A. Data

A.1. Data Collection

The video-text pairs for pre-training were obtained using Python implementation of YouTube API, youtube-search-python [1]. This API provides the exact query result as the YouTube webpage. We searched for keywords such as "TV series" and "TV shows," and filtered only those with English closed captions from the resulting videos. Next, we filtered out all the "TV" videos that were less than 40 minutes long to remove some false results. We then manually removed videos appearing in downstream datasets based on YouTube id and movie title. This process resulted in 3,613 filtered videos or about 1.1 million video clips. Due to resource limitations, we only processed and stored the videos at 8 FPS. We refer to this dataset as the TV dataset.

We further processed the video frames, and the corresponding closed captions to obtain additional information. We extracted all the frames and resized the smaller edge to 256 without changing the aspect ratio. All the closed captions were processed using FullStop [2] to produce complete sentences; each word obtained a punctuation label, and we split on the termination punctuation. Furthermore, we applied a sentiment model, a DistilRoBERTa-base [2], to generate sentiment scores of seven emotion categories (*i.e.*, anger, disgust, fear, joy, sadness, surprise, and neutral) for texts. Moreover, all the frames were processed using YOLOv7 [9] to generate bounding boxes for all humans. The detailed parameter for bounding boxes generation is in Tab. 1.

Image Size	Confidence Threshold	IOU Threshold
640×640	0.25	0.45

Table 1. The important parameters for YOLOv7 to generate human bounding boxes.

A.2. Data Exploration

We first explore the textural data in the TV dataset. We notice that many texts are not helpful for emotion understanding; they do not provide desired emotional signals. As shown in Tab. 2, the neutral score is the probability that the trained sentiment model predicts that the text is neutral.; we can see that the text expresses stronger emotion when the neutral score is low; the emotion signal is most apparent in the last three rows. This observation aligns with our intuition that instructional or descriptive language, such as those in [5] and [1], are not usually emotional and support our motivation for collecting the TV dataset. Based on the above observation, we believe that the model may be misled if too many samples with high neutral scores were used. Therefore, we must limit the number of samples with a high chance of neutrality to better direct the model's attention toward other more valuable emotional expressions. Moreover, the distribution of the neutral scores for the TV dataset is in Fig. 1; it forms a bimodal distribution where more data are closer to the left (non-neutral). Clearly, the left peak represents the desirable emotional samples, and the right peak represents the instructional or descriptive samples that can be discarded. To confirm our intuitions and to find a good threshold for filtering useless examples, we tested multiple neutral score thresholds on the TV dataset. As shown in Fig. 2, the model's performance on downstream tasks increases when more neutral examples are eliminated, supporting our conjecture that too many neutral samples are not helpful for emotion understanding. Furthermore, the performance peaks at around 0.05 and drops dramatically as too few samples were left when using a small threshold. Based on this observation, we keep only the samples with a neutral score of less than 0.05. This filtering process results in about 250k samples which is still much larger than the current emotion understanding datasets. Finally, we evaluate how the filtering process changed the probability distribution of other emotion labels. The comparison of the distribution before and after filtering for the other six emotion categories is in Fig. 3; the distribution of the original TV dataset is highly skewed where the majority of samples had probabilities close to zero for each of the six emotions. Following filtering, the skewness is reduced and the proportion of samples containing relevant information signals is enhanced. The filtered TV dataset is expected to provide better supervision for EmotionCLIP.

Some examples from the filtered TV dataset are shown in Fig. 4. It can be clearly felt that most of the examples showed strong emotional expression from both verbal and nonverbal cues. The word cloud in Fig. 5 is constructed based on the filtered TV dataset. We can see that some words related to emotional expression appear frequently in the dataset, such as *'sorry', 'happy', 'afraid', 'fear', 'angry', 'worried', 'love', etc*. In general, there are a large number of verbal communications with rich emotional expressions, which can hardly be covered by basic emotions.

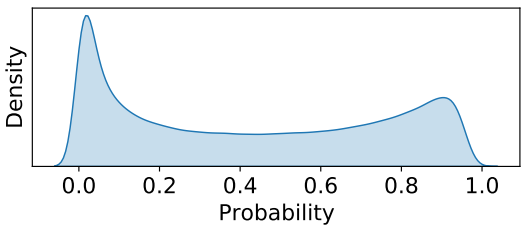

Figure 1. The distribution of the neutral scores on the TV dataset.

