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

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

A. Comparison with CLIP on Varied Prompts

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

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

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

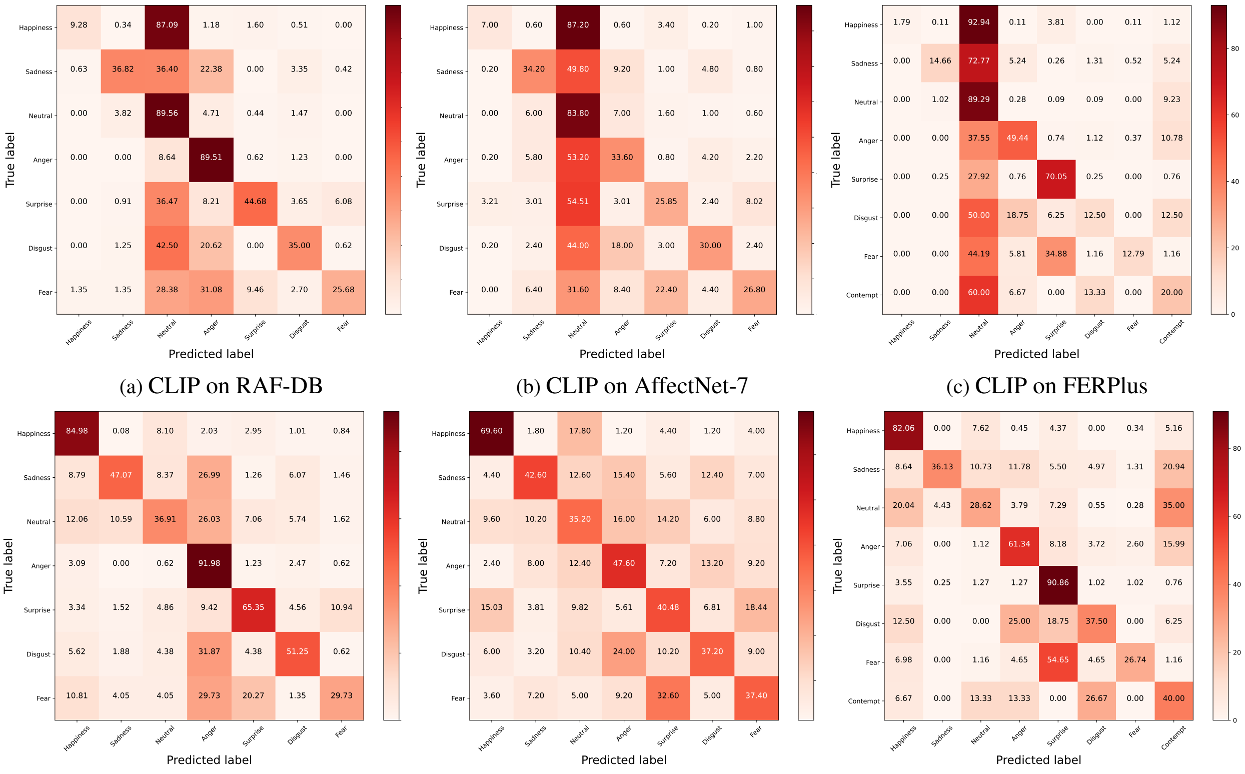
Table B. Zero-shot FER results of our method with different instructions. Both task-related and task-unrelated instructions are investigated. All models use the prompt “a photo of a face with an expression of [class].” for zero-shot prediction.

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

and 10-prompt configurations. All the prompt templates are outlined below:

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

For each prompt, we compare our method with the CLIP model one by one. As shown in Tab. A, except for seven



(a) CLIP on RAF-DB

(b) CLIP on AffectNet-7

(c) CLIP on FERPlus

(d) Ours on RAF-DB

(e) Ours on AffectNet-7

(f) Ours on FERPlus

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

values slightly lower than the CLIP model, our method outperforms the CLIP model on all the rest FER test sets, demonstrating the effectiveness of the proposed method. It should be noted that, compared with the CLIP model, only one extra projection matrix is added on top of the two encoders in our method and this projection matrix was trained in an unsupervised manner.

