Behavior-prompted Learning with Tree Attention for Advanced Facial Action Unit Detection

Hao Zou[1], Zheng Gao[1], Yante Li[2], Ce Li[3], Xiaobai Li[1,2*]
[1] The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[2] Center of Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
[3] Moore Threads Technology Co. Ltd., Beijing, China
zou_hao@zju.edu.cn, zhenggao@zju.edu.cn, Yante.li@oulu.fi, ce.li@mthreads.com, Xiaobai.li@zju.edu.cn

***Abstract*—Action Unit (AU) is a systematic coding of facial behaviors that plays a crucial role in facial expression recognition. AU detection faces significant challenges due to the fine-grained categorical differences and coexistence at varying intensities of the AU. To address these challenges, we propose a refined behavior-prompt AU detection model featuring a coarse-to-fine tree attention mechanism. Specifically, we introduce a learnable behavior-prompt approach that utilizes large vision-language models, harnessing their powerful semantic representation capabilities to encompass comprehensive prior knowledge of AU behaviors. Besides, considering AUs' diverse intensities and interactive nature, a coarse-to-fine tree attention module is customized to capture the fine-grained details of individual AUs and their long-range dependencies. To further mitigate vision-text bias, a feature interaction learning strategy is employed that progressively incorporates context-related visual information into prompts and decouples AU-specific representations. Extensive experiments demonstrate that our proposed method achieves state-of-the-art results on two widely used benchmarks, BP4D and DISFA. Our code is avaliable at https://github.com/ColinHaoZou/FAUD-CLIP.**

***Index Terms*—Action Unit; Vision-language model; Attention**

I. Introduction

Facial expressions are a natural and effective way to convey affective information in human non-verbal communication. To establish an objective and comprehensive framework of describing different expressions, Ekman and Friesen proposed the Facial Action Coding System (FACS) [3]. In FACS, various atomic muscle motions underlying facial skin are encoded as Action Units (AUs) and different combinations of AUs form a wide range of facial expressions. Accurate and effective facial AU detection can significantly benefit the downstream affective recognition tasks, including expression recognition [27], depression detection [23], and pain level analysis [31]. As a result, automatic facial AU detection has attracted increasing attention as a key component of affective computing.

Since individual AUs typically manifest in localized regions of the face, most AU detection methods [10], [18] attempt to leverage the location information provided by facial landmarks to extract AU-related features from the corresponding regions. Nevertheless, the location information provided by these facial landmarks is insufficient for facial action unit detection. According to the AU descriptions in FACS, AUs are not only related to the location but also to the motion they produce. For example, both AU12 and AU14 appear at the lip corners, but AU12 pulls the lip corners upward the cheek obliquely while AU14 tightens the lip corners, as shown in Fig. 1. Therefore, some methods [11], [21] have sought to incorporate temporal information by utilizing frame sequences as inputs to capture the dynamic features of AUs. However, such approaches often demand substantial computational resources and labeled data.

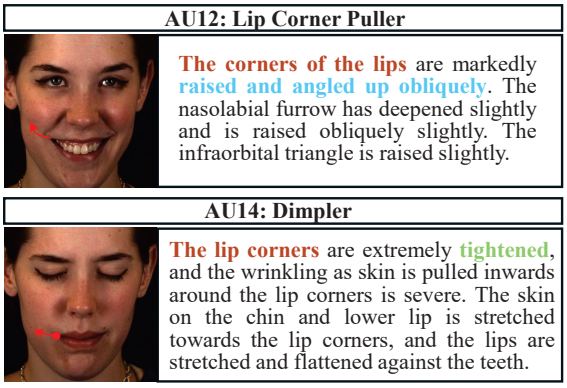

Fig. 1. The behavioral descriptions from the FACS [3] provide detailed information that could be leveraged to help the model better distinguish resembling AUs. For example, AU12 and AU14 are at the same facial location (in red) and with different motion appearances (in blue and green).

FACS provides detailed descriptions of AU behaviors, offering an efficient way to incorporate both spatial and motion information into models. Some methods [30], [34] have explored leveraging these descriptions to enhance AU representations. Yang et al. [30] proposed a cross-modality attention network that utilizes AU descriptions to generate attention maps. This approach relies solely on location information, overlooking motion cues for visual feature extraction. Recently, vision-language models such as Contrastive Language-Image Pretraining (CLIP) [16] have demonstrated significant capabilities in aligning image and text feature spaces by leveraging large-scale data. Zhang et al. [34] further extended CLIP by utilizing label names and descriptions as prompts to guide global feature extraction, showcasing the potential of CLIP for representation learning. Nevertheless, this method is not specifically designed for facial AU detection. *The low-*

* Corresponding Author

dimensional global image embeddings extracted by CLIP's image encoder cannot sufficiently capture the motion information of multiple AUs occurring in the face, thereby limiting effective interaction with the text embeddings. Furthermore, AU detection is characterized by locality and diverse intensity variations, while the interdependencies among AUs (e.g., cheek raising co-occurring with lip corner pull) are also crucial for accurate detection. *Although CLIP's image encoder can capture global relationships within a face through the self-attention mechanism, it still struggles to accurately establish long-range dependencies between local AU regions due to noises introduced by irrelevant facial areas.*

