

Supplementary material to *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

1 Descriptions of AU

The training AU captions are synthesized in two ways: the rule-based text generator, and the GPT-based generator with one-shot prompt. For the rule-based AU captions labeling, the brief descriptions of each AU are listed in Table 1 verb form. In order to ensure the syntactic correctness of synthesized captions, we appended s/es to the descriptions when merging them to the captions. If no AUs were labeled, we did not synthesize rule-based AU descriptions for the image. For GPT-based AU captions synthesizing, we first listed all 27 Action Units defined by the Facial Action Coding System (FACS), which were listed in Table 1 noun form. Consequently, we conducted one-shot prompting to achieve GPT-based AU captions. If no AUs were labeled, we used a predetermined GPT-synthesized description in Sec. 1.2 to avoid inaccurate outputs from GPT-3.5.

1.1 Brief descriptions of each AU

Table 1 shows the brief AU descriptions of noun form and verb form. The descriptions defined by FACS are of noun form. When synthesizing rule-based AU captions, the descriptions of verb form designed by us are more comprehensible for readers.

1.2 Predetermined description of no AUs

Due to the diversity of GPT-3.5's output and the high similarity of faces with no AUs labeled, we used a predetermined description which we selected from several outputs of GPT-3.5 for describing faces with no AUs. By conducting this strategy, we reduced the risk for GPT-3.5 to synthesize extra inaccurate descriptions for faces with no AUs. The predetermined description of no AUs is: *The eyes may be open and looking straight ahead, with the mouth closed or slightly open in a relaxed position. The forehead may be relatively smooth with minimal wrinkles or creases. The eyebrows may be in a natural position, not raised or furrowed.*

Table 1: Brief AU descriptions of noun form and verb form.

AU	Noun form	Verb form	AU	Noun form	Verb form
AU1	inner brow raiser	raise inner eyebrow	AU17	chin raiser	raise chin
AU2	outer brow raiser	raise outer eyebrow	AU18	lip pucker	pucker lips
AU4	brow lowerer	lower brow	AU19	tongue show	stick tongue out
AU5	upper lid raiser	raise upper lid	AU20	lip stretch	stretch lips
AU6	cheek raiser	raise cheek	AU21	neck tightener	tighten neck
AU7	lid tightener	tighten lids	AU22	lip funneler	show lip funneler
AU8	lips toward each other	pull lips toward each other	AU23	lip tightener	tighten lips
AU9	nose wrinkler	wrinkle nose	AU24	lip pressor	press lips
AU10	upper lip raiser	raise upper lip	AU25	lips part	separate lips
AU11	nasolabial deepener	deepen nasolabial	AU26	jaw drop	drop jaw
AU12	lip corner puller	pull lip corner	AU27	mouth stretch	stretch mouth
AU14	dimpler	tighten lip corner	AU28	lip suck	suck lips
AU15	lip corner depressor	depress lip corner	AU43	eyes closed	close eyes
AU16	lower lip depressor	depress lower lip			

2 Datasets

We used nearly 372k training image-text pairs in total. 72k of the training data are derived from AU datasets BP4D [10], DISFA [11], GFT [12], RAF-AU [13] and EmotioNet [14]. Considering the high cost of using GPT-3.5 API¹ 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 [15], RAF-DB [16] and FaceME [17]. The detail introduction for each dataset is in the following.

BP4D simultaneously records 2D and 3D facial expression videos in the lab. This dataset includes 41 participants (23 females and 18 males) with age ranging from 18 to 29 years. The released videos document the facial expression changes of the participants during eight different tasks. There are about 146,000 frames with 12 AU labels in the provided 2D videos. In our experiment, we split these frames into the training and test parts without overlapped subjects. In each video of training set, we selected one sample from every ten frames and also collected the samples which had different AU labels compared to the previous frame. Ultimately, there are 16627 frames of 28 subjects for training and 45805 frames of 13 subjects for testing. When evaluating the visual presentation in Tab.3 in the main text, we used the full training set.

DISFA collected spontaneous facial expressions of 27 participants while watching movie clips in the lab, annotating both the five intensity levels and 12 AU labels of each frame image. each video from the participant contains 4845 facial expression images and there are about 130,000 images in DISFA. We split these frames into the training and test parts without overlapped subjects. In each video of training set, we selected one sample from every ten frames and also collected the samples which had different AU labels compared to the previous frame. Ultimately, there are 14814 frames of 24 subjects for training and 14535 frames of 3 subjects for testing.

