

---

# EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition

Niki Maria Foteinopoulou<sup>1</sup> and Ioannis Patras<sup>1</sup>

<sup>1</sup> School of Electronic Engineering and Computer Science,  
Queen Mary University of London, London, United Kingdom

**Abstract**—Facial Expression Recognition (FER) is a crucial task in affective computing, but its conventional focus on the seven basic emotions limits its applicability to the complex and expanding emotional spectrum. To address the issue of new and unseen emotions present in dynamic in-the-wild FER, we propose a novel vision-language model that utilises sample-level text descriptions (i.e. captions of the context, expressions or emotional cues) as natural language supervision, aiming to enhance the learning of rich latent representations, for zero-shot classification. To test this, we evaluate using zero-shot classification of the model trained on sample-level descriptions on four popular dynamic FER datasets. Our findings show that this approach yields significant improvements when compared to baseline methods. Specifically, for zero-shot video FER, we outperform CLIP by over 10% in terms of Weighted Average Recall and 5% in terms of Unweighted Average Recall on several datasets. Furthermore, we evaluate the representations obtained from the network trained using sample-level descriptions on the downstream task of mental health symptom estimation, achieving performance comparable or superior to state-of-the-art methods and strong agreement with human experts. Namely, we achieved a Pearson’s correlation coefficient of up to 0.85 for schizophrenia symptom severity estimation, which is comparable to human experts’ agreement. The code is publicly available on <https://github.com/NickyFot/EmoCLIP.git>.

## I. INTRODUCTION

Facial Expression Recognition (FER) is a primary task of affective computing, with several practical applications in Human-Computer Interaction [6], education [40] and mental health [11], among others. To formalise the spectrum of human emotions, researchers have proposed several models. Ekman & Friesen [9] propose six emotions (later seven with the addition of “contempt”) as the basis of human emotional expression. This is the most widely accepted model for FER; however, human emotional experience is significantly more complex and varied than the seven basic categories, with up to 27 distinct categories for emotion reported in recent studies [7]. A continuous arousal-valence scale has been proposed [29] as an alternative to creating emotion categories; however, the scale is neither objective [12] nor self-explanatory to human readers. Therefore, as more fine-grained definitions of emotion are proposed and the categorical models expand, there is also a need for automated systems to adjust to new definitions, unseen emotions and mental states or compound emotions. Zero-shot methodologies in Facial Emotion Recognition (FER) specifically tackle these challenges, providing effective solutions to unseen emotions and mental states.

Zero-shot Learning (ZSL) is a machine learning paradigm where a model can recognise and classify objects or concepts it has never been trained on by leveraging auxiliary information or attributes associated with those unseen classes [36]. In the context of emotion recognition, ZSL has traditionally been achieved by some label semantic embedding to produce emotion prototypes, typically word2vec [44], [4]. The use of hard labels, however, is failing to include semantically rich information in the prototypes as well as ignoring the subtle differences in expression between subjects. Recent developments in Vision-Language Models (VLM) [28], [13], [1], using image-caption pairs instead of hard-labels have demonstrated superior generalisation and zero-shot abilities. Furthermore, CLIP [28] has been used as the basis for several methods in action recognition [43], [34] or video retrieval and captioning [31], [22], [41], [21]. However, the use of VLMs in dynamic FER remains relatively unexplored. Recent preliminary works in FER have used video-language contrastive training [16], [18], primarily relying on class-level prompt learning. Nonetheless, this approach is akin to class supervision and is not designed to generalise to unseen behaviours.

In this work, we propose a novel approach to zero-shot FER from video inputs by jointly learning video and text embeddings, utilising the contrastive learning framework with natural language supervision. The network architecture is simple and trains a video and text encoder concurrently, as shown in Fig. 1. The text and image encoders are initialised by leveraging the knowledge of large-scale pre-trained CLIP [28], as typical dynamic FER datasets do not have enough samples to train from scratch, and we train an additional temporal module from scratch, similar to previous works in action recognition [19], [34]. Contrary to previous VLM or ZSL works in emotion recognition [4], [44], [16], [18] we use sample-level descriptions during training i.e. captions of the subject’s facial expression and context, available in the MAFW [20] dataset. These act as soft labels and aim to achieve more semantically rich latent representations of the video samples. Then, during inference, we use class-level descriptions for each emotion. Specifically, we generate descriptions of each emotion in relation to the typical facial expressions associated with it. In the case of compound emotions, we propose manipulating the latent representation of the categories’ descriptions in the embedding space rather than creating additional prompts. Specifically, as compound emotions are combinations of basic emotions, we propose

averaging the embeddings of the components and adding them to the set of embeddings for each additional compound emotion. We show that the proposed methodology trained on sample-level descriptions shows generalisation capabilities invariant to domain shift as we perform zero-shot evaluation on multiple datasets. We also show that compared to the CLIP and FaRL baselines, the temporal information and domain-specific knowledge of the sample-level descriptions improve the zero-shot performance of FER. Finally, to show the generalisation ability of the representations we obtain from the video encoder, we adapt them to the domain of mental health. Using a simple MLP, trained in a fully supervised manner, we achieve results comparable to or outperforming previous state-of-the-art on estimating non-verbal symptoms of schizophrenia. We see a significant improvement compared to previous works, particularly on symptoms associated with affect and expression of emotions as well as total negative score, which is similar to human experts.

Our main contributions can be summarised as follows:

- Our paper introduces a novel zero-shot Facial Emotion Recognition (FER) paradigm from video input, employing sample-level descriptions and a dynamic model. This straightforward approach, which leverages CLIP [28], outperforms class-level descriptions and significantly improves zero-shot classification performance, particularly for under-represented emotions.
- We propose a novel method for representing compound emotions using average latent representations of basic emotions instead of concatenating or generating new prompts. This approach is more intuitive and efficient than prompt engineering and shows significant improvements across all metrics compared to prompt concatenation.
- Our proposed method, EmoCLIP, trained on the MAFW [20] dataset and evaluated on popular video FER datasets (AFEW [8], DFEW [14], FERV39K [35], and MAFW), achieves state-of-the-art performance on ZSL. Additionally, we evaluate the embeddings of the video encoder of EmoCLIP on the downstream task of schizophrenia symptom estimation using the NESS dataset [26]. We achieve results comparable to or better than previous state-of-the-art methods and comparable to human experts with a simple 2-layer MLP.

The remainder of this paper is organised as follows. Section II discusses related literature, Section III introduces our methodology, Section IV reports the results on the tasks of zero-shot FER for simple and compound emotions, and the downstream task of schizophrenia symptom estimation, and Section V concludes the paper.

## II. RELATED WORK

In this section, we review previous works on Vision and Language Models (VLM) and zero-shot learning in FER and mental health.

