

# 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



### Categorical Label:

Shocked Regretful  
 0 1 0 0 0 0 1 0 0 0

### Description:

An old man is chatting with his son while eating at a restaurant.

### Conversation:

- Dad. Who is Dede?  
 - Jesus. She was the love of my life and I was too stupid to realized it. I lost her because of something so dumb.

Figure 1. Emotions emerge naturally in human communication through verbal and nonverbal cues. The rich semantic details within the expression can hardly be represented by human-annotated categorical labels and descriptions in current datasets.

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 cap-

ture 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, 39]. 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-

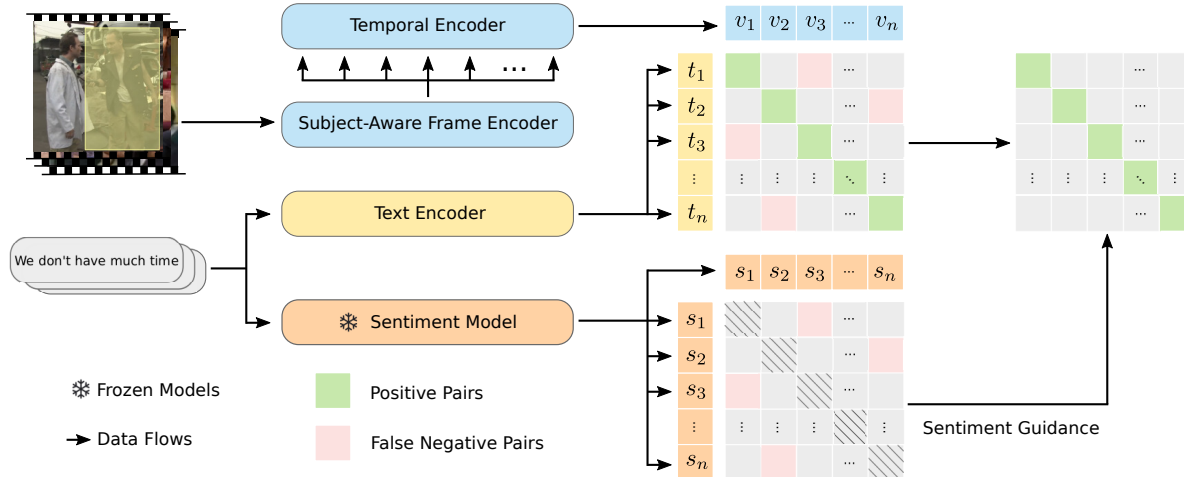


Figure 2. Illustration of **EmotionCLIP**. For *nonverbal communication*, subject and context information is modeled by a frame encoder and further aggregated into video-level representations by a temporal encoder. For *verbal communication*, textual information is encoded as text representations and 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.

sentations, are connected to ours. However, their primary focus is on the images’ stimuli rather than the recognition of human emotional expressions in the wild.

### 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 widely 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  $f_i$  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  $\mathbf{z} = [z_1, \dots, z_m, z_{cls}, z_{hmn}]$ , 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  $\mathbf{z}, P$  into a frame-level representation. All frame representa-

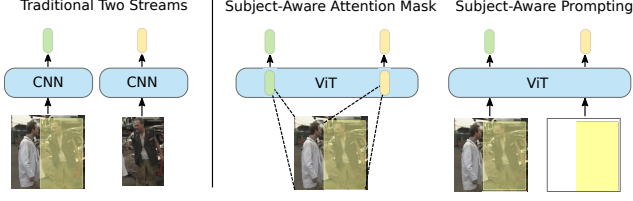


Figure 3. The traditional approaches (*left*) ignore the dependencies between context and subject, and encode a portion of the image redundantly. The proposed approaches (*right two diagrams*) efficiently model the subject and context in a synchronous way.

tions are then passed into the temporal encoder  $f_p$  to produce a video-level representation as  $v = f_v(\mathbf{z}, P)$ , where  $f_v = f_p \circ f_i$  and  $v \in \mathbb{R}^d$ .