[1] https://github.com/alexmercerind/youtube-search-python
[2] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

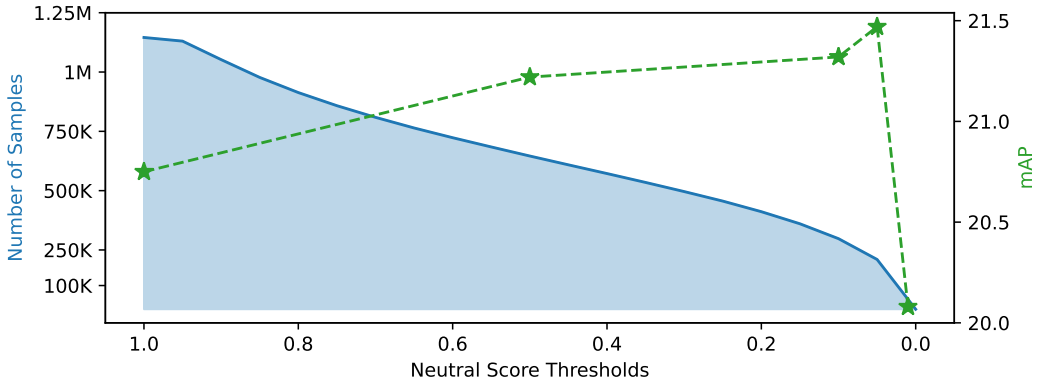

Figure 2. Effect of filtering with neutral scores on sample size and model performance.