¹<https://platform.openai.com/docs/guides/gpt/chat-completions-api>

GFT records facial expression variations in 32 groups of three individuals during social gatherings, including 96 participants. It is the first dataset to capture facial expressions of multiple individuals in natural communication and interactive scenarios. The dataset we prepared comprises approximately 133,000 frames annotated with 10 AUs. In our experiment, we split these frames into the training and test parts without overlapped subjects. In each video, we selected one sample from every ten frames and also collected the samples which had different AU labels compared to the previous frame. Ultimately, there are 17719 frames of 78 subjects for training and 4034 frames of 18 subjects for testing.

EmotionNet contains over 1,000,000 facial expression data downloaded from the internet. It includes manual annotations for 50,000 images, specifying 11 AUs. The manually annotated images are divided into two parts: a validation set (publicly released) and a test set (not publicly released). Additionally, approximately 900,000 facial expression images were annotated using automated methods for AU and emotion labels, but these labels may contain noise. To ensure the accuracy of the data, we used the manual annotated images (approximately 21,000 images) labeled with 11 AUs for our experiment. We randomly split these images into the training and test parts. As a result, there are 19046 images in the training dataset and 2117 images in the test dataset.

RAF-AU contains 4,601 facial images annotated with 32 AUs in the wild. Some AU annotations distinguish the AU of upper, lower, left, and right regions of the face, by adding "T", "B", "L", "R" to the AU numbers, such as "17+B22+T24". In our experiment, we did not distinguish this different regions AU and removed the region signs, suggesting that the presence of AU in any region of the face was regarded as the presence of that AU. In addition, we only used the AUs shown in Table 1 in this dataset. We randomly split all images into the training and test parts. As a result, there are 3733 images in the training dataset and 868 images in the test dataset.

AffectNet is a multi-class dataset collected from internet, including 287,618 training images annotated with 8 basic emotion classes: neutrality, happiness, sadness, anger, surprise, fear, disgust, contempt. Each AffectNet image only has one of the 8 emotion labels. In our experiment, we used all the official training and test images. The test dataset comprises 4000 images, with each of the 8 basic classes containing 500 images.

RAF-DB is a multi-class dataset containing of a 7-class basic emotion part and a 12-class compound emotion part. In our experiments, we used all the compound label part, where the labels were formed by combining 2 of the 6 classes: happy, sad, anger, surprise, fear, disgust. We randomly split all the 3954 images into the training and test parts. As a result, there are 3162 images for training and 792 images for testing.

FaceME is a multi-label facial expression dataset collected from the internet, containing 10,062 images and 85 labels. The labels not only include emotions, but also labels about action, health, and inward thoughts. Each image is labelled by 3 annotators. To enhance the credibility of the labels, we only used the labels that are annotated presence by at least 2 annotators. In our experiment, we used all images for training without dividing test dataset.

3 Visual representations

The first part of this section re-conducted the visual representations experiment introduced in Sec. 4.2 by combing all the five AU training sets and further conducted this experiment with a fine-tuning strategy. The second part discusses the comparison with SOTA AU detector (DGCN[1]) measured by *Acc.%* on RAF-AU.

Table 2: Performance on BP4D and AffectNet. (ViT-B/OPT-2.7B)

Models	BP4D (F1 score×100)													AffectNet (Acc.%)	
	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg	Emotion	
ViT-B	46.3	38.2	50.8	79.3	73.4	85.4	87.9	69.3	24.0	53.0	0.9	24.2	52.7	37.8	
ViT-B ^{fine-tune}	55.1	65.3	64.8	53.3	59.4	83.9	88.4	5.9	43.5	22.2	96.0	87.3	60.4	51.4	
ViT-B(Emot)	47.0	37.5	52.9	79.0	73.0	85.8	88.6	68.0	24.3	54.3	7.7	21.9	53.3	38.2	
QFormer(Emot)	52.1	44.7	66.7	80.8	74.8	85.2	88.6	50.4	36.9	59.2	14.3	28.5	56.9	54.3	
ViT-B(AU)	47.3	39.6	59.6	78.9	73.3	85.7	88.6	68.2	23.3	52.7	2.1	15.8	52.9	38.8	
QFormer(AU)	56.4	45.6	57.6	80.3	74.3	86.3	<u>89.9</u>	61.8	41.1	61.7	<u>33.2</u>	<u>42.1</u>	60.9	45.2	
ViT-B(Mix)	47.6	39.0	51.4	79.7	73.2	87.1	89.4	67.1	21.2	53.1	3.0	20.1	52.7	38.6	
QFormer(Mix)	53.0	<u>50.1</u>	52.4	<u>81.0</u>	74.8	87.1	90.3	<u>68.6</u>	49.8	56.6	25.0	35.0	60.3	52.1	
ViT-B(Cat)	48.5	36.6	55.6	78.2	73.3	86.8	88.6	64.2	24.9	52.9	6.2	16.6	52.7	39.8	
QFormer(Cat)	54.2	43.7	56.7	81.1	75.9	87.0	88.9	66.2	<u>47.5</u>	56.2	19.6	22.2	58.3	48.1	
ViT-B(Exp)	46.3	38.0	57.0	78.8	72.4	86.0	88.3	65.4	26.3	53.8	7.2	18.8	53.2	39.6	
QFormer(Exp)	<u>55.7</u>	44.8	58.3	80.5	<u>75.1</u>	86.9	89.6	67.4	43.7	56.4	24.0	41.1	60.3	48.2	