**Text Encoding.** The textual branch of EmotionCLIP takes sentences  $X_t$  as inputs. The text encoder  $f_t$  is a Transformer [16] with the architecture modification described in [63], and the sentiment model  $f_s$  is a pre-trained sentiment analysis model [68] that is frozen during training. The input text is encoded by both models as  $t = f_t(X_t)$  and  $s = f_s(X_t)$ , where  $t \in \mathbb{R}^d$  is the representation of the text and  $s \in \mathbb{R}^7$  is the pseudo sentiment score.

**Training Objective.** The training objective is to learn the correspondence between visual inputs and textual inputs by minimizing the sentiment-guided contrastive loss  $\mathcal{L}$ .

We discuss the details of the proposed subject-aware encoding approaches in Sec. 3.3 and the sentiment-guided contrastive learning framework in Sec. 3.4.

### 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:

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

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

$$\text{Attention}^*(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{U}) = \underbrace{\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)(\mathbf{J} - \mathbf{A})\mathbf{V}}_{\text{context}} + \underbrace{\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{A}\mathbf{U}\mathbf{V}}_{\text{subject}}, \quad (2)$$

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

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

The masking matrix  $\mathbf{M}$  is defined as:

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

where  $\mathbf{M}^{(1)} = \mathbf{0}_{(m+1) \times (m+1)}$ ,  $\mathbf{M}^{(4)} = \mathbf{0}_{1 \times 1}$ ,  $\mathbf{M}_{i \notin P}^{(2)} = \mathbf{M}_{i \notin P}^{(3)} = -\infty$ ,  $\mathbf{M}_{(m+1)}^{(3)} = -\infty$ , and all other entries are zero. Intuitively,  $\mathbf{M}^{(2)}$  and  $\mathbf{M}^{(3)}$  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 from non-human patches and the human token. Moreover, we mask out attention from  $z_{hmn}$  to  $z_{cls}$ ,  $\mathbf{M}_{(m+1)}^{(3)}$ , 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  $z_{hmn} = \sum_{i \in P} e_i$ .

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:

$$\text{SNCE}(v, t, s) = - \sum_{i \in B} \left( \log \frac{\exp(v_i \cdot t_i / \tau)}{\sum_{j \in B} \exp(v_i \cdot t_j / \tau - w_{i,j})} \right), \quad (5)$$

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

$$w_{i,j} = \begin{cases} \beta \cdot \text{KL}(s_i \| s_j)^{-1} & i \neq j \\ 0 & i = j \end{cases}, \quad (6)$$

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

$$\mathcal{L} = \frac{1}{2|B|} \left( \text{SNCE}(v, t, s) + \text{SNCE}(t, v, s) \right). \quad (7)$$

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.

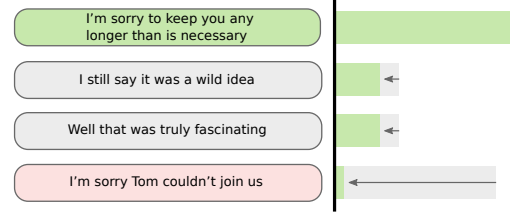


Figure 4. A sample batch where both the positive pair (*green*) and the false negative pair (*red*) exist. The green bar and gray bar represent the similarity between all text and the positive video before and after reweighting.

## 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

	mAP	AUC	$R^2$
EmotionCLIP (vanilla)	21.97	68.85	0.130
+ SAAM	21.53 $-0.44$	68.56 $-0.29$	0.137 $+0.007$
+ SAP	22.28 $+0.31$	69.06 $+0.21$	0.131 $+0.001$
+ SAP & SNCE	22.51 $+0.54$	69.30 $+0.45$	0.133 $+0.003$

Table 1. Component-wise analysis of our method on BoLD.

variety of themes. Valence and arousal scores are provided continuously (*i.e.*, every second) along movies.

