Describe Your Facial Expressions by Linking Image Encoders and Large Language Models

Yujian Yuan[1,2]
yuanyujian18@mails.ucas.ac.cn
Jiabei Zeng[1,2]
jiabei.zeng@ict.ac.cn
Shiguang Shan[1,2]
sgshan@ict.ac.cn

[1] Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China

[2] University of Chinese Academy of Sciences
Beijing, China

Abstract

This paper presents a novel task of describing human facial expressions of a facial image in natural language, which captures the nuances of facial actions and emotional states beyond traditional emotion categories or facial action units (AUs). To achieve the facial expression captioning model, we propose a three-stage training framework that trains a vision-to-language model using synthesized image-text pairs and the BLIP-2 pre-training techniques. To overcome the challenge of missing training image-text pairs for facial expression captioning, we propose a strategy that involves synthesizing and combining captions using GPT-3.5 and existing annotations on either emotion categories or AUs. Experiments demonstrate the effectiveness of our method in generating captions that describe details of facial actions and emotions, as well as the inferential relationship between them, even when those emotions are not present in the training data. It is also demonstrated that the vision-to-language task enhances the performance of the intermediate visual features on both AU detection and emotion classification tasks. The code and trained models are available at: https://github.com/Yujianyuan/Exp-BLIP.

1 Introduction

Facial expressions are a crucial form of nonverbal communication that convey a wide range of emotional states through subtle changes in facial muscle movements. The accurate perception and interpretation of facial expressions are critical for effective human-machine interactions. Existing works have focused on describing facial expressions by classifying them into pre-defined emotion categories (e.g., happiness, sadness, neutral, anger, disgust, surprise, fear) or detecting the appearance of facial action units (AUs) defined by the Facial Action Coding System (FACS)[6]. However, these methods have limitations. Pre-defined emotion categories cannot capture all the nuances of facial expressions, while AUs do not convey the indicated affective status. To overcome these limitations, we propose a novel facial expression captioning task that aims to describe facial actions and the corresponding

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

emotional states of a facial image in natural language. Our approach provides a comprehensive and detailed understanding of facial expressions, which has the potential to improve accuracy and effectiveness in understanding and interacting with humans.

It is non-trivial to accomplish facial expression captioning due to the lack of training image-text pairs that specifically describe nuances of facial expressions. Although existing image captioning models such as CoCa [42], BLIP [16], and Flamingo [1] demonstrate remarkable performance in describing the contents of an image, they fall short in describing the nuances of facial expressions. This limitation arises from the fact that the training image-text pairs typically do not provide the level of detail required to train a facial expression captioner. However, collecting such image-text pairs presents a significant challenge. Annotators must possess exceptional writing skills and the ability to recognize and describe the full range of facial expressions. Additionally, even with expert annotators, the process of annotation is time-consuming.

To overcome the scarcity of training image-text pairs for facial expression captioning, we propose a three-stage framework that leverages synthetic image-text pairs synthesized using GPT-3.5 and existing annotations on emotion categories and AUs. As data labeled with both emotion categories and AUs are rarely available, our proposed framework combines captions generated by the emotion-specific captioner (Emot-BLIP) and the AU-specific captioner (AU-BLIP) to train the final captioner (Exp-BLIP) using the combined captions and their corresponding images. Our contributions are summarised as follows:

(1) We propose facial expression captioning, a new task to describe nuances of facial expressions and infer the corresponding emotions in words. Compared with conventional emotion classification and AU detection tasks, facial expression captioning interprets facial expressions in a more detailed and comprehensive way.

(2) To train the captioner for facial expression captioning, we propose a three-stage framework that utilizes synthetic image-text pairs. This approach enables the training even when the training data that describe both facial actions and emotions are not available.

(3) Extensive experiments demonstrate the powerful ability of the trained facial expression captioner and the intermediate visual representation. Our method produces captions that describe details of the facial actions and emotions, even if the emotions are not in the training data. Experiments also demonstrate that vision-to-language task improves the representation ability of the intermediate visual features on AU detection and emotion classification tasks.

2 Related works

Facial expression description method: The current approaches to describing facial expressions primarily involve emotion classification[27, 37, 43, 48] and AU detection[11, 28, 36, 38] tasks. Emotion classification typically consists of categorizing emotions into eight basic categories: neutral, anger, disgust, fear, happiness, sadness, surprise, and contempt [29]. To address the limitations of describing facial expressions with single emotions, compound emotions [5] was proposed by combining several basic emotions. To overcome the shortcomings of archetypal expression descriptions, Ekman et al. [6] developed the facial action coding system (FACS) based on facial anatomy, which defines 44 AUs by focusing on different facial parts. However, it is important to note that both emotions and AUs alone cannot fully capture the complexity of facial status. Recently, Nezami et al. [33] have proposed a novel image captioning model that utilized facial expression features to generate image captions. Their work stated that the improvement in caption quality appeared to come not from

