Multimodal Prompt Alignment for Facial Expression Recognition

Fuyan Ma[1], Yiran He[2], Bin Sun[3]*, Shutao Li[3]
[1]Chinese Academy of Military Science
[2]Changchun University of Science and Technology, [3]Hunan University

Abstract

Prompt learning has been widely adopted to efficiently adapt vision-language models (VLMs) like CLIP for various downstream tasks. Despite their success, current VLM-based facial expression recognition (FER) methods struggle to capture fine-grained textual-visual relationships, which are essential for distinguishing subtle differences between facial expressions. To address this challenge, we propose a multimodal prompt alignment framework for FER, called MPA-FER, that provides fine-grained semantic guidance to the learning process of prompted visual features, resulting in more precise and interpretable representations. Specifically, we introduce a multi-granularity hard prompt generation strategy that utilizes a large language model (LLM) like ChatGPT to generate detailed descriptions for each facial expression. The LLM-based external knowledge is injected into the soft prompts by minimizing the feature discrepancy between the soft prompts and the hard prompts. To preserve the generalization abilities of the pretrained CLIP model, our approach incorporates prototype-guided visual feature alignment, ensuring that the prompted visual features from the frozen image encoder align closely with class-specific prototypes. Additionally, we propose a cross-modal global-local alignment module that focuses on expression-relevant facial features, further improving the alignment between textual and visual features. Extensive experiments demonstrate our framework outperforms state-of-the-art methods on three FER benchmark datasets, while retaining the benefits of the pretrained model and minimizing computational costs.

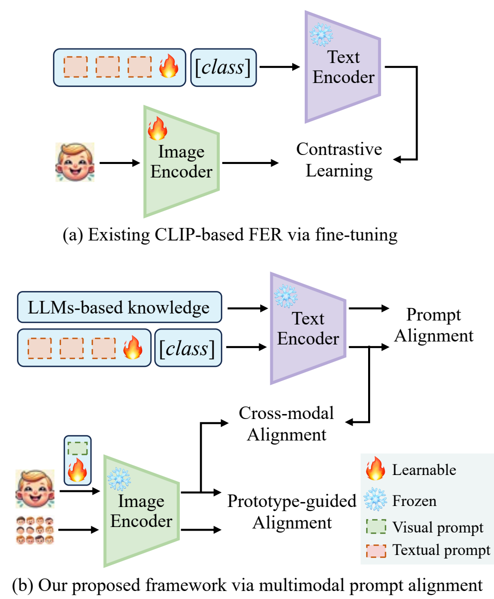


Figure 1. (a) Conventional CLIP-based FER paradigm that adapts the pretrained model by fine-tuning the image encoder, which may lead to overfitting and diminished generalization. (b) Our proposed multimodal prompt alignment approach for FER, where the CLIP backbone remains frozen and learnable visual and textual prompts are used to achieve fine-grained cross-modal alignment, leveraging external language priors to enhance task-specific adaptation.

1. Introduction

Facial expression is one of the most common and natural ways for people to convey different emotions. Understanding facial expressions has various emerging applications, such as human-robot interaction [38], psychology study [33] and healthcare [5]. Previous facial expression recognition (FER) methods have primarily been developed for highly controlled environments [22, 32, 45] with consistent lighting, neutral backgrounds, and limited head poses, which may lack the robustness needed for real-world applications. Despite the progress made by deep learning approaches [18, 42], in-the-wild FER remains challenging due to the subtle and nuanced differences among facial expressions, varying lighting conditions, occlusions, and the presence of irrelevant background information.

Mainstream in-the-wild FER methods [14, 17, 21, 23, 34, 35, 43] rely on Convolutional Neural Networks (CNNs) [12] or Vision Transformers (ViTs) [8] to extract features from facial images, using one-hot labels for supervision

*Bin Sun is the corresponding author.

during training. These methods have shown impressive performance, leveraging the power of deep networks to learn hierarchical spatial features from images. However, a major limitation of these approaches is their reliance on one-hot labels, which treats each expression as a separate class without explicitly integrating the semantic relationships between these classes. As a result, the lack of semantic guidance hinders the model's ability to generalize well, particularly when dealing with complex, real-world datasets where facial expressions might not always align with clear-cut class boundaries.

Advancements in vision-language models (VLMs) [2, 9, 11, 30], particularly CLIP [28], have demonstrated remarkable success in bridging the semantic gap between visual and textual modalities. CLIP's ability to learn powerful visual representations from large-scale data, coupled with its inherent generalization capabilities, has paved the way for adapting such models to specialized tasks like FER. Recent works have explored prompt learning strategies [10, 37, 46, 51] that adapt pretrained VLMs to downstream tasks (such as FER [15, 16, 20, 36, 49]) by learning task-specific textual prompts. However, these approaches still struggle to capture the fine-grained visual features necessary for accurate recognition. As shown in Fig. 1 (a), existing methods have the following limitations: integrating manually crafted text templates (e.g., "a photo of [class]") is often too coarse to capture the fine-grained visual cues that differentiate subtle facial expressions, and fine-tuning the entire image encoder risks overfitting and diminishing the generalization power of the pretrained model.