To address the aforementioned challenges, we propose a novel refined behavior-prompt approach featuring a coarse-to-fine tree attention mechanism. This model is specifically designed for facial AU detection task based on CLIP, enabling more effective utilization of semantic information from behavior prompts to extract AU-specific representations. Specifically, we leverage the powerful semantic representation capabilities of the large vision-language model to efficiently encode comprehensive prior knowledge of AU behaviors while introducing learnable prompts to better adapt to the AU detection task. Additionally, considering the diverse intensities and correlations of AUs, we introduce a coarse-to-fine tree attention module that utilizes a multi-scale and hierarchical attention mechanism to capture the fine-grained features and long-range dependencies of AUs. Finally, we deploy a novel feature interaction learning module that leverages behavioral prompts to decouple AU-specific representations and incorporates context-related visual information into the prompts during training, further mitigating vision-text bias and enhancing the guidance capability of the behavioral prompts.

The main contributions can be summarized as follows:

- We propose learnable behavior-prompts to encode location and motion cues, thereby adapting CLIP's text encoder for effective AU representation.
- A tree attention is explored to capture fine-grained AU features and long-range dependencies by leveraging the multi-scale and hierarchical mechanism.
- A feature interaction framework is designed to leverage behavior-prompts for disentangling AU-specific features while integrating contextual information to mitigate vision-text bias and enhance guidance capabilities of prompts.
- Extensive experiments demonstrate that the proposed method outperforms state-of-the-art (SOTA) methods on two widely-used benchmarks, namely BP4D and DISFA.

II. Related Work

A. Facial Action Unit Detection

Automatic facial AU detection plays a vital role in enabling computers to understand human emotions. As data-driven deep learning models have demonstrated powerful representational capabilities in computer vision tasks, many approaches have leveraged these models and incorporated specific designs to improve the performance of facial AU detection. For example, based on the traditional convolutional layer, Zhao et al. [36] redesigned a novel region layer that divides the entire feature into several uniform patches, learning these patches independently to capture features from different facial regions. Motivated by this work, Li et al. [10] proposed an enhancing layer and a cropping layer in EAC-Net, which is based on a pretrained VGG-19 [22]. In both layers, predefined facial landmarks were considered as AU-related regions to guide the network's attention to AU-specific areas and crop AU-related features for further learning. Since then, several studies [5], [7] have followed this cropping approach to explicitly learn AU features. However, significant challenges remain, s uch as variations in the relationship between facial landmarks and AUs across individuals, as well as differences in the size of regions of interest (ROIs) corresponding to different AUs. To address these issues, Shao et al. [18] explored a multi-task learning approach combining facial landmark detection and facial AU detection, leveraging their positional correlation to adaptively generate attention maps for each AU. Additionally, they further developed the region layer in DRML [36], proposing a hierarchical and multi-scale region layer. This layer consists of three hierarchical convolution layers, with each layer dividing features into several patches of varying sizes to capture AU-specific features. Although facial landmarks provide valuable information for AU detection, their performance remains limited. As a result, other studies have incorporated frame sequences to capture the temporal dynamic information of AUs. For instance, Shao et al. [21] introduced a temporal Graph Neural Network (GNN), which considers the temporal information of each node in a set of spatial graphs to capture AU dynamics. Li et al. [11] utilized a transformer to model the spatial relationships and inter-frame context of AUs. Other approaches [2], [6] have also been explored. Although these methods have achieved promising results, they often require substantial computational resources and labeled data.

More recently, two works in facial AU detection have incorporated textual descriptions as auxiliary information to help the model capture the dynamic features of AUs. For instance, Yang et al. [30] proposed a cross-modality attention module that combines semantic embeddings from AU descriptions in the FACS with visual features from input images to generate an attention map and capture discriminative AU-related features. However, this approach only utilizes location information to generate the attention map and weight the visual features, without fully exploiting the motion information from textual descriptions to guide the extraction of visual features. Zhang et al. [34] used label names and behavioral descriptions to fine-tune a pretrained model for both facial expression recognition and facial AU detection. Although this work aligns text and image embeddings using the pretrained model, it lacks designs specifically tailored for AUs. Therefore, leveraging textual descriptions to assist in facial AU detection deserves further exploration.

