Facial Affective Behavior Analysis with Instruction Tuning

Yifan Li[1], Anh Dao[1], Wentao Bao[1], Zhen Tan[2], Tianlong Chen[3,4,5], Huan Liu[2], and Yu Kong[1]

[1] Michigan State University, {liyifa11, anhdao, baowenta, yukong}@msu.edu
[2] Arizona State University, {ztan36, huanliu}@asu.edu
[3] University of North Carolina at Chapel Hill
[4] Massachusetts Institute of Technology, tianlong@mit.edu
[5] Harvard University

Abstract. Facial affective behavior analysis (FABA) is crucial for understanding human mental states from images. However, traditional approaches primarily deploy models to discriminate among discrete emotion categories, and lack the fine granularity and reasoning capability for complex facial behaviors. The advent of Multi-modal Large Language Models (MLLMs) has been proven successful in general visual understanding tasks. However, directly harnessing MLLMs for FABA is challenging due to the scarcity of datasets and benchmarks, neglecting facial prior knowledge, and low training efficiency. To address these challenges, we introduce (***i***) an instruction-following dataset for two FABA tasks, *i.e.*, facial emotion and action unit recognition, (***ii***) a benchmark *FABA-Bench* with a new metric considering both recognition and generation ability, and (***iii***) a new MLLM *EmoLA* as a strong baseline to the community. Our initiative on the dataset and benchmarks reveal the nature and rationale of facial affective behaviors, *i.e.*, fine-grained facial movement, interpretability, and reasoning. Moreover, to build an effective and efficient FABA MLLM, we introduce a facial prior expert module with face structure knowledge and a low-rank adaptation module into pre-trained MLLM. We conduct extensive experiments on FABA-Bench and four commonly-used FABA datasets. The results demonstrate that the proposed facial prior expert can boost the performance and EmoLA achieves the best results on our FABA-Bench. On commonly-used FABA datasets, EmoLA is competitive rivaling task-specific state-of-the-art models. The dataset and codes are available: https://johnx69.github.io/FABA/.

Keywords: Facial affective behavior analysis · Emotion recognition · AU recognition · Multi-modal large language model

1 Introduction

Facial affective behavior analysis (FABA) [44, 81] such as facial emotion recognition (FER) [53, 73], and action unit recognition (AUR) [25], aims to recognize

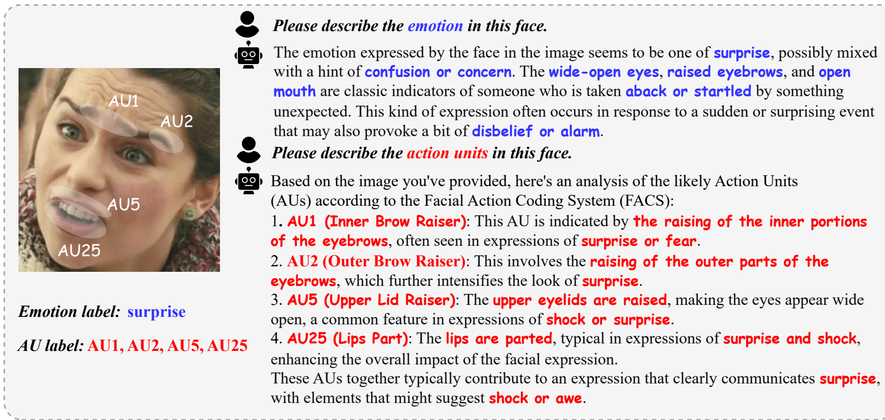


Fig. 1: An illustration of FABA-Instruct annotations. FABA-Instruct can provide fine-grained emotion and AU descriptions, which not only include the reasoning process about the facial movements but also present the inference to the emotion. Furthermore, compared to traditional category labels, FABA-Instruct has more abundant expressions to describe complex, nuanced, exaggerated, and undefined affective behaviors.

facial expressions and movements, which are critical to understanding an individual's emotional states and intentions [31]. FABA has emerged as a burgeoning field with potential across multiple domains. For example, in psychology, FABA can aid therapists by offering real-time insights into a patient's unspoken emotions through facial expression analysis, thereby enhancing therapeutic outcomes [89]. In education, it improves e-learning experiences by adjusting content delivery based on students' facial cues, indicating engagement or confusion [93].

Despite promising progress being made, most of the existing FABA approaches [62, 85, 96, 109] are based on discriminative models, which treat FER or AUR as a multi-class or multi-label classification task. Such approaches tend to induce shortcomings like coarse-grained emotional descriptions, inability to describe complex emotions, and lack of reasoning ability. Those limitations hinder the applications of FABA, for instance, in providing nuanced feedback to therapists about a patient's emotional nuances in psychology, or in accurately adapting educational content based on subtle student reactions in e-learning environments. To counteract these drawbacks, we are motivated by the success of recent multi-modal large language models (MLLMs) [113, 129], because of their evidenced ability to describe and reason over fine-grained and complex visual cues by instruction tuning after large-scale pre-training [69]. In practice, MLLMs transform the discriminative task into a sequence-to-sequence generative one [12, 15] based on large language models (LLMs) [41, 107, 108]. MLLMs have shown great capability on various visual understanding tasks such as visual question answering [68, 69, 144], captioning [50, 51], grounding [10, 126], segmentation [46], *etc.*

However, there exist three major challenges in FABA tasks when deploying MLLMs. (1) There is no suitable FABA dataset for MLLMs to perform instruction tuning. Existing FABA datasets have either coarse-grained annotations [27, 45, 52, 54, 82, 134, 135] or limited emotion descriptions [72] for video clips, since it's labor-intensive and expensive to manually annotate large-scale fine-grained FABA descriptions especially for AUR task. (2) There is an increasing number of MLLMs, how to select suitable MLLMs for FABA remains unknown. Existing metrics for evaluating MLLMs are language-oriented, without specific consideration of language usage in FABA tasks. (3) Furthermore, image features from vision encoders like CLIP [90] of current MLLMs struggle to capture facial structure information such as facial landmarks, leaving the impact of facial priors on FABA tasks unexplored. Fine-tuning the entire model hosting billions of parameters for these features leads to prohibitive computational costs.

To solve these challenges, we propose an instruction-following FABA dataset "FABA-Instruct". It includes $19K$ in-the-wild aligned face images with $30K$ fine-grained emotion and AU annotations using GPT4V enabling instruction tuning (Fig. 1). Based on this dataset, we propose a new benchmark "FABA-Bench" for evaluating both the visual recognition and text generation performance of various MLLMs on FABA tasks. Moreover, we introduce an efficient MLLM "EmoLA" for FABA tasks by incorporating a low-rank adaptation method and a facial prior expert to a pre-trained MLLM like LLaVA-1.5 [68]. Specifically, to obtain the facial prior knowledge, we utilize a pre-trained face alignment encoder to extract the facial landmark features, which are complementary to the vision encoder. To mitigate the computational cost, the LoRA [38] method is adopted in training such that only the parameter residual is learned through low-rank matrices. In this paper, we take the earliest trial of the FABA tasks with instruction tuning over MLLMs, which shed light on both FABA and MLLMs research community. Our contributions are summarized in three-folds:

- ***Instruction-following FABA data***. To our best knowledge, this is the first FABA dataset that enables instruction tuning. It reveals new aspects of FABA research topics and it will continuously bring the benefits of MLLMs to the FABA community.
- ***Instruction-following FABA benchmark***. To evaluate the recognition and reasoning ability of different models on instruction-following FABA tasks, we introduce the FABA-Bench benchmark with a unique metric, *i.e.*, REGE, allowing for both recognition and generation capability.
- ***MLLM-based FABA architecture***. To efficiently train on FABA tasks and utilize the facial prior knowledge, we introduce the EmoLA model which involves tuning LoRA on a pre-trained MLLM and incorporating a facial prior expert. We demonstrate the effectiveness of EmoLA on FABA-Instruct and four traditional FABA datasets. The results show that EmoLA achieves the best performance on FABA-Instruct and SOTA-comparable or even better results on the traditional FABA datasets.