### A. Vision and Language Models

The use of large contrastive pre-training for Vision-Language Models (VLM) has become increasingly popular, as these models have demonstrated impressive generalisation capabilities [28], [13], [45], [1]. As large VLM require a large number of data and very high computational resources to achieve those results, the latest research on VLM has been mostly concentrated on three paths: (a) latent space manipulation, (b) leveraging pre-trained spatial features and fine-tuning a temporal module for video input, and (c) prompt learning.

Menon & Vondrick [23] use an ensemble of prompts generated by Large Language Models (LLMs) containing descriptive features of each class and show significant improvements in terms of accuracy as well as explainability of decisions. Ouali *et al.* [24] propose a method for latent space feature alignment in a target domain without the need for additional training. Similarly, Bain *et al.* [3] propose several methods for temporal pooling of frames, using pre-trained VLMs with very little or no additional training. Such approaches, although effective, use no information from the temporal dimension in the video, which is essential in human FER to understand macro and micro-expressions.

Large VLMs trained on static images have been used for video classification, particularly for action recognition. Lin *et al.* [19] propose using a lightweight Transformer Decoder over the CLIP [28] spatial features for downstream classification. Similarly, ActionCLIP [34] class labels are used for natural language supervision; therefore, the open vocabulary capabilities are lost in both approaches. CLIP [28] has been used as a backbone in several video captioning works [21], [39], [22]. However, none of these works has been evaluated or trained on the domain of FER, where the behaviour is generally not as clearly defined.

To overcome the challenges of prompt engineering in VLMs, some works propose learning a set of tokens [48], [47], [25] to append to the class name, which can improve performance as the text-encoder acts more like a bag of words [42], [2]. Natural language supervision for facial expression recognition (FER) is a relatively unexplored idea, but there have been some preliminary works exploring this approach. For instance, CLIPER [16] has proposed prompt learning to improve closed dictionary FER. However, these tokens are class-specific and cannot be used in open dictionary settings for zero-shot classification.

### B. Zero-Shot Learning in Facial Expression Recognition

Several works [4], [27], [37], [38] have proposed zero-shot frameworks for emotion recognition or emotional response recognition [44] by aligning class name embeddings to multimedia embeddings and then evaluating the method on unseen emotions. However, these methods still rely on hard labels and simpler class prototype embeddings (such as word2vec). As such, they do not take into consideration the intra-class differences or the underlying concepts in each class and as such, do not capture semantically rich information in the latent representations.

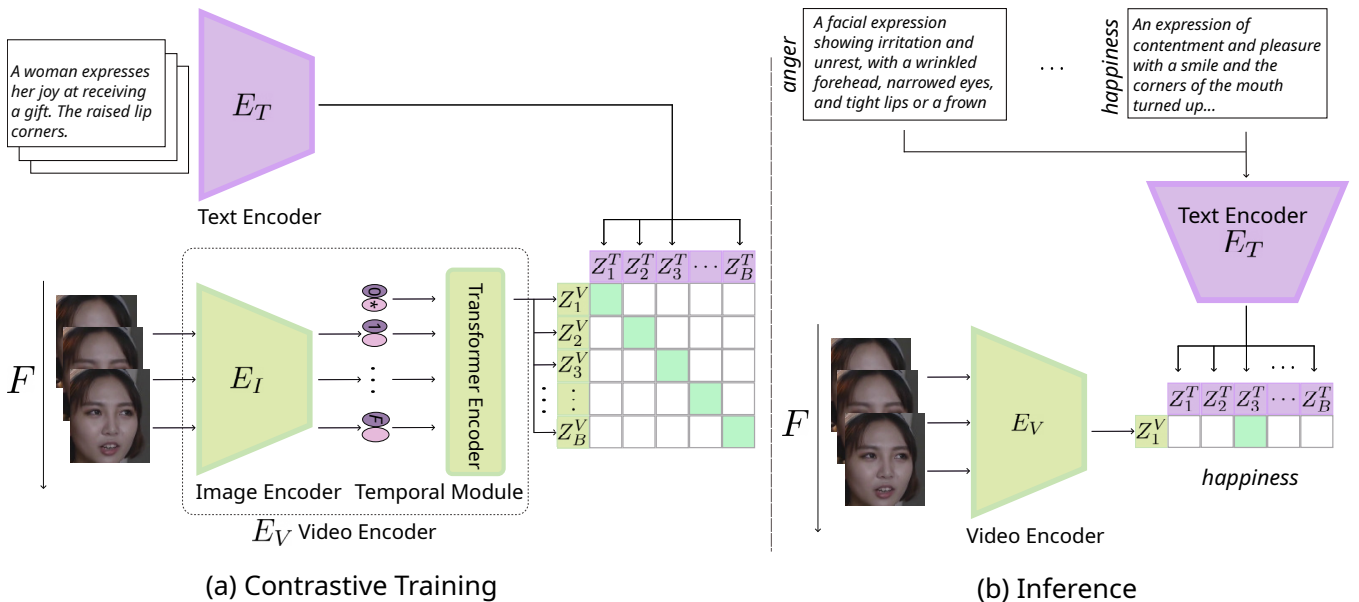


Fig. 1. Overview of our method, EmoCLIP. During training (a), we use joint training to optimise the cosine similarity of video-text embedding pairs in the mini-batch. Sample-specific descriptions of the subject’s facial expressions are used to train the model. During inference (b), we perform zero-shot classification using class-level descriptions for each of the emotion categories.

### III. METHODOLOGY

An overview of the proposed method can be seen in Fig. 1. In a nutshell, we follow the CLIP [28] contrastive training paradigm to optimise a video and a text encoder jointly. The video and text encoders of the network are jointly trained using a contrastive loss over the cosine similarities of the video-text pairings in the mini-batch. More specifically, the video encoder ( $E_V$ ) is composed of the CLIP image encoder ( $E_I$ ) and a Transformer Encoder to learn the temporal relationships of the frame spatial representations. The text encoder ( $E_T$ ) used in our approach is the CLIP text encoder. The weights of the image and text encoders in our model are initialised using the large pre-trained weights of CLIP [28] and finetuned on the target domain, as FER datasets are not large enough to train a VLM from scratch with adequate generalisation. Contrary to the previous video, VLM works in both action recognition [34], [19] and FER [16], [18], we propose using sample level descriptions for better representation learning, rather than embeddings of class prototypes. This leads to more semantically rich representations, which in turn allows for better generalisation.

We describe how we train the proposed method for dynamic FER in Section III-A, and how we use it at inference time for simple and compound emotions in Section III-C

#### A. Architecture and Training

The CLIP framework, proposed by Radford *et al.* [28], operates as a contrastive multi-modal system, encoding image and text features into a shared space and maximising the cosine similarity of matching image-text pairs (positives) while minimising the similarity of all other pairs (negatives) by optimising a cross-entropy loss over the similarity pairs.

We adopt the framework in the dynamic paradigm by introducing a transformer encoder over the spatial features to learn from the temporal dimension.