B. Vision-Language Models

Vision-language pre-training models are developed through contrastive learning on large-scale image-text pairs. Due to the convenience and cost-effectiveness of collecting such pairs from the internet, vision-language models have made significant progress. One of the most prominent models is CLIP [16], which leverages over 400 million image-text pairs collected from the internet for contrastive learning, achieving impressive zero-shot performance. This enables CLIP to be applied to a wide range of downstream computer vision tasks. Recently, inspired by CLIP, Li et al. [9] trained a multimodal mixture of encoder-decoder models that can flexibly transfer to vision-language understanding and generation tasks by effectively utilizing noisy network data. Li et al. [12] achieved simple and efficient training for CLIP by randomly masking or removing large portions of the image patches in image-text pairs. Furthermore, recent studies have explored applying vision-language models to facial expression-related tasks. For example, Li et al. proposed the CLIPER [8], which builds upon CLIP by designing multiple expression textual descriptors to learn fine-grained expression representations. Zhao et al. [37] further introduced a Transformer-based module to better capture temporal information for dynamic facial expression recognition. Zhang et al. [34] enhanced the CLIP model by integrating label names and behavioral descriptions to improve facial representation. However, existing approaches have not been specifically tailored for facial AU detection.

Compared with existing methods, our proposed methods customize a coarse-to-fine tree attention module to capture fine-grained image features and long-range dependencies for the facial AU detection task. Learnable behavioral prompts are also introduced to integrate AU motion information into the model. Additionally, we propose a novel feature interaction learning module that decouples AU-specific representations and further incorporates context-related visual information from the image features into the prompts for further refinement. Through these innovative designs, our model, built upon CLIP, achieves superior performance compared with previous SOTA methods.

III. METHODOLOGY

A. Overview

The primary task of facial AU detection is to identify the occurrence of all AUs in a given image. Since it is essential to accurately capture the appearance changes in the localized facial regions corresponding to different AUs, many methods rely on facial landmarks and frame sequences to incorporate location and motion information of AUs, respectively, thereby enhancing the model's representational capabilities. However, combining these data can lead to a computationally heavy model. Recently, with the rapid advancements in large visual-language pre-training models, textual description has emerged as a valuable resource to provide both location and motion details for AU detection. In this paper, to leverage large visual-language models for the AU detection task, this paper proposes a refined behavior-prompt AU detection method based on CLIP.

Fig. 2 illustrates the proposed method. Specifically, we introduce a coarse-to-fine tree attention module atop CLIP's image encoder, which utilizes a multi-scale and hierarchical attention mechanism to capture fine-grained AU features and their long-range dependencies. Additionally, we harness the powerful semantic representation capabilities of the large-scale vision-language model to encode learnable AU behavior prompts, ensuring that the text embedding incorporates comprehensive prior knowledge of AU behaviors, thus enhancing adaptation to the AU detection task. Finally, we propose a novel feature interaction module that capitalizes on the information interaction capabilities of the Transformer. This module not only decouples AU-specific representations from AU image features based on the behavior prompts but also incorporates context-related visual information into the learnable behavior prompts during training, further mitigating vision-text bias and enhancing performance for subsequent learning.

B. Image Encoder with Coarse-to-Fine Tree Attention Module

In this subsection, we first describe the image encoder used to extract AU-relevant features. Given a source facial image I from dataset D, AU-related features are extracted through a well-designed backbone network. The pre-trained CLIP image encoder, commonly used in general computer vision tasks, serves as the feature extraction backbone, denoted as $\mathcal{F}_{\mathcal{I}}$. Due to the subtle appearance changes associated with AUs, the low-dimensional image embeddings produced by the vanilla CLIP model are insufficient for capturing AU-specific features. Consequently, we transform the output of the image encoder to generate high-dimensional image features, denoted as $\mathcal{F}_{\mathcal{I}}(I) \in \mathbb{R}^{d \times h \times w}$, where d, h, w represent the dimension of channel, height, width of the image features, respectively.

Since CLIP is not specifically designed for the AU detection task, and AUs are characterized by localized facial muscle movements and coexistence at varying intensities, we propose a coarse-to-fine tree attention mechanism. This mechanism, inspired by [25], captures fine-grained features and their long-range dependencies through a multi-scale and hierarchical representation. Specifically, the tree attention recursively partitions the tokens generated by the image features into four uniform patches at each level, forming a token pyramid, as illustrated in Fig. 3. At each level, the top K patches with the highest attention scores are selected, and attention at the next level is computed only within the relevant regions corresponding to the selected top K patches.

The tree attention module consists of four Vision Transformer layers, each of which further contains a tree attention layer and an MLP layer. More specifically, for the tree attention layer, the image features $\mathcal{F}_{\mathcal{I}}(I)$ generate queries, keys, and values, denoted as $Q = \{q_1, q_2, \cdots, q_{hw}\}$, $K = \{k_1, k_2, \cdots, k_{hw}\}$, and $V = \{v_1, v_2, \cdots, v_{hw}\}$, respectively.