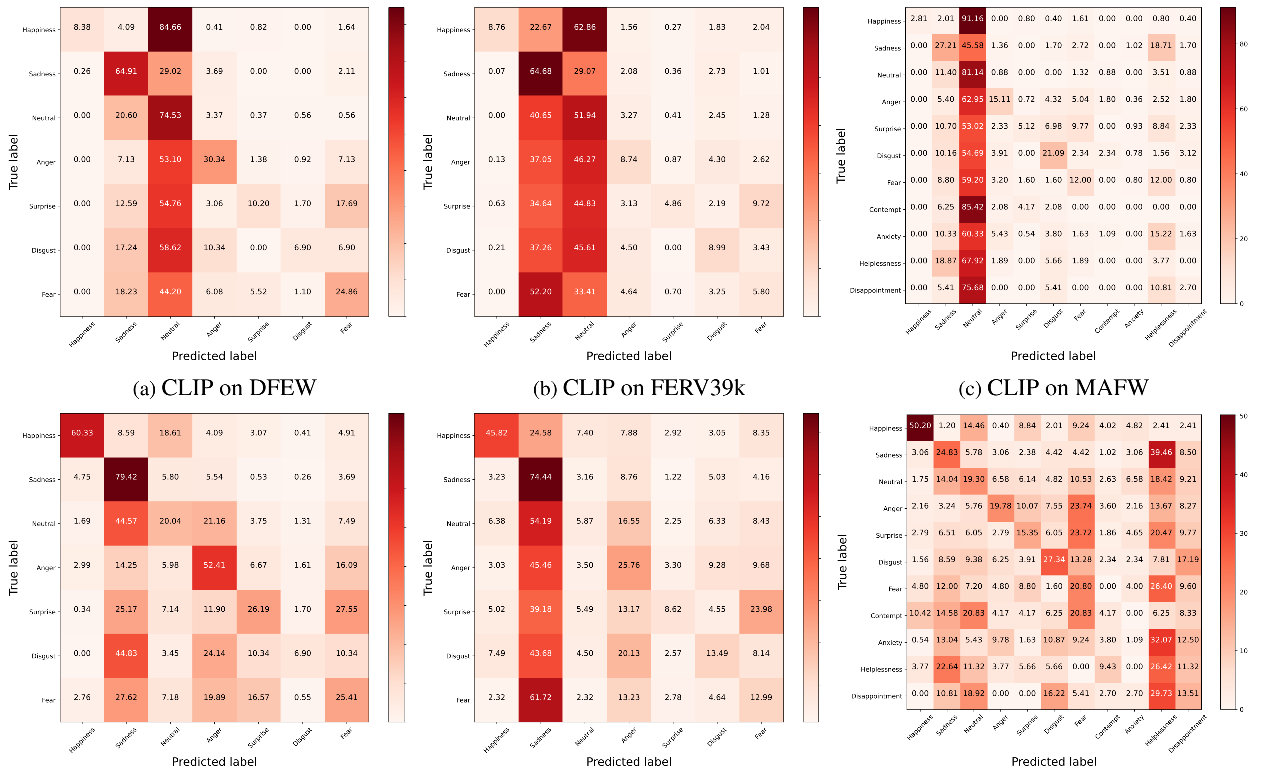
B. Evaluation of Different Instructions

To verify the effectiveness of the instruction adopted in our method, we conducted experiments on two types of instructions: task-related and task-unrelated. Regarding the task-unrelated instructions, we consider both empty and random text as input and take the average of them as the final results. For the task-related instruction, we provide results with three instructions and their average, including "*Please play the role of a facial action describer. Objectively describe the detailed facial actions of the person in the image.*", "*Please play the role of a facial expression recognition expert. Describe the facial expression of the person in the image.*", and "*Please describe the detailed facial actions of the person in the image.*".

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

C. Confusion Matrix Comparison with CLIP

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



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

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

Figure B. Confusion matrix compared with CLIP model on dynamic FER datasets.

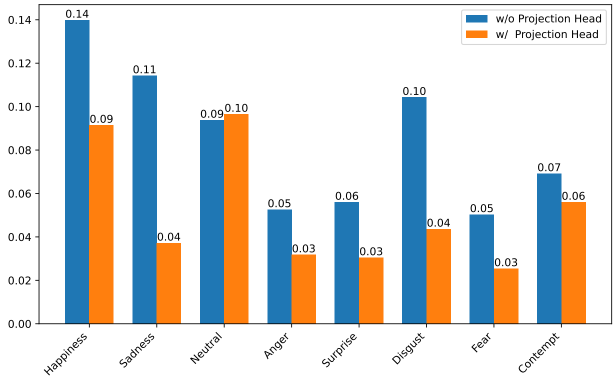


Figure C. Normalized CLIP latent visual feature variance of each emotion on FERPlus.