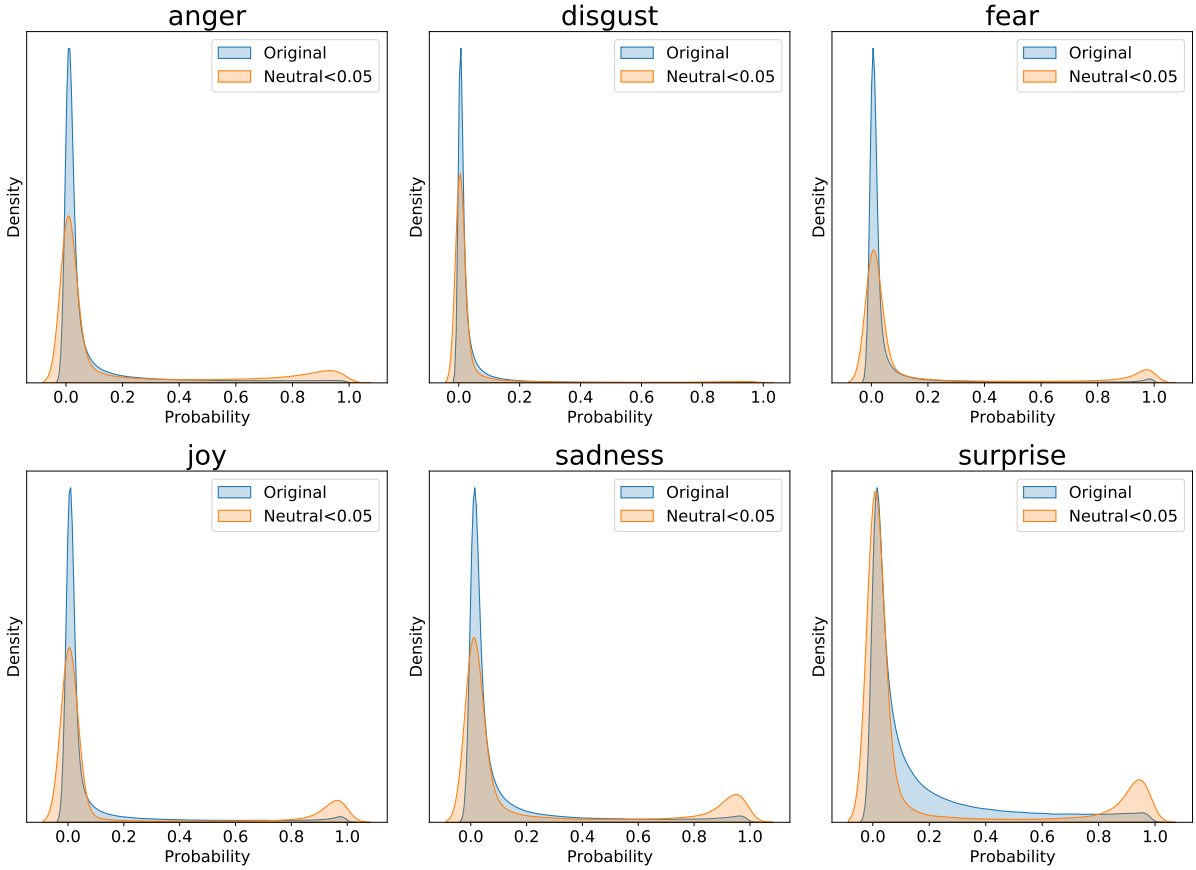

Figure 3. The effect of filtering out text with a neutral score greater than 0.05 on the distribution of the predicted probability of other emotion categories.

(a) Yes. Poor Georgie. He was dropped after he broke his hip.

(b) Fantastic rock and roll. Thank you.

(c) What happened to Danny? Why'd you let him die? I tried to help him.

(d) Our pleasure, Mr. Kyle. Our pleasure.

(e) She was suspicious of everything I did.

(f) Oh, why won't you? It's none of your business.

(g) No, I'm sorry.

(h) I know you saying that Miss. Bowden was deliberately seeking to endanger the life of the mother and child.

(i) I ought to punch you right in the nose.

(j) I find this rather embarrassing Mr. Barris. I don't see why.

(k) Okay. I can't keep running after you and cleaning up your mess.

(l) Worst case scenario a man can actively invite the demons in.

(m) I just think he'd be better off out of the ranch.

(n) After all I was persuaded. I can't let people down.

(o) Mrs. Matsen, I know this has been a terrible shock.

(p) Yeah. Happy in Jordan Hill's.

(q) Why did you do it? I must be punished.

(r) Be my pleasure, and I've enjoyed the evening.

(s) Oh, I felt so frightened. I was shaking.

(t) Oh, I'm ever so sorry.

(u) Oh, that's wonderful. I'll tell you they had a bunch of them.

(v) What do you want from me? Apologies? I don't apologize.

(w) I'm not gonna let those cattle die because of some fool notion in your head.

(x) Rhoda be thrilled to see you when she gets home.

Figure 4. Examples from TV Dataset.