To overcome such limitations, we propose a novel multimodal prompt alignment framework for FER (MPA-FER) that enhances the recognition performance by integrating external knowledge and aligning multi-granularity visual and textual representations, as shown in Fig. 1 (b). In this work, our approach leverages large language models (LLMs), such as ChatGPT-3.5[1], to generate detailed, class-specific descriptions that capture the critical visual patterns of facial expressions. These LLM-generated descriptions form multi-granularity hard textual prompts, which are then used to guide the learning of soft prompts. By aligning soft prompts with hard prompts at both token-level and prompt-level, our framework effectively incorporates external semantic knowledge into the adaptation process.

Furthermore, to preserve the generalization ability of the pretrained CLIP model while adapting it to the FER domain, we keep the CLIP model frozen and introduce a set of visual prompts to each frozen encoder layer, adding only negligible extra parameters. We employ class-specific prototypes, computed from the frozen visual features of the pretrained model, to steer the learning process and ensure that the prompted visual features remain aligned with the pretrained representation space. These prototypes act as anchors that maintain the alignment of the visual features with the true semantics of each class, thereby eliminating the need to fine-tune the image encoder's core parameters and reducing the risk of overfitting. Additionally, a cross-modal global-local alignment module is introduced to minimize the impact of class-irrelevant background regions. Our MPA-FER learns sparse, discriminative local features by aligning soft textual prompts with a sparse subset of facial regions, facilitating text-to-image matching that captures fine-grained semantics. By selectively focusing on these local features, we reduce noise from irrelevant regions and improve the model's ability to accurately distinguish subtle facial expressions, enhancing both its robustness and performance in the FER task.

Extensive experiments show that the proposed MPA-FER significantly boosts the FER performance by effectively integrating external semantic information and preserving the robust generalization capabilities of the pretrained CLIP model. In summary, the main contributions of this work are as follows:

- We are the first to explore the learning paradigm of using a fully frozen CLIP model for FER, as opposed to the fine-tuning approach that has been widely used in VLM-based FER methods.
- We propose the multimodal prompt alignment framework for FER that effectively incorporates external semantic knowledge into the adaptation process by aligning soft prompts with multi-granularity hard textual prompts.
- We develop a prototype-guided visual feature alignment technique that regularizes the prompted visual features by anchoring them to class-specific prototypes derived from frozen CLIP features.
- Through extensive experiments on three in-the-wild FER datasets, we demonstrate that our framework not only outperforms state-of-the-art methods but also retains the generalization benefits of the pretrained CLIP model, while operating at minimal computational overhead.

2. Related Work

Facial Expression Recognition Facial Expression Recognition (FER) has been an active research area for decades. Traditional FER methods largely rely on handcrafted features, such as Local Binary Patterns (LBP) [29] and Histogram of Oriented Gradients (HOG) [7], to encode the geometry and texture changes in the face induced by different expressions. However, these methods often fail to generalize well in unconstrained, in-the-wild settings where conditions such as lighting, occlusion, and varying head poses present significant challenges. With the rise of deep learning, CNNs and ViTs have become the dominant approaches in FER, benefiting from their ability to learn hierarchical features from large amounts of data [14, 17, 24, 40]. For instance, Li *et al.* [19] used CNN-based attention mechanisms to focus on key regions of interest based on facial landmarks. Zhao *et al.* [47] developed a multi-scale attention network to capture both global and local features, improving robustness to occlusion and pose variations. Transformer-based methods, such as Xue *et al.* [39] and Ma *et al.* [23], apply multi-head self-attention to model relationships between different facial regions, enabling more flexible feature extraction. While these methods have shown impressive results, they rely exclusively on visual features and one-hot labels, which often suffer from issues like redundant feature learning, overfitting, and a lack of interpretability.

Vision-Language Models Vision-Language Models (VLMs), such as CLIP, have shown great success in leveraging large-scale multimodal data to learn joint representations of images and text. For example, CoOp [51] and its variants [25, 31, 50] have used learnable prompt vectors to improve task-specific adaptations of VLMs. However, these methods often rely on generic textual prompts like "a photo of [class]", which lack the fine-grained detail necessary for distinguishing subtle facial expressions. Several prior studies have also introduced external knowledge to prompt learning to enhance the quality of semantic guidance [4, 6, 41]. For instance, Yao *et al.* [41] adopt the prompt template "a photo of [class]" to guide the learning of soft prompts at the global level. While these approaches have shown promise, they often suffer from insufficient knowledge and inadequate guidance. In the context of VLMs for FER [16, 20, 44, 49], the focus is mainly on exploiting multimodal fusion techniques and textual prompts to improve facial expression recognition. For example, Zhou *et al.* [49] introduce expression class tokens with a Conception-Appearance Tuner (CAT) to create trainable appearance prompts, and enhance visual embeddings via knowledge distillation from CLIP. Moreover, these VLM-based FER methods often adopt the fine-tuning paradigm for the pretrained CLIP image encoder to obtain better expression-related features and recognition performance. However, this fine-tuning paradigm increases training costs and the risk of overfitting. A more ideal approach should focus on improving the representation ability of the frozen backbone while enhancing cross-modal interaction as efficiently as possible. This principle underpins the approach we adopt in this work.

3. Methodology

3.1. Overview

In this work, we adopt the pretrained CLIP model as the vision-language foundation for FER, using a multimodal prompt alignment framework to preserve both strong generalization capability and high supervised performance. Our proposed method enables the use of an existing image-language pretrained model, eliminating the need to train one from scratch or fine-tune the entire encoders. An overview of the proposed MPA-FER is shown in Fig. 2. Our framework effectively integrates external semantic knowledge and aligns multi-granularity textual and visual representations to enhance FER performance.