**Emotic** [29] is an image dataset for emotion recognition in context, comprising 23,571 images of 34,320 annotated individuals in unconstrained real-world environments. Each subject is annotated with 26 discrete categories.

## 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



Figure 5. Attention weights for the HMN token from layer 1-4 (left to right) of the frame encoder in one trained network. Each row represents one frame. The green and yellow spots are the high-attention areas.

attention to all patches on the image. As shown in Fig. 5, HMN tokens first focus on random locations, but gradually turn their attention to the subject (*i.e.*, the person with a bounding box) as we move to later layers, demonstrating that SAP offers sufficient guidance to the network attention.

### 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.

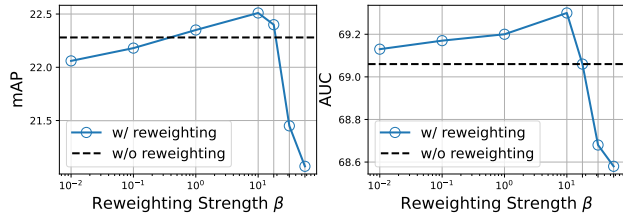


Figure 6. The effect of varying the strength of reweighting in SNCE. The larger the  $\beta$ , the stronger the suppression of negative samples.

Text	Logit
I'm sorry to keep you any longer than is necessary. (positive sample)	10.65
I'm sorry Tom couldn't join us.	12.6 $\rightarrow$ -697
I hate to say it guys but it's getting late.	9.81 $\rightarrow$ -97.6
I still say it was a wild idea.	13.8 $\rightarrow$ 13.6
Well, that was truly fascinating.	15.2 $\rightarrow$ 15.0

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.

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

Text Encoder	Frame Encoder	Temporal Encoder	mAP	AUC
✓	✓	-	22.51	69.30
-	✓	-	12.43 $-45\%$	54.96 $-21\%$
-	-	-	11.02 $-51\%$	50.28 $-27\%$
✗	✓	-	13.40 $-40\%$	57.38 $-17\%$
✗	✗	-	18.43 $-18\%$	65.08 $-6.0\%$
✓	✓	○	21.17 $-5.9\%$	68.74 $-0.8\%$

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.

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.

#### 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-

(a) BoLD [45]				(c) Emotic [29]			(e) Liris-Accede [5]		
Method	mAP $\uparrow$	AUC $\uparrow$	$R^2$ $\uparrow$	Method	mAP $\uparrow$	AUC $\uparrow$	Method	V. MSE $\downarrow$	A. MSE $\downarrow$
<i>Supervised</i>				<i>Supervised</i>			<i>Supervised</i>		
ST-GCN [94]	12.63	55.96	0.044	CAER-Net [34]	20.84	-	Quan <i>et al.</i> [61]	0.115	0.171
TSN [81]	17.02	62.70	0.095	Affective Graph [96]	28.42	-	Ko <i>et al.</i> [28]	0.102	<b>0.149</b>
Filntisis <i>et al.</i> [21]	16.56	62.66	0.092	Fusion Model [29]	29.45	-	*CERTH-ITI [4]	0.117	0.138
Pikoulis <i>et al.</i> [59]	19.29	66.82	<b>0.149</b>	EmotiCon (GCN) [53]	32.03	-	*THUHCSI [25]	0.092	0.140
<i>Linear-Eval</i>				<i>Linear-Eval</i>			<i>Linear-Eval</i>		
VideoCLIP [87]	11.19	51.23	-5.11	EmotiCon (Depth) [53]	<b>35.48</b>	-	*Yi <i>et al.</i> [91]	0.090	0.136
X-CLIP [55]	13.26	56.86	-0.03	VideoCLIP [87]	19.92	56.31	*GLA [75]	0.084	0.133
EmotionCLIP	<b>22.51</b>	<b>69.30</b>	0.133	X-CLIP [55]	22.80	61.31	*Zhao <i>et al.</i> [95]	0.071	0.137
				EmotionCLIP	32.91	<b>71.41</b>	*Affect2MM [54]	0.068	0.128
(b) MovieGraphs [77]			(d) MELD [60]						
Method	Val Acc $\uparrow$	Test Acc $\uparrow$	Method	Acc $\uparrow$	W. $F_1$ $\uparrow$	Abbreviation      Meaning			
<i>Supervised</i>			<i>Supervised</i>						
EmotionNet [84]	35.60	27.90	M2FNet (Visual) [12]	45.63	32.44	A. MSE	Arousal MSE		
*Affect2MM [54]	39.88	30.58	*M2FNet [12]	67.85	66.71	V. MSE	Valence MSE		
<i>Linear-Eval</i>			<i>Linear-Eval</i>			W. $F_1$	Weighted $F_1$		
VideoCLIP [87]	29.91	23.44	VideoCLIP [87]	45.19	32.06	Acc.	Top-1 Accuracy		
X-CLIP [55]	29.06	23.58	X-CLIP [55]	38.31	32.46	$\downarrow$	Lower is better		
EmotionCLIP	<b>41.60</b>	<b>32.35</b>	EmotionCLIP	<b>48.28</b>	<b>34.59</b>	$\uparrow$	Higher is better		