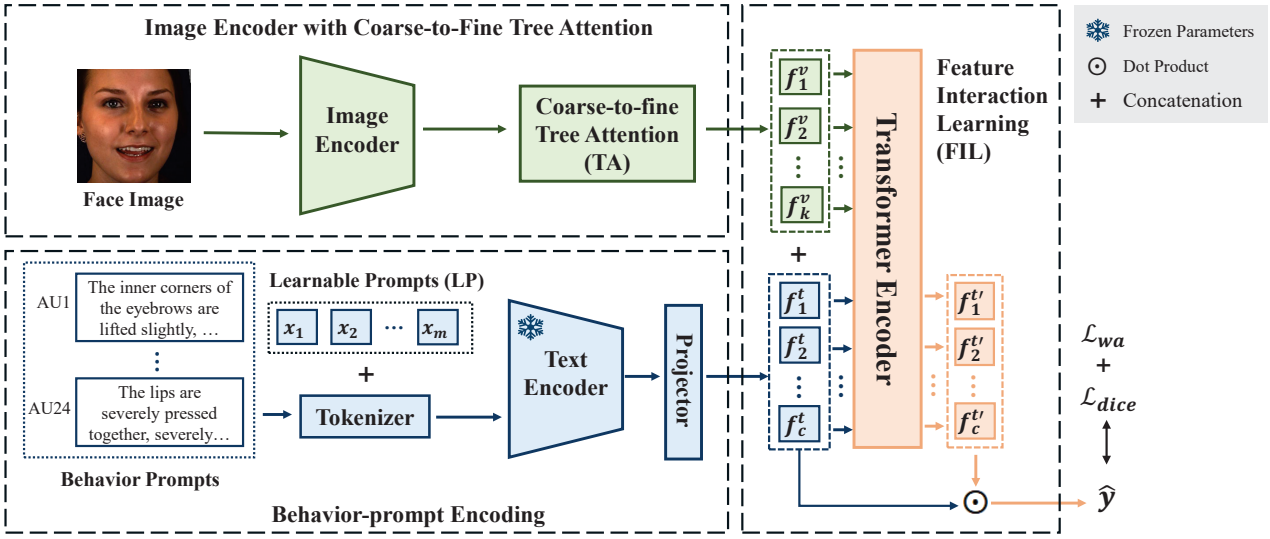

Fig. 2. An overview of the proposed method. The facial image is first passed through a pre-trained image encoder, and then input into a coarse-to-fine tree attention module to extract fine-grained AU image features. The behavioral prompts of the AUs are tokenized and concatenated with learnable prompts, and then fed into the pre-trained text encoder to obtain the corresponding behavioral text embeddings for the AUs. These embeddings are subsequently aligned with the image features using a projection layer. Finally, the image features and text embeddings are concatenated and passed to the feature interaction module, which decouples the AU-specific visual representations. The visual representations are then multiplied by the corresponding text embeddings for a dot-product similarity calculation to obtain the final prediction results.

For each q_i, it will compute a weighted average of the corresponding results from different levels:

$$s_i = \sum_{1 \le l \le L} w_i^l s_i^l, \tag{1}$$

$$s_i^l = Attention(q_i, K_{\Omega_i}^l, V_{\Omega_i}^l), \tag{2}$$

where w_i^l is a learnable weight, $K_{\Omega_i}^l$ and $V_{\Omega_i}^l$ are matrices composed of all keys and values within the region Ω_i^l, and the l-th level is denoted as $l \in \{1, 2, \cdots, L\}$. Attention represents standard attention computing. The whole process can be seen in Fig. 3. Finally, the output of the tree attention layer, denoted as $S = \{s_1, s_2, \cdots, s_{hw}\}$, is passed through the MLP layer to obtain updated image features, and this process is repeated until the final image feature $F^v \in \mathbb{R}^{d \times hw}$ is obtained.

C. Behavior-prompt Encoding

To incorporate AU-specific behavior information, our core idea is to leverage the powerful semantic representation capabilities of large vision-language models to explicitly encode comprehensive prior knowledge of AU behaviors. Specifically, the AU-specific behavioral descriptions, along with learnable prompts, are processed by a text encoder to generate the corresponding AU-specific text embeddings.

Similar to previous work [30], we adopt detailed AU-specific behavior descriptions from FACS [3] as prompts. These AU-specific behavior prompts are tokenized as the text input, represented as $Tokenizer(description)_k$ in (3). Meanwhile, following [38], we further introduce the learnable prompts as an additional component of the text input to enhance adaptation to the AU detection task. The final form of the prompts is given as follows:

$$p_k = [x]_1[x]_2 \cdots [x]_M[Tokenizer(description)]_k, \tag{3}$$

where $[x]_m$ represents a vector with the same dimension as word embeddings, M represents the number of the learnable

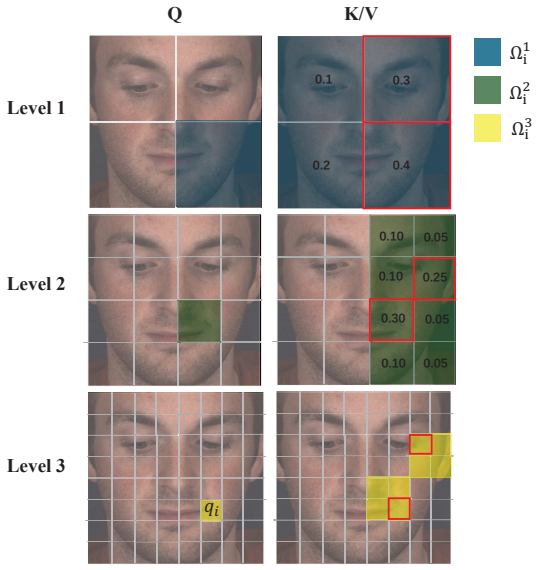

Fig. 3. An illustration of the tree attention mechanism for a query token q_i. The query token on the left is computed with the key/value tokens of the corresponding color on the right to obtain attention scores. The numbers on these patches indicate the attention scores, while the red boxes highlight the regions with the top-2 attention scores. Computation at the next level is carried out only within these highlighted regions.

