Enhancing Zero-Shot Facial Expression Recognition by LLM Knowledge Transfer

Zengqun Zhao, Yu Cao, Shaogang Gong, Ioannis Patras
Queen Mary University of London
{zengqun.zhao, yu.cao, s.gong, i.patras}@qmul.ac.uk

Abstract

Current facial expression recognition (FER) models are often designed in a supervised learning manner and thus are constrained by the lack of large-scale facial expression images with high-quality annotations. Consequently, these models often fail to generalize well, performing poorly on unseen images in inference. Vision-language-based zero-shot models demonstrate a promising potential for addressing such challenges. However, these models lack task-specific knowledge and therefore are not optimized for the nuances of recognizing facial expressions. To bridge this gap, this work proposes a novel method, Exp-CLIP, to enhance zero-shot FER by transferring the task knowledge from large language models (LLMs). Specifically, based on the pre-trained vision-language encoders, we incorporate a projection head designed to map the initial joint vision-language space into a space that captures representations of facial actions. To train this projection head for subsequent zero-shot predictions, we propose to align the projected visual representations with task-specific semantic meanings derived from the LLM encoder, and the text instruction-based strategy is employed to customize the LLM knowledge. Given unlabelled facial data and efficient training of the projection head, Exp-CLIP achieves superior zero-shot results to the CLIP models and several other large vision-language models (LVLMs) on seven in-the-wild FER datasets. The code is available at https://github.com/zengqunzhao/Exp-CLIP.

1. Introduction

Facial expressions are crucial for human communication, allowing individuals to convey a wide range of information non-verbally. In the field of Human Computer Interaction, Facial Expression Recognition (FER) often aims at the analysis of the underlying emotions or sentiments, and has various applications such as assisting people who are blind or have low vision [12], monitoring patients for better mental health assessment [3] and improving user expe-
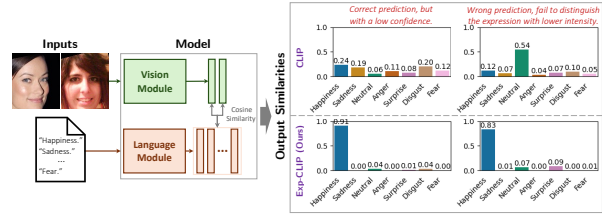

Figure 1. CLIP model learned more general feature representations, lacking task-specific knowledge and therefore are not optimized for the nuances of recognizing facial expressions.

rience when interacting with applications and devices [24]. FER can be categorized into two types based on input distinctions: Static FER (SFER) and Dynamic FER (DFER). SFER involves analyzing individual images or snapshots of faces while DFER focuses on analyzing facial expressions across a sequence of images or frames.

In the past decade, research in both SFER and DFER has shifted from controlled laboratory environments to more challenging in-the-wild conditions [22]. Given the availability of large-scale facial expression datasets captured in real-world scenarios [11, 11, 21, 25] and advancements in foundational models [7, 10], current FER models have attained high accuracy [17, 33, 42, 44]. However, these deep FER models were mainly trained in a supervised manner, relying heavily on large-scale training data with categorical annotations. The reliance on such labelled data presents significant challenges. Firstly, the process of acquiring and annotating large-scale training data is both labour-intensive and costly. Secondly, supervised FER models tend to struggle with generalization across different domains [4]. This means that while these models may perform well within the domain they were trained on, their performance often degrades when applied to new, unseen domains. Additionally, FER annotations frequently encounter challenges related to noise and inconsistent annotations caused by subjectivity, hindering the training process [34, 38]. Given these challenges, exploring alternative approaches becomes crucial. One promising direction is zero-shot learning, which uses semantic information to bridge the gap between seen and unseen classes. This approach is capable of generalizing
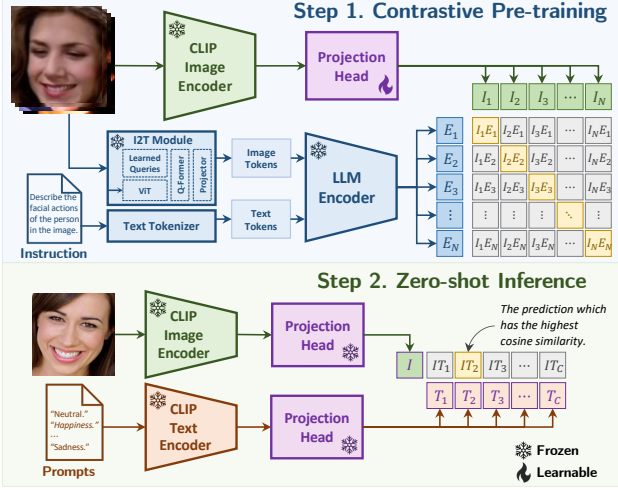
Figure 2. The proposed framework in this paper introduces contrastive pre-training and zero-shot FER. In the testing phase, a learned projection head is employed to enhance image-text representations for facial expressions. This projection head is learned in an unsupervised manner, leveraging the knowledge from LLMs. The I2T module consists of a ViT, a Q-Former [20], and a projector, which is adopted to map the images into LLM tokens.

to unseen expressions and removes the need for extensive annotation of training data [27].

Recently, large vision-language pre-training models have shown potential for achieving promising zero-shot performance [20, 28, 37, 49]. Those models are often designed using one of the following two strategies: (a) aligning large connections of images and raw texts using two separate encoders [28, 30, 49] or (b) incorporating visual modality to Large Language Models (LLMs) to create Large Vision-Language Models (LVLMs) [1, 16, 20]. Under the previous training paradigm, the CLIP model [28] was designed to learn from vast datasets containing millions of image-text pairs. This extensive training enables the model to achieve remarkable zero-shot capabilities, effectively categorizing images into various classes without specific training on those categories. Building on the foundations of the CLIP model, the FaRL model [49] extends this approach by focusing specifically on face-image-text pairs. This adaptation allows FaRL to develop a universal facial representation effective across a broad spectrum of face analysis tasks. Compared to previous models, FaRL demonstrates enhanced performance in accuracy and generalization. EmotionCLIP [41] is proposed to extract visual emotion representations from verbal and nonverbal communication. However, the original CLIP model learned more general feature representations may not be optimized for the nuances of recognizing facial expressions (see Fig. 1). Training texts in FaRL and EmotionCLIP are either general appearance descriptions or conversations, lacking detailed facial action of expression descriptions. Fine-tuning the CLIP model with facial expression visual data and their cor-

responding text descriptions, as discussed in [8], appears to be a more direct approach. However, this idea relies on facial images with detailed facial action descriptions and has the same manual labour costs as the supervised approaches. Whilst LVLM-based approaches appear to promise automation, prompting such models for zero-shot tasks demands substantial engineering efforts [9]. Additionally, these models suffer problems caused by hallucinations of LLM decoders [29]. It remains a significant challenge to achieve reliable zero-shot FER.

In this work, we present a novel approach to enhancing zero-shot FER model learning by leveraging LLMs knowledge transfer. Specifically, to improve the zero-shot capability of large pre-trained vision-language models for a given task, we propose to conduct projection on latent visual and textual representations to emphasize crucial task-specific elements while disregarding irrelevant ones. To this end, a shared projection head on top of the aligned image-encoder and text-encoder is introduced, serving the purpose of projecting the initial joint vision-language space into a task-specific joint space. To learn this projection head subject to optimising the downstream zero-shot FER prediction task, we consider aligning the projected visual embeddings of the image-encoder with task-specific semantic embeddings of the LLM encoder based on a contrastive training framework. By doing so, a text instruction-based strategy is utilized to customize the LLM knowledge. This alignment is beneficial to the image encoder by absorbing task-specific knowledge from LLMs. The framework of the proposed method is shown in Fig. 2. Our method involves contrastive pre-training and zero-shot inference in two parts: First, contrastive pre-training is introduced for training the projection head in an unsupervised manner, and second, the learned projection head is applied to both image and text encoder in inference for the zero-shot FER task. Our contributions are:

- We propose to align a general vision-language feature space into a task-aware LLM feature space by learning a lightweight projection head so to improve zero-shot FER model learning based on pre-trained vision-language models. The learnable projection head is optimized through contrastive learning on unlabeled facial data in an unsupervised fashion.

- We utilize a text instruction-based strategy to customize the LLMs knowledge for a downstream task and elicit the task-specific semantic features of LLM.

- The proposed method outperforms the various LVLMs on both image-based and video-based FER datasets.

2. Related Work

2.1. Facial Expression Recognition

Traditional FER Models: The traditional FER model relies on a feature extractor and a classifier. In the past

decade, research on static FER has primarily focused on addressing real-world challenges such as occlusion [35,46], pose variation [32], domain shift [4], and label noise [34,38, 47]. In the latest work, Wu *et al.* [38] introduce a landmark-aware net to address the ambiguity of expressions by leveraging facial landmarks to mitigate the impact of label noise. Lee *et al.* [14] proposed LatentOFER to improve FER accuracy under face occlusions. In their method, the model can detect occlusions, restore occluded parts of the face as if they were unoccluded, and recognize them. To address the imbalanced FER problem, Zhang *et al.* [43] propose to leverage re-balanced attention maps to regularize the model, enabling it to extract transformation invariant information about the minor classes from all training samples. In contrast, research on Dynamic FER primarily aims to develop models capable of generating robust spatial-temporal representations of facial expressions. The introduction of the Transformer to DFER, as seen in Former-DFER [45], marked a significant advancement, demonstrating superior performance compared to methods based on 3DCNNs or CNN-RNNs. Subsequent studies have further enhanced performance by addressing issues such as noisy frames [19] and the intensity of varied frames [17]. Moreover, Wang *et al.* [33] introduced M3DFEL to tackle the challenge of imbalances between short- and long-term temporal relationships in DFER. The concept of soft labelling [13] has also been applied to reduce data ambiguity.