More specifically, we utilise two separate encoders for processing video and text inputs. Given a video-text pair  $x = \{x^V, x^T\}$ , we obtain the video-text embeddings using the respective encoders so that  $\mathbf{z}^V = E_V(x^V)$  and  $\mathbf{z}^T = E_T(x^T)$ , where  $\mathbf{z}^V, \mathbf{z}^T \in \mathbb{R}^D$ . The video and text embeddings,  $\mathbf{z}^V$  and  $\mathbf{z}^T$ , respectively, obtained for each video-text pair in the mini-batch  $B$ , are utilised to generate a  $B \times B$  matrix of cosine similarities. The diagonal elements of the matrix correspond to the  $B$  positive pairings, while the remaining elements represent  $B^2 - B$  negative pairings. A cross-entropy loss is employed to maximise the similarity between the positives on the diagonal and minimise the similarity of the negatives.

The video encoder architecture is relatively simple and similar to architectures proposed for video captioning [22], [31]. Specifically, ( $E_V$ ) is composed of the CLIP image encoder ( $E_I$ ) that extracts frame-level features, which are then fed to a two-layer transformer encoder that acts as a temporal encoder. The state of the learnable classification token at the output of the transformer is used as the video embedding  $\mathbf{z}^V$ .

While the encoders  $E_V, E_T$  could be trained from scratch given a sufficiently large dataset of video-caption pairs, in the domain of FER, the only available dataset with such annotations that do not explicitly include emotional categories in the descriptions is the MAFW [20] dataset, which is relatively small. To this end, we leverage the pre-trained CLIP [28] image and text encoders to initialise the weights of  $E_I$  and  $E_T$  in our architecture and fine-tune on the target domain.

## B. Inference

During inference, the cosine similarity of text and image embedding in the joined latent space is used as the basis for the classification, as described by Radford *et al.* [28]. The prediction probability is then defined as:

$$P(y = i|x) = \frac{e^{\langle \mathbf{z}^V, \mathbf{z}_i^T \rangle / \tau}}{\sum_{j=1}^N e^{\langle \mathbf{z}^V, \mathbf{z}_j^T \rangle / \tau}}, \quad (1)$$

where  $\tau$  is a learnable temperature parameter in CLIP and  $\langle \cdot, \cdot \rangle$  is the cosine similarity.

## C. Class Descriptions

In the case of basic emotions, we provide class descriptions in the form of natural language obtained from LLMs, rather than using a prompt in the form of ‘*an expression of {emotion}*’, to match the information-rich sample level descriptions. The use of descriptions is intuitive in the context of emotions and FER, as using the class names would imply that the model has a deep understanding of emotional expression (e.g. in Fig. 1 “raised lip corners” and “happiness” are not directly associated). In addition, due to the very large intra-class variation of emotional expression, the definitions of emotions are fluid. Therefore, descriptions should be more fitted to differentiate fine-grained emotions. We note that these are different descriptions than the ones used during training, as in the latter, we use sample-level video-text pairs, as can be seen in Fig. 1. As such, our method is performing zero-shot classification during inference. Specifically, to obtain the class-level descriptions, we prompt ChatGPT with the input:

*Q: What are the facial expressions associated with {emotion}?*

We then curate the generated responses to exclude irrelevant information, such as body pose or emotional cues (such as “a sad expression” for helplessness). For example, “A facial expression showing irritation and unrest, with a wrinkled forehead, narrowed eyes, and tight lips or a frown” is the generated description for “*anger*”. Examples of prompts can be seen in Fig. 1 & 2; the full set of class descriptions used for inference are given in the Supplementary Material. This strategy is proposed over the prompt templates used in CLIP [28], as the prompt in the form of ‘*an expression of {emotion}*’ would imply a universal definition for each emotion, that the text encoder has learnt. Additionally, while the CLIP prompts have shown impressive results on clearly defined objects in images, emotions are significantly more vague and open to interpretation, both in terms of expression and understanding. The MAFW [20] dataset does not include sample-level descriptions for the neutral category; as such, we generate descriptions for neutral samples by randomly selecting and concatenating two prompts generated from ChatGPT for the neutral category. The full list of prompts for the neutral category is given in the Supplementary Material. The use of LLM, in this case, ChatGPT, over hand-crafted prompts is proposed to avoid introducing the authors’ bias in the prompts.

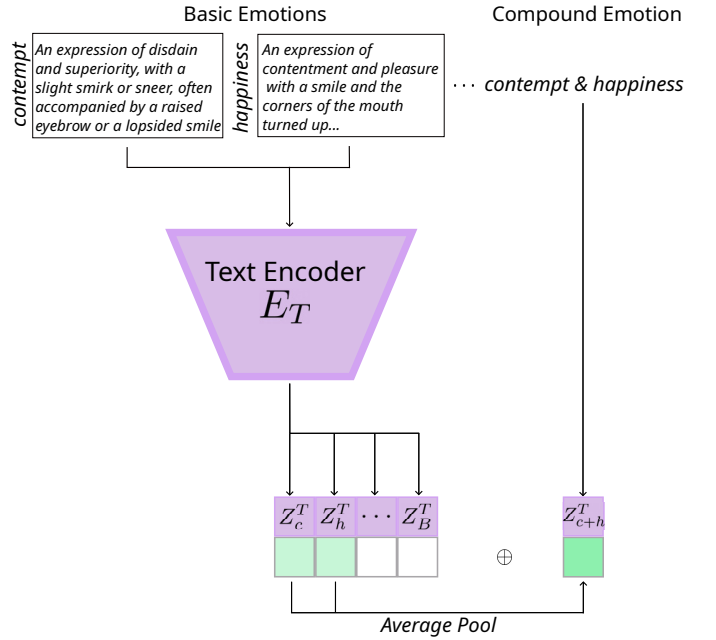


Fig. 2. EmoCLIP manipulates the latent space of basic emotions to create representations for compound emotions. We take the average latent representation of the components and concatenate it to the set of representations for each new compound emotion.

Compound emotions are a complex combination of basic emotions, such as “happily surprised” [7], [20], that are used to identify a wider range of human facial expressions. As, by definition, compound emotions are combinations of basic emotions, we propose a new approach to constructing the latent  $\mathbf{z}^T$  representation. Instead of treating them as independent emotional states and creating additional prompts, we use the pre-normalised latent representations of the components to compose the new compound emotion as shown in Fig. 2.

Formally, for any new compound emotion, we calculate its latent representation  $\mathbf{z}_n^T$  as the average of the latent representations of its  $C$  component emotions so that:

$$\mathbf{z}_n^T = \frac{1}{C} \sum_{c=0}^C \mathbf{z}_c^T \quad (2)$$

The resulting vector representations are then concatenated to the set of class representations, and inference is performed over the latent representations of basic and compound emotions.

## IV. EXPERIMENTAL RESULTS

In this section, we present the experimental set-up (Section IV-A), the ablation study (Section IV-B), and the experimental evaluation of the proposed framework in the Zero-Shot paradigm (Section IV-C) as well as the Downstream task of schizophrenia symptom estimation (Section IV-D).