Table 4. Comparisons to the state-of-the-art across multiple datasets. Methods marked with \* use multimodal inputs, *i.e.*, audio and text. Bold numbers indicate the best results achieved using visual inputs only.

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 directly from human communication, thereby considerably broadening the horizons of current AEI research. The emergence of this pre-training paradigm offers an alternative solution to data scarcity, paving the way for a myriad of potential applications. We anticipate that our work will stimulate further investigation in this area and contribute to 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

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. [7](#)
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. [2](#), [3](#)
- [3] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. *Human-Robot Interaction: An introduction*. Cambridge University Press, 2020. [1](#)
- [4] Elissavet Batziou, Emmanouil Michail, Konstantinos Avgerinakis, Stefanos Vrochidis, Ioannis Patras, and Ioannis Kompatsiaris. Visual and audio analysis of movies video for emotion detection@ emotional impact of movies task mediaeval 2018. In *MediaEval*, 2018. [8](#)
- [5] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015. [2](#), [5](#), [8](#)
- [6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. [7](#)
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. [1](#)
- [8] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, pages 13984–13993, 2020. [2](#)
- [9] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018. [5](#)
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. [1](#), [2](#), [6](#), [7](#)
- [11] Henrik Christensen, Nancy Amato, Holly Yanco, Maja Mataric, Howie Choset, Ann Drobnis, Ken Goldberg, Jessie Grizzle, Gregory Hager, John Hollerbach, S Hutchinson, V Krovi, D Lee, W Smart, and J Trinkle. A roadmap for US robotics—from internet to robotics 2020 edition. *Foundations and Trends® in Robotics*, 8(4):307–424, 2021. [1](#)
- [12] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2FNet: Multi-modal fusion network for emotion recognition in conversation. In *CVPR*, pages 4652–4661, 2022. [2](#), [8](#)
- [13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. Ieee, 2005. [2](#)
- [14] Dorottya Demszyk, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. [5](#)
- [15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. [2](#)
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [1](#), [4](#)
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [3](#)
- [18] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. [5](#)
- [19] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1(1):56–75, 1976. [2](#)
- [20] Vyvyan Evans. How does communication work? *Psychology Today*, Jan 2020. [1](#)
- [21] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos. Emotion understanding in videos through body, context, and visual-semantic embedding loss. In *ECCV Workshops, Part I 16*, pages 747–755. Springer, 2020. [2](#), [8](#)
- [22] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. [2](#)
- [23] Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans-Joachim Böhme. Fullstop: Multilingual deep models for punctuation prediction. In *SwissText*, 2021. [3](#)
- [24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. [1](#), [6](#), [7](#)
- [25] Zitong Jin, Yuqi Yao, Ye Ma, and Mingxing Xu. THUHCSI in mediaeval 2017 emotional impact of movies task. *MediaEval*, 17:13–17, 2017. [8](#)
- [26] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. [2](#)
- [27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [1](#), [2](#)
- [28] Tobey H Ko, Zhonglei Gu, Tiantian He, and Yang Liu. Towards learning emotional subspace. In *MediaEval*, 2018. [8](#)
- [29] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *CVPR*, July 2017. [1](#), [2](#), [5](#), [6](#), [8](#)