At the core of our method, we use the prompt learning scheme on both the text encoder and the image encoder. On the text side, we model the context words using trainable soft prompts, as opposed to handcrafting textual inputs based on class labels. To guide the learning of soft prompts, we use multi-granularity hard prompts, which consist of a generic template, class category names, and class-specific descriptions generated by a LLM. These detailed descriptions capture the unique visual features of each facial expression and provide task-specific semantic guidance.

On the vision side, we introduce a set of visual prompts into the output feature sequence from each encoder layer, learning face-specific context vectors. To ensure that the prompted visual features remain aligned with the pretrained representation space, we utilize frozen prototypes derived from the pretrained CLIP model. These class-specific prototypes act as anchors, ensuring that the prompted visual features keep consistent with the true semantics of each class. Additionally, our framework incorporates a cross-modal global-local alignment module to align the representations of soft textual prompts with the prompted visual features at both global and local levels. At the local level, we compute sparse cross-modal similarities to focus on the most relevant facial regions (such as the eyes, mouth, and brows), while filtering out irrelevant features from the background.

3.2. Fine-grained Prompt Generation

Although existing prompt-tuning schemes [15, 50, 51] can effectively adapt the pretrained CLIP model to downstream tasks, they often fail to capture the fine-grained visual features across different facial expressions, leading to underperformance compared to state-of-the-art FER methods. To address this limitation, we propose leveraging external knowledge as language priors and enhancing soft prompting by integrating multi-granularity hard prompts. The pipeline of our framework is illustrated in Fig. 2. To explicitly incorporate external knowledge into the soft prompting process, we use a large language model (LLM), such as ChatGPT-3.5, to generate detailed and comprehensive descriptions for each facial expression category. Specifically, we design a query to prompt the LLM: "What are the most useful visual features to distinguish the facial expression of $[class]$?". In response, the LLM generates descriptions in the format: "The most useful visual features for distinguishing the fa-

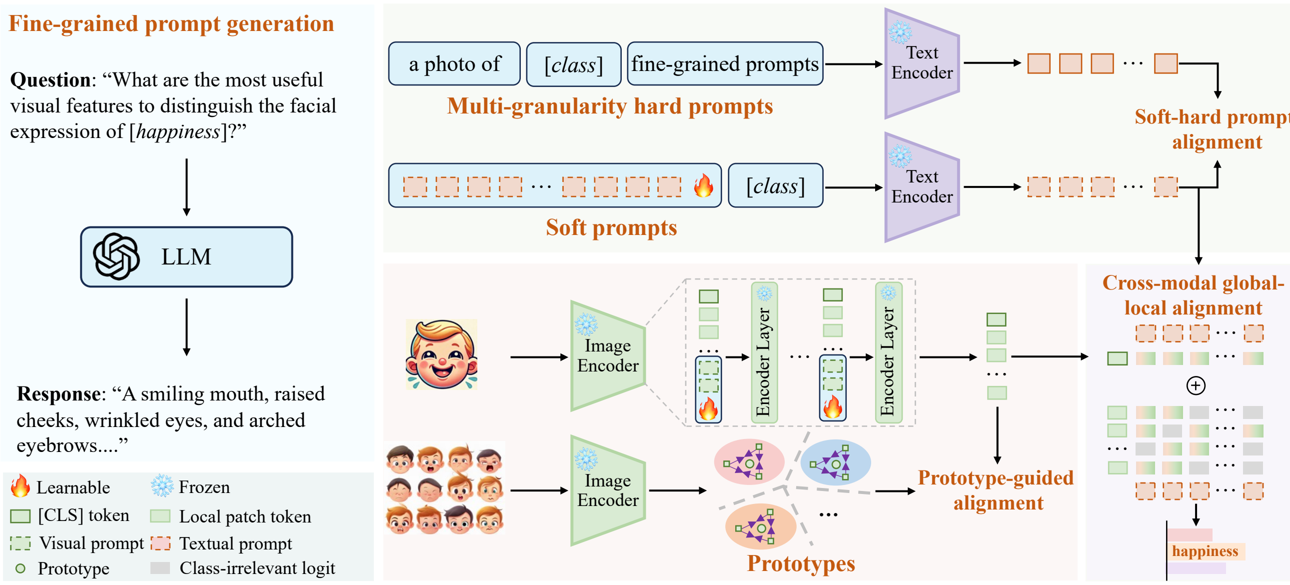


Figure 2. An overview of our proposed multimodal prompt alignment framework, MPA-FER. The multi-granularity hard prompts consist of a generic template, category names, and LLM-generated class-specific descriptions. These hard prompts and frozen prototypes guide the learning of soft prompts and prompted visual features, respectively. The textual features of the soft prompts are aligned with the prompted visual features at both global and local levels, where sparse cross-modal similarities are computed to filter out class-irrelevant local features.

cial expression of $[class]$ include: ...". We then construct class-specific, multi-granularity hard prompts by combining a generic text template with the facial expression name and the corresponding LLM-generated descriptions. This multi-granularity design ensures that both coarse-grained and fine-grained semantics are effectively integrated, further enhancing the adaptability and performance of our proposed framework.

3.3. Soft-hard Prompt Alignment

Previous prompt-tuning methods primarily focus on aligning visual-textual semantics and capturing cross-modal interactions via contrastive learning. In contrast, we additionally model text-text interactions and enhance soft prompt learning by integrating multi-granularity hard prompts through a soft-hard prompt alignment mechanism. Since the hard prompts contain fine-grained facial expression descriptions, this alignment explicitly injects external knowledge into the soft prompts by minimizing the discrepancy between the soft and hard prompts.