Vision-language-based FER Models: In contrast to traditional FER models, vision-language-based models do not rely on classifiers. Instead, they derive final classification results by comparing the similarity between visual features and diverse text features, selecting the highest similarity as the ultimate prediction. Based on the CLIP, Li *et al.* [18] introduced CLIPER, which incorporates multiple expression text descriptors to learn fine-grained expression representations, achieving state-of-the-art performance on various FER benchmarks. Zhao *et al.* [48] proposed DFER-CLIP, designed for dynamic FER with temporal learning, introducing textual descriptions of facial behaviour for contrastive learning. Zhang *et al.* [42] presented the CLEF model, which conducts text-driven contrastive learning for facial behaviour understanding, utilizing activity descriptions and coarse-grained information from certain datasets. However, the reliance on extensive labelled facial data in these models presents a challenge in data acquisition. The recent work by Foteinopoulou *et al.* [8] introduces a novel vision-language model utilizing sample-level text descriptions as natural language supervision for facial expression video to enhance zero-shot FER classification. Nevertheless, this approach faces the difficulty of labelling detailed facial action descriptions, requiring expertise in perceiving nuanced facial movements. In contrast, our method utilizes facial images without annotations in an unsupervised fash-

ion, presenting a more labour-efficient approach.

2.2. Large Vision-Language Models

The current landscape of LVLMs can be broadly categorized into two main types. The first type involves aligning large collections of images and raw texts by utilizing two distinct encoders [26, 28, 30, 40, 41, 49]. The second type integrates the visual modality into LLMs to form comprehensive LVLMs [1, 16, 20, 39]. In the context of the first type, the pioneer work CLIP [28] maps data of different modalities, text and images, into a shared embedding space. This shared space exhibits the capability to generalize to various image classification tasks with zero- and few-shot learning. Subsequently, VideoCLIP [40] and XCLIP [26] were introduced for zero-shot video-text understanding. FaRL [49] focuses on learning universal facial representations from face-image-text pairs, while EmotionCLIP [41] is designed to extract visual emotion representations from both verbal and nonverbal communication. The recent FLAVA [30] learned a foundational language and vision representation that enables unimodal vision and language understanding as well as multimodal reasoning, all within a single pre-trained model. Regarding the LLMs-based models, Flamingo [1] bridges the powerful pre-trained vision-only and language-only models, and BLIP2 [20] further improves the capacity of the LVLMs by bridging the modality gap with a lightweight Querying Transformer and adopting more powerful LLM. The recent model LLaVa [16] was trained by fine-tuning LLMs on GPT-generated multimodal instruction-following data, demonstrating impressive multimodel chat abilities. Nevertheless, the zero-shot FER performance of these models is challenged by substantial engineering efforts on prompting [9] and the hallucination issue inherited from LLMs decoders [29].

3. Methods

The core idea of our method is to map a general visual representation into a task-aware representation beneficial to facial expression recognition. A projection-based strategy is proposed to fulfil this purpose, in which a projection head is learned through leveraging the task-specific knowledge from the LLMs. We will first present an overview of our method in Sec. 3.1 and then illustrate the zero-shot FER task in Sec. 3.2, finally, we will show how to train the learnable projection head for zero-shot learning in Sec. 3.3.

3.1. Overview

Fig. 1 shows an overview of our approach. As we can see, there are two phases involved in our method: contrastive pre-training and zero-shot inference. For zero-shot inference, the network tries to recognize the unseen facial images or videos into one of the categorical emotions. Different from the supervised approaches, the network is not

---
Algorithm 1 Pseudo Code of the Proposed Method

Require: Pre-training Dataset X, Instruction c_s, Prompt List C_p, CLIP Image Encoder $E_I(\cdot)$, CLIP Text Encoder $E_T(\cdot)$, I2T Module $Q(\cdot)$, LLM-Encoder (T5) $G(\cdot)$, Text Tokenizer $T(\cdot)$

Notation: Sample $x \sim X$, Prompt $c_p^i \sim C_p$, Latent code $e_I \sim E_I(x)$, $e_P \sim E_T(c_p^i)$, $e_Q \sim Q(x)$, $e_G \sim G(e_Q, T(c_s))$

Stage 1: Pre-training Projection Head

Randomly initialize a projection head $P_\theta(\cdot)$ which project $e_I \to e_G$

while not converged **do**
    Generate embeddings: $e_I \leftarrow E_I(x), e_G \leftarrow G(Q(e_I), T(c_s))$
    Compute the prediction based on precise e_G: $e'_I \leftarrow P(e_I)$
    Update P_θ: $P_\theta \leftarrow P_\theta - \nu \nabla_{P_\theta} \mathcal{L}$ where $\mathcal{L} = \|e_G - e'_I\|^2$
end while
return P_θ

Stage 2: Zero-Shot Prediction

Trained projection head $P_\theta(\cdot)$

for each caption c_p^i in C_p **do**
    Generate embeddings: $e_I = E_I(x), e_P = E_T(c_p^i)$
    Compute cosine similarity: $\text{cos\_sim}(P_\theta(e_I), P_\theta(e_P)) = \frac{P_\theta(e_I) \cdot P_\theta(e_P)}{\|P_\theta(e_I)\| \|P_\theta(e_P)\|}$
end for
return c_p^* which yields the highest similarity score is the prediction.

---

trained with any labelled facial images or videos. In contrast to the plain vision-language pre-training model, we applied a shared projection head on top of both the image encoder and text encoder. The projection head is learned to project the general joint vision and language space to a task-aware joint vision and language space, boosting the zero-shot ability in the specific task. Regarding the contrastive pre-training stage, the unlabelled facial images are used as input for both the CLIP image encoder and the I2T module. The I2T module contains a ViT, a Q-Former, and a projector from BLIP2 [20], which is adapted to project the image representations into language tokens, and these tokens will feed into the LLM encoder along with the tokenized instructions. During the training, the outputs of the projection head and LLM encoder will be aligned by the contrastive loss, in which only the projection head will be optimised. The pseudo-code of the proposed method is shown in Algorithm 1.

3.2. Enhanced Zero-shot FER Model

Our approach leverages the pre-trained CLIP model, which is designed to align images with their corresponding textual descriptions, making it naturally fit for zero-shot tasks. This alignment is accomplished by comparing the features of the images with the classification weights generated by the text encoder. The text encoder utilizes textual descriptions that specify the classes of interest as input.

Formally, for SFER, given a facial image $x \sim X$ with size $H \times W$, the CLIP image encoder $E_I(\cdot)$ is used to get the CLIP visual embedding $e_I \in \mathbb{R}^L$, where L is the length of the feature vectors. For DFER, given a facial video $x \sim X$, we sample T frames of size $H \times W$ to form an input $x \in \mathbb{R}^{T \times 3 \times H \times W}$. For each frame x^t, we first utilize

the shared CLIP image encoder $E_I(\cdot)$ to extract frame-level feature vectors $e^t \in \mathbb{R}^L$, where $t \in \{1, 2, \cdots, T\}$. Then average pooling is applied along the temporal axes to get the CLIP visual embedding $e_I \in \mathbb{R}^L$. Regarding the text part, we feed the prompt list C_p^i into CLIP text encoder $E_T(\cdot)$ to get the CLIP textual embedding $e_P^i \in \mathbb{R}^L$, where $i \in \{1, 2, \cdots, U\}$ and U denotes the number of the categories. Different from the original CLIP, we adopt a shared projection head $P_\theta(\cdot)$ on both image and text encoders, then the final prediction can be formulated as:

$$p(y = k|x) = \frac{\exp(\cos(P_\theta(e_I), P_\theta(e_P^k))/\tau)}{\sum_{k'=1}^{U} \exp(\cos(P_\theta(e_I), P_\theta(e_P^{k'}))/\tau)} \quad (1)$$

where τ is a temperature parameter learned by CLIP and $\cos(\cdot, \cdot)$ denotes cosine similarity.

3.3. LLM Knowledge Transferring

The plain CLIP feature space learns general representations, lacking task-specific knowledge [26, 49]. Therefore, we leverage the task-specific knowledge generated from LLMs to improve the CLIP feature representations. This task-specific knowledge is elicited by an instruction-based strategy. To enhance the CLIP visual feature representation for understanding facial activities, we propose projecting the original CLIP visual features into a domain-aware feature space using a projection head, rather than fine-tuning the CLIP image encoder. The key design principle of this projection head is to maintain the strong vision-language alignment inherent in CLIP while emphasizing crucial task-specific elements and filtering out irrelevant information. To achieve this, we employ a linear layer as the projection head, implemented using a simple MLP layer. The empirical results in 4.3 demonstrate that this single MLP layer outperforms more complex modules, providing an efficient and effective way to improve zero-shot performance.