### A. Experimental Setup

**Datasets:** We train on the MAFW [20] dataset, as to our knowledge, it is the only FER dataset that includes sample-level descriptions of the subject’s context and facial expres-

Architecture	Temporal Layer	Prompt Ensemble	Frame Ensemble	Pre-training	UAR	WAR
CLIP [28] (Frozen backbone)				Laion-400m	16.97	21.69
CLIP [28] (Frozen backbone)		✓		Laion-400m	19.77	18.64
CLIP [28] (Frozen backbone)		✓		Laion-400m	19.46	17.61
CLIP [28] (Frozen backbone)			✓	Laion-400m	20.40	21.16
CLIP [28]			✓	MAFW	17.62	20.57
EmoCLIP (Frozen backbone)	✓			MAFW	23.60	31.36
EmoCLIP	✓			MAFW	<b>25.86</b>	<b>33.49</b>

TABLE I

PERFORMANCE OF THE PROPOSED METHOD, EMOCLIP, ON THE MAFW [20] DATASET ON 11-CLASS SINGLE EXPRESSION CLASSIFICATION AGAINST THE BASELINE, WITH DIFFERENT FRAME AGGREGATION AND PROMPTING STRATEGIES.

sions without explicitly mentioning the emotional state. The dataset contains  $10k$  audio-video clips with categorical annotations for 11 emotions, accompanied by short descriptions of the subject’s facial expressions in two languages, English and Mandarin Chinese; we use the former for all experiments in this work. The dataset is divided into five folds for evaluation.

We evaluate our method on three additional FER datasets on the seven basic emotions. The **AFEW** [8] dataset contains 1,809 clips from movies, divided into three subsets. We train the method from scratch and report results on the validation set, as the labels of the test set are not publicly available. **DFEW** [14] is composed of 16,000 videos from movies, split into five folds for cross-validation. **FERV39K** [35] is a large dataset with around 39,000 clips annotated by 30 annotators, and we report results on the test set.

Finally, we evaluate the video encoder of EmoCLIP as trained on the MAFW dataset, on the downstream task of schizophrenia symptom estimation using a subset of the **NESS**[26] dataset, as described in [5], [11]. The subset includes 113 in-the-wild baseline clinical interviews from 69 patients and two symptom scales, PANSS[15] and CAINS[10]. To ensure a fair comparison with previous works, we used leave-one-patient-out cross-validation. The values of “Total Negative” and “EXP - Total” in PANSS and CAINS scales respectively, were scaled during training to match the range of individual symptoms. As the videos in the dataset are very long, ranging between 40 and 120 minutes long, we average the clip-level predictions to obtain the final video-level prediction vector for each video and calculate the network’s performance over the video-level predictions during inference, as described in [11].

**Training Details:** The CLIP image encoders used in all experiments have a ViT-B/32 architecture. During training, we apply augmentation to the spatial dimensions of the video inputs for all datasets. Specifically, we apply a random horizontal flip to all frames in a sequence with a probability of 50%. We also apply a random rotation with a range of  $6^\circ$  and random centre crop. In the temporal dimension, we empirically trim all videos to  $T = 32$  number of frames and downsample by a factor of 4 (which results in the entire video being included for the majority of samples). In the downstream task, the random crop along the temporal dimension acts as an augmentation technique.

For all experiments, we use a Stochastic Gradient Descent optimiser (SGD) with a learning rate of  $10^{-3}$ . We finetune

the pre-trained parameters of the CLIP backbone using a different learning rate of  $10^{-6}$  for the image and text encoders.

### B. Ablation Study

In order to examine the effect of different components of EmoCLIP and evaluate the prompting strategy used in this work, we perform several experiments using baseline CLIP and our proposed method, EmoCLIP. More specifically, as discussed in the Methodology Section III, during inference, we use class descriptions generated with the help of ChatGPT; this is different to the prompt ensemble of Radford *et al.* [28] who use several prompt templates in the form of ‘*an expression of {emotion}*’ and average their latent representation vectors during inference. Furthermore, as CLIP [28] is a static model, we conduct two zero-shot evaluation strategies, first on the middle frame and then by averaging the latent representations of all frames, i.e. performing frame ensemble. To further show the necessity for a temporal layer, we finetune a CLIP [28] architecture on the MAFW [20] dataset using frame ensembling, which has a negative effect on CLIP’s performance on the task. We theorise that frame ensembling in FER tasks negatively affects training as averaging out frame representations in most cases will consider noisy and keyframes equally, thus muddling the video latent features [17], which confuses the network.

The proposed architecture, which incorporates a temporal layer, has significant performance improvements compared to the baseline CLIP architectures. Finally, we compare the performance of EmoCLIP with a frozen backbone to that of our proposed method. The results of the ablation study can be seen in Tab. I. As the prompt templates used in CLIP [28] are similar to the format “A photo of class name”, they are more suitable for static images rather than video. This is corroborated by the performance of the CLIP [28] baseline using prompt ensemble vs our class descriptions. Additionally, as emotional expression is not static, we also observe an increase in the baseline performance using frame ensembling compared to evaluating on the middle frame. Such an outcome is intuitive as dynamic FER has a wider temporal context that needs to be considered rather than a single frame. Furthermore, the temporal relationships between frames hold important information in FER tasks. This is also confirmed by the proposed architecture, in-

Mode	Architecture	Contrastive Pre-training	UAR	WAR
Supervised	C3D [20]	-	31.17	42.25
	Resnet18_LSTM [20]	-	28.08	39.38
	VIT_LSTM [20]	-	32.67	45.56
	C3D_LSTM [20]	-	29.75	43.76
	T-ESFL [20]	-	33.28	<b>48.18</b>
	EmoCLIP (LP)	-	30.26	44.231
	EmoCLIP (Frozen backbone)	MAFW [class descriptions]	<b>34.24</b>	41.46
Zero-shot	CLIP [28]	Laion-400m	20.40	21.16.
	FaRL - ViT-B/16 [45]	Laion Face-20M	14.07	7.70
	EmoCLIP	MAFW [sample descriptions]	<b>25.86</b>	<b>33.49</b>

TABLE II

PERFORMANCE OF THE PROPOSED METHOD ON THE MAFW [20] DATASET ON 11-CLASS SINGLE EXPRESSION CLASSIFICATION AGAINST OTHER SOTA ARCHITECTURES IN A SUPERVISED AND ZERO-SHOT SETTING.

cluding a temporal module, which offers a large increase in performance for both Weighted Average Recall (WAR) and Unweighted Average Recall (UAR). By finetuning the backbone to the FER domain, we see a further increase in performance by approximately 2% for each metric.