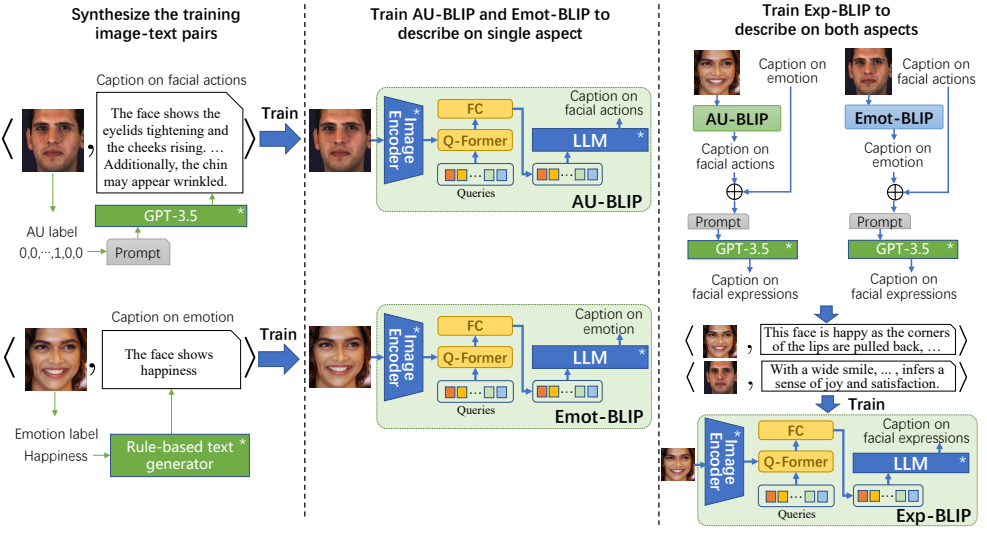

Figure 1: Three-stage training framework for facial expression captioning. **(Left)** Synthesize the training image-text pairs. **(Middle)** Train AU-BLIP and Emot-BLIP to respectively describe a single aspect of facial expression. **(Right)** Train Exp-BLIP utilizing synthesized data with fused captions to describe both facial actions and emotions.

the addition of adjectives linked to emotional aspects of the images, but from more variety in the actions described in the captions. Furthermore, Bryson [22] introduced a meticulous approach to generating weak prominent feature labels using semantic segmentation, demonstrating how these labels could enhance attribute-based face descriptions.

Application of visual and language pre-trained models: Recently, pre-trained vision models and large language models (LLMs) have made remarkable contributions to computer vision and natural language processing (NLP). Numerous studies [13, 21, 26, 44, 45, 50] have demonstrated the efficacy of directly applying the features of these large models to downstream tasks without fine-tuning. This approach has proven impactful in accomplishing multi-modal tasks such as image captioning and vision question answering. Recent multi-modal works [1, 8, 12, 15, 34, 39, 42] primarily focused on cross-modal alignment. For instance, CLIP[34] directly aligned visual and text features during training, while Flamingo[1] and BLIP-2[17] established a connection between pre-trained vision and language models.

Prompt engineering is a vital technique in utilizing LLMs. A text or template named prompt can be used to strongly guide the generation to output answers for desired tasks, thus beginning an era of "pre-train and prompt" [23]. Proper zero-shot or few-shot prompts [9, 31, 41] have the potential to effectively harness the capabilities of LLMs. For the recent studies, Sewon et al. [31] introduced a noisy channel approach for language model prompting in few-shot text classification. To deal with logical reasoning tasks, Takeshi Kojima et al. [14] proved that step prompting was a more powerful prompt than zero-shot prompts.

3 Facial Expression Captioning

This paper aims to develop an image captioner that describes the nuances of facial actions and the corresponding emotional states of a facial image in natural language. To achieve this goal, we propose a three-stage framework that leverages synthetic image-text pairs. Figure 1

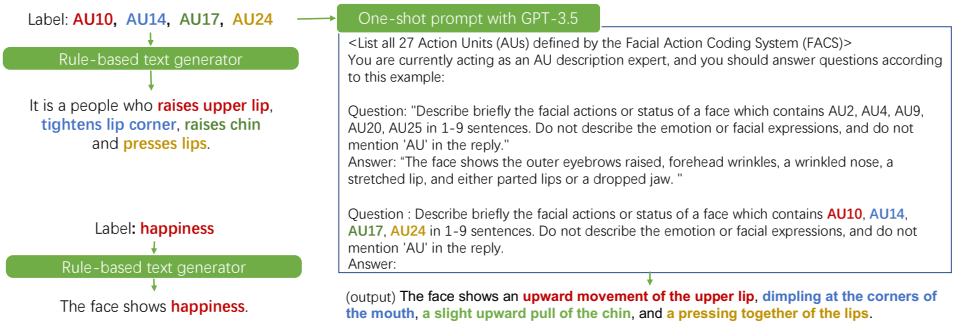

Figure 2: Inputs and outputs examples of the rule-based text generators for AUs and emotions(Left) and those of the GPT-3.5 based generator with one-shot prompt(Right) for AUs.

illustrates the proposed framework. In the first step, we synthesize the training image-text pairs using the prevalent large language model (e.g., GPT-3.5) or a rule-based text generator. In the second step, we train two separate captioners, AU-BLIP and Emot-BLIP, using synthetic pairs that describe facial expressions' specific aspects, including AUs and emotions. We do not directly train an image captioner that simultaneously describes both facial actions and emotions, as it is a challenge to derive training texts that correctly describe both two aspects due to few datasets being annotated with both emotions and AUs. In the third step, we combine the outputs of AU-BLIP and Emot-BLIP to synthesize descriptions for each image. With the combined captions, we train the final captioner (Exp-BLIP) to describe both aspects of facial expressions. Below, we present details of the three steps.