As shown in the contrastive pre-training phase in Fig. 2, regarding the left part, given an unlabelled image $x \sim X$ with size $H \times W$, the frozen CLIP image encoder $E_I(\cdot)$ is first applied to get the CLIP visual embedding $e_I \in \mathbb{R}^L$. Then this CLIP visual embedding will go through a learnable projection head $P_\theta(\cdot)$, and the output can be denoted as $P_\theta(e_I) \in \mathbb{R}^{L'}$. For the right part, the inputs contain the same unlabelled facial image as the CLIP image encoder and an additional instruction text. The image x will go into the I2T Module denoted as $Q(\cdot)$ including a ViT-L14, a Q-Former, and a projector, and the output embedding can be denoted as $Q(e_I) \in \mathbb{R}^H$, where H is the vector length of the output. The instruction c_s will be fed into the Text Tokenizer $T(\cdot)$ to obtain the instruction word embedding $T(c_s)) \in \mathbb{R}^W$, where W is the length of the instruction text. Then the image embedding $Q(e_I)$ will be concatenated with instruction text embedding $T(c_s)$ as input tokens for LLM encoder $G(x)$. Then the generated feature embed-

ding $\boldsymbol{e}_G \in \mathbb{R}^{L'}$ can be denoted as:

$$\boldsymbol{e}_G = G(concat(Q(\boldsymbol{e}_I), T(c_s))) \tag{2}$$

To align the CLIP feature space to the LLM task-specific feature space, we utilize the contrastive loss to optimize the learnable projection head:

$$p(y = k|x) = \frac{\exp(\cos(P_\theta(\boldsymbol{e}_I^k), \boldsymbol{e}_G^k)/\tau)}{\sum_{k'=1}^{N} \exp(\cos(P_\theta(\boldsymbol{e}_I^{k'}), \boldsymbol{e}_G^{k'})/\tau)} \tag{3}$$

where N is the number of the mini-batch.

During the training phase, only the learnable projection head $P_\theta(\cdot)$ which is a matrix with shape $L \times L'$ will be updated, and the rest of the models are all frozen. In contrast to fine-tuning the entire model end-to-end, this approach has the benefit of retaining the ability of the previous model to extract powerful general-purpose representations but adapt it to the FER domain with very few trainable parameters.

4. Experiments

4.1. Datasets

We conduct experiments on seven popular in-the-wild facial expression benchmarks, including three static FER datasets: RAF-DB [21], AffectNet [25], and FERPlus [2], and four dynamic FER datasets: DFEW [11], FERV39k [36], MAFW [23], and AFEW [6].

SFER Datasets: **RAF-DB** [21] contains 30,000 facial images annotated with basic or compound expressions. Consistent with most previous work, only images with six basic emotions (*i.e.*, happiness, sadness, anger, surprise, disgust, and fear) and neutral are used, including 3,068 images as test data. **AffectNet** [25] contains about 450,000 images that are manually annotated with 11 expression categories. The seven expression categories denoted by AffectNet-7 contain six basic emotions and neutral, including 3,500 images as test data. The eight expression categories denoted by AffectNet-8 contain six basic emotions, contempt, neutral. **FERPlus** [6] contains 35,887 images and is an extended version of FER2013. Same as AffectNet-8, eight expression categories are used, including 3,589 images as test data.

DFER Datasets: **DFEW** [11] contains 11,697 in-the-wild video clips, and all the samples have been split into five same-size parts without overlap. Each video is annotated to one of the six basic emotions and neutral. The 5-fold cross-validation is adopted as an evaluation protocol. **FERV39k** [36] contains 38,935 in-the-wild video clips, which is currently the largest in-the-wild DFER dataset. All the video clips were assigned to one of the six basic emotions and neutral. Video clips of all scenes are randomly shuffled and split into training (80%), and testing (20%) without overlapping. **MAFW** [23] contains 10,045 in-the-wild video clips,

which is the first large-scale, multi-modal, multi-label affective database with 11 single-expression categories (six basic emotions, contempt, anxiety, helplessness, disappointment, and neutral), 32 multiple-expression categories, and emotional descriptive texts. The 5-fold cross-validation is adopted as an evaluation protocol. **AFEW** [6] contains 1,809 in-the-wild video clips, and all the samples have been split into three splits: train (773 video clips), validation (383 video clips), and test (653 video clips). Since the test split is not publicly available, we report results on the validation split.

Experiment Settings: In the contrastive learning phase, the training is conducted on CAER-S [15] training set with 45K facial images. The projection head in our method refers to a matrix with size 512×4096 when utilising ViT-B-32/B-16 or size 768×4096 when utilising ViT-L-14. The learning rate for the learnable projection head is set as 1×10^{-3}. The learnable projection head is trained for 5 epochs in an end-to-end manner. The parameters of the projection head are optimized by the SGD optimizer with a mini-batch size of 512. Our models are trained on the Tesla A100 GPU based on the open-source PyTorch platform. We employ the Flan-T5-XXL [5] with 11B parameters as our LLM. During the zero-shot prediction, for the SFER, all the images are resized to 224×224 as an input. For the DEER, all the fixed 16 frames of sequence in our experiments follow the sampling strategy in [11, 17, 45], and then resized to 224×224 as inputs. To obtain more stable and reliable results, we train models three times with different random seeds and then use the average as the final result.

Evaluation Metrics: Consistent with most previous methods [17, 33, 42, 45], the weighted average recall (WAR, i.e., accuracy) and the unweighted average recall (UAR, i.e., the accuracy per class divided by the number of classes without considering the number of instances per class) are utilized for evaluating the performance of methods.

4.2. Comparisons with the State-of-the-Art

We first compared our method with several large-language models in a zero-shot way, then compared with some supervised SFER and DFER state of the arts. Our results are obtained using the instruction "*Please play the role of a facial action describer. Objectively describe the detailed facial actions of the person in the image.*", and the prompt "a photo of a face with an expression of [class]." for zero-shot prediction.

BLIP2 [20] is a state-of-the-art model on various vision-language tasks, to test its zero-shot performance on the FER task, inspired by work [37], we take the facial images as input and prompt BLIP2 by using: "*Question: Play the role of a facial expression classification expert and select one category with the highest similarity to the input image. Here are the optional categories: 'contempt', 'fear',*

Table 1. Zero-shot FER results in comparison with other visual language models. B/32, B/16, L/14 denote ViT-B/32, ViT-B/16, and ViT-L/14, respectively. *EmoCLIP was pre-trained on MAFW [23] with facial expression clips and action unit text descriptions without categorical labels. **Bold** donates the best, <u>Underline</u> denotes the second best.

Methods	Static FER (UAR/WAR)				Dynamic FER (UAR/WAR)			
	RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
BLIP2 (L/14) [20]	43.78/44.07	32.86/32.87	28.53/28.53	<u>43.40</u>/48.14	-	-	-	-
FaRL (B/16) [49]	24.98/38.53	26.95/26.95	23.50/23.51	24.27/35.26	23.99/25.73	20.71/22.87	14.98/17.08	28.17/30.45
FLAVA [30]	14.35/38.69	14.26/14.26	12.47/12.48	12.50/28.43	14.29/20.91	14.28/18.77	9.31/3.26	14.29/16.27
VideoCLIP [40]	-	-	-	-	18.15/17.52	15.35/15.57	10.26/11.70	16.32/17.32
XCLIP (B/32) [26]	-	-	-	-	20.03/16.06	16.80/12.50	11.51/13.53	21.81/22.05
EmotionCLIP [41]	-	-	-	-	13.77/19.89	14.79/19.54	9.20/11.65	14.86/17.06
EmoCLIP* [8]	-	-	-	-	<u>36.76</u>/46.27	<u>26.73</u>/**35.30**	**25.86**/**33.49**	<u>36.13</u>/<u>39.90</u>
Exp-CLIP B/32 (Ours)	39.56/42.14	31.73/31.73	27.63/27.63	37.83/37.04	24.25/25.87	21.48/21.25	17.53/20.27	29.72/31.23
Exp-CLIP B/16 (Ours)	48.96/54.50	<u>39.98</u>/39.98	<u>34.40</u>/34.40	40.81/<u>53.02</u>	27.12/31.75	24.78/27.99	19.17/23.35	32.84/33.86
Exp-CLIP L/14 (Ours)	**58.70/65.37**	**44.27/44.27**	**38.44/38.43**	**48.28/55.42**	**40.16/47.09**	**27.45**/<u>31.07</u>	<u>23.95</u>/<u>26.98</u>	**38.72/40.42**

Table 2. Zero-shot FER results in comparison with CLIP model.