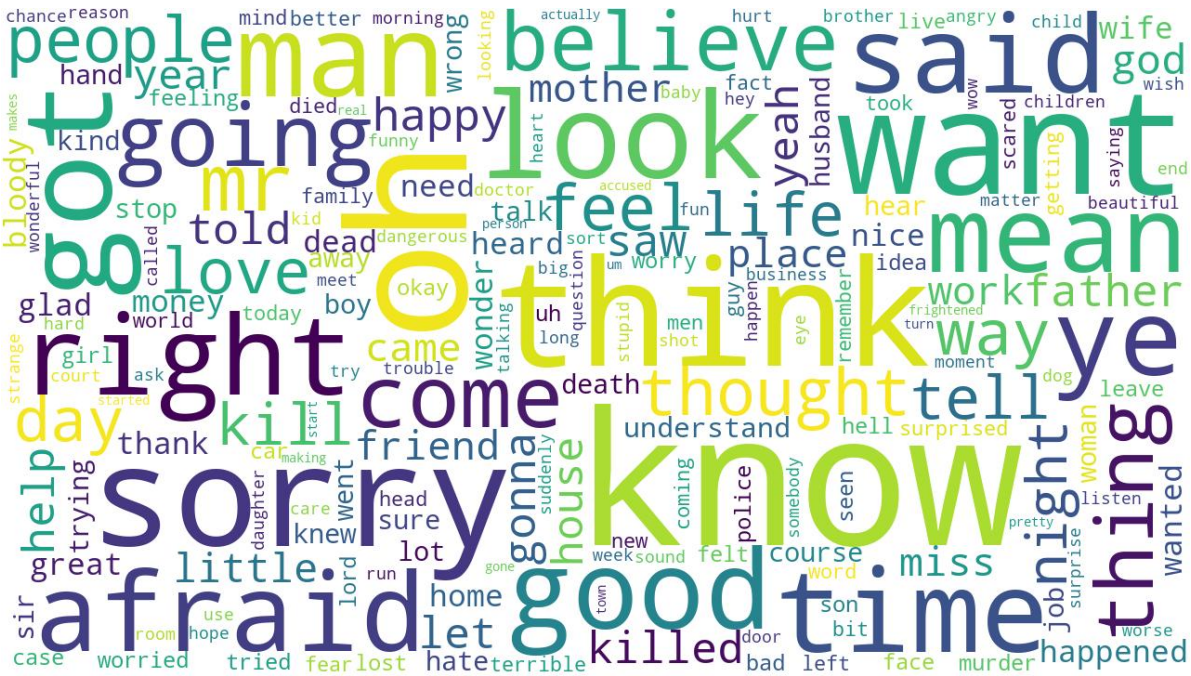

Figure 5. A word cloud generated from the collected TV dataset.

Text	Neutral Score
Only then it will become picture perfect	0.931
I do know they once owned the painting	0.924
Valerie had to wait a few months for hers and his mother	0.834
Tell him that we will meet at 6.	0.828
can you tell us what you know about that	0.734
They have also received a bill	0.717
So many of them did not clear	0.543
he’s the only collector	0.519
he was on his way to Pennsylvania the night	0.452
Do you want me to do it and bring a glass	0.351
I haven’t taken my vows and I probably never will	0.300
some of his new patients like three-year-old Harriet are definitely unusual	0.253
yes in your statement you said you saw the defendant running from the heat	0.234
I don’t want to lose the money any more than you do	0.185
any one of them might be guilty	0.164
I couldn’t have done it myself I didn’t even think of it	0.031
it’s a whole new world for me	0.027
I know our dream house yes David our dream house	0.009

Table 2. Example texts, and their corresponding neutral score predicted by the sentiment model.

B. Implementation Details

B.1. Model Details

EmotionCLIP adopts CLIP (ViT/B-32) [8] as part of the frame encoder and text encoder. Specifically, the frame encoder is a ViT ($L = 12, N_h = 12, d = 768, p = 32$), the text encoder is a Transformer ($L = 12, N_h = 8, d = 512$), and the temporal encoder is another Transformer ($L = 6, N_h = 8, d = 512$), where L is the number of layers, N_h is the number of attention heads, d is the embedding dimension, and p is the patch size. The sentiment model a fine-tuned checkpoint of DistilRoBERTa-base [3], which is frozen during training. Following the practice of CLIP, both the text encoder and sentiment model operate on a lower-cased byte pair encoding (BPE) representation of the text with a 49,152 vocab size. The max length of the text sequence is capped at 76 and bracketed with [SOS] and [EOS] tokens. The specific implementation of subject-aware context encoding in the frame encoder is as follows:

Subject-Aware Attention Masking. Follow the equation

$$\text{Attention}^*(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{U}) = \underbrace{\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)(\mathbf{J} - \mathbf{A})\mathbf{V}}_{\text{context}} + \underbrace{\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{A}\mathbf{U}\mathbf{V}}_{\text{subject}}, \tag{1}$$

defined in the main document, we make a few implementation choices to speed up the computation. First, we use $\mathbf{V}^{(l)}$ for the context encoding and $\mathbf{V}^{(l-1)}$ for the subject encoding, where l denotes the layer. Since each token at layer l is a weighted average of the token at layer $l-1$, the model is able to extract similar information to $\mathbf{V}^{(l)}$ by reweighting $\mathbf{V}^{(l-1)}$. Next, we set A to the attention from each token to the current layer HMN token. This modification ensures all entries in A are in $[0, 1]$ and values are automatically learned by the model. The above two modifications allow us to reuse the original multi-head attention layer by setting the attention mask to $\mathbf{M}$ as defined in the main document.