3.1 Synthesis of the training image-text pairs

To train the image captioners, AU-BLIP and Emot-BLIP that are respectively specialized for AUs and Emotions, we constitute the training image-text pairs by synthesizing descriptive texts based on annotations of AUs and emotion categories for the corresponding facial images. The texts are synthesized in two ways: the rule-based text generator, and the GPT-based generator with one-shot prompt. Examples of the two types of generators are shown in Figure 2. We use both generators to ensure diverse descriptions of facial actions. For synthesizing descriptions of emotions, we only use the rule-based text generator.

Rule-based labeling for AU captions: To synthesize concise and complete descriptions of facial actions, we leverage the annotations of AUs because most facial muscle movements are depicted as AUs by FACS[6]. As shown in Fig. 2, given a facial image and its label that indicates the appearance of each AU, we combine the brief descriptions of the appearing AU and make a sentence in the form of "It is a people who <action in appearing AUs>". The brief descriptions of each AU are listed in the supplemental materials. If no AUs are labeled as appearing in a face, we do not synthesize rule-based AU descriptions for the image.

GPT-based labeling for AU captions: To ensure diverse training texts, we also utilize GPT-3.5 to synthesize descriptions for facial actions. However, the outputs of GPT-3.5 are uncontrolled and it might include the re-introduction of AUs, the induction of different emotions, and other irrelevant contents. To achieve a clear and concise description, we design a one-shot prompting strategy that requires GPT-3.5 to act as an AU expert and provides it with an example. Fig. 2 shows an example. We first claim the list of AUs' definitions and then state the requirement through a Question-Answer example. We explicitly add two restric-

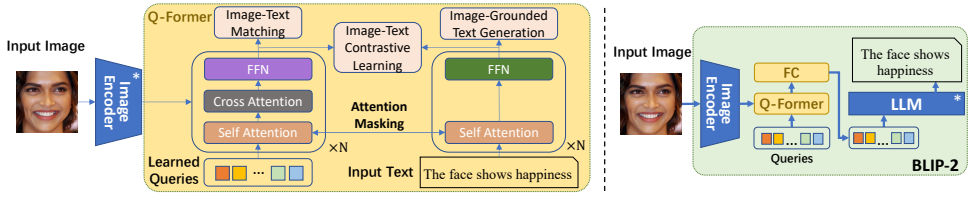

Figure 3: Two-step training procedure of BLIP-2[17]. The 1st step (**Left**): BLIP-2 enforces the queries to extract the visual representation most relevant to the text. The 2nd step(**Right**): BLIP-2 bootstraps from frozen LLMs to conduct vision-to-language generative training.

tions to regulate GPT-3.5's outputs, prohibiting the induction of emotion and the repetition of the word "AU" in the reply. Then, we throw out the question that asks GPT-3.5 to describe the face with the annotated AUs. If no AUs are labeled, we use a predetermined description to avoid inaccurate outputs from GPT-3.5, which is listed in the supplemental material.

Rule-based labeling for emotion captions: To keep the correct description of emotions, we synthesize the descriptions of emotions by simply generating a sentence in the form of "The face shows <emotion category>" according to the manual annotations in the existing datasets. If the images are labeled with more than one emotion category, e.g., compound emotions[18] or multi-label emotion category[25], we merge the involved emotion labels.

3.2 Training of AU-specific and emotion-specific image captioners

We use the synthetic pairs in Sec. 3.1 to train two image captioners: One describes facial actions (AU-BLIP), and the other describes emotions (Emot-BLIP). Both two captions are trained following the framework in BLIP-2[17]. In addition, other vision-to-language models can also be adopted in our three-stage facial expression captioning framework.

Revisiting BLIP-2: Fig. 3 illustrates the architecture of BLIP-2 and its two-step training procedure. As can be seen in Fig. 3 (Right), BLIP-2 takes an image as input, and then extracts the visual features from a pre-trained image encoder. The visual features are then passed through a Q-Former and a pre-trained large language model(LLM). Finally, BLIP-2 outputs a description of the input image. Q-Former is a trainable module that connects the image encoder and LLM. The detailed structure of Q-Former is illustrated in Fig. 3(Left). It is trained in two steps. In the first step, as shown in Fig. 3(Left), to guide the learned queries to aggregate the vision-language representation from the image encoder and training texts, BLIP-2 freezes the image encoder and optimizes Q-Former by minimizing the Image-Text Contrastive Learning (ITC) loss, Image-grounded Text Generation (ITG) loss and Image-Text Matching (ITM) loss. The self-attention layers for the queries and the texts share the same parameters. As BLIP-2[17] does, attention masking is applied to prevent information leakage. In the second step, to harvest the LLM's generative language capability and adjust the image encoder to extract features that are more related to the image captioning task, BLIP-2 connects Q-Former to a frozen LLM, as shown in Fig. 3 (Left), and jointly update the parameters of Q-Former and the image encoder by minimizing the discrepancy between the LLM-generated caption and the ground truth.