Methods	Static FER (UAR/WAR)				Dynamic FER (UAR/WAR)			
	RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
	ViT-B/32							
CLIP [28]	34.97/30.05	29.29/29.29	25.53/25.53	36.05/33.89	22.08/21.30	18.12/17.38	13.39/15.48	28.59/28.87
Exp-CLIP (Ours)	**39.56/42.14**	**31.73/31.73**	**27.63/27.63**	**37.83/37.04**	**24.25/25.87**	**21.48/21.25**	**17.53/20.27**	**29.72/31.23**
⇑	4.59/12.09	2.44/2.44	2.10/2.10	1.78/3.15	2.17/4.57	3.36/3.87	4.14/4.79	1.13/2.36
	ViT-B/16							
CLIP [28]	38.66/37.16	34.35/34.35	29.38/29.38	34.33/45.14	22.92/28.33	19.81/22.51	15.45/17.25	28.21/29.13
Exp-CLIP (Ours)	**48.96/54.50**	**39.98/39.98**	**34.40/34.40**	**40.81/53.02**	**27.12/31.75**	**24.78/27.99**	**19.17/23.35**	**32.84/33.86**
⇑	9.11/15.75	5.63/5.63	5.02/5.02	6.48/7.88	4.20/3.42	4.97/5.48	3.72/6.10	4.63/4.73
	ViT-L/14							
CLIP [28]	47.22/41.13	34.46/34.47	29.96/29.96	33.82/46.67	32.02/38.12	21.97/28.99	20.42/25.58	35.50/38.32
Exp-CLIP (Ours)	**58.70/65.37**	**44.27/44.27**	**38.44/38.43**	**48.28/55.42**	**40.16/47.09**	**27.45/31.07**	**23.95/26.98**	**38.72/40.42**
⇑	11.48/24.24	9.81/9.80	8.48/8.47	14.46/8.75	8.14/8.97	5.48/2.08	3.53/1.40	3.22/2.10

'disgust', 'surprise', 'anger', 'neutral', 'sadness', 'happiness'. Answer:''. We then calculate the WAR and UAR based on the output. As shown in Tab. 1, our results outperform the BLIP2 model on all the FER test sets. Specifically, our ViT-B-16-based zero-shot results can beat BLIP2 in most SFER test sets. Furthermore, we also compare our results with several other visual-language models. Video-CLIP [40], XCLIP [26], and EmotionCLIP [41] are proposed for video analysis, video recognition, and video emotion recognition, respectively, in which VideoCLIP was pre-trained on video clips and text transcriptions, XCLIP was pre-trained on action recognition dataset with action labels, and EmotionCLIP was pre-trained on the videos and conversations. These models all learn spatial-temporal features during the training. However, our method with the bare static model can still outperform them on four DFER test sets. Moreover, we also compared our method with CLIP betterment: FaRL and FLAVA, in which the FaRL model was pre-trained on LAION-FACE specializing in the face analysis task and FLAVA improves CLIP zero-shot performance by introducing both unimodal and multimodal architecture. The results in Tab. 1 demonstrate our results are superior to these two methods on all of the FER test sets. Compared with the EmoCLIP [8], which was trained on the MAFW clips and facial action descriptions, our method out-

performs it on DFEW, FERV39k, and AFEW test sets in five metrics without explicit text descriptions, indicating the effectiveness of the proposed method.

4.3. Ablation Analysis

Evaluation of the Projection Head Effectiveness: Our method is built on the CLIP model [28], to evaluate the effectiveness of the learned projection head, we compare our method to the CLIP model from two views: various ViT models and various zero-shot prompts. Our method uses the instruction "*Please play the role of a facial action describer. Objectively describe the detailed facial actions of the person in the image.*". As shown in Tab. 2, under different ViT models, our method outperforms the plain CLIP model on all the test sets. With ViT-B-32, our method improves the average UAR by 2.71% and average WAR by 4.42% over eight test sets. With ViT-B-16, our method improves the average UAR by 5.47% and average WAR by 6.75% over eight test sets. With ViT-L-14, our method improves the average UAR by 8.07% and average WAR by 8.23% over eight test sets. The comparison results on varied prompts are shown in the *Appendix Sec. A*.

Evaluation of Different Instructions: To verify the effectiveness of the instruction adopted in our method, we conducted experiments on two types of instructions: task-

Table 3. Zero-shot FER results of our method with different instructions.

	Static FER (UAR/WAR)				Dynamic FER (UAR/WAR)			
	RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
ViT-B/32								
Task-unrelated	33.70/34.77	30.08/30.09	26.10/26.11	33.03/35.43	21.57/20.74	19.02/17.69	14.71/17.51	29.42/30.88
Task-related	**38.86/43.98**	**31.38/31.38**	**27.75/27.75**	**39.44/37.01**	**24.94/28.52**	**21.84/23.32**	**17.79/20.85**	**30.73/32.55**
⇑	5.16/9.21	1.30/1.30	1.64/1.64	6.41/1.58	3.37/7.78	2.83/5.62	3.08/3.34	1.31/1.66
ViT-B/16								
Task-unrelated	41.51/40.97	36.27/36.28	31.29/31.30	38.70/48.27	24.69/28.20	21.67/23.86	17.19/19.69	30.20/31.06
Task-related	**48.84/54.68**	**39.46/39.46**	**34.04/34.04**	**40.19/53.82**	**26.40/31.64**	**25.11/28.57**	**18.98/23.64**	**31.50/32.63**
⇑	7.33/13.71	3.19/3.18	2.74/2.74	1.49/5.55	1.71/3.44	3.44/4.71	1.79/3.96	1.30/1.58
ViT-L/14								
Task-unrelated	57.32/61.26	43.14/43.14	37.20/37.20	46.79/54.16	37.94/43.40	25.86/30.23	23.54/26.47	38.08/40.33
Task-related	**58.08/65.73**	**44.41/44.41**	**38.24/38.24**	**47.70/55.56**	**40.59/47.19**	**27.90/31.51**	**24.16/27.27**	**39.81/41.21**
⇑	0.76/4.47	1.27/1.27	1.04/1.04	0.90/1.40	2.65/3.79	2.04/1.27	0.62/0.81	1.73/0.87

Table 4. Zero-shot FER results of our method with different projection heads. 2-MLP and 3-MLP denote 2-layer and 3-layer multilayer perceptron, respectively. TE denotes Transformer encoder. LM denotes the learnable linear matrix.

Projection Heads	Static FER (UAR/WAR)				Dynamic FER (UAR/WAR)			
	RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
2-MLP	45.05/59.49	35.46/35.47	29.73/29.73	39.66/46.61	31.61/37.35	25.21/26.91	17.62/18.92	29.92/31.76
3-MLP	42.28/61.05	34.95/34.95	30.18/30.18	36.60/51.04	29.67/36.56	25.30/27.25	17.81/19.47	31.91/33.07
TE	34.22/53.13	28.89/28.89	23.88/23.88	33.39/45.20	29.33/37.36	26.06/**32.31**	17.94/23.20	27.16/28.08
LM	**58.70/65.37**	**44.27/44.27**	**38.44/38.43**	**48.28/55.42**	**40.16/47.09**	**27.45**/31.07	**23.95/26.98**	**38.72/40.42**

related and task-unrelated. Regarding the task-unrelated instructions, we consider both empty and random text as input and take the average of them as the final results. For the task-related instruction, we provide results with three instructions and their average. As shown in Tab. 3, the task-related instructions are superior to task-unrelated instructions on both SFER and DFER, in which there are 3.14% average UAR improvements and 4.02% average WAR improvements over eight test sets with ViT-B-32, 2.87% average UAR improvements and 4.86% average WAR improvements over eight test sets with ViT-B-16, and 1.38% average UAR improvements and 1.87% average WAR improvements over eight test sets with ViT-L-14. We believe that task-related instructions facilitate LLMs to generate semantic features with task-aware ability, providing better objects for projection head optimisation. The detailed results on task-related and task-unrelated instructions are shown in the *Appendix Sec. B*.

Evaluation of Different Projection Heads: In our method, to align the original CLIP feature representations into task-specific LLM knowledge space, we propose to utilize a projection head as a linear mapping to fulfil this. To investigate if the improvement of the complexity of the projectors can further improve the alignment, we also conduct experiments with 2-layer MLP, 3-layer MLP, and Transformer Encoders. As shown in Tab. 4, even with a simply learnable matrix and fewer parameters, our strategy is more powerful and efficient. Compared to the Transformer Encoders, except for the WAR results on FERV39k, ours

shows much better results than more complex Transformer Encoders. We believe that more complex non-linear models will entail overfitting to the LLMs feature space, performing worse in downstream zero-shot FER tasks. The linear projection head adopted in our method is more suitable and interpretable for mapping the original CLIP feature space to a task-specific one, maintaining the capacity of the existing model to capture robust, versatile representations while tailoring it to the FER domain with minimal trainable parameters.

Evaluation of Varied Image and Text Encoders: To investigate if the LLM can do better than our method on zero-shot FER tasks, we conducted experiments using the LLM as either an image encoder (with I2T module), text encoder, or both. As shown in Tab. 5, adopting the LLM as either an image encoder or text encoder results in a worse zero-shot preference. Even when using the LLM as both an image and text encoder, the results remain undesirable and significantly lower than the plain CLIP model and our method. We believe the LLM excels in the semantic perception of images but performs poorly in recognition tasks. By contrast, based on the CLIP model, our method is more efficient and accurate in zero-shot FER tasks.