prompt tokens, and k represents the corresponding AU category. p_k represents the AU-specific prompt and The complete set of prompts is denoted as $P = \{p_1, p_2, \cdots, p_k\} \in \mathbb{R}^{d \times c}$, where d represents the channel dimension and c is the number of AU categories.

Additionally, the incorporation of the tree attention module for fine-grained AU feature extraction may lead to potential misalignment between the image features and the text embeddings. To address this issue, we first encode the AU-specific behavior prompts using CLIP's text encoder $\mathcal{F}_{\mathcal{T}}$, and then pass them through a projector consisting of two fully connected layers and a nonlinear activation function. This guarantees proper alignment between the text embeddings and the image features. The entire process is formulated as:

$$F^t = GELU(\mathcal{F}_{\mathcal{T}}(P)w_1 + b_1)w_2 + b_2, \tag{4}$$

where w_1, w_2, b_1, and b_2 are the weight matrices and bias vectors in the projector. $GELU$ is the non-linear activation function. $F^t \in \mathbb{R}^{d \times c}$ represents the aligned text embeddings.

D. Feature Interaction Learning

To decouple AU-specific representation from image features using behavior prompts, we exploit the information interaction capabilities of a Transformer-based architecture. Inspired by [28], the interaction process utilizes a transformer encoder $\mathcal{F}_{\mathcal{C}}$, which takes image features F^v and text embeddings F^t as inputs. The input is represented as $F = (F^v, F^t) \in \mathbb{R}^{d \times (hw+c)}$. Through the multi-head self-attention mechanism, the transformer encoder effectively decouples highly relevant AU-specific visual representations $F^{t'}$, leveraging the semantic information provided by behavior-prompts corresponding to AUs. Finally, the dot-product similarity between this AU-specific representations $F^{t'}$ and the corresponding text embeddings F^t are computed to obtain the predicted probability $\hat{y}$ for AUs. This process is expressed as:

$$F' = \mathcal{F}_{\mathcal{C}}(F), \tag{5}$$

$$\hat{y} = sigmoid(F^{t'} \cdot F^t), \tag{6}$$

where $F' = (F^{v'}, F^{t'}) \in \mathbb{R}^{d \times (hw+c)}$ represents the output of the Transformer encoder and $\hat{y}$ denotes the predicted probabilities. Furthermore, based on this design, the model can incorporate context-related visual information into the prompts through backpropagation during training, which further mitigate vision-text bias and enhance the guidance capability of the behavioral prompts for subsequent learning.

E. Loss Function

Facial AU detection, as a multi-label detection task, suffers from a significant label imbalance problem. To address this issue, we adopt a weighted asymmetric cross-entropy loss function [13] to improve the detection of both activated AUs and non-activated AUs that are challenging to distinguish. The formula is as follows:

$$\mathcal{L}_{wa} = -\frac{1}{N}\sum_{i=1}^{N} w_i[y_i log(\hat{y}_i) + (1 - y_i)\hat{y}_i log(1 - \hat{y}_i)], \tag{7}$$

where y_i is the ground truth, $\hat{y}_i$ is the predicted probability, and w_i is a weight of the i_{th} AU. N represents the total number of AUs. The w_i is computed by $w_i = N(1/r_i)/\sum_{j=1}^{N}(1/r_i)$, where r_i is the occurrence rate of i_{th} AU.

Additionally, considering that AU detection is often biased towards non-occurrence, we further introduce a weighted dice loss [18]. It can be formulated as:

$$\mathcal{L}_{dice} = \frac{1}{N}\sum_{i=1}^{N} w_i(1 - \frac{2y_i p_i + \epsilon}{y_i^2 + p_i^2 + \epsilon}), \tag{8}$$

where ϵ is a smooth term.

The overall loss function is formulated by integrating both components, denoted as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{wa} + \lambda_2 \mathcal{L}_{dice}, \tag{9}$$

where λ_1 and λ_2 are the trade-off parameters.

IV. EXPERIMENTS

A. Experiment Setting

1) Dataset: We evaluate the performance of the proposed method on two widely used benchmark datasets for facial AU detection: BP4D [35] and DISFA [14].