Subject-Aware Prompting. As described in the main document, we set HMN as $z_{hmn} = \sum_{i \in P} e_i$. Note that the indices in P represent the presence or absence of the subject in the non-overlapping image patches. This information is obtained from bounding boxes which may not align with the non-overlapping image patches. To address this issue, we add the indices of all tokens that have overlap $o_i > 0$ with the bounding boxes to P and compute $z_{hmn} = \sum_{i \in P} o_i e_i$

[3] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

B.2. Training Details

The frame encoder and text encoder are initialized using the pre-trained weights provided by OpenCLIP [4]. We use the AdamW optimizer to train the model, where $\beta_1 = 0.98, \beta_2 = 0.9, \epsilon = 1e{-}10, \lambda = 0.1$. The base learning rate of the parameters in the frame encoder and text encoder is set to 5e-5 for gains and biases, and 1e-8 for the remaining parameters. The learning rate of the parameters in the temporal encoder is set to 1e-6. The decoupled weight decay regularization is applied to all weights that are not gains or biases. Models are trained for 25 epochs with a batch size of 128. The learning rate is linearly warmed up for 2500 steps and decayed to 1e-10 following a cosine schedule for the rest of the training. For each video, we randomly sample 8 frames in each iteration to form an input sequence. The input frames have a spatial resolution of 224x224 and are obtained by random cropping. The sequence of the subject mask is obtained with the same operation as the corresponding frame.

B.3. Evaluation Details

We follow the linear-probe evaluation protocol in CLIP. Specifically, we uniformly sample 8 frames from each video to form an input sequence and extract video features using the pre-trained EmotionCLIP. For classification tasks, we train a logistic regression classifier using scikit-learn's implementation with sag solver. The maximum iteration is set to 2,000, and the regularization strength is determined by a random search on the validation sets. For the datasets that contain a validation split in addition to a test split, we use the provided validation set to perform the hyperparameter search, and for the datasets that do not provide a validation split or have not published labels for the test data, we split the training dataset to perform the hyperparameter search. For the regression tasks, we train a linear regression model using scikit-learn's Ridge implementation with default hyperparameters, followed by a Savgolet filter.

For the other two vision-language baseline models, we used the official implementations with pre-trained weights and ran them with their default settings. Specifically, for VideoCLIP [10], we use the pre-trained model provided in Fairseq [5]; for X-CLIP [7], we use the zero-shot X-CLIP-B/16 model trained on Kinetics-600 [6]. For other supervised learning methods, we use the scores reported in their papers.

[4] https://github.com/mlfoundations/open_clip
[5] https://github.com/facebookresearch/fairseq
[6] https://github.com/microsoft/VideoX/tree/master/X-CLIP

C. Detailed Results

C.1. Qualitative Results

Subject-Aware Prompting. We present additional qualitative results for SAP. As shown in Fig. 6, the attention of HMN changes according to the positional hint for the subject, which shows SAP is subject-aware. Moreover, Fig. 7 shows the exact same set of frames as Fig. 6 but the attention comes from CLS token; it is clear that the attention for CLS token tend to focus on the entire scene and does not change regardless of the positional hint. This result shows SAP behaves similarly to two stream approaches where CLS models the context and HMN models the subject but is less affected by the artifacts introduced in traditional manual subject cropping. Fig. 8 shows some examples where SAP fails to guide the attention. The majority of the failure cases are direct results of applying cropping during testing; some subjects are either entirely off the frame or partially off the frame. Moreover, there are cases where the bounding boxes are incorrect. Additionally, some subjects are too small compared to most of the subjects in the training dataset, leading to a large domain shift.

Sentiment-Guided Contrastive Learning. In this section, we demonstrate how the sentiment model guides the loss. Note that we use the inverse of the KL divergence between text from the positive sample and the negative samples to reweight the negative samples; the suppression strength is inversely proportional to the KL divergence. Tab. 4 shows some examples from the collected TV dataset; the text expressing similar emotion has a smaller KL divergence whereas the text expressing different emotion have a larger KL divergence. Since we treat the negative samples that express similar emotions to the positive samples as false negative samples, it is clear the proposed reweighting method suppresses the false negative samples.