Soft prompts consist of a small set of learnable parameters that are prone to overfitting on downstream tasks [6, 27]. To address this, we leverage the LLM-based hard prompts to regularize the soft prompts, encouraging them to remain close in the embedding space to the hard prompts. Specifically, for each class, the generated hard prompts are passed through the tokenizer to get token embeddings for each word in the hard prompt. Given the token embeddings of the soft prompts and the multi-granularity hard prompts, $\boldsymbol{t}_c$ and $\boldsymbol{t}_c^*$, we first align them at the token level via contrastive learning. Under the guidance of the hard prompts, the soft prompts are trained so that they can be correctly classified in the embedding space according to the class weights provided by the hard prompts. Consequently, the probability distribution of the soft prompt embeddings $\boldsymbol{t}_c$ over the class labels is defined as

$$P(y_d|\boldsymbol{t}_d) = \frac{\exp(sim(\boldsymbol{t}_d, \boldsymbol{t}_d^*)/\tau)}{\sum_{c=1}^{C} \exp(sim(\boldsymbol{t}_c, \boldsymbol{t}_d^*)/\tau)}, \quad (1)$$

where y_d denotes the binary label of class d, $sim(\cdot)$ represents the cosine similarity, C denotes the number of facial expressions, and τ is a temperature parameter. Finally, the token-level alignment loss $\mathcal{L}_{ta}$ can be optimized by minimizing the cross entropy loss:

$$\mathcal{L}_{ta} = -\sum_{c=1}^{C} y_c \log P(y_c|\boldsymbol{t}_d). \quad (2)$$

Furthermore, the discrepancy between learnable knowledge and LLM-based knowledge can be measured by the distance between the corresponding textual embeddings. The textual embeddings of the hard prompts and the soft prompts generated by the text encoder, θ, can be denoted as $\theta(\boldsymbol{t}_k^*)$ and $\theta(\boldsymbol{t}_k)$, respectively. To apply the soft-hard prompt alignment at the prompt level, we formulate it as follows:

$$P(y_d|\theta(\boldsymbol{t}_d)) = \frac{\exp(sim(\theta(\boldsymbol{t}_d), \theta(\boldsymbol{t}_d^*))/\tau)}{\sum_{c=1}^{C}\exp(sim(\theta(\boldsymbol{t}_c), \theta(\boldsymbol{t}_d^*))/\tau)},$$
$$\mathcal{L}_{pa} = -\sum_{c=1}^{C} y_c \log P(y_c|\theta(\boldsymbol{t}_d)), \tag{3}$$

where $\mathcal{L}_{pa}$ represents the prompt-level alignment loss. Thus, the soft-hard textual prompt alignment loss $\mathcal{L}_t$ can be computed as the sum of $\mathcal{L}_{pa}$ and $\mathcal{L}_{ta}$:

$$\mathcal{L}_t = \mathcal{L}_{ta} + \mathcal{L}_{pa} \tag{4}$$

Eq. (4) ensures that both token-level and prompt-level textual alignment are effectively optimized.

3.4. Prototype-guided Visual Feature Alignment

To better retain the general representation of the pretrained image encoder, we introduce the prompt learning scheme into the image encoder that keeps the pretrained model frozen. For prompting on the image encoder, we have two major objectives: 1) Providing additional parameters to adapt the CLIP image representations towards the FER dataset distribution, and 2) ensuring the prompted visual features align well with the pretrained CLIP features to address the prompt overfitting problem for better generalization.

In this regard, we introduce a set of visual prompts, which requires minimal additional parameters to train, into the image encoder ϕ that learns face-specific context vectors. The prompt tokens are randomly initialized learnable vectors, and they are designed to provide additional learning capacity to adapt the pretrained model to the FER dataset distribution. Specifically, the prompt tokens $\{p_i^l\}_{i=1}^{N_p}$ at l-th layer are appended to the visual token sequence $\boldsymbol{z}^{l-1}$ before applying the frozen pretrained encoder layer:

$$[\boldsymbol{z}^l, \_] = \phi_l(\left[\boldsymbol{z}^{l-1}, \{p_i^l\}_{i=1}^{N_p}\right]), l = 1, 2, 3, \ldots, K \tag{5}$$

where N_p and K are the number of visual prompts and the depth of the image encoder ϕ, respectively. During training, the visual prompts act as contextual modifiers to interact with these pretrained visual feature tokens. Finally, the prompted visual features $\boldsymbol{z}^K$ are utilized to learn a common feature space with textual labels.

To retain the generalized knowledge of CLIP features and further optimize the visual prompts more explicitly, we impose the regularization constraint on the prompted visual features and the CLIP visual prototypes. For the pretrained CLIP visual encoder, we define the prototypes $\boldsymbol{p}$ as the mean of the frozen feature representations from the subset of training dataset $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}}$ for each class c:

$$\boldsymbol{p}_c = \frac{1}{N_{subset}^c} \sum_{(x_i, y_i)\in\mathcal{D}_{subset}^c} \boldsymbol{z}_i^g, \tag{6}$$

where $\mathcal{D}_{subset}$ denotes a subset of $\mathcal{D}_{train}$, N_{subset}^c represents the number of samples belonging to class c in $\mathcal{D}_{subset}$. $\boldsymbol{z}^g$ denotes the global classification token from the output sequence of the final image encoder layer. The channel axis in $\boldsymbol{z}^g$ is omitted for simplicity.