- **BP4D** consists of 328 facial videos in both 2D and 3D formats, collected from 23 females and 18 males who were asked to respond to eight different emotion-eliciting tasks. The dataset contains approximately 140,000 frames, each annotated with the occurrence or non-occurrence of 12 AUs.
- **DISFA** recorded 27 facial image sequences from 12 females and 15 males while they watched emotion-eliciting video clips. The dataset comprises 130,815 frames, each annotated with six levels of intensity {0, 1, 2, 3, 4, 5} for eight AUs. Reference the settings of [24], [36], if the intensity is equal or greater than 2, it is considered to be present; otherwise, it is non-present.

Following the experimental setup of [18], [36], the proposed method employs subject-exclusive three-fold cross-validation on the BP4D and DISFA datasets, reporting the average results across the three folds.

2) Data Pre-processing: For all raw images in both benchmarks, we apply a similarity transformation to align the face region and crop them to a size of 256×256 based on landmarks generated by MTCNN [33]. Subsequently, random cropping is applied to the aligned face images to resize them to 224×224, enhancing data diversity. Additionally, we incorporate various data augmentation techniques, including random horizontal flipping and random color jittering in brightness, contrast, and saturation.

3) Implementation Details: We choose ViT-B/16-based CLIP pretrained on WIT as the backbone. During training, the text encoder of CLIP is kept frozen, while the initial learning rates for the image encoder and learnable prompts are set to 1×10^{-5} and 1×10^{-4}, respectively. For other modules, the initial learning rates are set to 1×10^{-3} for BP4D and 1×10^{-4} for DISFA, respectively. Additionally, a cosine decay learning

TABLE I

F1-SCORE FOR 12 AUS ON THE BP4D DATASET. THE TOP THREE RESULTS ARE HIGHLIGHTED WITH BOLD, UNDERLINE, AND BOX, RESPECTIVELY. % IS OMMITED.

Method		Year	AU												**Avg.**
			1	2	4	6	7	10	12	14	15	17	23	24	
Static image-based	DRML [36]	2016	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
	EAC-Net [10]	2017	39.0	39.0	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
	LP-Net [15]	2019	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
	SRERL [7]	2019	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	63.9	47.1	53.3	62.9
	JAA-Net [18]	2020	53.8	47.8	58.2	[78.5]	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
	UGN-B [24]	2021	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
	FAUDT [5]	2021	51.7	49.3	**61.0**	77.8	[79.5]	82.9	86.3	**67.6**	51.9	63.0	43.7	[56.3]	64.2
	KDSRL [1]	2022	53.3	47.4	56.2	79.4	80.7	85.1	**89.0**	67.4	**55.9**	61.9	48.5	49.0	[64.5]
	AC2D [20]	2024	54.2	**54.7**	56.5	77.0	76.2	84.0	**89.0**	63.6	54.8	[63.6]	46.5	54.8	64.6
Sequence-based	RTATL [29]	2021	57.1	[49.7]	60.5	77.9	76.1	[84.4]	87.2	64.3	[53.5]	**67.0**	48.9	48.6	64.6
	AAR [19]	2023	53.2	47.7	56.7	75.9	79.1	82.9	[88.6]	60.5	51.5	61.9	51.0	56.8	63.8
	KS [11]	2023	[55.3]	48.6	57.1	77.5	**81.8**	83.3	86.4	62.6	52.3	61.3	**51.6**	**58.3**	**64.7**
Textual description-Based	SEV-Net [30]	2021	**58.2**	50.4	[58.3]	**81.9**	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9
	Ours	2025	54.5	47.7	57.0	77.8	78.8	82.9	88.9	[66.7]	53.3	63.1	[49.7]	55.6	**64.7**

TABLE II

F1-SCORE FOR EIGHT AUS ON THE DISFA DATASET. THE TOP THREE RESULTS ARE HIGHLIGHTED WITH BOLD, UNDERLINE, AND BOX, RESPECTIVELY. % IS OMMITED.

Method		Year	AU								**Avg.**
			1	2	4	6	9	12	25	26	
Static image-based	DRML [36]	2016	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
	EAC-Net [10]	2017	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
	JAA-Net [18]	2019	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
	LP-Net [15]	2019	29.9	24.7	72.7	46.8	49.6	72.9	[93.8]	65.0	56.9
	SRERL [7]	2020	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
	UGN-B [24]	2021	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
	FAUDT [5]	2021	46.1	48.6	**72.8**	**56.7**	50.0	72.1	90.8	55.4	61.5
	KDSRL [1]	2022	[60.4]	59.2	67.5	52.7	51.5	76.1	91.3	57.7	[64.5]
	AC2D [20]	2024	57.8	59.2	70.1	50.1	54.4	75.1	90.3	[66.2]	65.4
Sequence-based	RTATL [29]	2021	57.8	52.8	70.8	53.2	52.7	74.5	91.5	51.9	63.1
	AAR [19]	2023	62.4	53.6	[71.5]	39.0	48.8	76.1	91.3	**70.6**	64.2
	KS [11]	2023	53.8	**59.9**	69.2	54.2	50.8	[75.8]	92.2	46.8	62.8
Textual description-Based	SEV-Net [30]	2021	55.3	53.1	61.5	53.6	38.2	71.6	**95.7**	41.5	58.8
	Ours	2025	**65.0**	[55.4]	71.0	[54.1]	[52.9]	74.3	94.4	64.5	**66.4**

rate scheduler is employed along with the AdamW optimizer, with β_1 = 0.9, β_2 = 0.99, and a weight decay of 5×10^{-4}. The model is trained for 12 epochs with a batch size of 64, and the first epoch is used for linear warm-up. Both parameters λ_1 and λ_2 in the overall loss funtion (9) are set to 1. The other parameters, K, L, and M are set to 2, 3, and 16, respectively. All our experiments are conducted using an Nvidia RTX 4090 GPU based on the open-source PyTorch platform.