Mode	Architecture	Repr.	Avg	UAR	WAR	F1	AUC
Supervised	C3D [20]	-	<b>9.51</b>	28.12	6.73	74.54	
	Resnet18_LSTM [20]	-	6.93	26.6	5.56	68.86	
	VIT_LSTM [20]	-	8.72	32.24	<b>7.59</b>	75.33	
	C3D_LSTM [20]	-	7.34	28.19	5.67	65.65	
	T-ESFL [20]	-	9.15	<b>34.35</b>	7.18	<b>75.63</b>	
Zero-shot	Random	-	2.38	7.72	0.34	50.00	
	CLIP [28]	✗	4.72	5.25	2.44	51.89	
	CLIP [28]	✓	4.14	5.35	2.46	<b>53.07</b>	
	FaRL [45]	✗	3.03	4.66	2.16	51.01	
	FaRL [45]	✓	4.00	5.75	2.56	51.10	
	EmoCLIP	✗	5.24	15.34	3.80	51.30	
	EmoCLIP	✓	<b>6.58</b>	<b>18.53</b>	<b>4.78</b>	52.59	

TABLE III

ZERO-SHOT CLASSIFICATION ON THE 43 COMPOUND EXPRESSIONS OF THE MAFW [20] DATASET. SUPERVISED METHODS ARE INCLUDED AS A REFERENCE.

### C. Zero-Shot Evaluation

To evaluate the effectiveness of the proposed method, we compare it with pre-trained CLIP [28] and FaRL [45] models in a Zero-shot setting. As both of these methodologies are trained on static images, we take the average of the latent representations of all frames in a video to compute the video embedding and use that to calculate the cosine similarity with the text description embeddings. We show the performance of our method against the CLIP and FaRL baselines, with a frozen CLIP backbone and the finetuned image-text encoders on the 11 class classification of MAFW [20] in Table II. We note that even though FaRL [45] is trained on a subset of the Laion dataset [30] filtered to include samples of faces, the model trained on FaRL performs significantly worse than both the CLIP baseline and our proposed method. We also note that the FaRL pre-trained weights are only available for the ViT-B/16 architecture, which may explain the difference in performance. Furthermore, while FaRL is trained with image-text pairs of faces, these are not necessarily related to facial expression or are not well grounded with text, as we

can see by visually inspecting the LAION-face dataset [46], so we hypothesise that the FaRL embedding space is significantly more niche than CLIP but not in a direction beneficial for zero-shot FER.

To further investigate the improvement of our method vs baseline CLIP, we use PCA to reduce the high-dimensional latent image vectors to three dimensions (which explain approximately 70% of the variance) and plot them in a 3D scatter plot, as shown in Fig. 3. We see that by fine-tuning to the FER domain, the categorical emotions form more distinct clusters than in CLIP, which is also reflected in the classification performance of our method. The more discriminate latent space is somewhat expected as the CLIP domain is much larger; thus, emotions and FER occupy a much smaller subspace, i.e. emotional categories will be closer to each other than to animals. By fine-tuning, the distances between dissimilar samples become larger, which allows for more discrete clusters.

For reference, we also train our architecture with class-level descriptions, and using an MLP head with two fully connected layers (Linear probe shown as LP on the table) over the EmoCLIP video encoder, we show the performance against several supervised architectures as reported in [20]. We see that the architecture trained using the class descriptions outperforms previous methods in terms of UAR and is comparable with others in terms of WAR. These results indicate that the contrastive vision-language approach leads to more semantically rich and discriminate latent representations even in a supervised setting. The difference in the two metrics is somewhat expected, as FER datasets are typically imbalanced.

Furthermore, we present experimental results on the classification of 43 compound emotions in the MAFW dataset in Table III. We evaluate the performance of our proposed method, EmoCLIP, against a baseline approach of using concatenated prompts, as well as CLIP and FaRL baselines. Specifically, we concatenate the class descriptions for each compound emotion and use this as class prompt input. We demonstrate that EmoCLIP outperforms the baseline approach for all metrics. Moreover, we note that in the 43 emotions classification, both CLIP and FaRL perform significantly worse than EmoCLIP and have performance

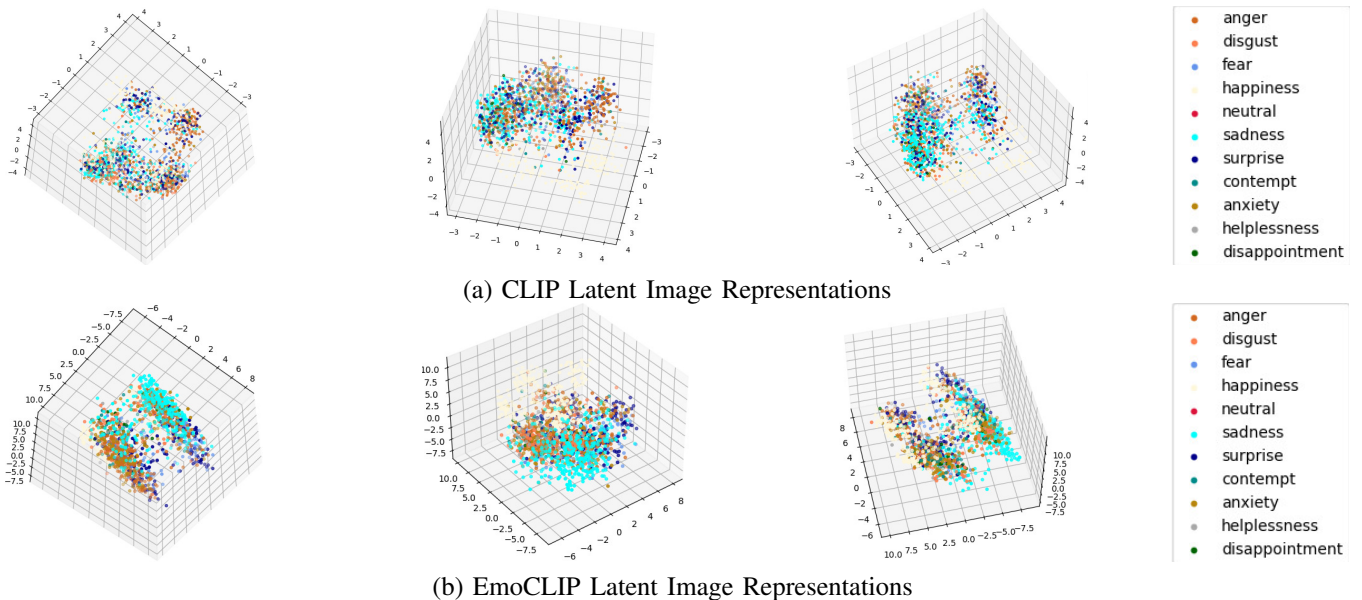


Fig. 3. Latent representation of the MAFW dataset using CLIP (a) and EmoCLIP (b) for each emotion category.