Therefore, we can explicitly condition the prompted visual features to be close to their corresponding prototype as follows,

$$\mathcal{L}_v = \sum_{c=1}^{C} \sum_{i}^{N_{train}} \mathbb{I}_{(y_i=c)} \mathcal{M}(\boldsymbol{z}^g, \boldsymbol{p}_c), \tag{7}$$

where $\mathcal{M}$ denotes the matching losses (such as L1 loss and cosine similarity) to measure the distance between the prompted classification token and its corresponding prototype. Since we do not introduce any trainable parameters for generating the frozen prototypes, the prototype-guided visual feature alignment not only regularizes the visual prompts to enhance generalization but also operates with minimal computational overhead.

3.5. Cross-modal Global-local Alignment

Different facial expressions are reflected in specific regions of the face, such as the eyes, mouth, and brow, while other areas, like the background, contribute little to the expression. To better model the relationship between textual representations and visual features, it is essential to focus on these expression-relevant local regions, rather than simply averaging similarities across all spatial locations. To this end, we propose a cross-modal global-local alignment module to align soft textual prompts and prompted visual features. Specifically, the global visual representation $\boldsymbol{z}^g$ and a set of local visual features $\boldsymbol{Z}^l = \{\boldsymbol{z}_i^l\}_{i=1}^{N_l}$ are used to align with the textual representation $\theta(\boldsymbol{t}_d)$. We further utilize the sparse local similarity to learn sparse discriminative local features, where only local features semantically related to the class are kept to perform classification:

$$sim_{top-k}(\boldsymbol{Z}^l, \theta(\boldsymbol{t}_d)) = \frac{1}{k}\sum_{i=1}^{N_l} \mathbb{I}_{top-k}(i) \cdot \langle \boldsymbol{z}_i^l, \theta(\boldsymbol{t}_d)\rangle, \tag{8}$$

$$\mathbb{I}_{top-k}(i) = \begin{cases} 1, & \text{if } \mathrm{rank}_i(\langle \boldsymbol{z}_i^l, \theta(\boldsymbol{t}_d)\rangle) \le k, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathbb{I}_{top-k}(l)$ denotes the indicator function for selecting the top-k local features with the highest similarities with the textual representation, and N_l is the number of local features. This selective focus reduces noise from irrelevant regions and ensures that the learning process emphasizes the facial features that are most indicative of the expression, thereby improving the accuracy and robustness of the model. The combination of global similarities and sparse local similarities produces the output logits:

$$logits = sim(\boldsymbol{z}^g, \theta(\boldsymbol{t}_d))) + sim_{top-k}(\boldsymbol{Z}^l, \theta(\boldsymbol{t}_d)). \tag{10}$$

Method	RAF-DB	AffectNet-7
Baseline CoOp [51]	86.15	62.27
+ Visual Prompts	88.47 (+2.32)	63.90 (+1.63)
+ Prototype-guided Align.	89.42 (+0.95)	65.02 (+1.12)
+ Soft-hard Prompt Align.	91.18 (+1.76)	66.58 (+1.56)
+ Cross-modal Align.	**92.51** (+1.33)	**67.85** (+1.27)

Table 1. Ablation study on the contribution of individual components in our MPA-FER framework for FER.

Thus, the cross-modal alignment can be optimized using cross-entropy loss by minimizing the discrepancy between the predicted results and the ground truth:

$$\mathcal{L}_{v\_t} = CE(logits, y_d), \tag{11}$$

where $\mathcal{L}_{v\_t}$ denotes the image-text alignment loss, and CE is the stand cross entropy loss.

Combining the textual prompt alignment loss in Eq. (4), the visual feature alignment loss in Eq. (7), the cross-modal alignment loss in Eq. (10), our method optimize the text prompts and the visual prompts by minimizing the following total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{v\_t} + \beta \times \mathcal{L}_t + \gamma \times \mathcal{L}_v, \tag{12}$$

where β, γ are hyper-parameters to balance the losses.

4. Experiments Settings

Datasets i) **RAF-DB** [17]: consists of 29,672 real-world facial images collected from the internet. Each image is annotated by 40 annotators with basic expressions or compound emotions. For our experiments, we focus on the common expression labels, which include 12,271 training samples and 3,068 testing samples. ii) **FERPlus** [3]: is an extended version of FER2013, containing 28,709 training images, 3,589 validation images, and 3,589 testing images. Each image in FERPlus is annotated by 10 workers across eight expression categories. For fairness, we use majority voting to generate the final labels. iii) **AffectNet** [26]: contains over one million facial images gathered from the internet. Previous studies have focused on AffectNet with either 7 or 8 facial expression categories. We refer to these versions as AffectNet-7 and AffectNet-8, respectively. For our experiments, we use approximately 280,000 images for training, along with 3,500 validation images for AffectNet-7 and 4,000 validation images for AffectNet-8.

Implementation Details In our experiments, both the text and image encoders remain frozen, with their weights directly inherited from the pretrained CLIP models (ViT-B/16 or ViT-L/14). The visual prompts and soft textual prompts are randomly initialized from a Gaussian distribution with a zero mean and 0.02 std. The visual prompt number N_p and the textual prompt number are set to 8 and 10, respectively. The training epochs are 120 for RAF-DB and FERPlus, 20 for AffectNet with a batch size of 32. We use SGD optimizer with a learning rate of 0.032 and adopt a cosine decay strategy for the learning rate. During the training process, all images are resized to 224 $\times$ 224 pixels, and several standard data augmentation techniques, such as random cropping and erasing, are employed on the training images. The hyper-parameter k used for local features selection in Eq. (9) is set to 16. In Eq. (12), we experimentally choose suitable loss balancing factors as $\beta = \gamma = 1$. We implement our method with the PyTorch framework and train the models on one NVIDIA V100 GPU with 32GB RAM.