4.4. Visualizations

To prove the benefit of the enhanced CLIP visual features, we utilize t-SNE [31] to illustrate the feature distribution of the original CLIP feature representations and projected feature representations. Different from the super-

Table 5. Zero-shot FER results of varied image and text encoders. PH denotes the projection head trained in our method.

| Image Encoder	Text Encoder	Static FER				Dynamic FER			
		RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
CLIP	CLIP	47.22/41.13	34.46/34.47	29.96/29.96	33.82/46.67	32.02/38.12	21.97/28.99	20.42/25.58	35.50/38.32
CLIP+PH	LLM	24.03/38.62	24.03/24.04	19.03/19.03	21.86/37.07	21.31/15.31	18.02/18.72	10.10/5.76	19.33/19.16
LLM	CLIP+PH	25.81/29.11	19.36/19.35	17.62/17.60	22.59/25.34	13.93/3.77	15.95/10.69	9.66/8.90	19.11/17.06
LLM	LLM	19.46/41.23	17.00/17.00	14.88/14.88	18.01/32.23	16.97/22.73	15.18/19.28	10.25/14.66	17.94/19.42
CLIP+PH	CLIP+PH	**58.70/65.37**	**44.27/44.27**	**38.44/38.43**	**48.28/55.42**	**40.16/47.09**	**27.45/31.07**	**23.95/26.98**	**38.72/40.42**
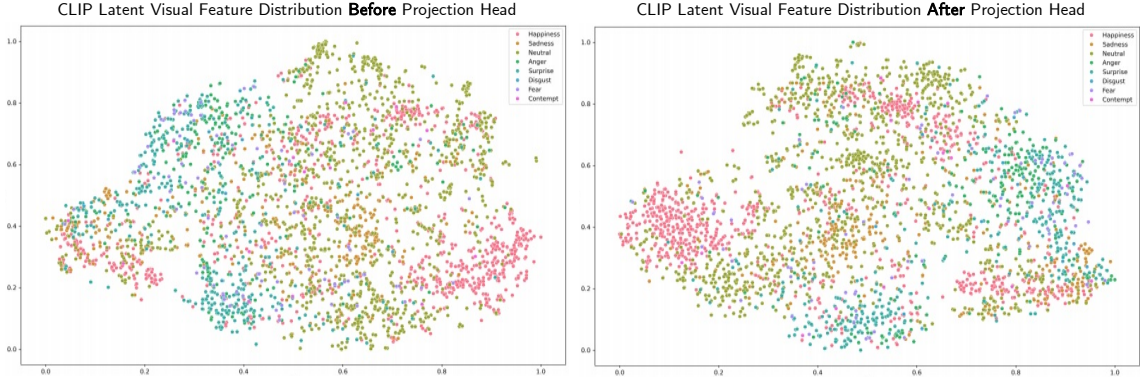
Figure 3. CLIP latent visual feature distribution on FERPlus dataset. Both results are based on ViT-L-14.

vised methods learned with annotations which can gather samples of the same category compactly, zero-shot visual feature representations are learned by aligning the cross-modality representations, presenting a weak performance in gathering samples compactly. Despite this, from the visual feature distributions in Fig. 3, we can see the samples of the same category can be better gathered after the projection, especially for emotions of happiness and disgust. This demonstrates the model learns more discriminative latent visual features after the projection head, and can also explain why our method obtains superior results to the original plain CLIP model. In addition, we calculate the normalized feature variance of each emotion on FERPlus in the *Appendix*. The projected features demonstrate a lower variance than the CLIP feature space, proving the projected features are more concentrated. We also provide the confusion matrix comparison in the *Appendix*, which shows that our method captures more nuanced facial expression features.

5. Limitations

The proposed Exp-CLIP improves zero-shot FER results by incorporating domain knowledge from LLMs. However, our method still faces several limitations due to the following assumptions: 1) LLMs can adequately capture and describe the subtle nuances of facial expressions, and 2) LLMs can generate outputs that are well-grounded in the input data, minimizing hallucinations. To address 1), we utilize the Flan-T5-XXL model with 11B parameters, which provides more reliable and detailed knowledge. For 2), we use feature representations extracted after the LLM encoder, bypassing the decoder to mitigate the influence of hallucinations. In addition, our method captures more nuanced facial

expression features derived from LLMs, e.g. performing better in distinguishing between neutral and other emotions. However, our zero-shot method still makes some incorrect predictions, particularly when distinguishing between similar emotions like Anger and Disgust, as well as Surprise and Fear, especially on hard or ambiguous samples.

6. Conclusion

To improve the zero-shot FER performance of the vision-language model, this paper proposes an Exp-CLIP boosted by transferring the task knowledge from large language models (LLMs). To achieve this knowledge transfer, a projection head is employed to align the projected visual representations with task-specific semantic meanings derived from the LLM encoder, and the text instruction-based strategy is employed to customize the LLM knowledge. As a result, the projected visual representations learn task-aware features, facilitating the downstream zero-shot FER task. The training was conducted with one learnable matrix on unlabelled facial data, which is efficient and labour-saving. Exhaustive experiments demonstrate the effectiveness of the proposed strategy. The comparison to several vision-language models shows the superior performance of the proposed method in the FER task. Although there is an obvious results gap between the proposed zero-shot methods and supervised methods, this paper shows a promising direction for more generalised and general FER models.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736, 2022. 2, 3

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, pages 279–283, 2016. 5

[3] Mina Bishay, Petar Palasek, Stefan Priebe, and Ioannis Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE TAFFC*, 12(4):949–961, 2019. 1

[4] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, Lingbo Liu, and Liang Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE TPAMI*, 44(12):9887–9903, 2021. 1, 3

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 25(70):1–53, 2024. 5

[6] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 9(3):34–41, 2012. 5

[7] 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. In *ICLR*, 2020. 1

[8] Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition. In *FG*, 2024. 2, 3, 6

[9] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 2, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[11] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *ACM MM*, pages 2881–2889, 2020. 1, 5

[12] Jinu Lilly Joseph and Santhosh P Mathew. Facial expression recognition for the blind using deep learning. In *IC2PCT*, pages 1–5, 2021. 1

[13] Ryosuke Kawamura, Hideaki Hayashi, Noriko Takemura, and Hajime Nagahara. Midas: Mixing ambiguous data with soft labels for dynamic facial expression recognition. In *WACV*, pages 6552–6562, 2024. 3

[14] Isack Lee, Eungi Lee, and Seok Bong Yoo. Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In *ICCV*, pages 1536–1546, 2023. 3

[15] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019. 5

[16] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 36, 2024. 2, 3

[17] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *AAAI*, pages 67–75, 2023. 1, 3, 5

[18] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *ICME*, pages 1–6, 2024. 3

[19] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022. 3

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 3, 4, 5, 6

[21] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE TIP*, 28(1):356–370, 2018. 1, 5

[22] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE TAFFC*, 13(3):1195–1215, 2020. 1

[23] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *ACM MM*, pages 24–32, 2022. 5, 6

[24] Ludo Maat and Maja Pantic. Gaze-x: Adaptive, affective, multimodal interface for single-user office scenarios. In *Artifical Intelligence for Human Computing*, pages 251–271. Springer, 2007. 1

[25] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE TAFFC*, 10(1):18–31, 2017. 1, 5

[26] 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, 2022. 3, 4, 6

[27] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE TPAMI*, 2022. 2

[28] 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 *ICML*, pages 8748–8763, 2021. 2, 3, 6

[29] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 2, 3

[30] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pages 15638–15650, 2022. 2, 3, 6

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 7

[32] Can Wang, Shangfei Wang, and Guang Liang. Identity-and pose-robust facial expression recognition through adversarial feature learning. In *ACM MM*, pages 238–246, 2019. 3

[33] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *CVPR*, pages 17958–17968, 2023. 1, 3, 5

[34] 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. 1, 3

[35] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE TIP*, 29:4057–4069, 2020. 3

[36] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos. In *CVPR*, pages 20922–20931, 2022. 5

[37] Wenhao Wu, Huanjin Yao, Mengxi Zhang, Yuxin Song, Wanli Ouyang, and Jingdong Wang. Gpt4vis: What can gpt-4 do for zero-shot visual recognition? *arXiv preprint arXiv:2311.15732*, 2023. 2, 5

[38] Zhiyu Wu and Jinshi Cui. La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In *ICCV*, pages 20698–20707, 2023. 1, 3

[39] Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. Vllms provide better context for emotion understanding through common sense reasoning. *arXiv preprint arXiv:2404.07078*, 2024. 3

[40] 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. In *EMNLP*, pages 6787–6800, 2021. 3, 6

[41] Sitao Zhang, Yimu Pan, and James Z Wang. Learning emotion representations from verbal and nonverbal communication. In *CVPR*, pages 18993–19004, 2023. 2, 3, 6

[42] Xiang Zhang, Taoyue Wang, Xiaotian Li, Huiyuan Yang, and Lijun Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *ICCV*, pages 20751–20762, 2023. 1, 3, 5

[43] Yuhang Zhang, Yaqi Li, Xuannan Liu, Weihong Deng, et al. Leave no stone unturned: Mine extra knowledge for imbalanced facial expression recognition. *NeurIPS*, 36, 2024. 3

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

[45] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *ACM MM*, pages 1553–1561, 2021. 3, 5

[46] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE TIP*, 30:6544–6556, 2021. 3

[47] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *AAAI*, pages 3510–3519, 2021. 3

[48] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. In *BMVC*, pages 1–14, 2023. 3

[49] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pages 18697–18709, 2022. 2, 3, 4, 6

Enhancing Zero-Shot Facial Expression Recognition by LLM Knowledge Transfer (Appendix)

A. Comparison with CLIP on Varied Prompts

We compare our method with CLIP using different prompts, in which both single prompt and prompt ensembles are considered. Specifically, for a single prompt, we utilize "{class name}.", "an expression of {class name}.", and "a photo of a face with an expression of {class name}.", respectively. For prompt ensembles, we utilize 5-prompt and 10-prompt configurations. All the prompt templates are outlined below:

- "{class name}.",
- "an expression of {class name}.",
- "a photo of a face exuding {class name}.",
- "a photo radiating {class name} in a person.",
- "a photo of a person embodying {class name}.",
- "a good photo capturing someone's {class name}.",
- "a photo showing someone immersed in {class name}.",
- "a photo capturing {class name} within an individual.",
- "a clean photo showcasing a person's {class name}.",
- "a photo of a face with an expression of {class name}.",

Table A. Zero-shot FER results with different prompts in comparison with CLIP model. Both single prompts and prompt ensembles are investigated. We employ three types for single prompts, labelled as P-1, P-2, and P-3 respectively. For prompt ensembles, we utilize 5-prompt and 10-prompt configurations, labelled as P-4 and P-5.