C.2. Quantitative Results

We reported detailed emotion classification performance on BoLD and Emotic in Tab. 5. Both datasets have fine-grained emotion annotations on 26 categories. We observed an intriguing phenomenon that EmotionCLIP performs quite differently on some emotion categories compared with prior approaches based on supervised learning. As shown in Tab. 3, EmotionCLIP with linear classifier achieves comparable mAP with two other supervised learning methods using RGB inputs on Emotic. However, we notice that EmotionCLIP performs significantly better than supervised learning methods in some categories (*e.g.*, *sadness, suffering*). The performance in the remaining categories is also different from that of supervised learning methods. This result shows that emotional representations learned from communication are different from those learned through annotations, which further demonstrates the complementarity of EmotionCLIP as a pre-training method to conventional supervised learning methods.

Categories	Kosti *et al.* [3]	Emoticon [6]	EmotionCLIP
Affection	27.85	36.78	**45.81**
Anger	9.49	14.92	**26.67**
Annoyance	14.06	18.45	**21.94**
Anticipation	58.64	**68.12**	58.07
Aversion	7.48	**16.48**	10.55
Confidence	**78.35**	59.23	76.94
Disapproval	14.97	**21.21**	19.23
Disconnection	21.32	25.17	**29.44**
Disquietment	16.89	16.41	**21.82**
Doubt/Confusion	29.63	**33.15**	22.70
Embarrassment	3.18	**11.25**	2.86
Engagement	87.53	**90.45**	87.79
Esteem	17.73	**22.23**	18.58
Excitement	77.16	**82.21**	71.05
Fatigue	9.70	19.15	**20.21**
Fear	**14.14**	11.32	12.08
Happiness	58.26	68.21	**78.44**
Pain	8.94	12.54	**16.73**
Peace	21.56	**35.14**	29.67
Pleasure	45.46	**61.34**	50.23
Sadness	19.66	26.15	**43.01**
Sensitivity	9.28	9.21	**9.53**
Suffering	18.84	22.81	**43.96**
Surprise	**18.81**	14.21	10.70
Sympathy	14.71	**24.63**	17.23
Yearning	8.34	**12.23**	10.29
mAP	27.38	32.03	**32.91**

Table 3. Per-category performance (AP) on Emotic.