Textual Prompt Configuration	RAF-DB	FERPlus
CoOp + (1)	86.15	87.48
CoOp + (2)	86.59	88.01
CoOp + (3)	87.26	88.30
CoOp + (1) + Soft-hard Prompt Align.	87.56	88.88
CoOp + (2) + Soft-hard Prompt Align.	87.81	89.21
CoOp + (3) + Soft-hard Prompt Align.	88.69	89.76
MPA-FER + (1)	89.60	88.77
MPA-FER + (2)	90.48	88.94
MPA-FER + (3)	91.01	89.69
MPA-FER + (1) + Soft-hard Prompt Align.	91.05	90.17
MPA-FER + (2) + Soft-hard Prompt Align.	91.87	90.43
MPA-FER + (3) + Soft-hard Prompt Align.	**92.51**	**91.15**

Table 2. Ablation study on different textual prompt configurations and the effect of soft-hard prompt alignment.

4.1. Ablation Study

Effect of Different Model Components. In our ablation study (see Tab. 1), we systematically evaluate the impact of each key component in our MPA-FER framework for FER. Starting with a baseline model that employs a frozen CLIP image encoder trained solely with contrastive loss, we first introduce the visual prompts, which improved accuracy by 2.32% on RAF-DB and 1.63% on AffectNet-7. Next, adding the prototype-guided visual alignment module yields further gains of 0.95% and 1.12%, respectively. The subsequent integration of soft-hard prompt alignment results in additional improvements of 1.76% on RAF-DB and 1.56% on AffectNet-7, while the final inclusion of the cross-modal alignment module achieves the best overall performance with further enhancements of 1.33% and 1.27%. These results demonstrate that each component contributes meaningfully to performance, validating the effectiveness of our multimodal prompt alignment strategy for FER.

Effect of Different Textual Prompts. We further evaluate the impact of various textual prompt configurations

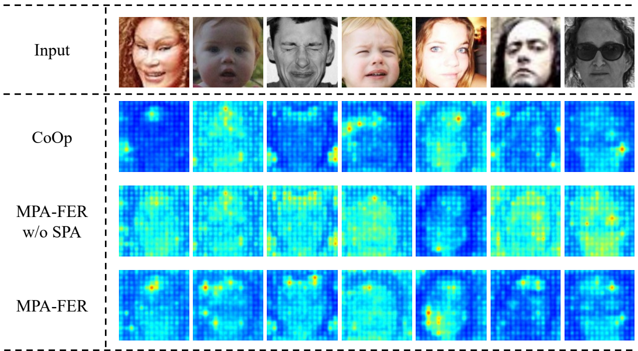


Figure 3. The attention visualization of CoOp, MPA-FER without soft-hard prompt alignment (SPA) and our MPA-FER. We obtain these visualizations by backpropagating gradients on the input images. The first row displays the input images, while the second row shows the attention maps from CoOp, which only fine-tunes language prompts but is relatively sensitive to background features. The third row illustrates the attention maps from our MPA-FER without SPA, where the introduction of visual prompts begins to shift the focus towards expression-relevant features. The last row presents the results of our MPA-FER, which further emphasizes discriminative local facial regions and significantly reduces the influence of irrelevant background areas.

on the performance of our proposed MPA-FER framework. In our experiments, we consider three types of hard prompts: (1) the conventional prompt "a photo of [class]", (2) a more semantically enriched prompt "a photo of a person making a facial expression of [class]", and (3) an augmented prompt "a photo of a person making a facial expression of [class], [LLM-based descriptions]", where the LLM-based descriptions provide detailed, class-specific visual cues. As shown in Tab. 2, both the baseline CoOp and our MPA-FER framework benefit from improved prompt configurations, with performance increasing from configuration (1) to (3) on both RAF-DB and FERPlus. Furthermore, incorporating soft-hard prompt alignment consistently enhances accuracy across all configurations. For example, our MPA-FER framework with configuration (3) and soft-hard prompt alignment achieves the highest accuracy of 92.51% on RAF-DB and 91.15% on FERPlus. These results demonstrate that explicitly incorporating the hard prompts with class-specific descriptions significantly improves semantic guidance, thereby enhancing overall performance. We also visualize the attention maps produced by different models (i.e., the baseline CoOp, our MPA-FER without SPA, and the MPA-FER) in Fig. 3 for an intuitive visual comparison.

Effect of Varying Training Data for Class-wise Prototypes. We conduct an ablation study by varying the number of training images per class used for generating class-wise prototypes from the training dataset. The experimen-

Number	1	4	16	32	64	Full
FERPlus	89.87	90.20	90.48	90.75	91.11	**91.15**
AffectNet-8	61.41	61.68	62.10	62.59	62.77	**62.80**

Table 3. Ablation study on the varying amounts of training images per class for generating class-wise prototypes. ("**Full**" indicates the utilization of the entire training dataset.)