4) Evaluation Metrics: Following previous works [10], [18], [36], We adopt the frame-based F1-score to evaluate the performance of the methods. The F1-score is formulated as $F1 = 2\frac{P \cdot R}{P+R}$, balancing precision P and recall R by jointly considering them.

B. Comparison with State-of-the-arts

We compared the proposed method with 13 SOTA AU detection methods under the same evaluation setup. These methods include DRML [36], EAC-Net [10], JÂA-Net [18], LP-Net [15], SRERL [7], UGN-B [24], FAUDT [5], KDSRL [1], AC^{2}D [20], RTATL [29], AAR [19], KS [11], and SEV-Net [30]. Among these methods, RTATL, AAR and KS take frame sequences as input and incorporate temporal information, while the remaining methods rely on static images as input. Additionally, compared with other methods, both SEV-Net and the proposed method integrate behavioral textual descriptions of AUs as additional input.

Table I presents the performance of our proposed method and these SOTA methods on the BP4D dataset. Overall, our method demonstrates competitive performance, achieving comparable or superior results in terms of the average F1-score across all SOTA methods. Specifically, compared with sequence-based methods such as RTATL, AAR, and KS, the proposed approach performs comparably by utilizing behavioral descriptions to incorporate motion information for AU

TABLE III
COMPONENT ABLATION STUDY ON THE BP4D DATASET USING F1-SCORE (%). BASELINE: PRE-TRAINED CLIP WITH LEARNABLE IMAGE ENCODER. COMPONENT DEFINITIONS: FIL (FEATURE INTERACTION LEARNING), TA (TREE ATTENTION), LP (LEARNABLE PROMPTS).

Baseline	FIL	TA	LP	F1-score
✓				62.9
✓	✓			63.5
✓	✓	✓		64.0
✓	✓	✓	✓	64.7

detection, rather than relying on extensive temporal inputs. Notably, for AUs such as AU12 and AU14, which occur in similar facial regions but exhibit distinct movement patterns, our method achieve competitive performance in terms of F1-score. The above results indicate that behavioral descriptions are beneficial in extracting AU-related motion information. Furthermore, compared with SEV-Net, which also incorporates AU behavioral descriptions as additional input, our method improves performance by 1.3%.

Table II summarizes the performance of all methods on the DISFA dataset. The proposed method outperforms all SOTA methods, achieving the highest average F1-score, with an improvement of at least 1.5%. Especially, compared with SEV-Net, our method boosts overall performance by 12.9%.

C. Ablation study

1) Effectiveness of Each Component: To evaluate the effectiveness of each component in our proposed method, we conduct a series of ablation experiments on the BP4D dataset. Initially, we use the ViT-B/16-based CLIP model as the baseline, with a frozen text encoder and a learnable image encoder. Building upon this baseline, we sequentially add the feature interaction learning module, the coarse-to-fine tree attention module, and the learnable prompts, performing training and validation at each stage. Table III demonstrates that each module contributes approximately 0.5 to 0.7 percentage points to the average F1-score.

2) Advantage of the Tree Attention: To capture the fine-grained features related to AUs, it is crucial to select an appropriate attention module. In this subsection, we compare four different attention modules. One of these is an activation-based spatial attention [32], which is computed by summing the absolute values raised to the power of p (where $p = 2.0$) in each channel. The Convolutional Block Attention Module (CBAM) [26] is a simple yet effective attention module that computes attention across both channel and spatial dimensions. Additionally, we also evaluate the original self-attention module from Vision Transformers [4] and tree attention module utilized in our approach. As shown in Fig. 4, the tree attention module achieves the highest F1-score. Based on these results, the tree attention network is adopted for extracting fine-grained AU features and long-range dependencies.

3) Visualization: In this subsection, we visualize the identified ROIs for different AUs using GRAD-CAM [17]. The comparative visualization results of the proposed method and baseline are shown in Fig. 5, demonstrating that our method achieves more precise localization of the target AU regions compared with the baseline, with a particularly notable improvement observed for AU10.