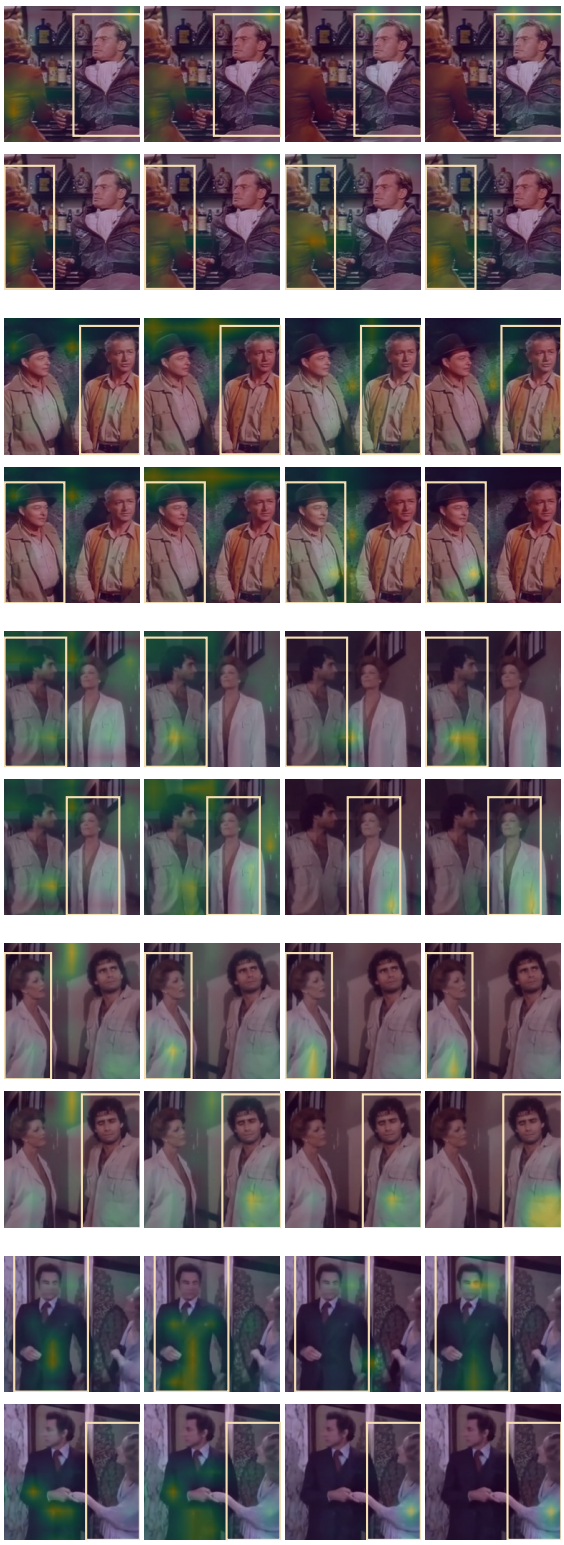

Figure 6. The attention weights for HMN token at layer 1 - 4 (left to right) for each frame. Note that changing the bounding box location causes the attention weights to change accordingly.

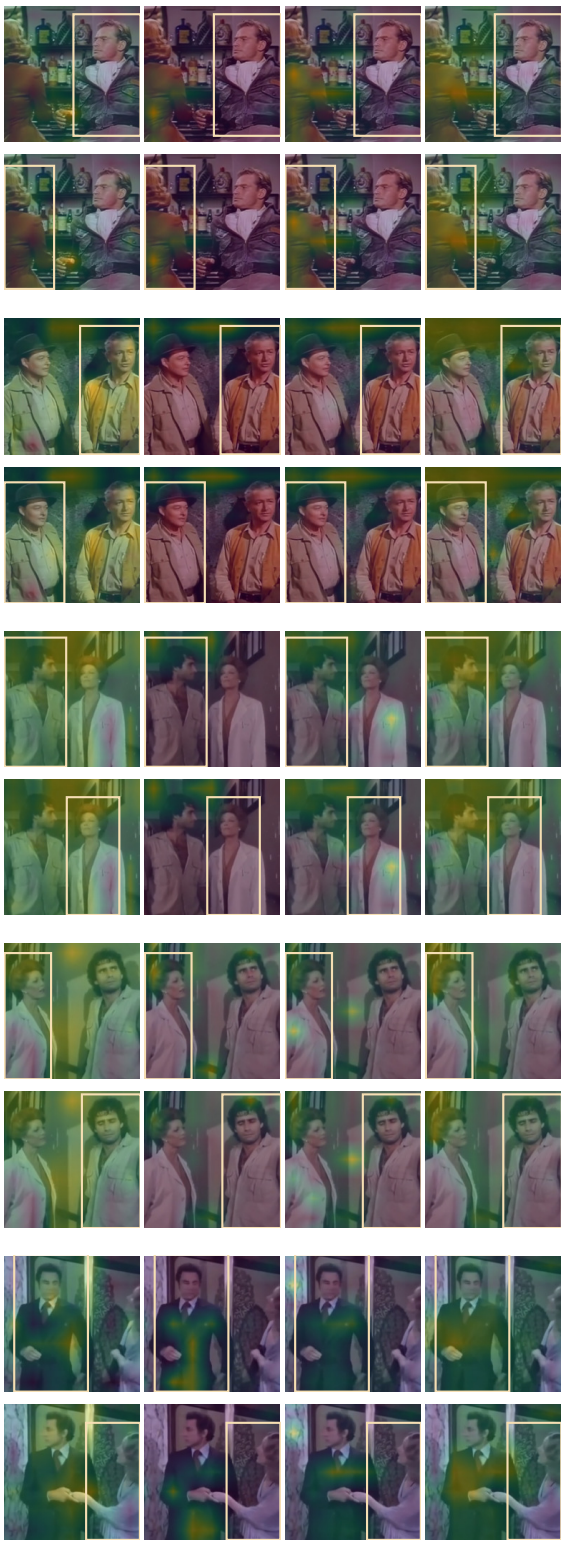

Figure 7. The attention weights for CLS token at layer 1 - 4 (left to right) for each frame. Note that changing the bounding box location does not change the attention weight.