2 Related work

2.1 Facial affective behavior analysis

Psychology perspective. According to psychology research [31], two mechanisms can be utilized to model facial affective behaviors, *i.e.*, emotion categories, and dimensional theory. According to Izard [39] and Ekman [26], the basic emotions can be categorized into one of the several prototypes. Ekman [25] also proposed to decompose the macroscopic affective behaviors into fine-grained Action Units (AUs) from the anatomical perspective. However, such discrete emotion representations may be not sufficient to capture complex and fine-grained emotions. Dimensional theory describes the continuous emotions from the Euclidean space perspective. Russel adopted two dimensions: arousal and valence [92] to represent pleasantness and degree of feelings. Although the dimensional theory allows for a more nuanced understanding of emotions, it is challenging to measure and recognize for humans. We argue that human-generated descriptions offer a superior way of characterizing facial affective behaviors (see Fig. 3 and Fig. 4), which not only capture the complexity and subtlety of affective behaviors but are also more accessible and quantifiable for humans.

Methodology perspective. To better recognize the affective behaviors, current research mainly focuses on deep-learning-based techniques. These approaches can be categorized into three streams according to the task types, *i.e.*, facial emotion recognition [7, 61, 101, 114, 131, 138], action unit recognition [5, 18, 40, 59, 99, 106, 123, 125, 133] and valance-arousal regression [60, 83, 95, 141]. Existing methods focus on capturing the fine-grained facial movement via attention mechanism [109,114,115,118,121,137], improving generalization ability by introducing auxiliary information like facial landmarks [85,106] or extra data [59,84], exploring the relationship among emotions or AUs [49, 97, 139], exploiting pre-trained model like self-supervised learning [5,60,63,103,117], and probing semantic information of affective behaviors [123] using CLIP [24, 90]. However, these methods are mainly discriminative-based, which fail to generate fine-grained descriptions. By contrast, our method can generate detailed descriptions based on the prior knowledge from MLLM and the facial prior expert.

2.2 Multi-modal LLMs and efficient LLM adaptation

Multi-modal LLMs. Multi-modal LLMs are getting popular in multi-modal content understanding [10, 23, 65, 110, 116]. They are built on top of LLMs [2, 14, 16, 41, 107, 132], and transform visual (videos and images) and text data into a sequence of *tokens* as input, resulting in generative modeling of downstream multi-modal understanding tasks by next token prediction. Specific to the image-based MLLMs, image tokens are typically encoded by CLIP vision transformer [24, 90]. Then, one of the major challenges in MLLM is how to project image token features into the language domain to better utilize the instruction-following capability of LLMs. In literature, Flamingo [1] is an early work that uses cross-attention to build the interaction between images and text tokens.

Table 1: Existing FABA datasets.

Datasets	AU	Emotion	Annotation	In-the wild
RAF-DB [54]	✘	✓	category	✘
CK+ [76]	✘	✓	category	✘
MMI [86]	✘	✓	category	✘
SFEW [42]	✘	✓	category	✓
AffectNet [83]	✘	✓	category	✓
MAFW [72]	✘	✓	category & short text	✓
DFEW [42]	✘	✓	category	✓
FERV39K [111]	✘	✓	category	✓
FER2013 [30]	✘	✓	category	✓
DISFA [82]	✓	✘	category	✘
GFT [29]	✓	✘	category	✘
CASME-II [122]	✓	✓	category	✘
BP4D [135]	✓	✓	category	✘
EmotioNet [27]	✓	✓	category	✓
AffWild2 [45]	✓	✓	category	✓
FABA-Instruct	✓	✓	instruction & description	✓

Table 2: FABA-Instruct statistics.

Statistics	Value
Total images	19877
Emotion training samples	19474
Emotion testing samples	403
Emotion description average length	50.47
AU training samples	15838
AU testing samples	325
AU description average length	207.35

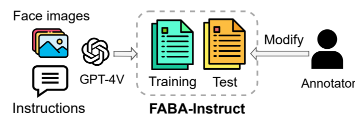


Fig. 2: FABA-Instruct annotation.

This design is followed by recent MLLMs [3, 13, 48, 102]. Based on the cross-attention mechanism, Q-Former [50] was proved to be a superior visual-text projector and inspires recent line of MLLM research [6, 19, 57, 144]. Recently, in contrast to the attention-based projector, LLaVA series [68, 69] propose to use a simple MLP as the projector and resort to the LLMs to handle the visual-text interaction. In this paper, based on LLaVA-1.5 [68], we developed an MLLM for affective behavior analysis and benchmarked multiple MLLM baselines.

Efficient LLM adaptation. Though most existing multimodal LLMs show impressive performance on general visual-language tasks, how to adapt them to downstream applications is still challenging [119], especially with limited instructional annotation data and computing resources. This raises the surge of research on parameter-efficient fine-tuning (PEFT) [37], which avoids fine-tuning the entire LLM in training. According to [119], existing PEFT methods can be categorized into addictive [34,37,47,56,66,67,78,88,91], partial [32,104,120,128], reparametrized [21, 38, 71, 130], hybrid [8, 34, 43, 80], and unified fine-tuning [36, 98, 112, 127]. In this paper, we are interested in the re-parameterized fine-tuning as it gains increasing research attention. In this line, LoRA [38] is a pioneering work, which is inspired by the fact that modal weight adaptation has a low *intrinsic rank*. In training, LoRA keeps most LLM parameters frozen while only optimizing the low-rank factorized matrices of dense layers' residual. In our work, we empirically found the effectiveness of LoRA on the adaptation of MLLMs toward affective behavior analysis tasks.

3 Dataset and Benchmark

3.1 Instruction-following FABA dataset

In this section, we will present our instruction-following FABA dataset **FABA-Instruct** with two different FABA tasks, *i.e.*, emotion recognition and AU recognition. Our dataset is different from existing FABA datasets w.r.t. tasks, annotation types, and the variety of images, which is presented in Tab. 1. This

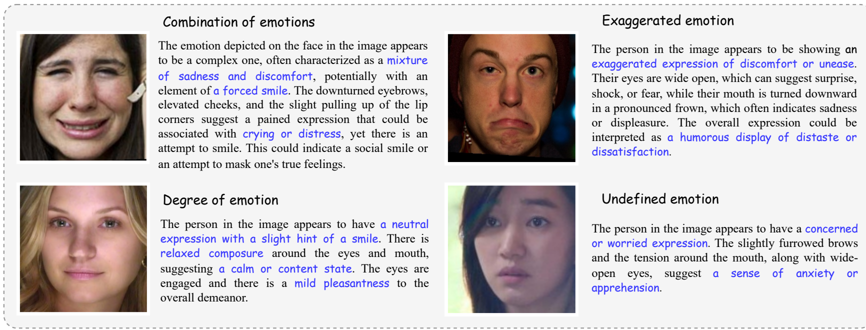


Fig. 3: Emotion description analysis. Emotion descriptions can express comprehensive emotion types like compositional emotions, exaggerated emotions, the degree of emotions, and undefined emotions, *etc*. In contrast, emotion categories struggle to capture such complex and nuanced emotional states.

table illustrates that most of previous FABA datasets are either task-specific [29,30,42,54,72,76,82,83,86,111,135], or laboratory collected [29,76,82,86,135]. More importantly, none of these FABA datasets have instruction-following annotations. FABA-Instruct is the inaugural dataset to offer instruction-following annotations of both AU and emotion, specifically for in-the-wild face images.