Architecture		Training labels	DFEW (7 classes)		AFEW (7 classes)		FERV39K (7 classes)		MAFW (11 classes)	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
Supervised	EmoCLIP	[class description]	<b>58.04</b>	<b>62.12</b>	<b>44.32</b>	<b>46.19</b>	<b>31.41</b>	36.18	<b>34.24</b>	41.46
	EmoCLIP (LP)	[class]	50.29	62.09	33.74	38.85	30.58	<b>43.54</b>	30.26	<b>44.21</b>
Zero-shot	CLIP [28]	[image caption]	19.86	10.60	23.05	11.80	20.99	17.10	20.04	21
	EmoCLIP (leave-one-class-out)	[class description]	22.85	24.96	35.11	27.57	<b>39.35</b>	<b>41.60</b>	24.12	24.74
	EmoCLIP	[video caption]	<b>36.76</b>	<b>46.27</b>	<b>36.13</b>	<b>39.90</b>	<u>26.73</u>	<u>35.30</u>	<b>25.86</b>	<b>33.49</b>

TABLE IV

EVALUATION OF EMOCLIP USING SAMPLE DESCRIPTIONS VS CLASS-LEVEL DESCRIPTION AS NATURAL LANGUAGE SUPERVISION, ON FOUR VIDEO FER DATASETS.

comparable to random (where only the majority class is predicted). We theorise this is due to the lack of temporal understanding of the static models. Furthermore, without fine-tuning on the target domain, the class descriptions are not discriminate enough for the zero-shot classification of emotions, as previously discussed. However, the representation average method on both baselines improves the performance of the static models on the compound emotions task. We also report the results of several supervised methods for reference, which perform significantly better than zero-shot approaches as expected.

Finally, we evaluate the performance of our proposed method using sample-level descriptions from MAFW [20] on four widely used video FER datasets and compare it with the CLIP baseline as shown in Table IV. Additionally, in line with previous works in zero-shot emotion classification [4], [37], [44], [27], we train our architecture using class-level descriptions and evaluate using leave-one-class-out (loco) cross-validation. We note that we cannot directly compare with these architectures, as they involve either different modalities (e.g. audio, pose) [4], [37], [27], [38] or a different task [44], we adopt however, their experimental set-up using our architecture to show how natural language supervision

and semantically rich class descriptions can help improve zero-shot FER performance.

We observe that the EmoCLIP trained on MAFW [20] sample-level descriptions show impressive generalisation ability on all datasets that we evaluate. Specifically, for AFEW [8], MAFW [20] and DFEW [14], we see that the EmoCLIP model is outperforming both the loco experiment and the CLIP [28] baseline. Furthermore, the generalisation of the method is resistant to domain shift from unseen datasets, as we observe from the significant performance increase between the CLIP [28] baseline and EmoCLIP. We note that for FERV39K [35], the loco experiment has a higher performance than the sample-wise training. However, it is very important to stress that the FERV39K [35] is significantly larger than the base dataset (over 3x more samples); therefore, methods trained on it would have an advantage, particularly as in the loco experiment, there is no domain shift.

For reference and to provide context, we include the performance of the architecture in a supervised setting on all four datasets, as there are not many methods performing zero-shot classification in FER. The supervised methods outperform zero-shot, as expected. However, we can com-

	N3: Poor Rapport			N6: Lack of Spont.			N1: Blunted Affect			Total Negative Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron <i>et al.</i> [32]	0.98	1.31	.20	1.37	1.69	.13	0.90	1.28	.37	-	-	-
Tron <i>et al.</i> [33]	1.01	1.26	.15	1.32	1.62	.09	0.99	1.36	.11	-	-	-
SchiNet [5]	<b>0.85</b>	1.20	.27	1.25	1.51	.25	0.84	1.18	.42	3.30	4.17	.29
Relational [11]	0.87	<b>1.16</b>	<b>.78</b>	<b>0.74</b>	<b>0.95</b>	<b>.47</b>	0.64	0.87	.56	2.80	3.78	.71
EmoCLIP (LP)	0.89	1.28	.72	0.91	1.15	.27	<b>0.56</b>	<b>0.79</b>	<b>.63</b>	<b>2.31</b>	<b>3.01</b>	<b>.85</b>

TABLE V  
PERFORMANCE ON THE DOWNSTREAM TASK AGAINST OTHER SoTA (PANSS-NEG).

	Facial Exp.			Vocal Exp.			Expr. Gestures			Quant. of Speech			EXP-Total Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron <i>et al.</i> [32]	0.80	1.03	.37	0.87	1.23	.23	0.85	1.19	.36	1.09	1.43	.27	-	-	-
Tron <i>et al.</i> [33]	0.75	1.07	.36	0.86	1.22	.26	0.91	1.22	.38	1.02	1.36	.25	-	-	-
SchiNet [5]	0.66	0.93	.46	0.77	1.10	.27	0.90	1.15	.36	0.98	1.30	.30	2.67	3.34	.45
Relational [11]	0.56	0.72	.75	0.65	0.89	.71	0.71	0.89	<b>.76</b>	<b>0.60</b>	<b>0.82</b>	<b>.54</b>	1.88	2.60	.77
EmoCLIP (LP)	<b>0.49</b>	<b>0.65</b>	<b>.77</b>	<b>0.59</b>	<b>0.83</b>	<b>.74</b>	<b>0.66</b>	<b>0.85</b>	<b>.74</b>	0.64	0.89	.50	<b>1.32</b>	<b>2.52</b>	<b>.83</b>

TABLE VI  
PERFORMANCE ON THE DOWNSTREAM TASK AGAINST OTHER SoTA (CAINS-EXP).

pare the performance of our proposed architecture trained contrastively with class-level descriptions and using an MLP head for multi-class classification (Linear Probe). We want to point out that the architecture trained using class descriptions is significantly outperforming the linear probe architecture, particularly in terms of UAR, showing again that natural language supervision can provide significant advantages even in a fully supervised setting. This is especially true for under-represented classes, as indicated by the increase in UAR, which is sensitive to the long tail of the distribution.

#### D. Downstream Task

We evaluate the representations obtained by the proposed method on the downstream task of schizophrenia symptom estimation on two scales, namely the PANSS [15] and CAINS [10] scales. As symptoms in both scales have ordinal labels, we address the problem of symptom estimation as a multi-label regression. We evaluate the proposed method in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Pearson’s Correlation Coefficient (PCC) for a fair comparison with previous works in the literature. We train the linear probe architecture, with the video encoder obtained through contrastive pre-training on the MAFW [20] frozen, and updating only the MLP weights. The results against previous state-of-the-art (SoTA) can be seen on Tables V & VI for the PANSS [15] and CAINS [10] scales respectively. The proposed method, pre-trained contrastively on sample level descriptions, outperforms or is comparable to previous methods on all symptoms, particularly for “N1: Blunted Affect” and “Facial Expression” on the PANSS and CAINS scales, which are by definition most related to apparent affect and FER.

