Multimodal Prompt Alignment for Facial Expression Recognition

Supplementary Material

1. LLM-based Facial Expression Descriptions

Tab. 1 provides a detailed list of facial expression categories along with their corresponding descriptions. The hard prompts in our MPA-FER consist of a basic prompt "a photo of a person making a facial expression of [class]", which is then enriched with detailed, class-specific descriptors generated by a large language model.

Category	Descriptions
Surprise	Widened eyes, an open mouth, raised eyebrows, and a frozen expression.
Fear	Raised eyebrows, parted lips, a furrowed brow, and a retracted chin.
Disgust	A wrinkled nose, lowered eyebrows, a tightened mouth, and narrow eyes.
Happiness	A smiling mouth, raised cheeks, wrinkled eyes, and arched eyebrows.
Sadness	Tears, a downward turned mouth, drooping upper eyelids, and a wrinkled forehead.
Anger	Furrowed eyebrows, narrow eyes, tightened lips, and flared nostrils.
Neutral	Relaxed facial muscles, a straight mouth, a smooth forehead, and unremarkable eyebrows.
Contempt	One side of its mouth raised, one eyebrow lower and one raised, narrowed eyes, and a raised chin.

Table 1. Facial Expressions and Their Descriptions

In our proposed MPA-FER framework, these LLM-generated descriptions serve as multi-granularity hard prompts. They provide explicit semantic guidance that, when aligned with trainable soft prompts, significantly enhances the model's ability to capture and discriminate subtle facial cues.

2. More Ablation Study

Effect of the Sparse Image-text Local Similarity. CLIP is pre-trained to align global visual features with textual representations, which works well for many image classification tasks. However, facial expressions often manifest in local regions of the face, and relying solely on global features may yield suboptimal results for FER. To investigate this, we conducted an ablation study within our cross-modal alignment module, comparing three configurations: MPA-FER with only global alignment, with only sparse local alignment, and with cross-modal global-local alignment.

Alignment Manner	RAF-DB	AffectNet-7
Global Align.	91.18	66.58
Sparse Local Align.	91.64	66.97
Global-local ALign.	92.51	67.85

Table 2. Ablation study on the effect of the sparse image-text local similarity.

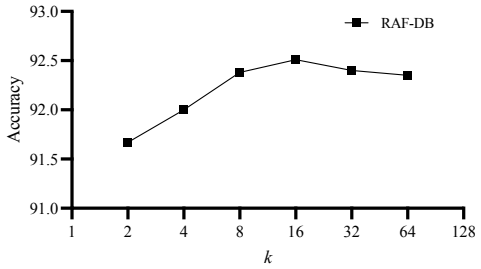

Figure 1. Ablation study of the hyperparameter k.

As shown in Tab. 2, our experiments reveal that only using global alignment is insufficient for capturing the discriminative, fine-grained features for FER. Incorporating sparse local features with the highest cross-modal similarities leads to a promising improvement in performance. Specifically, the use of sparse local alignment improves accuracy on RAF-DB and AffectNet-7, and when combined with global alignment, the performance further increases to 92.51% on RAF-DB and 67.85% on AffectNet-7. These findings confirm that local visual features are crucial for discriminating facial expressions and that the integration of sparse local alignment into the cross-modal module significantly enhances overall performance.

Additionally, we analyzed the effect of the hyperparameter k in the top-k operation when calculating the sparse local similarity. As shown in Fig. 1, our results indicate that an appropriate choice of k is essential; too small a value may discard informative local cues, while too large a value may reintroduce irrelevant background information. The best performance is achieved when $k = 16$, striking an effective balance between capturing overall context and emphasizing discriminative local features. These findings highlight the importance of incorporating sparse local similarity in the cross-modal alignment process for FER, as it enables the model to focus on the most relevant facial regions and enhances its overall accuracy and robustness.

Method	RAF-DB ⇒ CK+	RAF-DB ⇒ AffectNet-7
gACNN [3]	81.07	-
SPWFA-SE [4]	81.72	-
VTFF [6]	81.88	-
STSN [1]	-	48.49
KTN [1]	-	49.60
CRS-CONT [2]	84.43	50.71
MPA-FER	**86.55**	**54.90**

Table 3. Cross-dataset evaluation on CK+ and AffectNet-7.

3. Cross Dataset Evaluation

Following previous cross-dataset evaluation settings [2, 6], we train our MPA-FER on RAF-DB and evaluate it on CK+ [5] and AffectNet-7 [7] to verify its generalization performance. CK+ comprises 593 video sequences from 123 subjects. Following prior works, we treat the first frame of each video sequence as the neutral face and the last, peak frame as the facial expression.

As shown in Tab. 3, our experimental results demonstrate that MPA-FER generalizes well across datasets, achieving 86.55% accuracy on CK+ and 54.90% on AffectNet-7. We attribute this robust cross-dataset performance to our multimodal prompt alignment strategy, which effectively captures and transfers fine-grained facial expression features across different data distributions. It is noted that our approach does not fine-tune the parameters of the pretrained CLIP model. By keeping the CLIP backbone frozen, we preserve the robust, generalizable representations learned from large-scale multimodal data, thereby reducing the risk of overfitting on the FER-specific training data. This enables our model to adapt more effectively to unseen facial expressions and diverse datasets, underscoring the effectiveness of our method in real-world applications.

References

[1] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021.

[2] Hangyu Li, Nannan Wang, Xi Yang, and Xinbo Gao. Crs-cont: a well-trained general encoder for facial expression analysis. *IEEE Transactions on Image Processing*, 31:4637–4650, 2022.

[3] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28 (5):2439–2450, 2018.

[4] Yingjian Li, Guangming Lu, Jinxing Li, Zheng Zhang, and David Zhang. Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Transactions on Affective Computing*, 2020. doi: 10.1109/TAFFC.2020.3031602.

[5] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition-workshops*, pages 94–101, 2010.

[6] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021. doi: 10.1109/TAFFC.2021.3122146.

[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.