Backbone	Parameters (MB)	RAF-DB	AffectNet-7	AffectNet-8	FERPlus
ViT-B/16	0.218	92.51	67.85	62.80	91.15
ViT-L/14	0.443	93.74	68.89	63.74	91.81

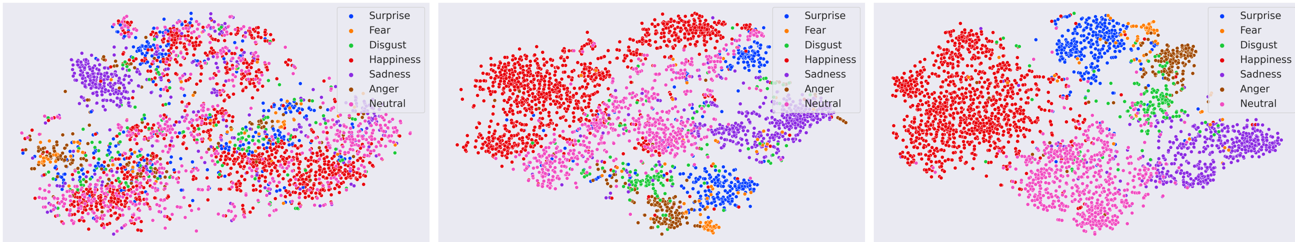
Table 4. Effectiveness of our proposed method with different CLIP backbone networks. We only count the number of parameters that require gradient updates.

tal results in Tab. 3 demonstrate that increasing the number of training images per class for generating class-wise prototypes consistently improves performance on both datasets. For example, on the FERPlus dataset, accuracy increases from 89.87% with a single image per class to 91.15% when using the full training set. These results indicate that using more training samples helps to capture more robust and representative class-specific features, although the improvements tend to plateau as the number of samples approaches the full dataset. In Fig. 4, we visualize the t-SNE embeddings of the learned feature representations from three variants of our proposed framework. As can be observed, the full model yields more distinct and tightly clustered features, highlighting the critical contributions of both components in enhancing the discriminative power of the learned representations for FER.

Effect of Working with Different CLIP Backbones. Tab. 4 shows the effectiveness of our proposed method using two different CLIP backbone networks: ViT-B/16 and ViT-L/14. With ViT-B/16, our model achieves 92.51% on RAF-DB, 67.85% on AffectNet-7, 62.80% on AffectNet-8, and 91.15% on FERPlus, with a total of 0.218 MB learnable parameters. When using the larger ViT-L/14 backbone, which requires only a modest increase in learnable parameters (0.443 MB), the performance improves consistently across all datasets (93.74% on RAF-DB, 68.89% on AffectNet-7, 63.74% on AffectNet-8, and 91.81% on FERPlus). These results demonstrate that our method scales well with larger CLIP backbones, yielding enhanced accuracy while maintaining a minimal increase in model complexity.

4.2. Comparison with the State-of-the-Arts

Tab. 5 and Tab. 6 present comprehensive comparisons of our proposed MPA-FER with state-of-the-art methods across multiple FER datasets. In Tab. 5, our method achieves



(a) MPA-FER w/o Visual Prompting (b) MPA-FER w/o Prototype-guided Alignment (c) Our MPA-FER

Figure 4. t-SNE visualizations of the learned feature representations of the test set of RAF-DB. Subfigure (a) illustrates the results when the CLIP image encoder is used without any visual prompts, while subfigure (b) shows the feature distribution when the prototype-guided alignment module is omitted. Subfigure (c) presents the embeddings from our complete MPA-FER.

Method	Publication	RAF-DB	FERPlus
CNN-based FER			
RAN[35]	TIP 2020	86.90	88.55
KTN[14]	TIP 2021	88.07	90.49
MA-Net[47]	TIP 2021	88.40	-
Transformer-based FER			
VTFF[23]	TAFFC 2021	88.14	88.81
TransFER[39]	ICCV 2021	90.91	90.83
APViT[40]	TAFFC 2022	91.98	90.86
TAN[24]	TAFFC 2023	89.12	90.67
CLIP-based FER			
CLEF [44]	ICCV2023	90.09	89.74
CLIPER [15]	ICME 2024	91.61	-
FER-former[20]	TMM 2024	91.30	90.96
E2NT[16]	arXiv 2024	92.63	91.18
CEPrompt [49]	TCSVT 2024	92.43	-
Our MPA-FER	ICCV 2025	**93.74**	**91.81**

Table 5. Comparison with state-of-the-art methods on RAF-DB and FERPlus.

Method	Publication	AffectNet-7	AffectNet-8
SCN [34]	CVPR 2020	63.40	60.23
MA-Net[47]	TIP 2021	64.53	60.29
VTFF[23]	TAFFC 2021	64.80	61.85
MVT [13]	arXiv 2021	64.57	61.40
POSTER[48]	ICCV 2023	67.31	63.34
CLEF [44]	ICCV 2023	65.66	62.77
CLIPER [15]	ICME 2024	66.29	61.98
CEPrompt [49]	TCSVT 2024	67.29	62.74
Our MPA-FER	ICCV 2025	**68.89**	**63.74**

Table 6. Comparison with state-of-the-art methods on AffectNet-7 and AffectNet-8.

93.74% on RAF-DB and 91.81% on FERPlus, outperforming traditional CNN-based approaches (e.g., RAN, KTN), recent Transformer-based methods (e.g., VTFF, TransFER, APViT, TAN), as well as existing CLIP-based methods (e.g., CLEF, CLIPER, FER-former, E2NT, CEPrompt). These results underscore the effectiveness of our multimodal prompt alignment framework, which leverages external semantic knowledge to enhance visual-textual representation without fine-tuning the pretrained CLIP model.