Data construction. The overall annotation pipeline is shown in Fig. 2. Specifically, we use 100 carefully designed templates as the instructions for querying GPT-4V [74,105] on the emotions and AUs in the face. For instance, the templates for querying emotion are like "What is the emotion in this face?", "What are the action units present in this face?". More details about these templates are in the Appendix. FABA-Instruct statistics are shown in Tab. 2. For the face images, we randomly sample 19,474 and 403 face images from the training and testing set of AffectNet [83] as the training and testing face images, respectively. AffectNet is a large-scale in-the-wild facial expression database, which crawls from the Internet by querying three search engines. We align and crop these face images using the Dlib library[6].

However, some of the annotations for these face images cannot be obtained due to the low resolution and occlusion issues, especially for the AU annotations. As a result, after filtering out the useless annotations, we obtain 19,474 and 15,835 instructions in terms of emotions and AUs, respectively. Moreover, since some of the annotations are inaccurate or inexact, we carefully check and revise each annotation for the test sets of two tasks.

Emotion description analysis. Existing emotion datasets like RAF-DB [54], AffectNet [83], CK+ [76], FER2013 [30], MMI [86], SFEW [22], mainly adopt the seven emotion categories, *i.e.*, `happiness`, `sadness`, `anger`, `fear`, `disgust`, `surprise`, `neutral`. However, we argue that classifying emotions into one of the several discrete emotion categories is limited in practice since emotions are

[6] http://dlib.net/

The facial action units (AUs) visible in this image suggest a variety of facial muscle movements. Here's a breakdown:
- AU4: Brow Lowerer - The person's brows are drawn together, causing vertical lines to appear between the eyebrows, which is indicative of a frown or a scowl.
- AU6: Cheek Raiser - The muscles around the eye appear to be contracting, creating a small squint which can indicate a smile or a similar expression.
- AU7: Lid Tightener - The tightening around the eyes can be seen, which often accompanies AU6.
- AU12: Lip Corner Puller - The corners of the mouth are drawn back towards the ears which is the primary indicator of a smile.
- AU23: Lip Tightener - The lips appear tight and stretched across the face.
- AU25: Lips part - The lips are parted, which may be due to the facial expression of the smile or speaking.
- AU26: Jaw Drop - The jaw is lowered, contributing to the open mouth appearance.

The overall configuration of these facial action units seems to create an intense or exaggerated expression, which could be associated with strong emotions such as excitement, aggression, or intense joy.

In this image, we can analyze the facial expression by looking at the action units (AUs), which are part of the Facial Action Coding System (FACS):
- AU6: Cheek Raiser – There is evidence of the orbicularis oculi muscle contracting, which produces crow's feet and causes the cheeks to move upwards. This is often associated with expressions of happiness or genuine smiles.
- AU12: Lip Corner Puller – The lip corners are pulled upwards, which is a characteristic of smiling.
- AU25: Lips Part – The lips are slightly parted, which can happen naturally during a relaxed, happy facial expression.

Based on these AUs, the expression can be interpreted as a happy or content one, commonly associated with a genuine smile (also known as a Duchenne smile). This type of smile typically involves both the muscles around the eyes and the mouth, which are activated in this facial expression.

Fig. 4: AU description analysis. AU descriptions give not only the AU labels, but also provide explanations on the cause (which muscle movement) and effect (which emotion it will lead to) w.r.t. each AU, and the relationship between current AU and other AUs or emotions.

subjective [4], complex [25], continuous [70], and contextual [33]. For instance, as shown in Fig. 3, it is inaccurate to classify the compositional expression, `sadness with a forced smile`, into one of the emotion categories. For some exaggerated emotions, *e.g.*, `making a mouth`, the actual emotion is unknown from the face image and it is not reasonable to classify the emotion according to its superficial facial movements. Also, discrete emotions cannot express the degree of emotion, and it is inexact to classify the girl's face as either `happy` or `neutral`. Furthermore, since emotions are complex, basic emotions cannot cover undefined emotions like `worry`, `skeptical`, *etc.* By contrast, the emotion descriptions can address these issues due to their expressive nature.

Action Unit description analysis. The annotations from traditional AUR datasets, *e.g.*, BP4D [135], DISFA [82], EmotioNet [27], GFT [29], mainly adopt a string of binary vector to denote whether each AU is activated or not. However, the representation capability of this way is limited. It cannot indicate the degree of AU's activation and cannot provide any explanations and analysis for the prediction. For instance, in Fig. 4, merely stating that AU6 (cheek raiser) is activated cannot depict the degree of cheek muscle raising. In contrast, the AU descriptions can provide more reasoning cues like "`small squint`" or "`crow's feet`", which can capture the activated degree of AU6. Furthermore, descriptions can also show the inference ability by providing the relationship between current AU and other AUs or emotions, *e.g.*, "`indicative of a frown or a scowl`", "`associated with strong emotions`". Such descriptions can not only provide more nuances but also improve the interpretability [11,58] of the model.

3.2 Instruction-following FABA benchmark

Evaluation. Given the fact that our FABA-Instruct uses free-form textual descriptions to represent emotions and AUs, its model evaluation stands distinct

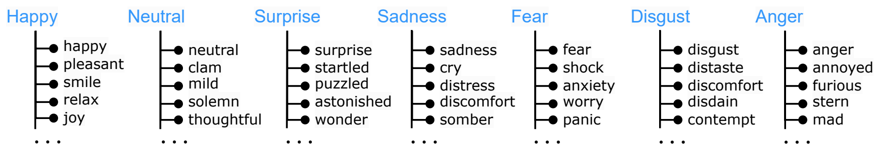
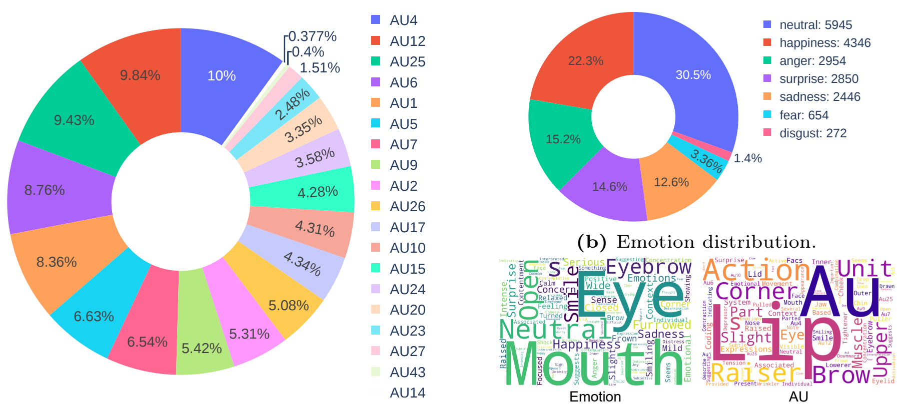


Fig. 5: The synonyms of emotions for classifying the text.



(b) Emotion distribution.

(a) AU distribution. **(c)** Word clouds.