Models	Prompts	Static FER (UAR/WAR)				Dynamic FER (UAR/WAR)			
		RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
		ViT-B/32							
CLIP	P-1	28.62/23.57	26.09/26.09	21.86/21.86	**30.29**/19.76	21.07/**26.94**	19.34/18.82	14.99/16.35	25.74/26.51
	P-2	40.04/35.76	32.84/32.84	27.86/27.86	37.40/27.03	23.22/19.85	20.71/16.99	18.26/18.85	30.97/29.66
	P-3	34.97/30.05	29.29/29.29	25.53/25.53	36.05/33.89	22.08/21.30	18.12/17.38	13.39/15.48	28.59/28.87
	P-4	35.58/25.36	29.69/29.69	25.53/25.53	34.30/25.28	20.58/18.36	18.56/15.83	15.59/17.16	28.87/28.87
	P-5	37.73/30.28	31.61/31.61	27.31/27.31	**39.85**/29.45	23.75/22.48	20.90/17.84	17.51/19.73	27.26/27.30
Ours	P-1	**35.88/41.59**	29.89/29.90	**25.93/25.93**	28.81/**38.38**	21.15/26.73	**20.03/19.76**	**15.77/19.84**	**29.60/31.32**
	P-2	**42.12/49.12**	**33.61/33.61**	**29.31/29.31**	39.82/**45.55**	**26.38/25.90**	**22.53/20.16**	**18.99/22.14**	**31.98/31.93**
	P-3	**39.56/42.14**	**31.73/31.73**	**27.63/27.63**	**37.83/37.04**	**24.25/25.87**	**21.48/21.25**	**17.53/20.27**	**29.72/31.23**
	P-4	**41.04/45.47**	**32.56/32.56**	**28.43/28.43**	36.24/**42.93**	**23.57/24.00**	**21.19/19.98**	**18.14/21.74**	**33.69/35.35**
	P-5	**41.13/47.63**	**33.62/33.63**	**29.29/29.29**	38.38/**43.89**	**24.48/24.85**	**21.98/20.46**	**18.40/22.97**	**32.91/34.65**
		ViT-B/16							
CLIP	P-1	32.78/29.27	30.41/30.41	26.26/26.26	31.98/27.16	23.87/30.44	19.23/21.13	17.56/**22.61**	27.85/30.71
	P-2	53.37/50.52	38.61/38.61	34.03/34.03	**45.10**/39.02	**31.48**/26.45	24.80/22.26	20.58/23.22	35.44/34.91
	P-3	39.85/38.75	34.35/34.35	29.38/29.38	34.33/45.14	22.92/28.33	19.81/22.51	15.45/17.25	28.21/29.13
	P-4	42.69/34.71	35.83/35.84	30.48/30.48	39.25/35.64	25.71/28.85	20.19/20.31	17.60/20.52	32.56/32.81
	P-5	41.74/37.42	34.49/34.50	29.70/29.71	34.42/36.95	27.89/29.38	23.44/23.37	19.98/22.24	33.12/34.12
Ours	P-1	**41.80/48.58**	**35.43/35.44**	**30.52/30.52**	**38.72/45.29**	**24.65/31.10**	**24.42/27.31**	**19.43**/22.37	**37.52/39.02**
	P-2	**55.88/60.21**	**39.75/39.76**	**34.44/34.44**	42.42/**53.25**	30.78/**31.52**	**26.72/28.26**	**21.74/26.31**	**35.98/35.52**
	P-3	**48.96/54.50**	**39.98/39.98**	**34.40/34.40**	**40.81/53.02**	**27.12/31.75**	**24.78/27.99**	**19.17/23.35**	**32.84/33.86**
	P-4	**52.42/57.26**	**41.10/41.10**	**35.05/35.05**	**43.30/52.94**	**28.52/32.55**	**26.19/29.03**	**21.35/26.84**	**36.85/37.71**
	P-5	**50.82/57.66**	**40.76/40.76**	**34.88/34.88**	**43.05/54.30**	**30.79/34.59**	**27.53/30.05**	**22.17/27.40**	**37.02/38.50**
		ViT-L/14							
CLIP	P-1	48.20/38.14	36.24/36.24	29.76/29.76	46.59/32.99	**27.74**/24.43	19.75/15.59	19.77/16.79	29.93/28.87
	P-2	49.37/39.34	37.35/37.35	33.86/33.86	42.06/26.49	32.68/20.83	20.95/14.54	20.02/16.46	33.82/32.28
	P-3	47.22/41.13	34.46/34.47	29.96/29.96	33.82/46.67	32.02/38.12	21.97/28.99	20.42/25.58	35.50/38.32
	P-4	49.79/36.34	39.41/39.41	33.53/33.53	43.65/30.22	35.39/29.31	21.97/18.85	21.84/22.86	31.61/31.50
	P-5	52.21/41.36	39.72/39.73	32.98/32.98	43.28/31.40	35.85/35.40	23.54/24.10	23.20/26.99	34.08/34.65
Ours	P-1	**49.08/53.13**	**38.05/38.06**	**32.93/32.93**	**48.55/49.76**	26.76/**31.65**	**22.31/25.27**	**20.69/23.92**	**32.45/33.95**
	P-2	**56.17/60.77**	**42.68/42.68**	**37.46/37.46**	**51.54/55.15**	**39.04/38.48**	**26.45/25.10**	**24.81/28.47**	**38.50/38.58**
	P-3	**58.70/65.37**	**44.27/44.27**	**38.44/38.43**	**48.28/55.42**	**40.16/47.09**	**27.45/31.07**	**23.95/26.98**	**38.72/40.42**
	P-4	**56.15/60.23**	**43.48/43.48**	**38.03/38.03**	**50.97/53.30**	**38.15/39.75**	**27.14/27.57**	**25.14/27.72**	**40.17/41.29**
	P-5	**56.52/61.41**	**43.68/43.68**	**38.11/38.11**	**49.93/53.37**	**39.02/43.57**	**28.70/31.70**	**25.94/30.99**	**40.31/42.00**

Table B. Zero-shot FER results of our method with different instructions. Both task-related and task-unrelated instructions are investigated. All models use the prompt "a photo of a face with an expression of [class]." for zero-shot prediction.

