Transformer. It shows that the performance change does not correlate to the fusion layer change, and SAAM consistently underperforms SAP, irrespective of the fusion location. This finding implies that imposing hard masks on the model’s attention may introduce unanticipated biases, while adaptively modeling context-subject interaction is more reasonable. In subsequent experiments and discussions, we use SAP as the standard implementation, unless otherwise stated.

Qualitative Analysis. SAP offers merely a positional hint, as opposed to the mandatory attention-shifting in SAAM. Since the purpose of SAP is to ensure subject-aware encoding, it is necessary to understand if the attention guidance is appropriate. We analyze SAP by plotting HMN token’s

4.2.2 Analysis of Sentiment-Guided Contrastive Learning

We first compare models trained with different $\beta$, the hyperparameter used to control the strength of reweighting in SNCE. Note that the training objective is equivalent to the vanilla infoNCE loss [56] when $\beta$ is set to zero. As $\beta$ increases, more negative samples within the batch are suppressed. As shown in Table 1 and Fig. 6, reweighting with appropriate strength can significantly increase the performance of the model as it guides the direction of learning by eliminating some significant false negatives. However, an excessively large $\beta$ can hinder the training of the model, which is within expectation. First, the sentiment scores used in the reweighting process are weak pseudo-labels provided by a pre-trained sentiment analysis model, which is not entirely reliable and accurate. Second, previous work has clearly demonstrated that batch size has a decisive impact on self-supervised learning [10, 24, 62]. A too-large $\beta$ will cause too many negative samples to be suppressed, reducing the effective batch size and thus hindering the learning process.

Qualitative Analysis. We show how the text expressing different emotions are treated by our sentiment-guided loss. Given a positive pair, the logits are the scaled similarities between texts and the positive video; the model is penalized on large logits unless it is associated with the positive text.
Table 2. An example batch containing both false negative and true negative samples. The logits represent the similarity between every text input and the positive video from a random epoch during training. \(\rightarrow\) represents the sentiment-guided reweighting process.

<table>
<thead>
<tr>
<th>Text</th>
<th>Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'm sorry to keep you any longer than is necessary. (positive sample)</td>
<td>10.65</td>
</tr>
<tr>
<td>I'm sorry Tom couldn't join us.</td>
<td>12.6</td>
</tr>
<tr>
<td>I hate to say it guys but it's getting late.</td>
<td>9.81</td>
</tr>
<tr>
<td>I still say it was a wild idea.</td>
<td>13.8</td>
</tr>
<tr>
<td>Well, that was truly fascinating.</td>
<td>15.2</td>
</tr>
</tbody>
</table>

As shown in Table 2, the texts in the second and third rows provide undesired contributions to the loss as they express similar emotions as the positive sample. After reweighting, false negatives (2nd and 3rd) are effectively eliminated while true negatives (4th and 5th) are negligibly affected.

### 4.2.3 Analysis of Model Implementation

The frame and text encoders of our model are initialized with the image-text pre-training weights from CLIP [62]. Research in neural science has demonstrated the necessity of basic visual and language understanding capabilities for learning high-level emotional semantics [58, 70]. To validate the effectiveness of using image-text pre-training for weight initialization, we evaluate variants with different implementations.

We first consider encoders with random initialization. As shown in Table 3, the model’s performance drops sharply when training from scratch. This result is within expectation for two reasons. From an engineering perspective, previous works have demonstrated the necessity of using pre-training weights for large video models [1, 6] and vision-language models [35, 55]. From a cognitive point of view, it is nearly infeasible to learn abstract concepts directly without basic comprehension skills [58]; if the model cannot recognize people, it is impossible to understand body language and facial expressions properly.

We then consider frozen encoders with pre-training weights. This is a standard paradigm for video-language understanding that trains models with offline extracted features [36, 87]. As shown in Table 3, our model with variants using fixed encoders performs worse compared with the model using trainable encoders. This reflects the fact that affective tasks rely on visual and verbal semantics differently from low-level vision tasks, which is what CLIP and its successors overlooked [40].

We study the effect of the temporal encoder. We consider a variant where the temporal encoder is replaced by a mean pooling layer that simply averages the features of all frames. As shown in Table 3, the performance gap is obvious compared with the baseline. This phenomenon suggests that temporal dependency plays a vital role in emotion representations.