As the NESS [26] dataset is annotated by multiple health-care professionals, we can compare the PCC achieved by the method to the annotators’ PCC which has a mean value of 0.85 across all symptoms [5], [26]. The proposed method on the downstream task achieves performance comparable to

that of human experts on the total scores of both scales, as we see in Tables V & VI.

#### V. CONCLUSION

In this work, we presented a novel contrastive pre-training paradigm for video FER, trained on video-text pairs with sample-level descriptions without any class-level information. While contrastive learning and natural language supervision have been used in other domains, zero-shot emotion recognition remains surprisingly unexplored, with works focusing on creating class prototypes with simpler word encoding methods [4], [27], [37], [38]. Emotional prototypes, however, disregard the intra-class variation that is inherently present in FER tasks. To overcome the limitations of training on coarse emotional categories, EmoCLIP is trained on sample-level descriptions. We evaluate our method on four popular FER video datasets [8], [14], [35], [20] and test using zero-shot evaluation on the basic emotions as well as compound emotions. Our method outperforms the CLIP baseline by a large margin and shows impressive generalisation ability on unseen datasets and emotions. To our knowledge, this is the first work to train with sample-level descriptions for FER and to propose zero-shot evaluation using semantically rich class descriptions in the domain. We also evaluated the EmoCLIP video encoder features on schizophrenia symptom estimation, outperforming previous state-of-the-art methods and achieving performance comparable to human experts in terms of PCC.



## APPENDIX

In this supplementary document, we present additional qualitative results that complement and extend the findings outlined in the main text. More specifically, in Appendix , we show the detailed class descriptions used by our methods. Then, in Appendix , we show examples of correctly and incorrectly classified samples.

### CLASS DESCRIPTIONS

As described in Section III-C, we prompt a popular Large Language Model (LLM), specifically ChatGPT, to obtain the descriptions for each class. The descriptions are then curated manually to exclude information unrelated to the facial expression (e.g. body pose). The final list of descriptions can be seen in Table VII.

Class Name	Description
anger	A facial expression showing irritation and unrest, with a wrinkled forehead, narrowed eyes, and tight lips or a frown
disgust	An expression of repulsion and displeasure, with a raised upper lip, a scrunched nose, and a downturned mouth
fear	An expression of tension and withdrawal, with wide-open eyes, raised eyebrows, and a slightly open mouth. The face may appear physically tense or frozen in fear
happiness	An expression of contentment and pleasure, with a smile and the corners of the mouth turned up, often accompanied by crinkling around the eyes. The face may appear relaxed and at ease
neutral	An expression of calm and neutrality, with a neutral mouth and no particular indication of emotion. The eyebrows are usually not raised or furrowed
sadness	An expression of sadness and sorrow, with a downturned mouth or frown, and sometimes tears or a tightness around the eyes. The face may appear physically withdrawn or resigned
surprise	An expression of shock and astonishment, with wide-open eyes and raised eyebrows, sometimes accompanied by a gasp or an open mouth
contempt	An expression of disdain and superiority, with a slight smirk or sneer, often accompanied by a raised eyebrow or a lopsided smile
anxiety	An expression of worry and apprehension, with furrowed eyebrows and a tight mouth. The eyes may appear wide and darting, and the face may appear physically tense or worried
helplessness	An expression of defeat and resignation, with the eyes looking down and the mouth turned down. The eyebrows may be furrowed, and the face may appear resigned or resigned
disappointment	An expression of frustration and disillusionment, with a slight frown or drooping of the mouth. The eyebrows may be lowered or furrowed, and the face may appear physically drawn or tired

TABLE VII

CLASS DESCRIPTIONS FOR EACH EMOTION USED DURING INFERENCE

The MAFW [20] dataset does not include sample-level descriptions for the neutral category; as such, we use variations of the neutral description randomly for each sample in the neutral category. The list of neutral descriptions for samples with no description during training can be seen in Table. VIII.

1.	A lack of emotional expression, as if the person’s face is in a resting state.
2.	The facial muscles are generally relaxed, creating a smooth and even appearance.
3.	The mouth is typically closed or slightly open, with the lips not turned up or down.
4.	The eyebrows are in a neutral position, not furrowed or raised, and the eyes are generally looking straight ahead or slightly down.
5.	While the face may not show any specific emotions, the expression can still convey a sense of attentiveness or alertness.

TABLE VIII

DESCRIPTIONS USED FOR NEUTRAL CATEGORY SAMPLES DURING TRAINING

As AFEW [8], DFEW [14] and FERV39K [35] only have seven emotional classes, i.e. *anger*, *disgust*, *fear*, *happiness*, *neutral*, *sadness* & *surprise*, during inference we exclude descriptions for the additional classes (*contempt*, *anxiety*, *helplessness* & *disappointment*).

### QUALITATIVE ANALYSIS

Figure 4 showcases a selection of correctly and incorrectly classified instances from the MAFW dataset. As the interpretation of emotion for humans is dependent on not only the facial expression but also the context, we see that some samples are harder than others to classify. It appears that examples with higher emotional intensity and more animated subjects appear to be easier to classify. Conversely, subtle expressions of emotion are prone to misclassification. For instance, in the anger examples depicted in Figure 4, the man displays multiple anger-associated expressions, such as a furrowed brow, whereas the woman exhibits a calmer demeanour. Similarly, in the happiness examples, the appropriately classified sample features a smiling man, while the incorrectly classified example shows a man with a subtle facial expression while speaking.

### REFERENCES

- [1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.
- [2] P. Bagad, M. Tapaswi, and C. G. M. Snoek. Test of Time: Instilling Video-Language Models with a Sense of Time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Apr. 2023.
- [3] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. A CLIP-Hitchhiker’s Guide to Long Video Retrieval, May 2022.
- [4] A. Banerjee, U. Bhattacharya, and A. Bera. Learning Unseen Emotions from Gestures via Semantically-Conditioned Zero-Shot Perception with Adversarial Autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):3–10, June 2022. Number: 1.
- [5] M. Bishay, P. Palasek, S. Priebe, and I. Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Transactions on Affective Computing*, 12(4):949–961, 2019.
- [6] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, pages 1–18, 2021.