Table 3: Performance on RAF-AU. (ViT-B/OPT-2.7B)

Models	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU15	AU16	AU17	AU22	AU24	AU25	AU26	AU27	Avg
ViT-B	48.4	49.9	67.9	23.8	32.4	19.7	25.1	47.5	69.1	15.5	24.5	18.4	2.0	16.9	86.7	32.3	58.9	37.6
ViT-B ^{fine-tune}	66.3	53.9	79.5	<u>67.1</u>	36.7	32.5	79.6	72.9	66.4	46.8	57.7	68.0	19.6	26.5	<u>94.7</u>	67.7	81.8	59.9
ViT-B(Emot)	48.1	50.5	68.9	27.2	34.8	23.0	28.0	50.2	71.1	16.4	27.6	19.8	2.2	18.7	87.4	29.2	65.3	39.3
QFormer(Emot)	54.2	53.9	72.8	58.3	35.7	<u>26.7</u>	53.4	62.8	62.4	26.1	49.5	49.9	<u>5.3</u>	11.7	92.5	42.1	<u>76.6</u>	49.1
ViT-B(AU)	50.3	52.6	71.0	30.8	29.6	22.2	35.5	52.4	70.8	16.1	29.8	28.1	4.4	<u>19.2</u>	87.6	36.2	67.1	41.4
QFormer(AU)	<u>66.6</u>	<u>66.1</u>	78.0	63.3	31.5	25.0	69.7	54.5	61.5	19.6	39.6	<u>62.5</u>	3.6	9.7	<u>94.7</u>	<u>55.3</u>	69.6	51.2
ViT-B(Mix)	47.9	53.3	68.5	32.0	34.4	25.9	29.3	51.6	69.7	15.9	27.1	27.4	2.4	18.3	88.7	33.6	65.4	40.7
QFormer(Mix)	67.9	68.4	<u>79.2</u>	62.5	39.1	18.0	77.2	<u>65.9</u>	<u>72.3</u>	<u>33.6</u>	<u>54.0</u>	45.9	0.0	8.7	92.5	53.9	<u>76.6</u>	<u>53.9</u>
ViT-B(Cat)	49.5	52.0	69.1	33.8	33.2	22.4	35.1	48.8	71.1	16.5	22.2	26.5	2.5	18.0	88.8	31.6	65.8	40.4
QFormer(Cat)	66.1	65.3	77.2	60.9	30.3	8.4	<u>78.2</u>	63.6	59.6	25.7	38.8	60.6	0.0	12.6	<u>94.7</u>	51.9	71.8	50.9
ViT-B(Exp)	48.7	52.7	68.6	30.2	32.9	25.9	32.0	52.8	70.8	16.1	23.5	23.2	4.6	<u>19.2</u>	88.4	35.4	66.4	40.7
QFormer(Exp)	66.2	63.6	76.1	68.2	<u>38.2</u>	12.2	74.3	60.2	72.7	22.3	38.2	54.9	2.2	9.1	95.5	46.1	65.4	50.9

Table 4: Performance on BP4D and AffectNet. (ViT-G/OPT-6.7B)