Fig. 6: The distribution of FABA-Instruct on AU (a) and emotion (b) tasks, and the word clouds (c). We extract the emotion labels using the synonyms of each emotion.

from the traditional FABA tasks and Natural Language Generation (NLG) [28] tasks. Specifically, for one thing, the traditional NLG metrics such as BLEU [87] or ROUGE [64] scores mainly concentrate on the coherence and fluency of the generative text, ignoring the FABA relevant consideration in the evaluation. For another, existing FABA metrics, *e.g.*, accuracy or F1 score, focus only on the recognition performance of the model, lacking the evaluation on textual aspects of the model such as the reasoning and explanation. To compensate for the drawbacks of these two metrics, we introduce a new metric REGE to evaluate the **RE**cognition and **GE**neration performance of models on our FABA-Instruct.

Our REGE score is defined to consider both text generation and image recognition aspects of FABA models. For text generation, we choose to use the ROUGE score, which is a generally used metric in NLG by comparing the overlap of n-grams, word sequences, and word pairs between the generated texts and the reference ones. For recognition, we adopt the recognition accuracy for the multi-class performance of facial emotion recognition, and the F1 score for the multi-label performance of AU recognition. Denote the recognition performance as S_{re} and the generation metric as S_{ge}, our REGE metric is computed by taking their sum: $S_{rege} = S_{re} + S_{ge}$.

Calculation of S_{re}. For the FER task, we propose to classify a face image into one of the aforementioned seven basic emotion categories. In order to classify the description, we manually select some synonyms from the training descriptions for each emotion category (see Fig. 5), and the emotion distribution of the training descriptions is shown in Fig. 6b. In practice, since the descriptions incorporate negative sentences, we first deleted these sentences. After that, we count the frequency of emotion synonyms in the sentences and treat the emotion with the highest count as the emotion label for this description. After obtaining these *emotion labels* of texts, we can calculate the accuracy score to evaluate the recognition performance on the FER task.

Similarly, for the AUR task, we propose to classify a face image into one or multiple AU categories. We show the distribution of AUs in Fig. 6a. Following existing literature [27, 135], we choose 12 AUs for evaluation. Subsequently, we can calculate the F1 scores to represent the recognition performance for AUR.

4 EmoLA: An Instruction-tuned MLLM for FABA

In this Section, we introduce a novel MLLM designed for FABA tasks. The overall framework of EmoLA is illustrated in Fig. 7. Its important components include two image experts (a visual and a facial prior expert), a language expert (a tokenizer with the word embedding) and a language decoder (LLM) with a LoRA module. Specifically, a face image X_V is encoded by a visual expert known as a pre-trained CLIP-L/14 [90] with a two-layer Multi-Layer Perceptron (MLP), which generates the visual embedding tokens H_v. Similarly, for the input instruction X_Q, the language expert can provide the language tokens H_q.

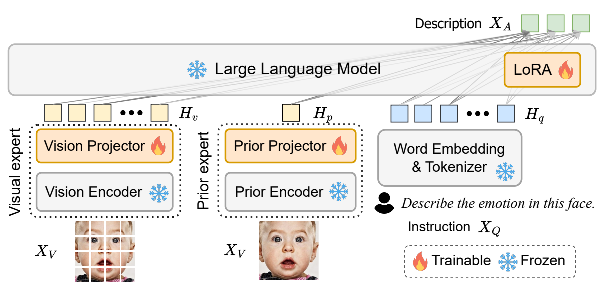


Fig. 7: EmoLA architecture.

Notably, the visual tokens H_v may fail to capture the facial structure information since CLIP is trained with general image-text pairs rather than FABA datasets, which may make the visual expert focus more on general semantic information and overlook task-specific features. Hence, we suggest employing an extra facial prior encoder $f_p(\cdot)$, trained on facial-related datasets, to better capture the facial prior knowledge and enhance the recognition ability for face images. Specifically, we adopt a pretrained facial landmark detector from Insightface[7] to extract the landmark feature. Other face priors, such as recognition features [20], parsing, *etc*., can also be considered. We leave exploring these additional priors as future work for EmoLA. Therefore, the prior feature extracted by f_p can be expressed as: $Z_P = f_p(X_V)$. After obtaining Z_P, we utilize the MLP $g_\theta(\cdot)$ to project the facial prior feature Z_P to the token embedding space:

$$H_p = g_\theta(Z_P), \tag{1}$$

[7] https://github.com/deepinsight/insightface/tree/master

where H_p is the facial prior token. This token can provide prior knowledge of face structure, which may be ignored by the visual expert.

After obtaining the visual embedding tokens H_v, facial prior token H_p and the language embedding tokens H_q, we concatenate them together and feed them into the LLM decoder. Here we use Vicuna [14] as the LLM decoder. As shown in Fig. 7, the visual encoder, prior encoder, word embedding and the LLM decoder are frozen. Except for the prior encoder $f_p(\cdot)$, the initial weights of the other modules come from a pretrained MLLM. In this paper, we utilize the LLaVA-1.5 [68] as the backbone model. In order to train efficiently without tuning the entire MLLM, we propose to tune an extra LoRA module $h_\phi(\cdot)$, a visual projector $h_\gamma(\cdot)$, and the prior projector $g_\theta(\cdot)$. Compared to finetuning the entire LLM, tuning LoRA reduces both memory and computation costs during training without inducing additional expenses for inference. Our experiments demonstrate the effectiveness of such a design in Tab. 7. As a result, the overall parameters to be optimized are $\boldsymbol{\Theta} = \{\theta, \gamma, \phi\}$. We optimize these parameters following the auto-regressive way, and the likelihood of the generated description X_A conditioned on image X_V, prior facial feature Z_P and instructions X_Q is given by:

$$p\left(X_A|X_V, Z_P, X_Q\right) = \prod_{i=1}^{L} p_{\boldsymbol{\Theta}}\left(x_i|X_V, Z_P, X_Q, X_{A,<i}\right), \tag{2}$$

where L is the length of the token sequence, x_i is the current token that needs to be predicted. $X_{A,<i}$ are the previous answer tokens.

5 Experiments

Implementation details. We initialize all the frozen weights of EmoLA with LLaVA-1.5 7b [68], and we only tune the prior projector and LoRA during the training stage. We train our EmoLA for one epoch optimized by AdamW with an initial learning rate of 1e-4 for all the datasets. The rank of LoRA is set to 128. We conduct all the experiments on 8 A6000 GPUs.

Table 3: Comparison on RAF-DB.

Methods	Accuracy (%)
RAN [109]	86.9
MA-Net [142]	87.22
EfficientFace [143]	88.36
RUL [136]	88.98
DAN [115]	89.70
EAC [137]	90.35
APViT [121]	91.98
POSTER [79]	**92.05**
EmoLA (**Ours.**)	**92.05**

Database and protocols. We perform experiments on four traditional FER and AUR datasets, including one emotion dataset (RAF-DB [54]) and three AU datasets (BP4D [135], DISFA [82] and GFT [29]). We turn the annotations of these datasets into instruction-following ones by adding instructions. Instead of predicting the class index or binary vectors in discriminative models, our EmoLA directly outputs the corresponding emotions (*i.e.*, "Happy") or AU labels (*i.e.*, "AU1, AU4"). We adopt accuracy and F1 score to evaluate the recognition performance for FER and AUR tasks, respectively. More details about these datasets can be found in the Appendix. For BP4D and DISFA, following [59, 85, 96], we also perform a subject-exclusive

Table 4: F1 score (in %) of 8 AUs on DISFA. All the results are from original papers.