Pre-trained image encoder and LLM: For the image encoder, we utilize ViT-G/14 from EVA-CLIP [8] and a ViT-Base trained on AffectNet [32] dataset under the training framework of MAE [10]. Consistent with BLIP-2, we choose the unsupervised-trained OPT model family [46] as our language model. Model details are listed in supplementary material.

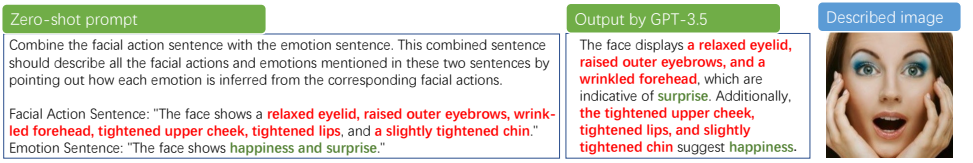

Figure 4: Inputs and outputs examples of the GPT-3.5 based generator with zero-shot prompt for fusing AU and emotion captions. The image described by the fused caption is on the right.

Table 1: Statistics of training and test data.(#sub:number of subjects; *:sampled set)

	Originally annotated with AUs					Originally annotated with emotions		
	BP4D	DISFA	GFT	RAF-AU	EmotioNet	AffectNet	RAF-DB	FaceME
train(#image/#sub)	16627*/28	14814*/24	17719*/78	3733/-	19046/-	287618/-	3162/-	10052/-
test(#image/#sub)	45805/13	14535/3	4034*/18	868/-	2117/-	4000/-	792/-	-

3.3 Exp-BLIP with fused captions

We train the final facial expression captioner (Exp-BLIP) with the synthesized descriptions by GPT-3.5 and AU-BLIP/Emot-BLIP. Fig. 1 (Right) shows how we train Exp-BLIP. To assign each training image a description of both AU and emotion, we use AU-BLIP/Emot-BLIP to generate a pseudo-caption if it lacks the original annotation on AU/Emotion. Then, we utilize GPT-3.5 to fuse the two captions with zero-shot prompting, leveraging the powerful inferring ability of GPT-3.5. Fig.4 shows an example of zero-shot prompting in fusing captions. With the fused captions, we train Exp-BLIP following the architecture and training procedure of BLIP-2[17] in Sec. 3.1.

4 Experiment

We evaluate the performance of AU-BLIP, Emot-BLIP, and Exp-BLIP from two perspectives: the quality of generated captions, and the ability of visual representation.

4.1 Experiment settings

Training data: We used nearly 372k training image-text pairs in total. 72k of the training data are derived from AU datasets BP4D [47], DISFA [30], GFT [9], RAF-AU [40] and EmotioNet [7] using the rule-based text generator and GPT-based generator in Sec. 3.1. Texts from the two generators were both used to train AU-BLIP. Only the GPT-3.5 generated ones were used to synthesize training pairs for Exp-BLIP. Considering the high cost of using GPT-3.5 API[1] and the significant similarity among the consecutive frames in video-based AU datasets, we first select one sample from every ten frames. Then, we selected the frames which had different AU labels compared with the previous frame to fully utilize the datasets. This selection was conducted in training sets of video-based AU datasets (BP4D, DISFA, and GFT) and test set of GFT. 300k of the training data are derived from AffectNet, RAF-DB [18] and FaceME [25]. All of the datasets except FaceME are split into training and test parts without overlapped subjects. Tab. 1 shows the statistics of training and test data.

Evaluation protocol for image captioning: We evaluated our image captioning models in two ways: (1) computing metrics for AU captioning, and (2) conducting manual evaluation to assess the completeness and correctness of facial expression captioning. The former was conducted on both in-the-lab data (test set of GFT) and in-the-wild data (test set of RAF-AU). we chose Meteor [2], Rouge [19] and Ada-similarity (Ada-sim) between the generated captions and the ground truths as our metrics. Ada-similarity calculated the cosine similarity of two features extracted from text-embedding-ada-002[2]. Since each sample has multiple

[1] https://platform.openai.com/docs/guides/gpt/chat-completions-api
[2] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Table 2: Metrics of AU captioning with different models (ViT-B/G denotes the used image encoder, 2.7B/6.7B denotes the used language model OPT-2.7B/6.7B)

Models	RAF-AU			GFT		
	Meteor ↑	Rouge ↑	Ada-sim(%) ↑	Meteor ↑	Rouge ↑	Ada-sim(%) ↑
BLIP-2(ViT-G,6.7B)	0.031	0.112	76.32	0.033	0.118	77.58
BLIP-2(ViT-G,6.7B)^{COCO}	0.034	0.129	77.31	0.034	0.131	78.47
AU-BLIP(ViT-B,2.7B)	0.236	0.457	90.13	0.278	0.509	93.11
AU-BLIP(ViT-G,2.7B)	0.254	0.478	90.23	0.279	0.487	93.45
AU-BLIP(ViT-G,6.7B)	**0.263**	**0.505**	**90.55**	**0.295**	**0.538**	**93.97**
Mix-BLIP(ViT-B,2.7B)	0.158	0.324	86.00	0.273	0.459	92.55
Mix-BLIP(ViT-G,6.7B)	0.204	0.364	87.87	0.279	0.489	93.00
Cat-BLIP(ViT-B,2.7B)	0.211	0.374	87.38	0.240	0.355	92.02
Cat-BLIP(ViT-G,6.7B)	0.235	0.426	89.76	0.248	0.385	92.88
Exp-BLIP(ViT-B,2.7B)	0.164	0.225	83.86	0.215	0.279	90.72
Exp-BLIP(ViT-G,6.7B)	0.184	0.229	84.65	0.217	0.318	90.75