Models	BP4D (F1 score×100)													AffectNet (Acc.%)	
	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg	Emotion	
ViT-G	34.2	16.7	48.6	80.1	70.0	85.9	89.6	59.6	15.7	51.8	19.1	17.1	49.0	44.8	
ViT-G ^{fine-tune}	44.5	30.6	57.4	79.4	69.2	86.0	88.8	62.1	39.2	59.9	41.7	30.9	57.5	50.3	
ViT-G(Emot)	35.5	16.2	53.9	78.1	73.8	87.5	89.7	70.5	35.0	56.7	33.0	21.1	54.3	50.9	
QFormer(Emot)	5.9	3.8	43.6	71.7	74.4	82.9	87.7	30.0	4.3	9.8	0.1	12.4	35.6	52.7	
ViT-G(AU)	56.3	49.7	<u>66.2</u>	79.5	75.3	85.9	89.0	59.9	<u>46.0</u>	70.4	<u>41.2</u>	40.0	<u>63.3</u>	50.1	
QFormer(AU)	<u>58.2</u>	<u>50.6</u>	67.2	79.3	75.8	<u>86.6</u>	89.8	63.9	50.8	<u>65.1</u>	36.5	42.4	63.9	52.2	
ViT-G(Mix)	55.1	41.2	56.4	76.9	73.8	83.2	84.5	57.7	43.7	59.0	34.3	36.4	58.5	48.2	
QFormer(Mix)	54.5	43.8	56.2	79.9	75.7	84.8	89.5	<u>68.3</u>	44.9	57.2	37.4	33.8	60.5	49.3	
ViT-G(Cat)	52.5	40.1	55.0	73.6	73.8	82.4	87.7	62.9	45.3	58.9	30.8	<u>43.3</u>	58.9	48.9	
QFormer(Cat)	54.9	44.4	49.8	80.1	76.0	85.8	89.8	67.2	42.9	58.0	33.2	36.0	59.8	49.5	
ViT-G(Exp)	59.0	50.2	59.5	78.5	<u>77.7</u>	86.5	89.4	64.3	36.0	62.5	39.9	55.6	<u>63.3</u>	51.0	
QFormer(Exp)	57.5	54.1	58.7	79.4	78.1	86.5	89.3	64.2	39.9	62.9	32.1	40.5	61.9	51.6	

Table 5: Performance on RAF-AU. (ViT-G/OPT-6.7B)

Models	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU15	AU16	AU17	AU22	AU24	AU25	AU26	AU27	Avg
ViT-G	48.9	50.0	71.0	65.4	26.1	12.1	64.5	53.1	60.7	9.8	50.0	48.9	10.4	9.7	90.0	28.4	75.1	45.5
ViT-G ^{fine-tune}	67.4	64.5	78.7	64.5	34.6	<u>29.5</u>	67.2	68.7	67.2	45.8	61.2	52.4	25.2	<u>23.7</u>	94.2	65.6	81.5	<u>58.3</u>
ViT-G(Emot)	49.8	29.9	74.7	66.5	14.9	9.9	59.3	59.4	58.7	28.9	51.8	47.6	5.3	14.0	88.7	41.0	74.2	45.6
QFormer(Emot)	30.4	9.6	73.1	58.9	8.4	16.5	47.1	47.0	49.9	0.0	34.0	5.0	0.0	0.0	86.9	16.1	73.4	32.7
ViT-G(AU)	72.4	65.4	80.7	<u>75.9</u>	47.8	27.3	76.2	60.3	63.2	35.4	53.1	55.4	30.4	15.9	93.6	58.1	77.3	58.1
QFormer(AU)	74.6	72.1	81.5	71.2	<u>46.5</u>	40.5	75.3	<u>64.0</u>	<u>66.8</u>	45.7	46.7	69.0	28.3	32.4	94.3	<u>61.5</u>	78.3	61.7
ViT-G(Mix)	76.1	66.7	83.6	77.3	29.8	26.7	73.4	49.1	57.6	48.2	36.8	36.1	2.8	10.5	94.9	38.1	74.7	51.9
QFormer(Mix)	<u>79.9</u>	66.4	<u>84.4</u>	74.4	25.9	24.3	75.5	56.2	62.9	36.5	37.4	54.5	0.0	3.5	94.6	32.4	<u>79.2</u>	52.2
ViT-G(Cat)	77.4	66.4	85.3	75.2	28.0	23.7	75.0	49.1	65.2	<u>46.0</u>	30.8	51.7	2.8	3.4	<u>95.3</u>	43.6	74.8	52.6
QFormer(Cat)	82.0	<u>71.3</u>	<u>84.4</u>	72.8	24.6	21.4	74.7	54.9	63.7	38.1	40.9	58.0	0.0	11.4	95.4	36.1	74.2	53.2
ViT-G(Exp)	70.6	66.9	80.1	71.9	40.9	21.4	75.8	59.5	58.2	45.5	<u>57.8</u>	<u>58.8</u>	22.5	9.8	93.4	60.4	74.7	57.0
QFormer(Exp)	71.8	70.9	81.8	70.6	32.8	25.6	75.0	63.6	58.7	26.5	49.5	56.5	22.7	7.0	93.5	56.8	73.2	55.1

Table 6: Performance of fine-tuning on BP4D and AffectNet. (ViT-B/OPT-2.7B)