- [30] Marina Krakovsky. Artificial (emotional) intelligence. *Communications of the ACM*, 61(4):18–19, 2018. **1**
- [31] Sven Kraus, Matthias Althoff, Bernd Heiing, and Martin Buss. Cognition and emotion in autonomous cars. In *2009 IEEE Intelligent Vehicles Symposium*, pages 635–640. IEEE, 2009. **1**
- [32] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. **2**
- [33] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019. **1**
- [34] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019. **2, 4, 8**
- [35] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: ClipBERT for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. **2, 7**
- [36] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. **2, 7**
- [37] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. **2**
- [38] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. **2**
- [39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. **2**
- [40] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen CLIP models are efficient video learners. *arXiv preprint arXiv:2208.03550*, 2022. **2, 7**
- [41] Kristen A Lindquist, Jennifer K MacCormack, and Holly Shablack. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, 6:444, 2015. **1, 3**
- [42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. **1**
- [43] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. **2**
- [44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. **2**
- [45] Yu Luo, Jianbo Ye, Reginald B. Adams, Jia Li, Michelle G. Newman, and James Z. Wang. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *IJCV*, 128(1):1–25, Jan 2020. **1, 2, 5, 8**
- [46] Yoshihiro Maruyama. The conditions of artificial general intelligence: logic, autonomy, resilience, integrity, morality, emotion, embodiment, and embeddedness. In *International Conference on Artificial General Intelligence*, pages 242–251. Springer, 2020. **1**
- [47] James K McNulty and Frank D Fincham. Beyond positive psychology? toward a contextual view of psychological processes and well-being. *American Psychologist*, 67(2):101, 2012. **4**
- [48] Albert Mehrabian. *Silent Messages*. Wadsworth Belmont, CA, 1971. **1**
- [49] Albert Mehrabian. *Nonverbal Communication*. Transaction Publishers, 1972. **1**
- [50] Albert Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. The MIT Press, Cambridge, 1980. **1**
- [51] B. Mesquita, L.F. Barrett, and E.R. Smith. *The Mind in Context*. Guilford Publications, 2010. **4**
- [52] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019. **2, 3**
- [53] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. EmotiCon: Context-aware multimodal emotion recognition using frege’s principle. In *CVPR*, pages 14234–14243, 2020. **2, 4, 8**
- [54] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2mm: Affective analysis of multimedia content using emotion causality. In *CVPR*, pages 5661–5671, 2021. **2, 5, 8**
- [55] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18. Springer, 2022. **2, 7, 8**
- [56] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. **2, 6**
- [57] Yimu Pan, Alison D. Gernand, Jeffery A. Goldstein, Leena Mithal, Delia Mwinyelle, and James Z. Wang. Vision-language contrastive learning approach to robust automatic placenta analysis using photographic images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, page 707–716, Berlin, Heidelberg, 2022. Springer-Verlag. **2**
- [58] Luiz Pessoa and Ralph Adolphs. Emotion processing and the amygdala: from a ‘low road’ to ‘many roads’ of evaluating biological significance. *Nature Reviews Neuroscience*, 11(11):773–782, 2010. **1, 7**