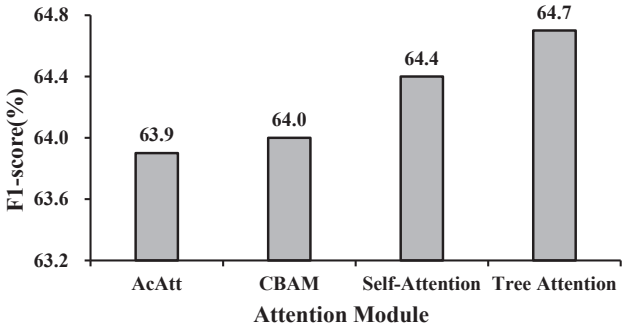

Fig. 4. F1-score(%) of different attention modules. AcAtt denotes the activation-based spatial attention. CBAM denotes the convolutional block attention module. Self-attention denotes the traditional attention module. Tree Attention denotes our proposed coarse-to-fine tree attention module.

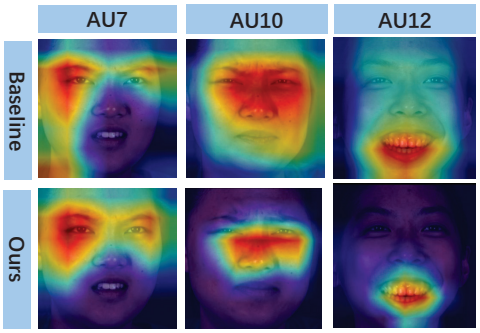

Fig. 5. Visulization of AU7, AU10, and AU12 for the baseline and our proposed methods. (Best viewed in color).

V. CONCLUSIONS

In this paper, we propose a refined behavior-prompt facial AU detection model featuring a coarse-to-fine tree attention mechanism. Specifically, we introduce learnable behavior prompts to leverage a large visual-language model, effectively capturing comprehensive prior knowledge of AU behavior. Moreover, due to the varying intensities and interactive nature of AUs, we propose a coarse-to-fine tree attention module to more effectively capture fine-grained visual features of individual AUs and their long-range dependencies. Additionally, a feature interaction module is employed to decouple AU-specific representations and incorporate content-related visual information into the prompts during training, further mitigating vision-text bias and enhancing the guidance capability of the behavioral prompts for subsequent learning. Extensive experiments on benchmark datasets demonstrate that our proposed method achieves competitive performance compared with SOTA methods.

ACKNOWLEDGMENT

This work was partially supported by the Eudaimonia Institute at the University of Oulu, and by the Research Council of Finland under the High-Performance Computing (HPC) project "FaceCanvas: Action Unit Palette for Facial Transformations (Grant No. 364905)". The authors gratefully acknowledge the computational resources provided by the HPC infrastructure and the foundational support from both institutions.

REFERENCES

[1] Y. Chang and S. Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022.

[2] W.-S. Chu, F. De la Torre, and J. F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017.

[3] J. F. Cohn, Z. Ambadar, and P. Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[5] G. M. Jacob and B. Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7680–7689, 2021.

[6] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016.

[7] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8594–8601, 2019.

[8] H. Li, H. Niu, Z. Zhu, and F. 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.

[9] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[10] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 103–110. IEEE, 2017.

[11] X. Li, X. Zhang, T. Wang, and L. Yin. Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20979–20989, 2023.

[12] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[13] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1239–1246, 2022.

[14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[15] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11917–11926, 2019.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[18] Z. Shao, Z. Liu, J. Cai, and L. Ma. Jaa-net: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021.

[19] Z. Shao, Y. Zhou, J. Cai, H. Zhu, and R. Yao. Facial action unit detection via adaptive attention and relation. *IEEE Transactions on Image Processing*, 32:3354–3366, 2023.

[20] Z. Shao, H. Zhu, Y. Zhou, X. Xiang, B. Liu, R. Yao, and L. Ma. Facial action unit detection by adaptively constraining self-attention and causally deconfounding sample. *International Journal of Computer Vision*, pages 1–16, 2024.

[21] Z. Shao, L. Zou, J. Cai, Y. Wu, and L. Ma. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv e-prints*, pages arXiv–2001, 2020.

[22] K. Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] S. Song, S. Jaiswal, L. Shen, and M. Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2020.

[24] T. Song, L. Chen, W. Zheng, and Q. Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5993–6001, 2021.

[25] S. Tang, J. Zhang, S. Zhu, and P. Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[27] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2871–2880, 2020.

[28] J. Yan, S. Huang, N. Mu, L. Huangfu, and B. Liu. Category-prompt refined feature learning for long-tailed multi-label image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2146–2155, 2024.

[29] J. Yan, J. Wang, Q. Li, C. Wang, and S. Pu. Self-supervised regional and temporal auxiliary tasks for facial action unit recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1038–1046, 2021.

[30] H. Yang, L. Yin, Y. Zhou, and J. Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10482–10491, 2021.

[31] Z. Zafar and N. A. Khan. Pain intensity evaluation through facial action units. In *2014 22nd International Conference on Pattern Recognition*, pages 4696–4701. IEEE, 2014.

[32] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[34] X. Zhang, T. Wang, X. Li, H. Yang, and L. Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20751–20762, 2023.

[35] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[36] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[37] Z. Zhao and I. Patras. Prompting visual-language models for dynamic facial expression recognition. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[38] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.