Models	BP4D (F1 score×100)														AffectNet (Acc.%)	
	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg		Emotion	
ViT-B ^{fine-tune}	55.1	65.3	64.8	53.3	59.4	83.9	88.4	5.9	43.5	22.2	96.0	87.3	60.4		51.4	
ViT-B(Emot) ^{fine-tune}	55.2	44.0	60.0	78.4	73.2	85.6	89.4	62.0	44.9	61.5	40.2	44.1	61.5		52.9	
QFormer(Emot) ^{fine-tune}	55.1	47.5	62.4	77.8	71.0	86.1	89.5	60.2	47.2	61.6	43.9	47.4	62.5		54.7	
ViT-B(AU) ^{fine-tune}	55.6	44.4	62.5	77.8	72.6	85.5	89.3	58.3	45.5	61.6	39.2	47.8	61.7		51.9	
QFormer(AU) ^{fine-tune}	54.5	48.8	62.4	<u>80.1</u>	<u>75.2</u>	85.8	90.0	66.9	49.6	64.5	46.5	45.0	<u>64.1</u>		53.6	
ViT-B(Mix) ^{fine-tune}	<u>55.8</u>	43.6	59.2	78.5	73.8	86.1	89.5	<u>64.7</u>	43.6	60.6	39.9	47.0	61.9		52.6	
QFormer(Mix) ^{fine-tune}	56.9	50.9	<u>65.1</u>	78.6	74.1	85.3	90.9	60.5	<u>49.8</u>	62.6	44.8	47.2	63.9		54.3	
ViT-B(Cat) ^{fine-tune}	54.1	45.8	61.8	78.3	72.9	85.4	89.1	62.0	44.4	60.4	42.5	46.2	61.9		52.1	
QFormer(Cat) ^{fine-tune}	54.6	47.3	63.4	78.8	74.0	85.4	89.7	63.0	48.9	64.1	<u>46.8</u>	47.2	63.6		54.3	
ViT-B(Exp) ^{fine-tune}	55.6	43.4	60.0	77.8	73.2	86.0	89.9	61.9	44.1	61.6	42.5	41.7	61.5		52.1	
QFormer(Exp) ^{fine-tune}	<u>55.8</u>	<u>51.3</u>	67.0	80.2	75.5	85.9	<u>90.1</u>	62.1	50.3	64.5	45.7	43.8	64.4		54.8	

Table 7: Performance of fine-tuning on RAF-AU. (ViT-B/OPT-2.7B)



Models	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU15	AU16	AU17	AU22	AU24	AU25	AU26	AU27	Avg
ViT-B ^{fine-tune}	66.3	53.9	79.5	67.1	36.7	32.5	79.6	72.9	66.4	46.8	57.7	68.0	19.6	26.5	94.7	67.7	81.8	59.9
ViT-B(Emot) ^{fine-tune}	68.5	67.3	81.6	62.5	28.2	27.6	76.6	69.9	66.3	49.1	51.1	64.9	9.6	26.8	95.4	56.2	78.4	57.6
QFormer(Emot) ^{fine-tune}	75.0	71.4	83.6	74.0	46.6	31.2	81.8	71.9	<u>74.0</u>	60.3	65.7	72.9	19.8	23.5	95.8	72.9	86.5	65.1
ViT-B(AU) ^{fine-tune}	69.3	66.7	82.3	67.5	28.3	32.2	75.5	71.3	64.6	49.1	46.8	62.4	18.2	20.9	95.3	54.6	77.5	57.8
QFormer(AU) ^{fine-tune}	73.9	69.1	83.9	74.5	<u>48.9</u>	<u>36.8</u>	<u>81.3</u>	77.2	70.8	<u>63.5</u>	66.4	<u>74.9</u>	26.2	39.0	95.7	<u>74.7</u>	84.8	67.2
ViT-B(Mix) ^{fine-tune}	70.4	66.9	82.9	68.4	23.6	28.6	77.2	71.3	65.0	48.1	45.8	63.5	17.6	34.2	95.0	49.8	75.2	57.9
QFormer(Mix) ^{fine-tune}	79.1	80.4	83.9	77.6	51.5	30.8	81.1	74.9	73.2	68.8	<u>69.4</u>	75.0	34.3	37.6	96.6	77.8	<u>85.3</u>	69.3
ViT-B(Cat) ^{fine-tune}	71.2	67.4	81.6	67.8	25.8	22.2	77.0	70.1	66.7	54.5	49.3	66.3	18.4	34.8	95.0	54.8	75.7	58.7
QFormer(Cat) ^{fine-tune}	<u>75.2</u>	70.2	82.5	75.1	42.3	30.4	79.6	76.4	71.2	62.5	70.5	74.0	<u>27.4</u>	<u>37.8</u>	95.2	73.1	84.3	66.4
ViT-B(Exp) ^{fine-tune}	70.4	66.9	82.8	65.9	26.8	33.9	76.9	68.5	64.3	48.1	44.6	62.2	20.2	28.0	95.0	50.0	74.2	57.6
QFormer(Exp) ^{fine-tune}	71.7	<u>72.3</u>	<u>82.2</u>	<u>77.0</u>	47.5	44.1	80.1	74.3	75.1	62.1	68.7	74.8	25.3	35.5	<u>96.1</u>	74.1	85.0	<u>67.4</u>