Similarly, Tab. 6 demonstrates that our MPA-FER consistently outperforms other methods on AffectNet-7 and AffectNet-8. Our approach achieves 68.89% on AffectNet-7 and 63.74% on AffectNet-8, thereby confirming its robustness and generalizability under challenging, in-the-wild conditions. Overall, these comparisons validate that MPA-FER not only improves performance over traditional and recent deep learning-based FER methods but also sets a new state-of-the-art in the field by efficiently leveraging multimodal information. See more experiments in the Supplementary Material.

5. Conclusion

In this paper, we introduce the multimodal prompt alignment framework for facial expression recognition (MPA-FER) that leverages external semantic knowledge without requiring extensive fine-tuning of the pretrained CLIP model. By incorporating LLM-based hard prompts and aligning them with trainable soft prompts, our approach captures fine-grained facial cues to enhance visual-textual representation learning. Additionally, our prototype-guided feature alignment and cross-modal alignment modules ensure that the prompted visual features remain robust and discriminative by focusing on the most expression-relevant regions while suppressing irrelevant background information. Extensive experiments on in-the-wild FER datasets demonstrate that MPA-FER outperforms state-of-the-art methods, achieving superior accuracy with adding few learnable parameters. The ablation studies further validate the contributions of each component within our framework, highlighting its effectiveness and efficiency. Future work will explore further integration of multimodal cues and the extension of our framework to other fine-grained visual recognition tasks.

Acknowledgement. This work is supported by the National Natural Science Fund of China (62221002, 62171183, U24A20327).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 6

[4] Yequan Bie, Luyang Luo, Zhixuan Chen, and Hao Chen. Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 773–783. Springer, 2024. 3

[5] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 2022. doi: 10.1109/TII.2022.3141400. 1

[6] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23232–23241, 2023. 3, 4

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[10] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[12] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 1

[13] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvit: Mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021. 8

[14] 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. 1, 3, 8

[15] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2, 3, 8

[16] Hangyu Li, Yihan Xu, Jiangchao Yao, Nannan Wang, Xinbo Gao, and Bo Han. Knowledge-enhanced facial expression recognition with emotional-to-neutral transformation. *arXiv preprint arXiv:2409.08598*, 2024. 2, 3, 8

[17] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018. 1, 3, 6

[18] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020. doi: 10.1109/TAFFC.2020.2981446. 1

[19] 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. 3

[20] Yande Li, Mingjie Wang, Minglun Gong, Yonggang Lu, and Li Liu. Fer-former: Multimodal transformer for facial expression recognition. *IEEE Transactions on Multimedia*, 2024. 2, 3, 8

[21] Hanwei Liu, Rudong An, Zhimeng Zhang, Bowen Ma, Wei Zhang, Yan Song, Yujing Hu, Wei Chen, and Yu Ding. Norface: Improving facial expression analysis by identity normalization. In *European Conference on Computer Vision*, pages 293–314. Springer, 2024. 1

[22] 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. 1

[23] 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. 1, 3, 8

[24] Fuyan Ma, Bin Sun, and Shutao Li. Transformer-augmented network with online label correction for facial expression recognition. *IEEE Transactions on Affective Computing*, 15 (2):593–605, 2023. 3, 8

[25] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023. 3

[26] 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. 6

[27] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024. 4

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 2

[29] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing*, pages II–370, 2005. 2

[30] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[31] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022. 3

[32] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of International Workshop on EMOTION: Corpora for Research on Emotion and Affect*, pages 65–70, 2010. 1

[33] Simon Wallace, Michael Coleman, and Anthony Bailey. An investigation of basic facial expression recognition in autism spectrum disorders. *Cognition and Emotion*, 22(7):1353–1380, 2008. 1

[34] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 1, 8

[35] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 1, 8

[36] Shangfei Wang, Yi Wu, Yanan Chang, Guoming Li, and Meng Mao. Pose-aware facial expression recognition assisted by expression descriptions. *IEEE Transactions on Affective Computing*, 15(1):241–253, 2023. 2

[37] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044, 2023. 2

[38] Matthias Wimmer, Bruce A MacDonald, Dinuka Jayamuni, and Arpit Yadav. Facial expression recognition for human-robot interaction–a prototype. In *International Workshop on Robot Vision*, pages 139–152. Springer, 2008. 1

[39] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 3, 8

[40] Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, and Guodong Guo. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 2022. doi: 10.1109/TAFFC.2022.3226473. 3, 8

[41] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 3

[42] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European Conference on Computer Vision*, pages 222–237, 2018. 1

[43] Feifei Zhang, Mingliang Xu, and Changsheng Xu. Weakly-supervised facial expression recognition in the wild with noisy data. *IEEE Transactions on Multimedia*, 24:1800–1814, 2021. 1

[44] Xiang Zhang, Taoyue Wang, Xiaotian Li, Huiyuan Yang, and Lijun Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20751–20762, 2023. 3, 8

[45] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 1

[46] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023. 2

[47] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 3, 8

[48] Ce Zheng, Matias Mendieta, and Chen Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3155, 2023. 8

[49] Haoliang Zhou, Shucheng Huang, Feifei Zhang, and Changsheng Xu. Ceprompt: Cross-modal emotion-aware prompting for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 3, 8

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition, pages 16816–16825, 2022. 3

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 3, 6