(a) The bounding boxes are absent.

(b) The subjects are partially off the frame.

(c) The subjects bounding boxes are incorrect.

(d) The subjects are too small.

Figure 8. The different failure cases for the attention weights of HMN token at layer 1 - 4 (left to right) for each frame in BoLD dataset.

Source	Target	KL Divergence
it would baffle the police	I don't think that you could understand	9.8e-4
he'll be glad to hear	it's a very nice church	5.2e-4
you seem awfully anxious to make it look like suicide	you're scared of them	9.8e-4
I had a great childhood	I just can't wait to get back into some action	9.7e-4
I wonder if you would look at your passport and find a visa for perco for me	I just I was wondering if he was here	9.4e-4
a man is dead	I have lost my son	5.0e-4
i guess i must be the luckiest man around these parts	I'm very glad	4.8e-4
oh my goodness I didn't know I had such a devoted fans	oh my god Jimbo look who's here	4.7e-4
oh really oh yes yes yes miss Travers I'm surprised	I can't say I'm surprised though	4.3e-4
I'm sorry to keep you any longer than is necessary	I'm sorry to have to keep requesting you like this	3.6e-4
she were my nurse and after that sickness come the greatest happiness of my life	I never thought I'd ever be cold again	5.01
i have consulted with them	she doesn't think she'll ever see him again	5.04
I'm sorry to keep you any longer than is necessary	I don't think this is something you can come out of	5.48
it would baffle the police	i promised i wouldn't harm him	5.88
well that was truly fascinating	I'm afraid of you	5.92
alright I'm sorry about yesterday	you'll be surprised	6.00
she let me down	I'm Tarsus glad to meet you	6.03
i've got to get my crew out of here	i think for nancy the thrill of the chase was half of the fun	6.29
i should have known he'd be all right	the army turned me down	7.44
I like the way you laugh	he would never let you down deliberately	7.25

Table 4. Examples of false negative targets (with low KL divergence) and true negative targets (with high KL divergence). The KL divergence is calculated using sentiment scores. The proposed sentiment-guided contrastive learning method will down-weight the target if the KL divergence between the source and target is relatively low, thereby eliminating emotional false negatives.

Categories	BoLD [4]		Emotic [3]	
	AP	AUC	AP	AUC
Affection	42.06	84.53	45.81	79.47
Anger	15.24	71.93	26.67	76.76
Annoyance	18.78	61.56	21.94	74.04
Anticipation	32.23	60.45	58.07	61.94
Aversion	9.08	63.45	10.55	72.09
Confidence	40.33	66.63	76.94	76.51
Disapproval	14.12	57.62	19.23	79.77
Disconnection	11.08	56.86	29.44	71.33
Disquietment	23.75	68.04	21.82	63.73
Doubt/Confusion	22.82	63.28	22.70	60.82
Embarrassment	2.29	70.70	2.86	56.76
Engagement	44.54	64.34	87.79	70.34
Esteem	20.66	63.67	18.58	57.24
Excitement	28.04	73.08	71.05	73.57
Fatigue	13.17	71.04	20.21	68.88
Fear	19.41	71.74	12.08	73.06
Happiness	48.59	80.53	78.44	78.60
Pain	14.58	77.17	16.73	84.11
Peace	28.09	65.09	29.67	70.58
Pleasure	37.87	76.60	50.23	69.87
Sadness	25.85	82.49	43.01	85.05
Sensitivity	14.81	72.48	9.53	74.28
Suffering	26.61	80.15	43.96	88.04
Surprise	11.91	63.62	10.70	59.26
Sympathy	12.60	67.02	17.23	68.59
Yearning	6.66	67.65	10.29	61.88
Average	22.51	69.30	32.91	71.41

Table 5. Emotion classification performance on BoLD and Emotic. AP: average precision. AUC: ROC-AUC.

References

[1] 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. 1

[2] Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans-Joachim Böhme. Fullstop: Multilingual deep models for punctuation prediction. In *SwissText*, 2021. 1

[3] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *CVPR*, July 2017. 6, 10

[4] 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. 10

[5] 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. 1

[6] 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. 6

[7] 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. 5

[8] 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. 5

[9] 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. 1

[10] 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. 5