3.1 Combing AU training sets

Tables 2 and 3 report the performance of visual representations, including F1 scores on BP4D and RAF-AU datasets, and classification accuracy on AffectNet dataset. Different from the experiment in Sec. 4.2, these classifiers were trained on the combination of full EmotionNet, RAFAU training sets and sampled BP4D, DISFA and GFT training sets, rather than single datasets. The sampling strategy is introduced in Sec. 2. In addition, we selected more AUs in RAF-AU for the experiment. ViT-B stands for the pre-trained image encoder. The values are reported under a linear probe strategy. ViT-B^{fine-tune} is ViT-B 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-B/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). For these representations, linear probe strategies were applied. The language model trained with ViT-B/QFormer is OPT2.7B. All the models were trained for 20 epochs except for ViT-B^{fine-tune} which was trained for 50 epochs. Tables 4 and 5 report the performance of classifiers trained following the same settings above, except for the use of image encoder ViT-G and language model OPT-6.7B.

Tables 2 and 3 show that all of the ViT-B and Q-Former models with linear probe strategy in our approach outperform the ViT-B baseline. 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 superior to those of the corresponding image encoder, suggesting a stronger visual representation of Q-Former than the image encoder. This observation is consistent with the main text and Tables 4 & 5.


Table 8: Performance of visual representation on RAFAU. (Acc.%; *: original values.)

Models	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU12	AU17	AU23	AU24	AU25	Avg	Avg(F1)
FaRL()	69.0	71.0	76.6	70.6	61.6	68.8	69.2	75.0	70.3	66.7	81.1	69.2	70.8	55.7
DGCN*()	79.8	87.1	68.5	85.5	91.9	93.4	81.1	79.8	86.5	98.4	96.6	65.9	<u>84.5</u>	-
ViT-G	72.7	76.2	80.5	76.6	69.5	70.6	77.6	81.2	74.4	76.4	71.2	89.3	76.3	53.5
ViT-G ^{fine-tune}	83.5	89.3	80.9	<u>87.3</u>	<u>89.7</u>	<u>82.3</u>	75.7	<u>84.6</u>	73.5	85.5	36.1	58.3	77.2	45.2
ViT-G(Emot)	73.4	72.5	82.3	77.6	71.0	65.2	81.3	80.6	76.6	75.0	68.9	87.0	76.0	53.0
QFormer(Emot)	70.6	74.8	80.8	80.1	73.0	73.0	81.5	77.1	73.5	84.9	72.6	85.1	77.2	52.2
ViT-G(AU)	85.5	84.1	85.1	87.8	85.5	72.4	91.9	83.5	83.9	76.8	76.2	90.1	83.6	<u>59.9</u>
QFormer(AU)	87.2	87.4	84.1	84.1	82.9	70.7	<u>90.2</u>	84.0	<u>85.8</u>	81.9	<u>82.6</u>	94.7	84.6	60.2
ViT-G(Mix)	77.6	78.0	84.8	80.4	78.1	72.5	85.0	82.3	74.0	93.8	74.2	94.5	81.3	57.4
QFormer(Mix)	82.8	80.3	<u>85.1</u>	82.8	83.3	75.8	87.3	83.3	80.2	<u>94.0</u>	77.5	94.5	83.9	59.4
ViT-G(Cat)	80.9	79.8	84.9	81.7	77.8	73.7	85.0	82.0	75.1	92.3	74.2	<u>95.2</u>	81.9	57.2
QFormer(Cat)	<u>86.3</u>	84.2	85.4	85.1	81.5	73.3	86.4	<u>84.6</u>	83.1	85.6	79.5	95.9	84.2	59.5
ViT-G(Exp)	82.4	83.3	84.4	85.6	76.8	69.4	86.2	84.4	80.8	72.8	72.1	91.5	80.8	57.5
QFormer(Exp)	85.7	<u>87.8</u>	85.3	83.9	79.6	71.8	89.2	84.8	79.3	77.0	76.3	93.3	82.8	59.7

Compared with ViT-B baseline with linear probe strategy, ViT-B^{fine-tune} with fine-tuning strategy achieves better performance on both AU detection and emotion classification tasks. To explore the advantages of fine-tuning strategy, we fine-tuned each model in Tables 2 and 3 and the results are shown in Tables 6 and 7. The settings of the models in Tables 6 and 7 are the same as those in Tables 2 and 3 except for the fine-tuning strategy.

