Prompting Visual-Language Models for Dynamic Facial Expression Recognition (Supplementary Material)

Zengqun Zhao
zengqun.zhao@qmul.ac.uk
Ioannis Patras
i.patras@qmul.ac.uk

School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London, E1 4NS, UK

A Facial Expression Descriptions

Expressions	Descriptors
happiness	a smiling mouth, raised cheeks, wrinkled eyes, and arched eyebrows.
sadness	tears, a downward turned mouth, drooping upper eyelids,and a wrinkled forehead.
neutral	relaxed facial muscles, a straight mouth, a smooth forehead, and unremarkable eyebrows.
anger	furrowed eyebrows, narrow eyes, tightened lips, and flared nostrils.
surprise	widened eyes, an open mouth, raised eyebrows, and a frozen expression.
disgust	a wrinkled nose, lowered eyebrows, a tightened mouth, and narrow eyes.
fear	raised eyebrows, parted lips, a furrowed brow, and a retracted chin.
contempt	one side of its mouth raised, one eyebrow lower and one raised, narrowed eyes, and a raised chin.
anxiety	a tensed forehead, tightly pressed lips, pupil dilation, and tensed facial muscles.
helplessness	drooping eyebrows, a downward gaze, a downturned mouth, and lacking expression.
disappointment	a downturned mouth, lowered eyebrows, narrowed eyes, and a sighing face.

descriptions number	DFEW		FERV39k		MAFW	
	UAR	WAR	UAR	WAR	UAR	WAR
1	59.61	71.25	**41.27**	**51.65**	39.89	52.55
2	**60.42**	**72.01**	40.84	51.60	**40.42**	**52.92**

Table A: Evaluation of the prompt ensembling.

descriptiors number	DFEW		FERV39k		MAFW	
	UAR	WAR	UAR	WAR	UAR	WAR
2	**59.65**	**71.91**	40.79	51.57	39.65	52.26
4	59.61	71.25	**41.27**	**51.65**	**39.89**	**52.55**
6	59.59	71.87	40.54	51.46	39.25	52.37

Table B: Evaluation of the different number of descriptors.

B Additional Ablation Studies

The descriptions used in our method are class-level instead of sample-level, and our point is to create discriminative class-level text embedding for supervising the visual part. However, we should admit that the unique description has limitations. To this end, we also conducted experiments with prompt ensembles; the results are shown in Tab. A and indicate that employing more descriptions is beneficial.

In our method, we prompt a large language model such as ChatGPT to automatically generate useful visual descriptors for each facial expression with the process described in Section 3.3. Furthermore, we also investigated the effect of the different descriptions. Specifically, we study the effect of the different numbers of the facial-action-unit-level descriptors. We select the top-2, top-4, and top-6 descriptors generated from the LLMs – the results are shown in Tab. B. We believe fewer descriptors cause the lack of correlation among different expressions and more descriptors result in diminishing the discriminative features.

C Confusion Matrix on Three Benchmarks

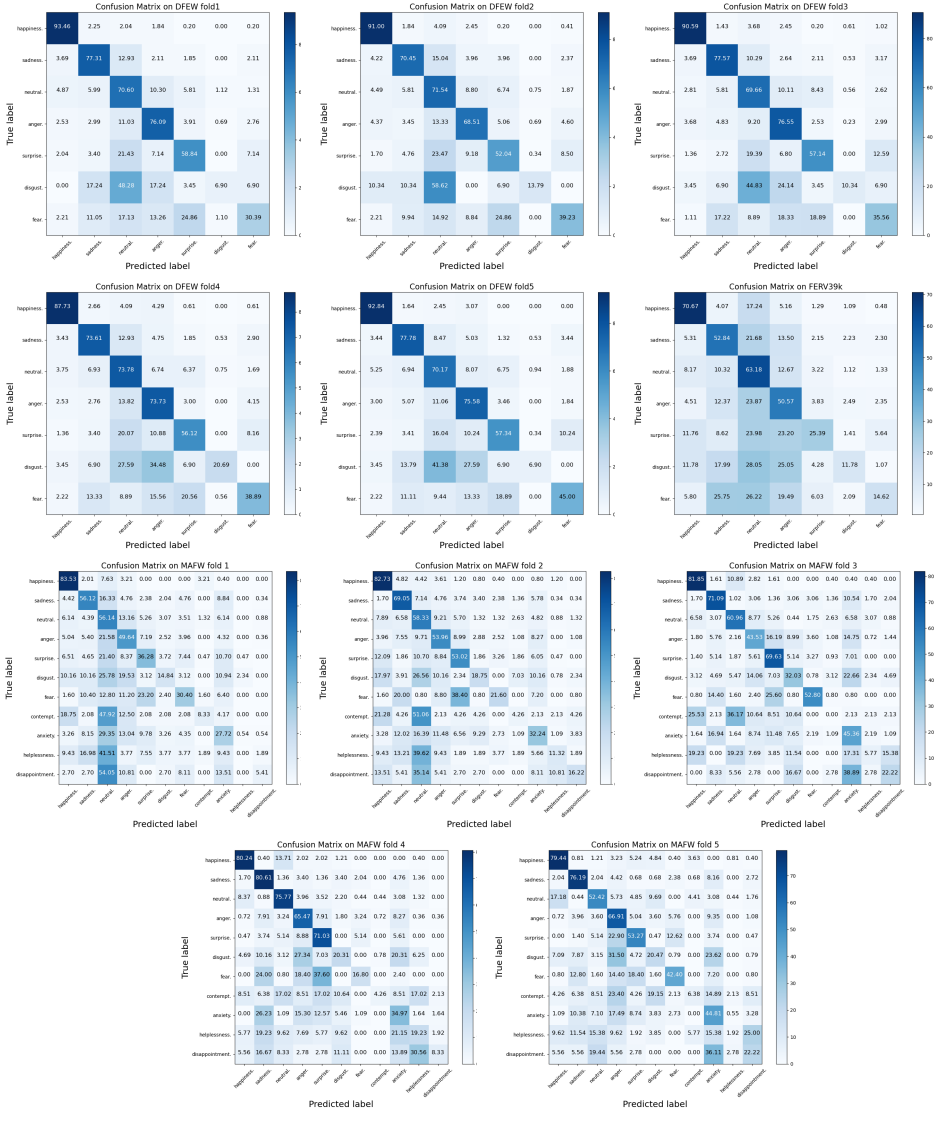

Figure A: Confusion matrix on 5-fold DFEW, FERV39k test set and 5-fold MAFW.