ground truths (synthesized by GPT-3.5 and the rule-based generator), we computed the metrics between the predicted caption and each ground truth, and then we chose the highest score. The latter was conducted on 505 samples randomly chosen from all the test data in Tab.1. Then, we invited 9 human labelers to rate the correctness of each predicted caption from 1 to 10 and to choose the most/least complete one from the three predictions.

Evaluation protocol for visual presentation: We evaluated the visual representations from the image encoder and Q-Former by investigating their performance in AU detection and emotion classification tasks using a linear probe or a fine-tuning strategy. We added a fully connected layer after the representation as the classifier. The AU classifier was trained on full BP4D and RAF-AU training set without sampling, while the emotion classifier was trained on AffectNet. In the linear probe, the image encoder and Q-Former were frozen. In the fine-tuning strategy, the parameters of the image encoder and Q-Former were updated. F1 score and classification accuracy were used to measure the performance of AU detection and emotion classification, respectively. To compare with state-of-the-art (SOTA) method in RAF-AU, we selected the same AUs of AU-CNN[40] for training.

Implementation: During the training of AU-BLIP, Emot-BLIP, and Exp-BLIP, we used a batch size of 28 for the first step and 32 for the second step. We used the AdamW [24] optimizer and a weight decay of 0.05. A cosine learning rate decay with a peak learning rate of 1e-4 and a linear warm-up of 2k steps was adopted. The minimum learning rate in the second step was 5e-5. The input images were of size 224×224, augmented with random resized cropping and horizontal flipping. All the experiments run on 4 Nvidia A100 (40G).

4.2 Experiment results and analysis

Evaluation on image Captioning: Table 2 reports the metrics on AU captioning using the original BLIP-2[17], AU-BLIP, Exp-BLIP, Mix-BLIP, Cat-BLIP with different versions of image encoder and LLMs. Mix-BLIP and Cat-BLIP adopt different approaches to fuse AU caption and emotion caption(Sec. 4.3). BLIP-2 is the model provided by [17] that was trained on 129M images. BLIP-2^{COCO} is the fine-tuned BLIP-2 on COCO [20] for general image captioning. As shown in Table 2, our models trained on AU-specific data achieve consistently larger values than the two models trained for general image captioning, because the general image captioner describes limited information on facial expressions, as shown in Fig. 5. It indicates the necessity to train a task-specific vision-to-language model for describing facial expressions even though the general captioning model has shown powerful ability.

It is also observed in Table 2 that AU-BLIP has larger values in metric than the fused-caption models. It is because the ground truths are synthetic AU captions, while Mix-BLIP tends to randomly generate AU or emotion captions, and Cat/Exp-BLIP describe extra information about emotions, which degrades the metrics to AU captioning.

Figure 5 shows examples of the generated captions by different models. All the models are with image encoder ViT-G and language model OPT-6.7B. As can be seen, BLIP-2 describes very few facial expressions but describes other attributes of faces. Emot-BLIP describes emotions with limited facial actions. AU-BLIP only describes facial actions. Exp-BLIP not only describes both facial actions and emotions, but also describes the inferential relation between them. Our models produce inaccurate results in some cases. For example, in the failure case of figure 5, our Emot-BLIP and Exp-BLIP infer different emotions from the ground truth. Although labeled as contempt, this picture also slightly shows happiness as we predict. More failure cases for different reasons are listed in the *suppl. material*.

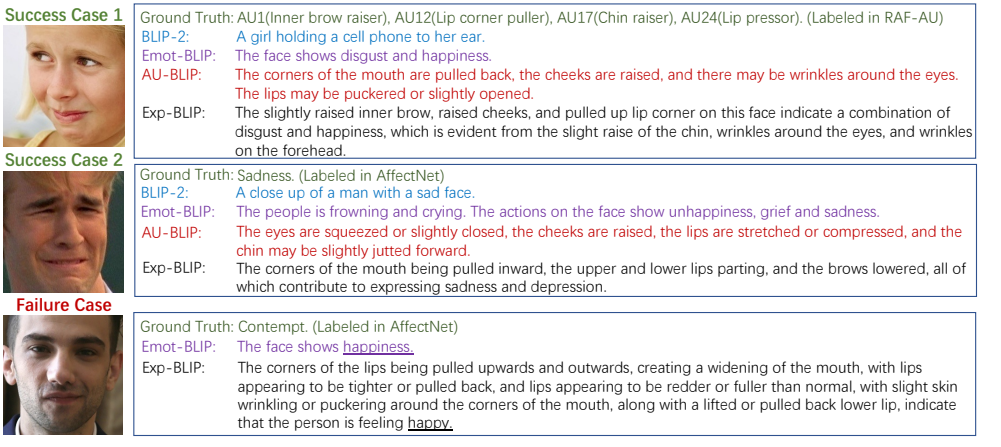

Figure 5: Examples of the generated captions by different models with ViT-G/OPT-6.7B