It is observed that most of the ViT-B and Q-Former models with fine-tuning strategy in our approach perform better than the ViT-B^{fine-tune} baseline, further indicating the superiority of our approach in visual representation. Comparing the linear probe results in Tab. 2 and 3 with the corresponding fine-tuning results in Tab. 6 and 7, all the fine-tuning results show an improvement over the linear probe results, indicating that fine-tuning is a more effective strategy for extracting excellent features when the diversity of training data is large.

3.2 Comparison with DGCN measured by Accuracy

Table 8 reports the classification accuracy on RAF-AU dataset, including the comparison with the SOTA AU detector (DGCN()). It is observed that the performance of QFormer(AU) is compatible with that of DGCN. However, classification accuracy is not a prevalent metric for AU detection evaluation, because of the unbalanced presence of each AU. Consequently, we add the Average F1 score of each method to Table 8. The observation of performance evaluated by F1 score is consistent with those in the main text.

4 Examples

Due to the length constraints of the main text, we are unable to provide examples of Mix-BLIP, Cat-BLIP and Exp-BLIP in the main text. In order to intuitively present the difference of the outputs of Mix-BLIP, Cat-BLIP and Exp-BLIP, We show several examples of the three models' outputs in Figure 1. It can be observed that Mix-BLIP describes only one respect of AU and emotion randomly. Cat-BLIP describes both AU and emotion while no connection between them. Exp-BLIP describes not only both facial actions and emotions, but also the inferential relation between them.

Although our models demonstrate promising potential in describing facial expressions, they occasionally produce inaccurate results. Figure 2 illustrates three kinds of bad cases of our models and the bad sentences are underlined. The first case is caused by large area

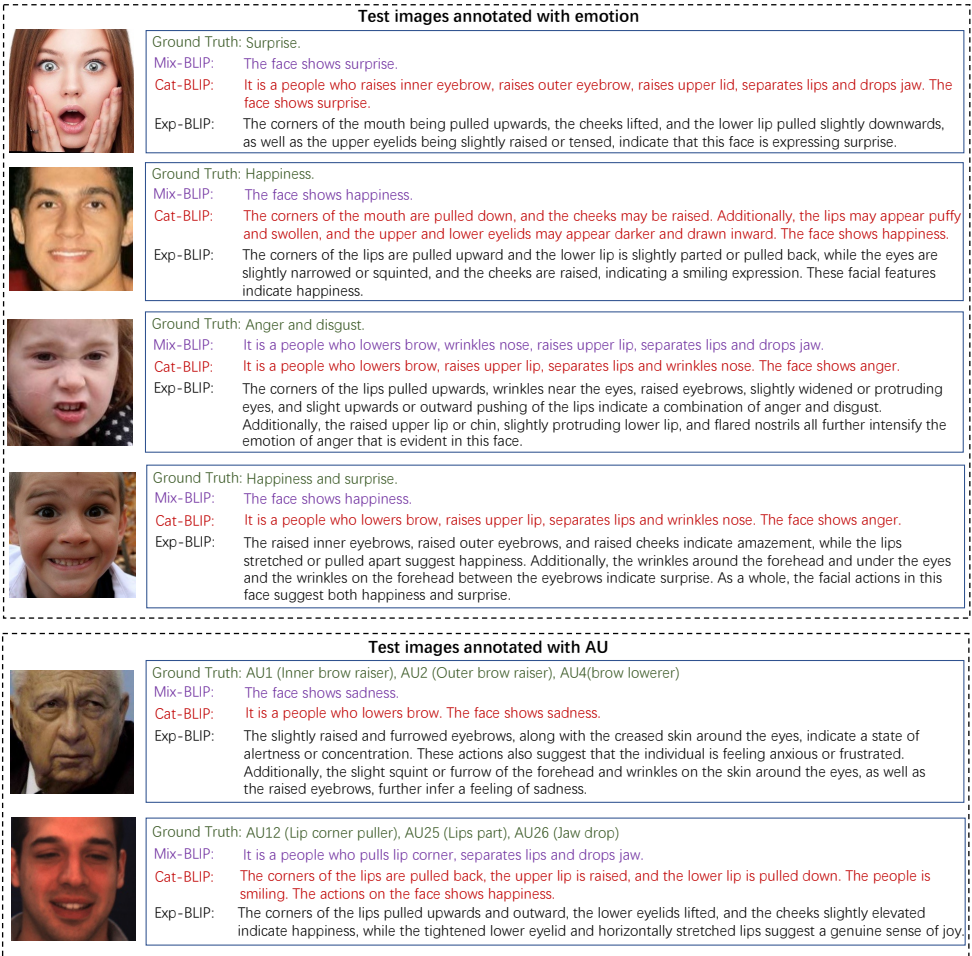


Figure 1: Examples of the generated captions by different models with ViT-G/OPT-6.7B