	Static FER (UAR/WAR)				Dynamic FER (UAR/WAR)			
	RAF-DB	AffectNet-7	AffectNet-8	FERPlus	DFEW	FERV39k	MAFW	AFEW
ViT-B/32								
Task Unrelated	33.53/43.42 30.49/24.19 37.09/36.70	31.24/31.24 27.77/27.78 31.23/31.24	27.18/27.18 24.40/24.41 26.73/26.73	30.43/36.69 30.54/33.44 38.12/36.15	22.41/23.60 20.03/16.19 22.26/22.43	20.28/19.59 18.38/15.42 18.39/18.07	16.10/20.41 13.21/14.83 14.82/17.29	28.28/29.66 29.33/29.66 30.65/33.33
Mean	33.70/34.77	30.08/30.09	26.10/26.11	33.03/35.43	21.57/20.74	19.02/17.69	14.71/17.51	29.42/30.88
Variance	7.28/63.49	2.67/2.66	1.48/1.47	12.96/2.02	1.18/10.58	0.80/2.97	1.40/5.21	0.94/2.99
Task Related	39.56/42.14 38.63/45.11 38.39/44.69	31.73/31.73 31.06/31.07 31.35/31.35	27.63/27.63 27.58/27.58 28.03/28.03	37.83/37.04 39.46/37.17 41.03/36.82	24.25/25.87 25.18/29.48 25.39/30.22	21.48/21.25 21.94/24.09 22.11/24.61	17.53/20.27 17.96/21.08 17.89/21.21	29.72/31.23 30.87/32.81 31.61/33.60
Mean	38.86/43.98	31.38/31.38	27.75/27.75	39.44/37.01	24.94/28.52	21.84/23.32	17.79/20.85	30.73/32.55
Variance	0.26/1.73	0.08/0.07	0.04/0.04	1.70/0.02	0.25/3.61	0.07/2.18	0.04/0.17	0.61/0.97
⇑ (mean)	**5.16/9.21**	**1.30/1.30**	**1.64/1.64**	**6.41/1.58**	**3.37/7.78**	**2.83/5.62**	**3.08/3.34**	**1.31/1.66**
ViT-B/16								
Task Unrelated	41.68/41.53 37.23/34.75 45.61/46.64	35.01/35.01 34.57/34.58 39.23/39.24	30.55/30.56 29.73/29.73 33.60/33.61	39.36/51.16 36.25/43.13 40.49/50.53	23.48/27.22 24.25/28.85 26.35/28.54	22.07/24.20 20.61/23.44 22.34/23.95	16.30/19.67 17.13/19.68 18.15/19.71	30.65/31.23 29.09/30.18 30.87/31.76
Mean	41.51/40.97	36.27/36.28	31.29/31.30	38.70/48.27	24.69/28.20	21.67/23.86	17.19/19.69	30.20/31.06
Variance	11.72/23.72	4.41/4.42	2.77/2.78	3.21/13.29	1.47/0.50	0.58/0.10	0.57/0.00	0.63/0.43
Task Related	48.96/54.50 48.61/54.14 48.94/55.41	39.98/39.98 39.12/39.13 39.27/39.27	34.40/34.40 33.63/33.63 34.08/34.08	40.81/53.02 39.85/53.94 39.91/54.51	27.12/31.75 25.82/31.21 26.27/31.97	24.78/27.99 25.28/28.69 25.28/29.03	19.17/23.35 18.88/23.62 18.90/23.96	32.84/33.86 30.95/32.02 30.71/32.02
Mean	48.84/54.68	39.46/39.46	34.04/34.04	40.19/53.82	26.40/31.64	25.11/28.57	18.98/23.64	31.50/32.63
Variance	0.03/0.29	0.14/0.14	0.10/0.10	0.19/0.38	0.29/0.10	0.06/0.19	0.02/0.06	0.91/0.75
⇑ (mean)	**7.33/13.71**	**3.19/3.18**	**2.74/2.74**	**1.49/5.55**	**1.71/3.44**	**3.44/4.71**	**1.79/3.96**	**1.30/1.58**
ViT-L/14								
Task Unrelated	55.83/62.32 58.02/62.39 58.12/59.06	41.41/41.41 44.78/44.78 43.24/43.24	35.76/35.76 38.36/38.36 37.49/37.48	44.86/54.72 48.17/54.42 47.35/53.33	36.39/43.82 38.80/44.34 38.63/42.03	26.29/31.92 26.24/31.11 25.05/27.67	23.94/26.94 23.34/27.02 23.35/25.44	37.89/40.16 37.92/40.94 38.44/39.90
Mean	57.32/61.26	43.14/43.14	37.20/37.20	46.79/54.16	37.94/43.40	25.86/30.23	23.54/26.47	38.08/40.33
Variance	1.12/2.41	1.90/1.90	1.17/1.17	1.98/0.36	1.21/0.98	0.33/3.39	0.08/0.53	0.06/0.20
Task Related	58.70/65.37 58.35/66.07 57.19/65.74	44.27/44.27 44.52/44.53 44.44/44.44	38.44/38.43 38.18/38.18 38.11/38.11	48.28/55.42 46.93/55.75 47.88/55.50	40.16/47.09 40.73/47.19 40.88/47.28	27.45/31.07 28.03/31.62 28.23/31.83	23.95/26.98 24.27/27.47 24.26/27.37	38.72/40.42 40.00/41.21 40.72/41.99
Mean	58.08/65.73	44.41/44.41	38.24/38.24	47.70/55.56	40.59/47.19	27.90/31.51	24.16/27.27	39.81/41.21
Variance	0.42/0.08	0.01/0.01	0.02/0.02	0.32/0.02	0.10/0.01	0.11/0.10	0.02/0.04	0.69/0.41
⇑ (mean)	**0.76/4.47**	**1.27/1.27**	**1.04/1.04**	**0.90/1.40**	**2.65/3.79**	**2.04/1.27**	**0.62/0.81**	**1.73/0.87**

For each prompt, we compare our method with the CLIP model one by one. As shown in Tab. A, except for seven values slightly lower than the CLIP model, our method outperforms the CLIP model on all the rest FER test sets, demonstrating the effectiveness of the proposed method. It should be noted that, compared with the CLIP model, only one extra projection matrix is added on top of the two en-coders in our method and this projection matrix was trained in an unsupervised manner.

B. Evaluation of Different Instructions

To verify the effectiveness of the instruction adopted in our method, we conducted experiments on two types of in-structions: task-related and task-unrelated. Regarding the
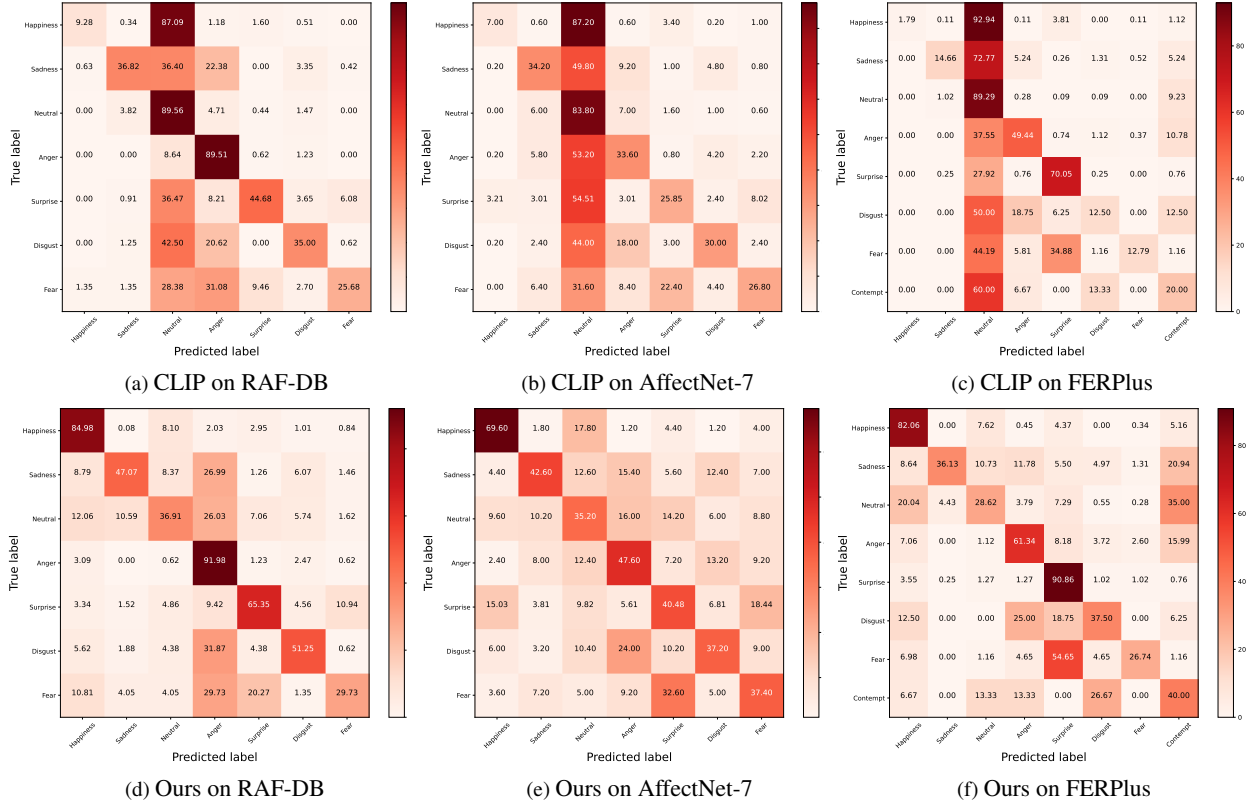
12

Figure A. Confusion matrix compared with CLIP model on static FER datasets.

task-unrelated instructions, we consider both empty and random text as input and take the average of them as the final results. For the task-related instruction, we provide results with three instructions and their average, including "*Please play the role of a facial action describer. Objectively describe the detailed facial actions of the person in the image.*", "*Please play the role of a facial expression recognition expert. Describe the facial expression of the person in the image.*", and "*Please describe the detailed facial actions of the person in the image.*".

As shown in Tab. B, the task-related instructions are superior to task-unrelated instructions on both SFER and DFER, in which there are 3.14% average UAR improvements and 4.02% average WAR improvements over eight test sets with ViT-B-32, 2.87% average UAR improvements and 4.86% average WAR improvements over eight test sets with ViT-B-16, and 1.38% average UAR improvements and 1.87% average WAR improvements over eight test sets with ViT-L-14. In addition, the three task-related instructions show a small variance in downstream zero-shot prediction. We believe that task-related instructions facilitate LLMs to generate semantic features with task-aware ability, providing better objects for projection head optimisation.

C. Confusion Matrix Comparison with CLIP

To offer a more extensive comparison with the CLIP model, we include the learned confusion matrices of both our method and the CLIP model, displayed in Fig. A and Fig. B. From the confusion matrix on both static FER and dynamic FER datasets, we can see that our method learned better representations for each expression category, resulting in a more balanced accuracy across all categories. Specifically, our method can better distinguish between neutral expressions and other expressions. This suggests that our method captures more nuanced facial expression features derived from LLMs.

D. Normalized Feature Variance

The normalized feature variance of each emotion on FERPlus is shown in Fig. C. The projected features demonstrate a lower variance than the CLIP feature space, proving the projected features are more concentrated.