Evaluation on visual representation: Tables 3 and 4 report the performance of visual representations, including F1 scores on BP4D and RAF-AU datasets, and classification accuracy on AffectNet dataset. In addition, the comparisons with SOTA facial representation (FaRL[49]), AU detectors (Me-GraphAU[28], AU-CNN[40]), emotion classifier (EfficientNet-B2[35]) are included. ViT-G stands for the pre-trained image encoder mentioned in Sec. 3.2. The values are reported under a linear probe strategy. ViT-G$^{\text{fine-tune}}$ is ViT-G fine-tuned on the training data for AU detection or emotion classification, using a fine-tuning strategy. Other models with the names formatted as <ViT-G/QFormer>(AU/Emot/Mix/Cat/Exp) denote the visual representations (e.g., the image encoder or Q-Former) from the varied models (i.e., AU-BLIP, Emot-BLIP, Mix-BLIP, Cat-BLIP, Exp-BLIP) with LLM OPT-6.7B. For these representations and FaRL, linear probe strategies were applied. Other AU detectors and emotion classifiers used the values reported in their papers. It is noting that Me-GraphAU was evaluated under a 3-fold protocol on BP4D, while our models were evaluated on one of the three folds, the one that was not used in training AU/Exp-BLIPs. All the models were trained for 100 epochs on BP4D and 400 epochs on RAF-AU.

Tables 3 and 4 show that most of the ViT-G and Q-Former models with linear probe strategy in our approach achieve higher F1 scores than ViT-G and ViT-G$^{\text{fine-tune}}$. This suggests that incorporating language tasks improves the visual representations and enhances performance on related downstream tasks. It is also observed that the features of Q-Former are

Table 3: Performance of visual representation on BP4D and AffectNet. (*: original values)

Models	BP4D (F1 score×100)													AffectNet (*Acc.*%)
	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg	Emotion
FaRL[45]	48.4	45.9	47.5	78.9	69.8	82.8	86.0	58.3	42.0	55.9	36.1	38.0	57.5	38.8
Me-GraphAU*[28]	52.7	44.3	60.9	79.9	**80.1**	85.3	89.2	**69.4**	**55.4**	64.4	49.8	55.1	65.5	-
EfficientNet-B2*[65]	-	-	-	-	-	-	-	-	-	-	-	-	-	**63.0**
ViT-G	53.2	34.2	58.7	81.1	71.1	85.7	90.0	56.1	31.2	58.3	43.4	51.1	59.5	44.8
ViT-G$^{\text{fine-tune}}$	32.8	18.4	38.1	77.2	68.6	84.0	85.1	60.1	32.3	61.6	35.7	32.9	52.2	50.3
ViT-G(Emot)	44.1	28.7	54.0	80.5	76.2	88.1	90.5	64.9	44.9	55.2	43.6	50.5	60.1	50.9
QFormer(Emot)	43.7	21.0	54.8	80.7	78.8	87.6	**91.4**	58.7	43.4	52.7	34.8	52.1	58.3	52.7
ViT-G(AU)	57.7	54.8	**67.0**	81.1	75.6	86.6	90.1	60.2	51.8	64.6	49.4	56.8	66.3	50.1
QFormer(AU)	**59.9**	**55.9**	65.7	81.2	76.7	87.7	90.8	60.4	50.8	**64.8**	49.4	55.7	**66.6**	52.2
ViT-G(Mix)	59.0	47.6	62.3	80.8	76.3	86.4	87.9	64.0	45.6	57.2	44.3	53.3	63.7	48.2
QFormer(Mix)	58.7	46.3	61.7	81.5	77.0	85.1	90.2	64.2	43.7	59.0	45.2	55.7	64.0	49.3
ViT-G(Cat)	58.9	46.9	60.7	80.8	76.9	87.2	88.0	61.5	45.3	60.3	43.0	55.6	63.8	48.9
QFormer(Cat)	57.8	48.4	57.3	**82.1**	77.2	85.3	90.6	64.4	44.0	58.2	46.1	55.3	63.9	49.5
ViT-G(Exp)	59.2	51.7	62.5	80.9	78.1	88.1	90.2	63.4	48.5	62.8	47.4	**61.7**	66.2	51.0
QFormer(Exp)	**59.9**	52.9	61.3	81.3	78.7	**88.2**	90.7	64.5	47.6	61.7	**50.5**	57.0	66.2	51.6

Table 4: Performance of visual representation on RAF-AU. (*: original values)