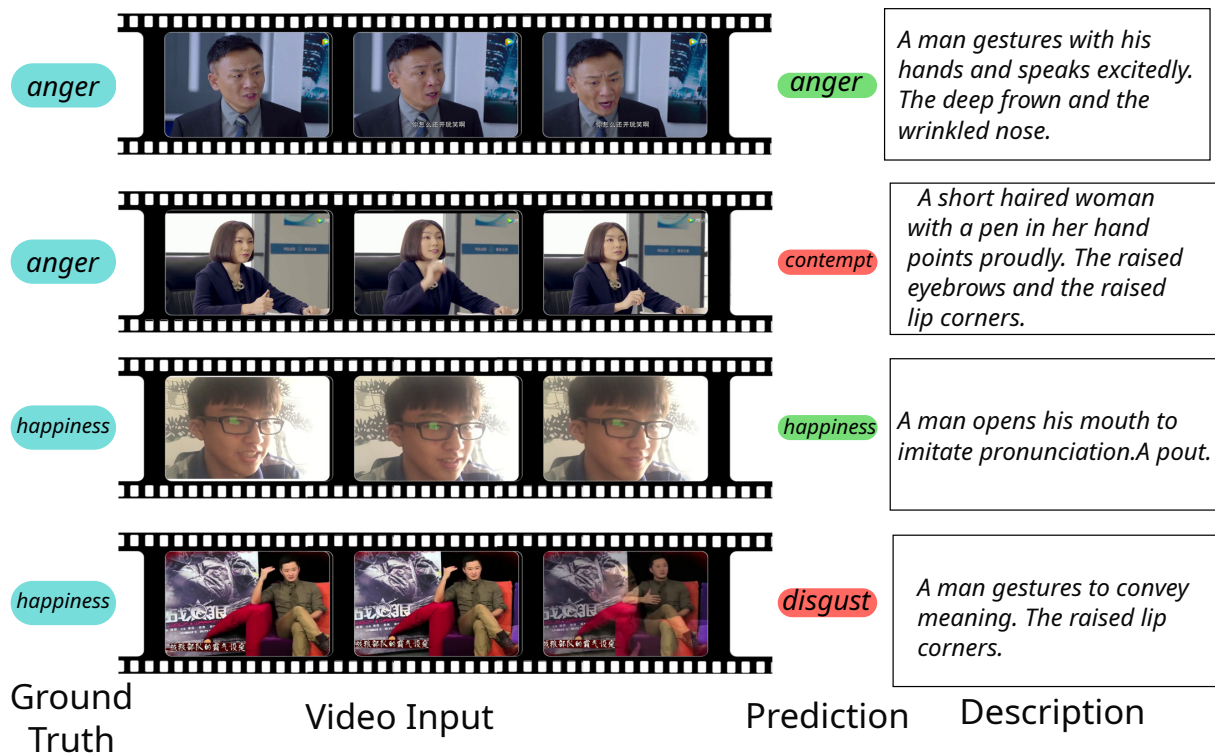


Fig. 4. Example frames of correctly and incorrectly classified samples from the MAFW [20] dataset, for anger (top) and happiness (bottom). The sample-level descriptions are included for reference but are not used during inference.

[7] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

[8] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012.

[9] P. Ekman and W. V. Friesen. *Facial action coding system: Investigator’s guide*. Consulting Psychologists Press, 1978.

[10] C. Forbes, J. J. Blanchard, M. Bennett, W. P. Horan, A. Kring, and R. Gur. Initial development and preliminary validation of a new negative symptom measure: The Clinical Assessment Interview for Negative Symptoms (CAINS). *Schizophrenia Research*, 124(1-3):36–42, Dec. 2010.

[11] N. M. Foteinopoulou and I. Patras. Learning from Label Relationships in Human Affect. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 80–89, Lisboa Portugal, Oct. 2022. ACM.

[12] N. M. Foteinopoulou, C. Tzelepis, and I. Patras. Estimating continuous affect with label uncertainty. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Nara, Japan, Sept. 2021. IEEE.

[13] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[14] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu. Dfnet: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.

[15] S. R. Kay, A. Fiszbein, and L. A. Opler. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276, Jan. 1987.

[16] H. Li, H. Niu, Z. Zhu, and F. Zhao. CLIPER: A Unified Vision-Language Framework for In-the-Wild Facial Expression Recognition, Feb. 2023. arXiv:2303.00193 [cs].

[17] H. Li, M. Sui, Z. Zhu, et al. Nr-dfnet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022.

[18] Y. Li, M. Wang, M. Gong, Y. Lu, and L. Liu. FER-former: Multi-modal Transformer for Facial Expression Recognition, Mar. 2023. arXiv:2303.12997 [cs].

[19] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li. Frozen CLIP Models are Efficient Video Learners. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 388–404, Cham, 2022. Springer Nature Switzerland.

[20] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, pages 24–32, New York, NY, USA, Oct. 2022. Association for Computing Machinery.

[21] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, Oct. 2022.

[22] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, pages 638–647. Association for Computing Machinery, Oct. 2022.

[23] S. Menon and C. Vondrick. Visual classification via description from large language models. *ICLR*, 2023.

[24] Y. Ouali, A. Bulat, B. Martinez, and G. Tzimiropoulos. Black Box Few-Shot Adaptation for Vision-Language models, Apr. 2023.

[25] S. Parisot, Y. Yang, and S. McDonagh. Learning to Name Classes for Vision and Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Apr. 2023.

[26] S. Priebe, M. Savill, T. Wykes, R. Bentall, U. Reininghaus, C. Lauber, S. Bremner, S. Eldridge, and F. Röhricht. Effectiveness of group body psychotherapy for negative symptoms of schizophrenia: multicentre randomised controlled trial. *The British Journal of Psychiatry*, 209(1):54–61, 2016.

[27] F. Qi, X. Yang, and C. Xu. Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM ’21, pages 1074–1083, New York, NY, USA, Oct. 2021. Association for Computing Machinery.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings*

- of *Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [29] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
- [30] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, Nov. 2021. arXiv:2111.02114 [cs].
- [31] M. Tang, Z. Wang, Z. LIU, F. Rao, D. Li, and X. Li. CLIP4Caption: CLIP for Video Caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 4858–4862, New York, NY, USA, Oct. 2021. Association for Computing Machinery.
- [32] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 72–81. Springer, 2015.
- [33] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 220–223. IEEE, 2016.
- [34] M. Wang, J. Xing, and Y. Liu. ActionCLIP: A New Paradigm for Video Action Recognition, Sept. 2021. arXiv:2109.08472 [cs].
- [35] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang. Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20922–20931, 2022.
- [36] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.
- [37] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller. Exploring Zero-Shot Emotion Recognition in Speech Using Semantic-Embedding Prototypes. 24, 2022. Conference Name: IEEE Transactions on Multimedia.
- [38] X. Xu, J. Deng, Z. Zhang, Z. Yang, and B. W. Schuller. Zero-shot speech emotion recognition using generative learning with reconstructed prototypes. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [39] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, and J. Luo. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment, Sept. 2022. arXiv:2209.06430 [cs].
- [40] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin. Affective computing in education: A systematic review and future research. *Computers & Education*, 142:103649, 2019.
- [41] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning, Feb. 2023. arXiv:2302.14115 [cs].
- [42] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In *International Conference on Learning Representations*, 2023.
- [43] G. Zara, S. Roy, P. Rota, and E. Ricci. AutoLabel: CLIP-based framework for Open-set Video Domain Adaptation, Apr. 2023. arXiv:2304.01110 [cs].
- [44] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang. Zero-shot emotion recognition via affective structural embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1151–1160, 2019.
- [45] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen. General Facial Representation Learning in a Visual-Linguistic Manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18676–18688, June 2022.
- [46] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.
- [47] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.