Method/AU	1	2	4	6	9	12	25	26	Avg.
DRML [140]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-Net [55]	41.5	26.4	66.4	50.7	80.5	89.3	88.9	15.6	48.5
DSIN [17]	42.4	39.0	68.1	28.6	46.8	70.8	90.4	42.2	53.6
SRERL [49]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net [85]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
CMS [94]	40.2	44.3	53.2	57.1	50.3	73.5	81.1	59.7	57.4
ARL [97]	43.9	42.1	63.6	41.8	40.0	76.2	95.2	66.8	58.7
SEV-Net [123]	55.3	53.1	61.5	53.6	38.2	71.6	95.7	41.5	58.8
HMP-PS [100]	38.0	45.9	65.2	50.9	50.8	76.0	93.3	67.6	61.0
ATCM [40]	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
ReCoT [59]	51.3	36.2	66.8	50.1	52.4	78.8	95.3	69.7	62.6
JÂA-Net [96]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
PIAP [106]	50.2	51.8	71.9	50.6	54.5	79.7	94.1	57.2	63.8
GraphAU (R-50) [77]	54.6	47.1	72.9	54.0	55.7	76.7	91.1	53.0	63.1
EmoLA (**Ours.**)	50.5	56.9	83.5	55.2	43.1	80.1	91.6	60.0	**65.1**

3-fold cross-validation. For GFT and RAF-FB, we train and test according to the original dataset division protocol.

We also perform experiments on our FABA-Instruct dataset w.r.t. FER and AUR tasks. We compare with other MLLMs using our SEGE metric (Sec. 3.2). We train and test all the models according to our dataset division protocol.

5.1 Comparison on traditional FER and AUR datasets

Comparison on traditional FER datasets. We conduct a comparative experiment with the latest state-of-the-art (SOTA) methods on RAF-DB, as presented in Tab. 3. Our EmoLA achieves the best results compared with previous SOTA methods. The results demonstrate the significant potential of MLLMs in addressing the FER problem. It is worth noting that most of these methods are specifically tailored for FER tasks, while our EmoLA is easy to adapt to other tasks (*e.g.*, AUR) due to the high flexibility brought by instruction tuning.
Comparison on traditional AUR datasets. Tab. 5, Tab. 4 and Tab. 6 show the comparison results of EmoLA with other SOTA methods on AUR datasets. It can be observed that EmoLA outperforms all the other SOTA methods on DISFA and GFT datasets, achieving a 1.3% improvement over PIAP [106] on DISFA, and a 3.5% increase over EmoCo [103] on GFT. Additionally, it closely competes with ReCoT on BP4D, with a marginal gap of only 0.6%. The gap stems from the benefits of the consistency regularization and co-training in ReCoT to the BP4D dataset. We can observe that EmoLA excels in multi-label classification tasks, potentially due to its strategy of exclusively predicting the positive labels, thereby mitigating the imbalanced issue caused by negative labels.

5.2 Comparison on FABA-Bench

We compare with current MLLMs, *i.e.*, LLaVA-1.5 [68], MiniGPT4-V2 [9], Shikra [10], and mPLUG-Owl2 [124], on our FABA-Bench in Tab. 7. These baselines

Table 5: F1 score (in %) of 12 AUs on BP4D. The results with * are taken from [59]. All the other results are taken directly from their original papers.

Method/AU	1	2	4	6	7	10	12	14	15	17	23	24	Avg.
DRML [140]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-Net [55]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
DSIN [17]	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
CMS [94]	49.1	44.1	50.3	79.2	74.7	80.9	88.3	63.9	44.4	60.3	41.4	51.2	60.6
LP-Net [49]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
ARL [97]	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	55.4	61.1
JÂA-Net* [96]	47.2	41.6	49.1	77.2	77.5	82.9	85.8	63.4	50.8	62.5	47.2	52.7	61.5
SRERL [49]	46.9	45.3	55.6	77.1	78.4	83.5	87.6	60.6	52.2	63.9	47.1	53.3	62.9
HMP-PS [100]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
SEV-Net [123]	58.2	50.4	58.3	81.9	73.9	87.8	87.5	61.6	52.6	62.2	44.6	47.6	63.9
PIAP [106]	54.2	47.1	54.0	79.0	78.2	86.3	89.5	66.1	49.7	63.2	49.9	52.0	64.1
ATCM [40]	51.7	49.3	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	56.3	64.2
GraphAU (R-50) [77]	53.7	46.9	59.0	78.5	80.0	84.4	87.8	67.3	52.5	63.2	50.6	52.4	64.7
ReCoT* [59]	51.5	47.8	58.9	79.2	80.2	84.9	88.4	61.6	53.3	64.6	51.8	55.4	**64.8**
EmoLA (**Ours.**)	57.4	52.4	61.0	78.1	77.8	81.9	89.5	60.5	49.3	64.9	46.0	52.4	64.2

Table 6: F1 score (in %) results of 10 AUs on GFT. The results with * are taken from [103]. The others are taken directly from original papers. “ft” stands for finetune.

Method/AU	1	2	4	6	10	12	14	15	23	24	Avg.
EACNet [55]	15.5	56.6	0.1	81.0	76.1	84.0	0.1	38.5	57.8	51.2	46.1
TCAE [63]	43.9	49.5	6.3	71.0	76.2	79.5	10.7	28.5	34.5	41.7	44.2
ARL [97]	51.9	45.9	13.7	79.2	75.5	82.8	0.1	44.9	59.2	47.5	50.1
MoCo (ft)* [35]	45.3	48.2	20.3	80.7	78.8	78.1	22.6	46.0	53.9	50.3	52.4
Temporal Ranking (ft)* [75]	58.8	56.8	33.2	72.5	76.2	80.8	19.9	46.8	55.2	47.3	54.7
JÂANet [96]	46.5	49.3	19.2	79.0	75.0	84.8	44.1	33.5	54.9	50.7	53.7
EmoCo [103]	65.9	55.9	40.7	83.1	75.1	81.4	21.3	48.5	58.0	56.5	58.6
EmoLA (**Ours.**)	69.8	59.1	52.8	85.3	73.0	85.3	32.3	47.6	63.1	52.2	**62.1**

mainly employ Vicuna [14] as LLM decoder and incorporate instruction tuning. We reproduce all the baselines using their open-sourced codes. For fair comparison, all the models are trained for 1 epoch and train on two tasks individually. Except for our EmoLA, all the other baselines are finetuned based on the pre-trained MLLMs. More details about baselines can refer to the Appendix.

It is noteworthy that EmoLA achieves the best results in two tasks with fewer tuning parameters compared to other MLLMs. EmoLA finetunes merely 10% of the parameters compared to LLaVA-1.5, yet it achieves better performance on FABA-Bench, which can be attributed to two aspects. For one thing, EmoLA only finetunes LoRA which is much more efficient than tuning the entire LLM decoder. For another, EmoLA incorporates a facial prior expert to extract the facial structure knowledge. This feature compensates for the information overlooked by the visual encoder in FABA tasks. Moreover, we can also see that due to the strong capability of LLM, the language generation ability of Shikra, LLaVA-1.5 and EmoLA is comparable, making it unreasonable to evaluate these models merely based on the NLG metrics. Our metric, REGE, accounts for both

Table 7: Comparison on FABA-bench. All the baselines are reproduced by their open-sourced codes. S_{re} in emotion and AU tasks stands for accuracy and average F1 scores, respectively. S_{ge} and S_{rege} are ROUGE and our REGE score. All scores are in %.