Models	AU1	AU2	AU4	AU5	AU6	AU9	AU10	AU12	AU16	AU17	AU25	AU26	AU27	Avg
FaRL[45]	57.8	56.3	75.2	56.2	32.8	50.2	61.5	68.9	49.7	38.8	84.4	55.0	60.6	57.5
AU-CNN*[40]	60.5	65.6	73.4	69.7	**58.2**	67.4	68.4	69.6	59.4	25.6	92.1	64.7	**82.7**	65.9
ViT-G	60.0	57.6	77.2	62.3	35.8	59.7	70.6	73.6	53.6	45.5	88.7	56.3	70.1	62.4
ViT-G$^{\text{fine-tune}}$	27.6	34.5	59.1	52.5	23.4	38.5	48.9	51.0	41.9	25.0	86.1	46.5	62.6	46.0
ViT-G(Emot)	61.0	55.3	78.4	62.2	37.5	66.8	70.8	71.5	53.8	48.6	88.8	55.9	69.5	63.1
QFormer(Emot)	57.8	53.6	78.8	62.8	36.1	62.9	67.2	68.4	52.7	42.4	86.6	56.1	62.0	60.6
ViT-G(AU)	74.9	68.2	82.3	**73.9**	50.2	**79.0**	71.1	74.0	58.6	59.4	91.5	68.0	78.0	71.5
QFormer(AU)	75.2	**70.9**	82.2	71.7	47.2	76.1	**73.2**	74.3	59.3	**61.4**	95.6	**72.3**	79.1	**72.2**
ViT-G(Mix)	69.1	60.5	**85.2**	68.7	43.8	69.9	69.0	72.9	56.9	47.7	94.5	61.5	66.7	66.6
QFormer(Mix)	74.9	64.6	83.8	69.6	47.8	72.1	68.6	74.7	**59.6**	53.5	95.4	65.5	69.8	69.2
ViT-G(Cat)	71.8	64.4	84.6	70.4	39.2	70.0	67.9	74.2	56.0	49.4	95.8	60.9	65.8	67.0
QFormer(Cat)	**76.8**	67.1	**85.2**	72.4	44.1	72.3	66.6	**75.0**	57.6	58.4	**96.2**	62.3	68.7	69.4
ViT-G(Exp)	72.6	65.4	82.5	68.9	40.8	70.6	72.9	74.0	57.6	52.4	92.3	70.1	78.8	69.1
QFormer(Exp)	74.0	68.5	82.5	70.3	42.9	74.9	72.9	73.1	57.4	52.8	93.5	68.9	71.7	69.5

superior to those of the corresponding image encoder, except for Emot-BLIP for AU detection. This may be due to Q-Former's ability to encode extra discriminative information about facial actions or emotions through the learned queries. Q-Former(Emot) has a larger F1 score than ViT-G(Emot). This may be because Q-Former(Emot)'s features align more closely with its emotional semantics, resulting in reduced representation ability for capturing AU details. It is worth mentioning that ViT-G$^{\text{fine-tune}}$ performs less powerful than ViT-G on BP4D and RAF-AU, potentially because ViT-G$^{\text{fine-tune}}$ suffers from over-fitting. When trained on larger dataset (i.e., AffectNet), ViT-G$^{\text{fine-tune}}$ shows better performance.

It is also observed that AU-BLIP and Emot-BLIP reach the best performance respectively on AU detection and emotion classification. They outperform all the models with combined AUs and emotions. One possible reason is that the representation from AU-BLIP and Emot-BLIP contains the least irrelevant information, making it easier to fit a linear classifier.

In the AU detection task, the performance of QFormer(AU/Exp) is compatible to SOTA facial representation and AU detectors. However, in the emotion classification task, there is a gap between the performance of QFormer(Emot) and SOTA emotion classifier (i.e., EfficientNet-B2), which is worth further exploring. This may be attributed to the simpleness of synthetic emotion captions and the limited ability of the LLM (OPT-6.7B), resulting QFormer(Emot) benefiting little from the emotion captions.

Table 5: Correctness and completeness for different combining strategies

Models	Correctness ↑	The highest completeness(ratio) ↑	The lowest completeness(ratio) ↓
Mix-BLIP(ViT-B,2.7B)	**7.61**	3.43%	91.17%
Cat-BLIP(ViT-B,2.7B)	7.27	28.85%	7.91%
Exp-BLIP(ViT-B,2.7B)	**7.61**	**67.72%**	**0.92%**

4.3 Ablation study

Strategies for combining AU and emotion captions: Table 5 displays the assessments of the correctness and completeness of captions generated by three models, which fuse AU and emotion captions using different strategies. Mix-BLIP trains the facial expression captioner using the mixed data that with only AU captions or emotion captions. Cat-BLIP differs from Exp-BLIP by directly concatenating the two captions rather than fusing them with GPT-3.5 as Exp-BLIP does (Sec. 3.3). Mix-BLIP achieves the highest correctness while the least completeness among the three methods, which betrays our motivation of generating detailed facial descriptions. In Table 2, Mix-BLIP and Cat-BLIP have higher metrics for AU captioning than Exp-BLIP, which may be explained by the different formats of training texts. The AU captions in the training data of Mix-BLIP and Cat-BLIP are synthesized by rules, the same as the ground truths used in computing the metrics. Exp-BLIP is trained with the descriptions generated by GPT-3.5. In Table 3 and 4, the visual representations from Exp-BLIP show superior performance than those from the other two, indicating the advantage of our proposed combination strategy.