E. Evaluation of Vision-to-Text-Description Understanding Ability

To verify the vision-to-text-description ability of our method, we conducted experiments using facial videos and their corresponding facial action descriptions. Specifically,

(a) CLIP on DFEW   (b) CLIP on FERV39k   (c) CLIP on MAFW

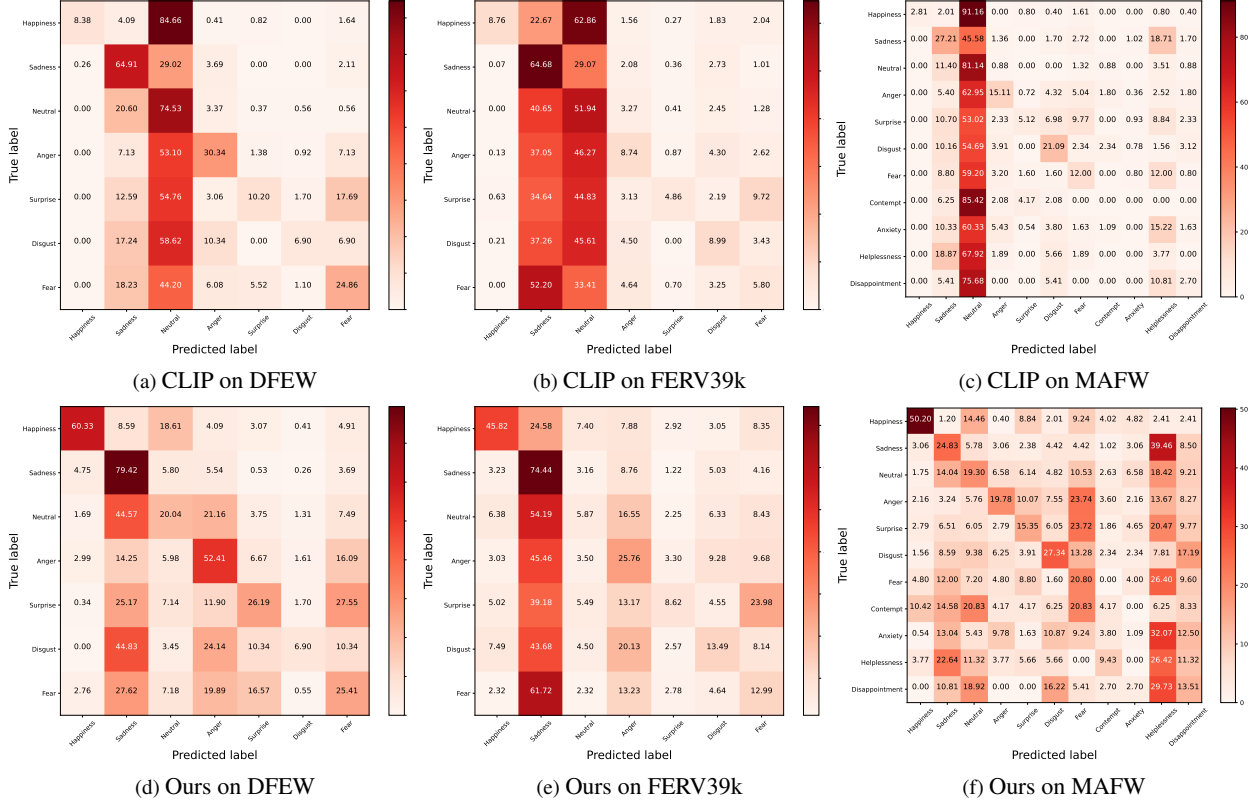(d) Ours on DFEW   (e) Ours on FERV39k   (f) Ours on MAFW

Figure B. Confusion matrix compared with CLIP model on dynamic FER datasets.
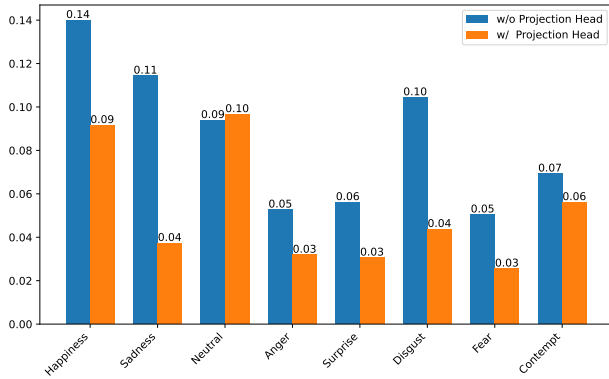
Figure C. Normalized CLIP latent visual feature variance of each emotion on FERPlus.

Figure D. Comparison of the video-to-text retrieval performance on the subset of the MAFW.

we performed a video-text retrieval task on a subset of the MAFW dataset. This subset comprises 8,034 sample-level text descriptions, each detailing specific facial actions. We utilized Precision@k as the metric for video-to-text retrieval and compared the performance of our method against several other vision-language models. As illustrated in Fig. D, our Exp-CLIP model outperforms the other methods, which means our model retrieves more relevant text descriptions
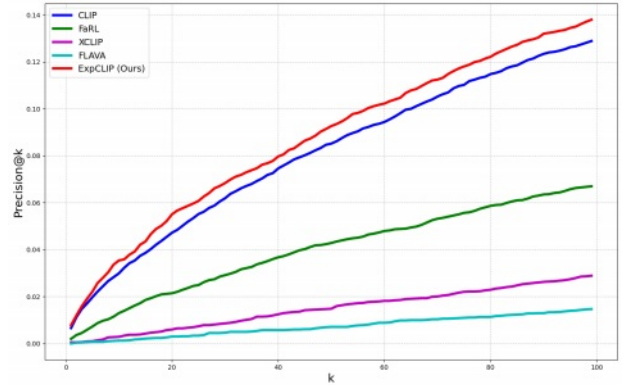
for a given facial video compared to other models and excels in understanding and retrieving complex and detailed facial action descriptions.

F. More Sample-Level Predictions and Analysis

We provide more sample-level predictions in the Fig. E. From the probability predictions on top images, we observe that our Exp-CLIP model correctly recognizes facial

Correct Predictions

Surprise

Surprise

Anger

Anger

Incorrect Predictions (Fail to Distinguish Anger & Disgust and Surprise & Fear on Hard or Ambiguous Samples)

Anger
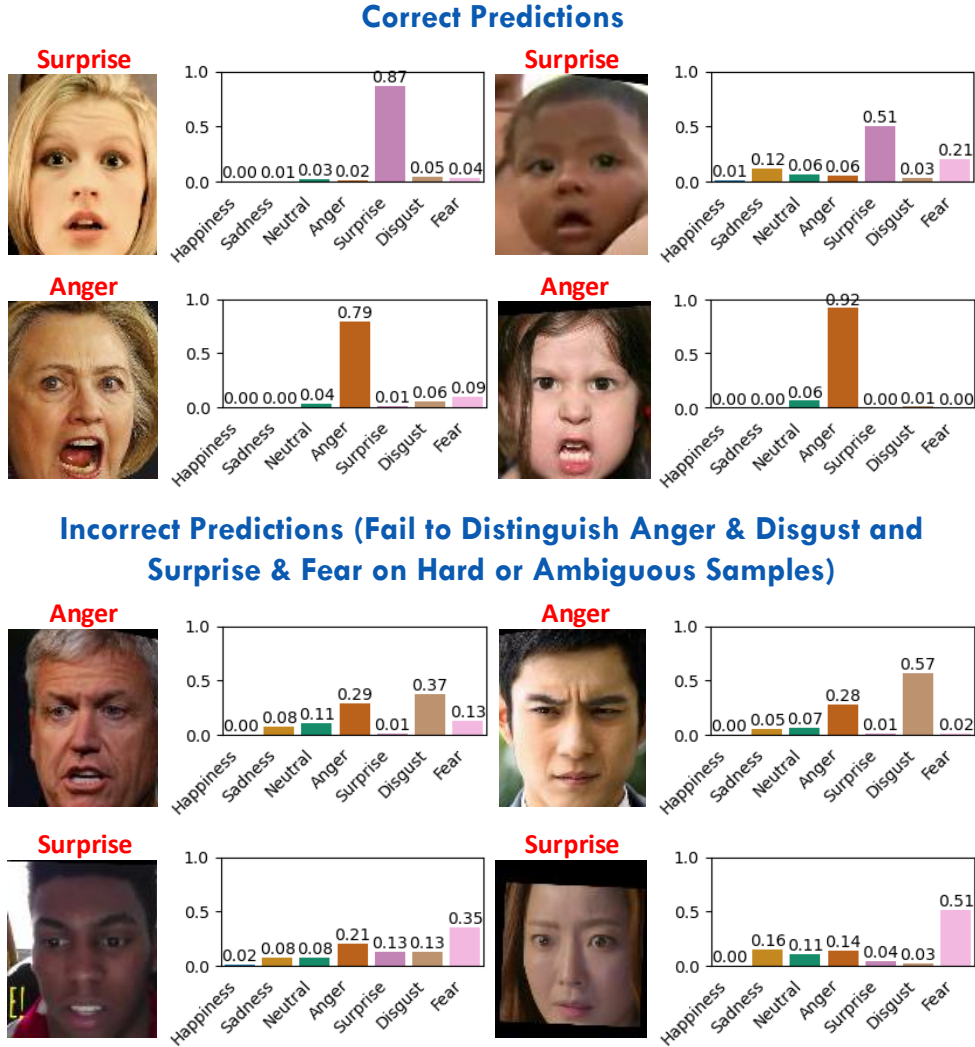

Anger

Surprise

Surprise

Figure E. Zero-shot facial expression predictions of the proposed Exp-CLIP.

expressions with high confidence. However, our zero-shot method still makes some incorrect predictions, particularly when distinguishing between similar emotions like Anger and Disgust, as well as Surprise and Fear, especially on hard or ambiguous samples. These misclassifications likely arise from the subtle differences in facial expressions between these emotions, which can be difficult even for supervised models. Further refinement of our model or additional fine-tuning on few-shot labelled data may help mitigate these specific shortcomings.