Table 3. Ablation study on different model implementations. ✓ means trainable, initialization with pre-training weights. ✗ means frozen, initialization with pre-training weights. - means trainable, random initialization. ○ means no parameters.

<table>
<thead>
<tr>
<th>Text Encoder</th>
<th>Frame Encoder</th>
<th>Temporal Encoder</th>
<th>mAP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>22.51</td>
<td>69.30</td>
</tr>
<tr>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>12.43</td>
<td>54.96</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.02</td>
<td>50.28</td>
</tr>
<tr>
<td>✗</td>
<td>✓</td>
<td>-</td>
<td>13.40</td>
<td>57.38</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>-</td>
<td>18.43</td>
<td>65.08</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>○</td>
<td>21.17</td>
<td>68.74</td>
</tr>
</tbody>
</table>

4.3. Comparison with the State of the Art

Based on our previous ablation experiments, we choose the model with SAP and SNCE as the default implementation and compare it with the state-of-the-art. In addition, we also compare with VideoCLIP [87] and X-CLIP [55], both of which are state-of-the-art vision-language pre-training models for general video recognition purposes. To evaluate the quality of learned representations, we follow the practice in CLIP [62] and use linear-probe evaluation protocol [10, 24] for vision-language pre-training models.

**BoLD.** As shown in Table 4a, EmotionCLIP substantially outperforms the state-of-the-art supervised learning methods on the challenging 26-class emotion classification task and achieves comparable results on continuous emotion regression. It is worth noting that a complex multi-stream model is used in [59] to integrate the human body and context information, while we achieve better results with a single-stream structure using RGB information only. This difference reflects that the subject-aware approach we designed models the relationship between the subject and context. We also notice that other vision-language-based meth-
ods perform poorly on emotion recognition tasks, although they are designed for general video understanding purposes. This phenomenon is largely attributed to the lack of proper guidance; the model can only learn low-level visual patterns and fails to capture semantic and emotional information. **MovieGraphs.** As shown in Table 4b, EmotionCLIP substantially outperforms the best vision-based method and even surpasses Affect2MM [54], a powerful multimodal approach that uses audio and text descriptions in addition to visual information. Instead, other vision-language pre-training models are still far from supervised methods.

**MELD.** EmotionCLIP performs well on MELD as shown in Table 4d; it achieves comparable results to the state-of-the-art vision-based methods. It is worth noting that this dataset is extended from an NLP dataset, so the visual data is noisier than the original text data. In fact, according to the ablation experiments in [12], it is possible to achieve an accuracy of 67.24% using only text, while adding visual modality information only improves the accuracy by about 0.5%. This result explains why our method significantly lags behind multimodal methods using text inputs.

**Liris-Accede.** As shown in Table 4e, EmotionCLIP achieves promising results using visual inputs only. It even competes with many multimodal approaches that are benefited from the use of audio features [93].

**Emotic.** As shown in Table 4c, EmotionCLIP outperforms all RGB-based supervised methods while other vision-language models perform poorly. The improvement of [53] is attributable to the use of additional depth information. This result demonstrates the capability of EmotionCLIP in learning relevant features from complex environments.

### 5. Conclusion

The pre-training methodology, which has brought about significant advancements in numerous CV and NLP domains, has not yet been employed in AEI research. We address this void by introducing EmotionCLIP, the first vision-language pre-training framework that circumvents the need for curated data and annotations. Our study establishes the viability of acquiring generalized emotion representations for curated data and annotations. Our study establishes the evolution of more adaptable and efficacious approaches for affective computing.

**Acknowledgments.** This research was supported by generous gifts from the Amazon Research Awards program. The work used computational resources from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation.
References


[91] Yun Yi, Hanli Wang, and Qinyu Li. CNN features for emotional impact of movies task. In *MediaEval*, 2018. 8


[95] Jie Zhang, Yin Zhao, Longjun Cai, Chaoping Tu, and Wu Wei. Video affective effects prediction with multimodal fusion and shot-long temporal context. *CoRR*, abs/1909.01763, 2019. 8