Scaling language and visual model: The comparison between methods with (ViT-B, OPT-2.7B), (ViT-G, OPT-2.7B), and (ViT-G, OPT-6.7B) in Table 2 suggests that in AU captioning, a stronger image encoder or a stronger LLM both lead to better performance, which is consistent with the conclusion in [17].

Zero-shot ability: Fig. 6 shows two examples of Exp-BLIP-generated descriptions of facial expressions. Exp-BLIP describes the unseen emotions (red) and facial actions (blue) within training data. The zero-shot ability is ascribed to the language model.

The corners of the mouth pulled upwards and outwards, causing wrinkles around the eyes, the cheeks being pushed up and outwards causing wrinkles around the jawline, and the lower eyelids appearing slightly puffed out, all indicate a happy emotion. Additionally, the slightly raised cheeks and tight lower eyelids can also imply happiness, while the presence of creases around the mouth and eyes may also be indicative of **a smirk** or slight smile, as well as possible dimples appearing around the cheeks. These facial actions combined with the upward and outwards pulling of the corners of the mouth suggest a feeling of **satisfaction**, and the raised cheeks and tight lower eyelids may also indicate a slight laugh. All of these facial actions combine to create an overall impression of happiness on this face.

The slight raising or opening of the eyelids and slight wrinkling of the eyebrows, along with the slight **elevation of the forehead**, indicate a feeling of sadness. The slightly raised upper lip and slightly lowered lower lip, along with the slight tension in the mouth, indicate a sense of **caution** or **worry**. The slight **furrow** between the eyebrows and the **raised skin around the eye** area indicate a sense of **fatigue** or **lack of energy**.

Figure 6: Examples of zero-shot ability of Exp-BLIP(ViT-G,OPT-6.7B)

5 Conclusion

We introduce facial expression captioning as a novel task that aims to capture the nuanced facial actions and emotional states of a given facial image in natural language. To tackle this task, we propose a three-stage training framework that employs synthetic image-text pairs to train a facial expression captioner called Exp-BLIP. The experimental results showcase the powerful ability of Exp-BLIP and the intermediate visual representation. Nonetheless, Exp-BLIP suffers from certain limitations inherited from the pre-trained large models, such as generating repeated, syntax errors, or harmful content, which are inevitable in a frozen language model. Exp-BLIP can be further enhanced by adopting more powerful LLMs and image encoders, and leveraging more data from other related tasks.

Acknowledgement

This work is supported by National Natural Science Foundation of China (No. 62176248). We also thank ICT computing platform for providing GPUs.

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. In *Advances in Neural Information Processing Systems(NeuriPS)*, 2022.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Annual Meeting of the Association for Computational Linguistics (ACL) workshops*, 2005.

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems(NeuriPS)*, 2022.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems(NeuriPS)*, 2020.

[5] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[6] P Ekman and W V Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[7] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.

[8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023.

[9] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *International Conference on Automatic Face and Gesture Recognition(FG)*, 2017.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022.

[11] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2021.

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning(ICML)*, 2021.

[13] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision(ECCV)*, 2022.

[14] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems(NeuriPS)*, 2022.

[15] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems(NeuriPS)*, 2021.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning(ICML)*, 2022.

[17] 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 *International Conference on Machine Learning(ICML)*, 2023.

[18] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision(ECCV)*, 2014.

[21] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision(ECCV)*, 2022.

[22] Bryson Lingenfelter. *Face Captioning Using Prominent Feature Recognition*. PhD thesis, University of Nevada, Reno, 2021.

[23] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Deep Learning Inside Out(DeeLIO)*, 2022.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations(ICLR)*, 2019.

[25] Zijia Lu, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Zero-shot facial expression recognition with multi-label label propagation. In *Asian Conference on Computer Vision(ACCV)*, 2019.

[26] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022.

[27] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *European Conference on Computer Vision(ECCV)*, 2022.

[28] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *International Joint Conference on Artificial Intelligence(IJCAI)*, 2022.

[29] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992.

[30] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[31] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In *Annual Meeting of the Association for Computational Linguistics(ACL)*, 2022.

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

[33] Omid Mohamad Nezami, Mark Dras, Stephen Wan, and Cecile Paris. Image captioning using facial expression and attention. *Journal of Artificial Intelligence Research(JAIR)*, 68:661–689, 2020.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning(ICML)*, 2021.

[35] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

[36] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaa-net: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision(IJCV)*, 129:321–340, 2021.

[37] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2020.

[38] Lingfeng Wang, Jin Qi, Jian Cheng, and Kenji Suzuki. Action unit detection by exploiting spatial-temporal and label-wise attention with transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022.

[39] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023.

[40] Wen-Jing Yan, Shan Li, Chengtao Que, Jiquan Pei, and Weihong Deng. Raf-au database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations. In *Asian Conference on Computer Vision(ACCV)*, 2020.

[41] Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. Ontology-enhanced prompt-tuning for few-shot learning. In *the ACM Web Conference*, 2022.

[42] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research(TMLR)*, 2022.

[43] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022.

[44] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision(ECCV)*, 2022.

[45] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023.

[46] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[47] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *International Conference on Automatic Face and Gesture Recognition(FG)*, 2013.

[48] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision(ECCV)*, 2022.

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022.

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision(IJCV)*, 130(9): 2337–2348, 2022.