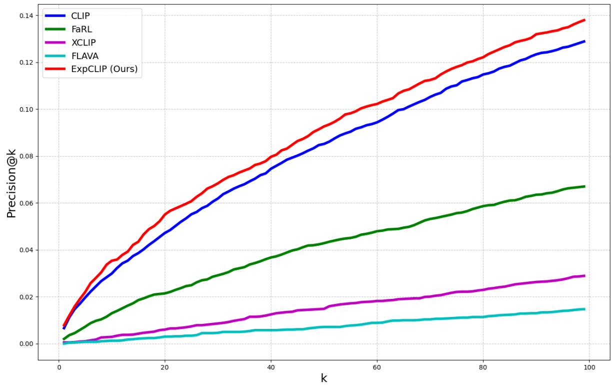


Figure D. Comparison of the video-to-text retrieval performance on the subset of the MAFW.

D. Normalized Feature Variance

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

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

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

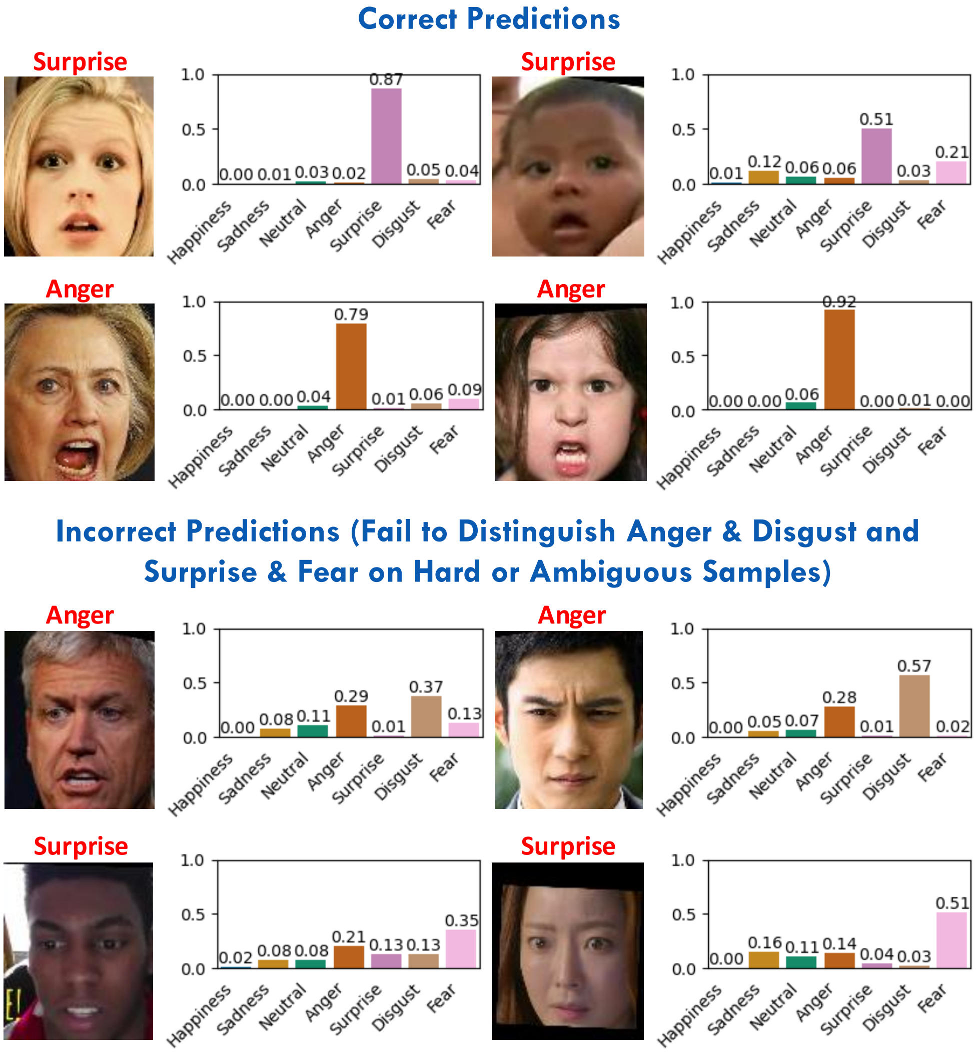


Figure E. Zero-shot facial expression predictions of the proposed Exp-CLIP.

facial action descriptions.

F. More Sample-Level Predictions and Analysis

We provide more sample-level predictions in the Fig. E. From the probability predictions on top images, we observe that our Exp-CLIP model correctly recognizes facial expressions with high confidence. However, our zero-shot method still makes some incorrect predictions, particularly when distinguishing between similar emotions like Anger and Disgust, as well as Surprise and Fear, especially on hard or ambiguous samples. These misclassifications likely arise from the subtle differences in facial expressions between these emotions, which can be difficult even for supervised models. Further refinement of our model or additional fine-tuning on few-shot labelled data may help mitigate these specific shortcomings.