- [59] Ioannis Pikoulis, Panagiotis Paraskevas Filntisis, and Petros Maragos. Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild. *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08, 2021. [2](#), [4](#), [7](#), [8](#)
- [60] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. [2](#), [5](#), [8](#)
- [61] Khanh-An C Quan, Vinh-Tiep Nguyen, and Minh-Triet Tran. Frame-based evaluation with deep features to predict emotional impact of movies. In *MediaEval*, 2018. [8](#)
- [62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. [1](#), [2](#), [3](#), [6](#), [7](#)
- [63] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. [4](#)
- [64] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. [2](#)
- [65] Michael David Resnik. The context principle in Frege’s philosophy. *Philosophy and Phenomenological Research*, 27(3):356–365, 1967. [4](#)
- [66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. [1](#)
- [67] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977. [5](#)
- [68] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. [1](#), [3](#), [4](#), [5](#)
- [69] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE TPAMI*, 37(6):1113–1133, 2014. [2](#)
- [70] Ajay B Satpute and Kristen A Lindquist. At the neural intersection between language and emotion. *Affective Science*, 2(2):207–220, 2021. [7](#)
- [71] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. [2](#)
- [72] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. [2](#), [3](#)
- [73] Dagmar Schuller and Björn W Schuller. The age of artificial emotional intelligence. *Computer*, 51(9):38–46, 2018. [1](#)
- [74] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. [2](#)
- [75] Jennifer J Sun, Ting Liu, and Gautam Prasad. Gla in mediæval 2018 emotional impact of movies task. *arXiv preprint arXiv:1911.12361*, 2019. [8](#)
- [76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. [3](#), [4](#)
- [77] Paul Vicol, Makarand Tapaswi, Lluís Castrejon, and Sanja Fidler. MovieGraphs: Towards understanding human-centric situations from videos. In *CVPR*, pages 8581–8590, 2018. [2](#), [5](#), [8](#)
- [78] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. On position embeddings in bert. In *ICLR*, 2020. [5](#)
- [79] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. [3](#)
- [80] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020. [2](#)
- [81] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 41(11):2740–2755, 2018. [8](#)
- [82] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Action-CLIP: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. [2](#)
- [83] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. HairCLIP: Design your hair by text and reference image. In *CVPR*, pages 18072–18081, 2022. [2](#)
- [84] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning visual emotion representations from web data. In *CVPR*, June 2020. [2](#), [8](#)
- [85] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016. [2](#)
- [86] Benjamin Wortman and James Z Wang. HICEM: A high-coverage emotion model for artificial emotional intelligence. *arXiv preprint arXiv:2206.07593*, 2022. [2](#)
- [87] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and

- Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2, 7, 8
- [88] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *ICCV*, pages 3601–3610, 2021. 2
- [89] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. CPT: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 4
- [90] Jianbo Ye, Jia Li, Michelle G Newman, Reginald B Adams, and James Z Wang. Probabilistic multigraph modeling for improving the quality of crowdsourced affective data. *IEEE Transactions on Affective Computing*, 10(1):115–128, 2017. 2
- [91] Yun Yi, Hanli Wang, and Qinyu Li. CNN features for emotional impact of movies task. In *MediaEval*, 2018. 8
- [92] Richard Yonck. *Heart of the Machine: Our Future in a World of Artificial Emotional Intelligence*. Arcade, 2020. 1
- [93] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE, 2018. 8
- [94] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017. 8
- [95] Jie Zhang, Yin Zhao, Longjun Cai, Chaoping Tu, and Wu Wei. Video affective effects prediction with multi-modal fusion and shot-long temporal context. *CoRR*, abs/1909.01763, 2019. 8
- [96] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *ICME*, pages 151–156, 2019. 2, 8
- [97] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free CLIP-Adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2
- [98] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *CVPR*, pages 8552–8562, 2022. 2
- [99] Yue Zhang, Wanying Ding, Ran Xu, and Xiaohua Hu. Visual emotion representation learning via emotion-aware pre-training. In Lud De Raedt, editor, *IJCAI*, pages 1679–1685. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 2
- [100] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2