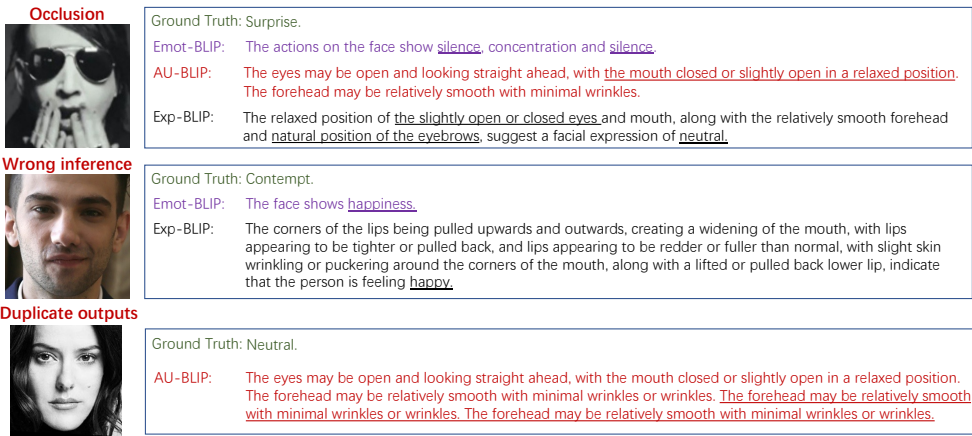


Figure 2: Bad cases of the generated captions for emotion-annotated images by different models with ViT-G/OPT-6.7B

of facial occlusion. This occlusion restricts the ability of image encoder to extract detail facial information, resulting a wrong description for facial expressions. The second case is the wrong inference. Our Emot-BLIP and Exp-BLIP infer different emotions from the ground truth, which may be due to the limited capabilities of QFormer and the language model (OPT). Although labeled as contempt, this picture also slightly shows happiness as we predict. The third case is generating duplicate outputs. The outputs of our models inherit the drawbacks of language model. OPT model sometimes tends to generate repeated and syntax error sentences, which will appear in the outputs of our models.

5 Hyperparameters

Table 9 and Table 10 show the Hyperparameters of image encoders and LLMs used in the main text, separately. ViT-Base was trained on AffectNet dataset for 200 epochs under the training framework of MAE and a cosine learning rate decay with a peak learning rate of $1e-4$ was adopted. The architecture of Q-Former in Figure 3 in the main text is BERT-Base, with cross-attention layer inserted in each block. FFN means Feed Forward Network in the transformer block. The Hyperparameters of BERT-Base is listed in Table 11. Table 11 shows the Hyperparameters of stage2 of fine-tuning BLIP-2 introduced in Sec. 3.2 of the main text.

Table 9: Details of image encoder model (Vision Transformer) variants.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base(ViT-B)	12	768	3072	12	86M
ViT-Giant(ViT-G)	39	1408	6144	16	1843M

Table 10: Hyperparameters for Large Language Model OPT2.7/OPT6.7 architecture.

LLM	$OPT_{2.7B}$	$OPT_{6.7B}$
FFN dim	10240	16384
Hidden size	2560	4096
Word embedding project dim	2560	4096
Attention heads	32	
Layers	32	
Dropout	0.1	
Max position embeddings	2048	

Table 11: Hyperparameters for BERT-Base architecture.

Model	$BERT_{Base}$
Hidden size	768
Intermediate size	3072
Attention heads	12
Layers	12
max position embeddings	512

Table 12: Hyperparameters for fine-tuning BLIP-2 with ViT-B/ViT-G on image captioning.

LLM	$OPT_{2.7B}$	$OPT_{6.7B}$
Fine-tuning epochs	20	
Warmup steps	1000	
Learning rate	$1e-5$	
Batch size	32	
AdamW β	(0.9,0.999)	
Weight decay	0.05	
Drop path	0	
Image resolution	224	
Prompt	""	
Inference beam size	5	
Layer-wise learning rate decay for ViT	1	0.95

References

- [1] 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.
- [2] 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.
- [3] Xibin Jia, Shaowu Xu, Yuhan Zhou, Luo Wang, and Weiting Li. A novel dual-channel graph convolutional neural network for facial action unit recognition. *Pattern Recognition Letters*, 166:61–68, 2023.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.