Methods	Emotion			AU														
	S_{re}	S_{ge}	S_{rege}	1	2	4	5	6	10	12	17	24	25	26	43	S_{re}	S_{ge}	S_{rege}
MiniGPT4-v2 [9]	58.2	19.6	77.8	47.9	35.5	42.3	32.7	29.2	6.6	10.3	0.0	2.5	0.1	0.0	0.0	17.9	19.9	37.8
mPLUG-Owl2 [124]	53.6	28.4	82.0	72.3	17.5	75.2	54.2	75.6	0.0	13.0	0.0	0.0	3.9	18.2	0.0	27.5	28.2	55.7
Shikra [10]	62.5	32.1	94.6	70.6	33.9	76.6	63.3	57.8	43.4	58.0	53.0	54.1	68.5	42.4	0.0	51.8	34.8	86.6
LLaVA-1.5 [68]	62.3	31.6	93.9	74.2	32.7	76.5	67.9	63.6	41.0	61.0	53.4	54.1	67.5	43.5	50.0	57.1	34.3	91.4
EmoLA (**Ours.**)	64.5	31.7	**96.2**	72.8	37.3	79.9	67.3	69.9	41.7	63.6	56.8	55.6	73.4	56.8	0.0	56.3	35.2	**91.5**

the recognition and generation abilities of MLLMs in FABA, offering a more comprehensive evaluation of their performance.

5.3 Ablation study

We explore the effectiveness of the facial landmark token in Tab. 8, and the tuning strategies in Tab. 9. More ablations can be found in Appendix, *e.g.*, multi-task training, the position of the prior token, *etc*. Fig. 8 shows the capability of EmoLA on FABA tasks, with additional examples available in the Appendix.

Table 8: Influence of prior token. H_v and H_p indicate visual tokens and the facial prior token, respectively.

Tokens		Emotion		AU	
H_v	H_p	S_{re}	S_{ge}	S_{re}	S_{ge}
	✓	41.2	29.9	40.5	33.7
✓		62.5	32.1	55.3	34.8
✓	✓	64.5	31.7	56.3	35.2

Effectiveness of prior token. As shown in Tab. 8, if we input only the facial prior token H_p to EmoLA without visual tokens, there will be a significant decline in recognition performance. However, thanks to the powerful capability of LLM decoder, the generalization ability does not suffer greatly. Furthermore, even with just a single token fed into the decoder, the model retains a certain level of recognition capability. This demonstrates that the landmark prior contains representative features for FABA tasks.

Compared to merely inputting the prior token H_p, incorporating visual tokens H_v will enhance the recognition performance since these visual tokens carry more detailed face information. Incorporating both visual tokens and prior tokens leads to further improvements with eligible performance dropping in text generation. This suggests that the face prior token supplies complementary information of visual tokens. We suppose this additional information includes facial structure details introduced by landmark facial prior extractor.

Influence of tuning strategy. We also investigate the impact of tuning strategies on two modules requiring optimization, h_ϕ and g_θ, as shown in Tab. 9. When only the prior projector g_θ is fine-tuned, with all other parameters kept frozen, we observe that while the overall recognition performance is relatively low, it still surpasses that achieved by using only a single prior token. This suggests the effectiveness of tuning the prior projector g_θ, since the parameters of g_θ only take a small portion of the entire parameters.

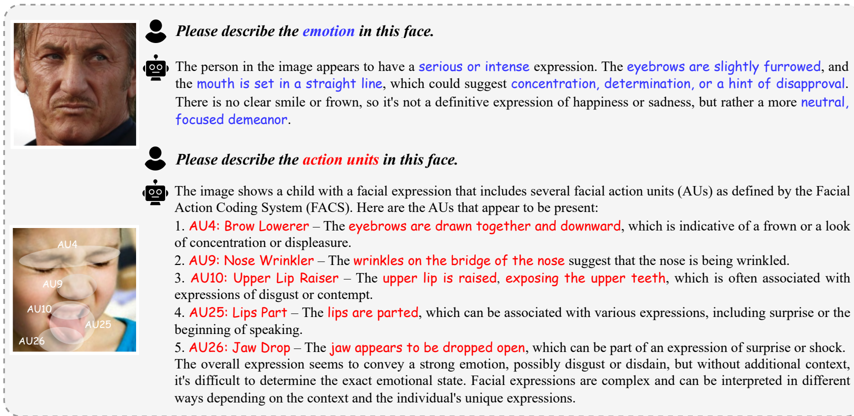


Fig. 8: An example of EmoLA's capability on FABA tasks.

We also tried exclusively tuning LoRA with visual projector $h_\phi + h_\gamma$ while keeping other parameters frozen. This is different from merely inputting the visual tokens H_v since it also includes a prior token H_p generated by the randomly initialized prior projector g_θ. The results indicate that tuning $h_\phi + h_\gamma$ significantly improves the performance by better aligning the output of the LLM. It can also be observed that the performance slightly exceeds that achieved using only the visual tokens, which also emphasizes the value of the prior token. Again, the efficacy is further augmented by simultaneously fine-tuning both $h_\theta + h_\gamma$ and the prior projector g_θ.

Table 9: Influence of the tuning strategy. $h_\phi + h_\gamma$ and g_θ indicate the LoRA module with visual projector and the prior projector, respectively.

Modules		Emotion		AU	
$h_\phi + h_\gamma$	g_θ	S_{re}	S_{ge}	S_{re}	S_{ge}
	✓	44.9	29.6	47.7	34.0
✓		63.0	32.1	55.6	34.9
✓	✓	64.5	31.7	56.3	35.2

6 Conclusion

In this paper, to address the challenges raised in FABA tasks when employing MLLMs, we proposed an instruction-following FABA dataset *FABA-Instruct* by means of GPT4. Based on this dataset, we introduced a benchmark *FABA-Bench* to comprehensively evaluate the FABA models on instruction-following data. Furthermore, we presented an instruction-tuned MLLM *EmoLA* for FABA, which is efficient and effective by tuning LoRA on a pre-trained MLLM and incorporating a facial prior expert. Extensive experiments across four traditional FABA datasets and our FABA-Bench demonstrate the effectiveness of EmoLA. In the future, we intend to broaden our method to additional facial-related tasks, *e.g.*, face detection, face generation, *etc*. Moreover, incorporating other facial prior features holds the potential for performance improvement. Expanding our EmoLA from 2D face images to video streams also presents a promising avenue for future research.

Acknowledgement: Yifan Li, Wentao Bao, and Yu Kong are partially supported by NSF Awards 1949694 and 2040209. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF. We also would like to express our deepest gratitude to Dr. Junwen Chen for her invaluable support and contribution to this research.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: Adv. Neural Inform. Process. Syst. (2022) 4
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 4
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023) 5
4. Barrett, L.F., Mesquita, B., Ochsner, K.N., Gross, J.J.: The experience of emotion. Annu. Rev. Psychol. **58**, 373–403 (2007) 7
5. Chang, Y., Wang, S.: Knowledge-driven self-supervised representation learning for facial action unit recognition. In: CVPR. pp. 20417–20426 (2022) 4
6. Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., Xu, B.: X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160 (2023) 5
7. Chen, F., Shao, J., Zhu, S., Shen, H.T.: Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In: CVPR. pp. 10761–10770 (2023) 4
8. Chen, J., Zhang, A., Shi, X., Li, M., Smola, A., Yang, D.: Parameter-efficient fine-tuning design spaces. arXiv preprint arXiv:2301.01821 (2023) 5
9. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 11, 13
10. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 2, 4, 11, 13
11. Chen, R., Zhang, H., Liang, S., Li, J., Cao, X.: Less is more: Fewer interpretable region via submodular subset selection. In: ICLR (2024) 7
12. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. In: ICLR (2021) 2
13. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023) 5
14. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/` 4, 10, 12

15. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: ICML. pp. 1931–1942 (2021) 2
16. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 4
17. Corneanu, C., Madadi, M., Escalera, S.: Deep structure inference network for facial action unit recognition. In: ECCV. pp. 298–313 (2018) 11, 12
18. Cui, Z., Kuang, C., Gao, T., Talamadupula, K., Ji, Q.: Biomechanics-guided facial action unit detection through force modeling. In: CVPR. pp. 8694–8703 (2023) 4
19. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) 5
20. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699 (2019) 9
21. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Adv. Neural Inform. Process. Syst. **36** (2024) 5
22. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: IEEE Int. Conf. Comput. Vis. Worksh. pp. 2106–2112 (2011) 6
23. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023) 4
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020) 4
25. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978) 1, 4, 7
26. Ekman, P., et al.: Basic emotions. Handbook of cognition and emotion **98**(45-60), 16 (1999) 4
27. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: CVPR. pp. 5562–5570 (2016) 3, 5, 7, 9
28. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Jour. Art. Intel. Resea. **61**, 65–170 (2018) 8
29. Girard, J.M., Chu, W.S., Jeni, L.A., Cohn, J.F.: Sayette group formation task (gft) spontaneous facial expression database. In: IEEE FG. pp. 581–588 (2017) 5, 6, 7, 10
30. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: Adv. Neural Inform. Process. Syst. pp. 117–124 (2013) 5, 6
31. Grandjean, D., Sander, D., Scherer, K.R.: Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Consciousness and cognition **17**(2), 484–495 (2008) 2, 4
32. Guo, D., Rush, A.M., Kim, Y.: Parameter-efficient transfer learning with diff pruning. arXiv preprint arXiv:2012.07463 (2020) 5
33. Haidt, J., Keltner, D.: Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. Cognition & Emotion **13**(3), 225–266 (1999) 7

34. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366 (2021) 5
35. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020) 12
36. He, S., Ding, L., Dong, D., Zhang, M., Tao, D.: Sparseadapter: An easy approach for improving the parameter-efficiency of adapters. arXiv preprint arXiv:2210.04284 (2022) 5
37. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799 (2019) 5
38. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 3, 5
39. Izard, C.E.: Human emotions. Springer Science & Business Media (2013) 4
40. Jacob, G.M., Stenger, B.: Facial action unit detection with transformers. In: CVPR. pp. 7680–7689 (2021) 4, 11, 12
41. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023) 2, 4
42. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: ACM MM. pp. 2881–2889 (2020) 5, 6
43. Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: Efficient low-rank hypercomplex adapter layers. Adv. Neural Inform. Process. Syst. **34**, 1022–1035 (2021) 5
44. Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first abaw 2020 competition. In: IEEE FG. pp. 637–643 (2020) 1
45. Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018) 3, 5
46. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 2
47. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021) 5
48. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 5
49. Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. In: AAAI. pp. 8594–8601 (2019) 4, 11, 12
50. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 2, 5
51. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022) 2
52. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE TIP **28**(1), 356–370 (2019) 3
53. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Trans. Affect. Comput. **13**(3), 1195–1215 (2020) 1

54. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593 (2017) 3, 5, 6, 10
55. Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: Deep nets with enhancing and cropping for facial action unit detection. IEEE TPAMI **40**(11), 2583–2596 (2018) 11, 12
56. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021) 5
57. Li, X., Behpour, S., Doan, T.L., He, W., Gou, L., Ren, L.: Up-dp: Unsupervised prompt learning for data pre-selection with vision-language models. In: Adv. Neural Inform. Process. Syst. vol. 36 (2024) 5
58. Li, X., Pan, D., Li, C., Qiang, Y., Zhu, D.: Negative flux aggregation to estimate feature attributions. In: IJCAI (2023) 7
59. Li, Y., Han, H., Shan, S., Ji, Z., Bai, J., Chen, X.: Recot: Regularized co-training for facial action unit recognition with noisy labels. In: BMVC (2023) 4, 10, 11, 12
60. Li, Y., Sun, H., Liu, Z., Han, H., Shan, S.: Affective behaviour analysis using pretrained model with facial prior. In: Eur. Conf. Comput. Vis. Worksh. pp. 19–30 (2022) 4
61. Li, Y., Wang, Y., Cui, Z.: Decoupled multimodal distilling for emotion recognition. In: CVPR. pp. 6631–6640 (2023) 4
62. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE TIP **28**(5), 2439–2450 (2018) 2
63. Li, Y., Zeng, J., Shan, S., Chen, X.: Self-supervised representation learning from videos for facial action unit detection. In: CVPR. pp. 10924–10933 (2019) 4, 12
64. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 8
65. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 (2023) 4
66. Lin, Z., Madotto, A., Fung, P.: Exploring versatile generative language model via parameter-efficient transfer learning. arXiv preprint arXiv:2004.03829 (2020) 5
67. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.A.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Adv. Neural Inform. Process. Syst. **35**, 1950–1965 (2022) 5
68. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 2, 3, 5, 10, 11, 13
69. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Adv. Neural Inform. Process. Syst. **36** (2024) 2, 5
70. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: CVPR. pp. 1749–1756 (2014) 7
71. Liu, Q., Wu, X., Zhao, X., Zhu, Y., Xu, D., Tian, F., Zheng, Y.: Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. arXiv preprint arXiv:2310.18339 (2023) 5
72. Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., Shan, S.: Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In: ACM MM. pp. 24–32 (2022) 3, 5, 6
73. Lopes, A.T., De Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. PR **61**, 610–628 (2017) 1

74. Lu, H., Niu, X., Wang, J., Wang, Y., Hu, Q., Tang, J., Zhang, Y., Yuan, K., Huang, B., Yu, Z., et al.: Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. arXiv preprint arXiv:2403.05916 (2024) 6
75. Lu, L., Tavabi, L., Soleymani, M.: Self-supervised learning for facial action unit recognition through temporal consistency. In: BMVC (2020) 12
76. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: CVPRW. pp. 94–101 (2010) 5, 6
77. Luo, C., Song, S., Xie, W., Shen, L., Gunes, H.: Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. arXiv preprint arXiv:2205.01782 (2022) 11, 12
78. Mahabadi, R.K., Ruder, S., Dehghani, M., Henderson, J.: Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. arXiv preprint arXiv:2106.04489 (2021) 5
79. Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A.: Poster v2: A simpler and stronger facial expression recognition network. arXiv preprint arXiv:2301.12149 (2023) 10
80. Mao, Y., Mathias, L., Hou, R., Almahairi, A., Ma, H., Han, J., Yih, W.t., Khabsa, M.: Unipelt: A unified framework for parameter-efficient language model tuning. arXiv preprint arXiv:2110.07577 (2021) 5
81. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE Trans. Affect. Comput. **10**(3), 325–347 (2017) 1
82. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013) 3, 5, 6, 7, 10
83. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017) 4, 5, 6
84. Niu, X., Han, H., Shan, S., Chen, X.: Multi-label co-regularization for semi-supervised facial action unit recognition. In: Adv. Neural Inform. Process. Syst. pp. 909–919 (2019) 4
85. Niu, X., Han, H., Yang, S., Huang, Y., Shan, S.: Local relationship learning with person-specific shape regularization for facial action unit detection. In: CVPR. pp. 11917–11926 (2019) 2, 4, 10, 11
86. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: ICME. pp. 5–pp (2005) 5, 6
87. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Asso. Comput. Ling. pp. 311–318 (2002) 8
88. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: Adapterfusion: Non-destructive task composition for transfer learning. arXiv preprint arXiv:2005.00247 (2020) 5
89. Pixton, T.S.: Happy to see me, aren't you, sally? signal detection analysis of emotion detection in briefly presented male and female faces. Scandinavian Journal of Psychology **52**(4), 361–368 (2011) 2
90. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) 3, 4, 9
91. Rücklé, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., Gurevych, I.: Adapterdrop: On the efficiency of adapters in transformers. arXiv preprint arXiv:2010.11918 (2020) 5

92. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980) 4
93. Saneiro, M., Santos, O.C., Salmeron-Majadas, S., Boticario, J.G., et al.: Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. The Scientific World Journal **2014** (2014) 2
94. Sankaran, N., Mohan, D.D., Setlur, S., Govindaraju, V., Fedorishin, D.: Representation learning through cross-modality supervision. In: IEEE FG. pp. 1–8 (2019) 11, 12
95. Savchenko, A.V.: Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. arXiv preprint arXiv:2203.13436 (2022) 4
96. Shao, Z., Liu, Z., Cai, J., Ma, L.: Jaa-net: Joint facial action unit detection and face alignment via adaptive attention. IJCV **129**(2), 321–340 (2021) 2, 10, 11, 12
97. Shao, Z., Liu, Z., Cai, J., Wu, Y., Ma, L.: Facial action unit detection using attention and relation learning. IEEE Trans. Affect. Comput. (2019) 4, 11, 12
98. Shen, J., Wang, H., Gui, S., Tan, J., Wang, Z., Liu, J.: {UMEC}: Unified model and embedding compression for efficient recommendation systems. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=BM---bH_RSh 5
99. Song, T., Cui, Z., Wang, Y., Zheng, W., Ji, Q.: Dynamic probabilistic graph convolution for facial action unit intensity estimation. In: CVPR. pp. 4845–4854 (2021) 4
100. Song, T., Cui, Z., Zheng, W., Ji, Q.: Hybrid message passing with performance-driven structures for facial action unit detection. In: CVPR. pp. 6267–6276 (2021) 11, 12
101. Sun, B., Cao, S., Li, D., He, J., Yu, L.: Dynamic micro-expression recognition using knowledge distillation. IEEE Trans. Affect. Comput. **13**(2), 1037–1043 (2020) 4
102. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023) 5
103. Sun, X., Zeng, J., Shan, S.: Emotion-aware contrastive learning for facial action unit detection. In: FG. pp. 01–08 (2021) 4, 11, 12
104. Sung, Y.L., Nair, V., Raffel, C.A.: Training neural networks with fixed sparse masks. Adv. Neural Inform. Process. Syst. **34**, 24193–24205 (2021) 5
105. Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., Karami, M., Li, J., Cheng, L., Liu, H.: Large language models for data annotation: A survey. arXiv preprint arXiv:2402.13446 (2024) 6
106. Tang, Y., Zeng, W., Zhao, D., Zhang, H.: Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In: ICCV. pp. 12899–12908 (2021) 4, 11, 12
107. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2, 4
108. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 2
109. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE TIP **29**, 4057–4069 (2020) 2, 4, 10

110. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023) 4
111. Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., Zhang, W.: Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In: CVPR. pp. 20922–20931 (2022) 5, 6
112. Wang, Y., Agarwal, S., Mukherjee, S., Liu, X., Gao, J., Awadallah, A.H., Gao, J.: Adamix: Mixture-of-adaptations for parameter-efficient model tuning. arXiv preprint arXiv:2210.17451 (2022) 5
113. Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., Zhai, B., Yuan, J., You, Q., Yang, H.: Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. arXiv preprint arXiv:2401.06805 (2024) 2
114. Wang, Z., Zeng, F., Liu, S., Zeng, B.: Oaenet: Oriented attention ensemble for accurate facial expression recognition. PR **112**, 107694 (2021) 4
115. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: Multi-head cross attention network for facial expression recognition. Biomimetics **8**(2), 199 (2023) 4, 10
116. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: NExT-GPT: Any-to-any multimodal llm. CoRR **abs/2309.05519** (2023) 4
117. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR. pp. 10687–10698 (2020) 4
118. Xie, S., Hu, H., Wu, Y.: Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. PR **92**, 177–191 (2019) 4
119. Xu, L., Xie, H., Qin, S.Z.J., Tao, X., Wang, F.L.: Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148 (2023) 5
120. Xu, R., Luo, F., Zhang, Z., Tan, C., Chang, B., Huang, S., Huang, F.: Raise a child in large language model: Towards effective and generalizable fine-tuning. arXiv preprint arXiv:2109.05687 (2021) 5
121. Xue, F., Wang, Q., Tan, Z., Ma, Z., Guo, G.: Vision transformer with attentive pooling for robust facial expression recognition. IEEE Trans. Affect. Comput. (2022) 4, 10
122. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. PloS one **9**(1), e86041 (2014) 5
123. Yang, H., Yin, L., Zhou, Y., Gu, J.: Exploiting semantic embedding and visual feature for facial action unit detection. In: CVPR. pp. 10482–10491 (2021) 4, 11, 12
124. Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257 (2023) 11, 13
125. Yin, Y., Chang, D., Song, G., Sang, S., Zhi, T., Liu, J., Luo, L., Soleymani, M.: Fg-net: Facial action unit detection with generalizable pyramidal features. In: Wint. Appl. Comput. Vis. pp. 6099–6108 (2024) 4
126. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In: ICLR (2023) 2
127. Yu, S., Chen, T., Shen, J., Yuan, H., Tan, J., Yang, S., Liu, J., Wang, Z.: Unified visual transformer compression. arXiv preprint arXiv:2203.08243 (2022) 5

128. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021) 5
129. Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., Yu, D.: Mm-llms: Recent advances in multimodal large language models. In: arXiv preprint arXiv:2401.13601 (2024) 2
130. Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T.: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512 (2023) 5
131. Zhang, S., Pan, Y., Wang, J.Z.: Learning emotion representations from verbal and nonverbal communication. In: CVPR. pp. 18993–19004 (2023) 4
132. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022) 4
133. Zhang, X., Yang, H., Wang, T., Li, X., Yin, L.: Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection. In: Wint. Appl. Comput. Vis. pp. 6077–6086 (2024) 4
134. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high-resolution spontaneous 3d dynamic facial expression database. In: IEEE FG. pp. 1–6 (2013) 3
135. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. Ima. Vis. Comput. **32**(10), 692–706 (2014) 3, 5, 6, 7, 9, 10
136. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. In: Adv. Neural Inform. Process. Syst. pp. 17616–17627 (2021) 10
137. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: ECCV. pp. 418–434 (2022) 4, 10
138. Zhang, Z., Wang, L., Yang, J.: Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In: CVPR. pp. 18888–18897 (2023) 4
139. Zhao, K., Chu, W.S., Martinez, A.M.: Learning facial action units from web images with scalable weakly supervised clustering. In: CVPR. pp. 2090–2099 (2018) 4
140. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: CVPR. pp. 3391–3399 (2016) 11, 12
141. Zhao, S., Li, Y., Yao, X., Nie, W., Xu, P., Yang, J., Keutzer, K.: Emotion-based end-to-end matching between image and music in valence-arousal space. In: ACM MM. pp. 2945–2954 (2020) 4
142. Zhao, Z., Liu, Q., Wang, S.: Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE TIP **30**, 6544–6556 (2021) 10
143. Zhao, Z., Liu, Q., Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: AAAI. pp. 3510–3519 (2021) 10
144. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 2, 5