

Supplementary Materials: Domain Knowledge Enhanced Vision-Language Pretrained Model for Dynamic Facial Expression Recognition

In the Appendix, we provide detailed textual descriptions of each facial expression used in this paper. Meanwhile, we present the fine-grained performance of the proposed method for each expression on FERV39K and MAFW. In addition, we provide more recognition results of the model under occlusion and low light conditions.

1 The Detailed Descriptions of Expressions

According to the prior relationships between the expressions and AUs shown in Tab. 1, we use AUs to enrich initial textual prompts for the basic expressions, which include *Happiness*, *Sadness*, *Anger*, *Surprise*, *Disgust*, *Fear*. The textual descriptions of AUs are shown in Tab. 7. Specifically, we integrate the class name of each expression with the textual descriptions of the primarily relevant AUs to form the textual prompt. Meanwhile, since the primarily relevant AUs are not always activated simultaneously, we use "or" in the text prompt to distinguish their semantics. In addition, we simplify the textual descriptions of AUs that describe movements within the same facial region when they are associated with the same expression. For instance, since both AU12 and AU25 describe lip movements, we simplify their combined textual description from "lip corner puller, lip part" to "lip corner puller or part". Since AU1 and AU2 both describe brow movements, their combined textual description is simplified from "inner brow raiser, outer brow raiser" to "brow raiser". As for other expressions without relevant AUs, we use the class names as initial textual prompts. The detailed textual prompts for expressions are presented in Tab. 8.

Table 7: The textual descriptions of AUs.

AU	textual description	AU	textual description
1	Inner brow raiser	2	Outer brow raiser
4	Brow lowerer	5	Upper lid raiser
6	Cheek raiser	7	Lid tightener
9	Nose wrinkler	10	Upper lip raiser
12	Lip corner puller	15	Lip corner depressor
17	Chin raiser	20	Lip stretcher
24	Lip pressor	25	Lips part
26	Jaw drop		

2 The Proposed Method’s Fine-grained Performance

We further present fine-grained results on FERV39K and MAFW in Tab. 9 and Tab. 10, respectively. From these two tables, we observe that DK-CLIP at various scales outperforms state-of-the-art supervised methods (e.g., M3DFEL and T-ESFL). Moreover, compared with self-supervised methods and other CLIP-based methods, our method achieves comparable or even better results than models of the same scale for most facial expressions. For instance, on

Table 8: The detailed textual prompts of expressions.

Expressions	Textual Prompts
Happiness	Happiness with cheek raiser, lip corner puller or part
Sadness	Sadness with inner brow raiser or brow lowerer, lip corner depressor, chin raiser
Anger	Anger with brow lowerer, lid tightener, chin raiser, lip pressor
Surprise	Surprise with brow raiser, upper lid raiser, lip part, jaw drop
Disgust	Disgust with nose wrinkler, upper lip raiser, chin raiser
Fear	Fear with brow raiser or brow lowerer, upper lid raiser, lip stretcher or part
Neutral	Neutral
Contempt	Contempt
Anxiety	Anxiety
Helplessness	Helplessness
Disappointment	Disappointment

FERV39K, our proposed DK-CLIP outperforms DFER-CLIP by 5.9% for *Happiness*, 1.79% for *Sadness*, 1.49% for *Anger*, and 2.98% for *Disgust*. On MAFW, our method surpasses MAE-DFER by approximately 9.9% for *Happiness*, 10.16% for *Disgust*, 5.27% for *Fear*, and 7.07% for *Anxiety*. These results indicate that our DK-CLIP can effectively adapt CLIP to DFER, thereby addressing issues related to limited training data and complex time-dependent modeling.

3 Recognition Results Visualization

We provide more detailed recognition results to demonstrate the recognition performance of our proposed method under occlusion conditions. As shown in Fig. 6, our method not only accurately recognizes expressions under occlusion but also provides information on the changes in expressions within the video sequences. For instance, as depicted in the second and third rows of Fig. 6, the similarity scores decrease with increased facial occlusion. The reason is that occluded faces provide less information about the expressions, resulting in the model treating the corresponding snippets as noisy frames. Furthermore, while snippets in the video sequences that are inconsistent with video-level labels can be detrimental to global predictions, their recognition results are critical for weakly supervised fine-grained expression analysis. Our method can effectively identify non-target snippets and accurately recognize their expression classes, as shown in the fourth row of Fig. 6.

Table 9: Performance comparison of our DK-CLIP with the state-of-the-art methods on FERV39K.

Methods	Accuracy of Each Emotion(%)							Metrics(%)	
	Hap.	Sad.	Neu.	Ang.	Sur.	Dis.	Fea.	UAR	WAR
FormerDFER	65.65	51.33	56.74	43.64	21.94	8.57	12.53	37.20	46.85
NR-DFERNet	69.18	54.77	51.12	49.70	13.17	0.00	0.23	33.99	45.97
IAL	-	-	-	-	-	-	-	35.82	48.54
M3DFEL	-	-	-	-	-	-	-	35.94	47.67
MARLIN (ViT-B/16)	67.73	47.09	<u>62.16</u>	43.85	13.17	4.28	7.66	35.13	46.64
MAE-DFER (ViT-B/16)	73.05	53.98	59.14	50.44	30.09	17.99	<u>17.17</u>	<u>43.12</u>	<u>52.07</u>
CLIPER (ViT-B/32)	-	-	-	-	-	-	-	41.23	51.34
DFER-CLIP (ViT-B/32)	70.67	52.84	63.18	50.57	25.39	11.78	14.62	41.27	51.65
DK-CLIP (ViT-B/32)	76.57	<u>54.63</u>	59.55	52.06	14.97	14.76	12.80	40.76	51.58
DK-CLIP (ViT-B/16)	<u>73.50</u>	49.17	61.08	<u>50.91</u>	<u>27.90</u>	<u>16.35</u>	27.05	43.71	52.14

Table 10: Performance comparison of our DK-CLIP with the state-of-the-art methods on MAFW.

Methods	Accuracy of Each Emotion(%)											Metrics(%)	
	Hap.	Sad.	Neu.	Ang.	Sur.	Dis.	Fea.	Con.	Anx.	Hel.	Dis.	UAR	WAR
FormerDFER	-	-	-	-	-	-	-	-	-	-	-	31.16	43.27
T-ESFL	<u>83.82</u>	67.98	<u>61.16</u>	62.70	48.50	2.51	29.90	0.00	9.52	0.00	0.00	33.28	48.18
MARLIN (ViT-B/16)	76.18	61.23	50.97	60.58	53.85	21.12	23.20	0.00	28.60	6.89	0.56	34.83	48.05
MAE-DFER (ViT-B/16)	77.13	<u>71.09</u>	58.26	67.77	57.46	25.35	34.88	8.90	33.08	11.83	12.09	<u>41.62</u>	53.41
DFER-CLIP (ViT-B/32)	81.56	70.61	60.72	55.90	<u>56.65</u>	21.28	32.80	4.65	<u>37.02</u>	<u>7.65</u>	14.88	39.89	52.55
DK-CLIP (ViT-B/32)	81.72	72.11	60.63	<u>66.19</u>	54.74	<u>27.67</u>	<u>36.31</u>	4.66	34.94	6.85	7.11	41.17	54.93
DK-CLIP (ViT-B/16)	87.03	<u>70.61</u>	64.77	61.87	54.40	35.51	40.15	<u>6.81</u>	40.17	5.72	6.03	43.01	56.56

**Figure 6: The recognition results of five video sequences under occlusion conditions.**