2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Cognition-Inspired Dual-Stream Semantic Enhancement for Vision-Based Dynamic Emotion Modeling

Huanzhen Wang[1], Ziheng Zhou[1], Zeng Tao[1], Aoxing Li[1], Yingkai Zhao[1], Yuxuan Lin[2], Yan Wang[3,*], Wenqiang Zhang[1,2,*]

***Abstract*— The human brain constructs emotional percepts not by processing facial expressions in isolation, but through a dynamic, hierarchical integration of sensory input with semantic and contextual knowledge. However, existing vision-based dynamic emotion modeling approaches often neglect emotion perception and cognitive theories. To bridge this gap between machine and human emotion perception, we propose cognition-inspired Dual-stream Semantic Enhancement (DuSE). Our model instantiates a dual-stream cognitive architecture. The first stream, a Hierarchical Temporal Prompt Cluster (HTPC), operationalizes the cognitive priming effect. It simulates how linguistic cues pre-sensitize neural pathways, modulating the processing of incoming visual stimuli by aligning textual semantics with fine-grained temporal features of facial dynamics. The second stream, a Latent Semantic Emotion Aggregator (LSEA), computationally models the knowledge integration process, akin to the mechanism described by the Conceptual Act Theory. It aggregates sensory inputs and synthesizes them with learned conceptual knowledge, reflecting the role of the hippocampus and default mode network in constructing a coherent emotional experience. By explicitly modeling these neuro-cognitive mechanisms, DuSE provides a more neurally plausible and robust framework for dynamic facial expression recognition (DFER). Extensive experiments on challenging in-the-wild benchmarks validate our cognition-centric approach, demonstrating that emulating the brain's strategies for emotion processing yields state-of-the-art performance and enhances model interpretability.**

I. INTRODUCTION

Facial expressions serve as the core cue for conveying human emotions, capable of rapidly activating emotion-processing brain regions such as the amygdala to enable emotional salience detection and empathy inference [1]. They also serve as a universal channel for emotional communication and are widely applied in fields such as healthcare, robotics, and human-computer interaction [2]. While dynamic facial expression recognition (DFER) leverages temporal cues for improved emotional interpretation [3], unimodal approaches often falter under occlusion or noise due to their limited perceptual scope [4]. To enhance robustness, recent research has turned to multimodal strategies integrating visual and auditory signals [5], with multiphysical methods emerging as a promising direction [6]. The emergence of multimodal and cross-modal approaches also poses interpretability challenges for vision-based emotion modeling that align with human cognition. This paper is dedicated to exploring the design of bionic algorithm models that align with the theoretical foundations and practical requirements of affective cognitive science.

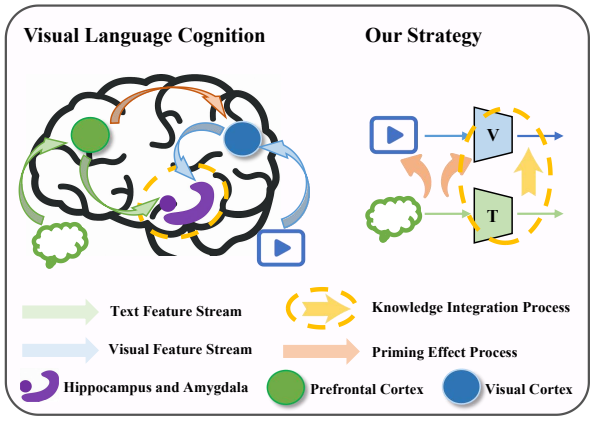

Fig. 1. The priming effect and knowledge integration mechanisms in human emotional cognition inspire us to explore the progressive complementary relationship between prompts and knowledge in visual language sentiment modeling.

When the brain recognizes emotions, it integrates inputs from multiple sources such as vision and language, relying on core regions including the amygdala, insula, and prefrontal cortex [7]. Research indicates that emotional stimuli—whether visual facial expressions or auditory cues—activate regions including the amygdala and insula [8]. Several key points warrant attention in the process of emotional transmission. On one hand, the brain exhibits a hierarchical structure across temporal scales. Experimental evidence indicates that lower-level sensory cortices possess intrinsic short time scales, whereas higher-level cross-modal networks operate on longer time scales [9]. On the other hand, the brain's perception of emotion is profoundly influenced by natural language and contextual semantics. Research reveals that when individuals receive information through narrative or linguistic channels, the hippocampus and default mode network rapidly retrieve relevant conceptual memories to interpret emotional cues [10]. Subsequently,

[1]Huanzhen Wang, Ziheng Zhou, Zeng Tao, Aoxing Li and Yingkai Zhao are with College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China. {hzwang24, zhouzh24, axli24, ykzhao24}@m.fudan.edu.cn, ztao19@fudan.edu.cn

[2]Yuxuan Lin and Wenqiang Zhang are with Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China. yuxuanlin24@m.fudan.edu.cn, wqzhang@fudan.edu.cn

[3]Yan Wang is with School of Data Science and Engineering, East China Normal University, Shanghai, China. yanwang@dase.ecnu.edu.cn

*Corresponding author: Wenqiang Zhang and Yan Wang.

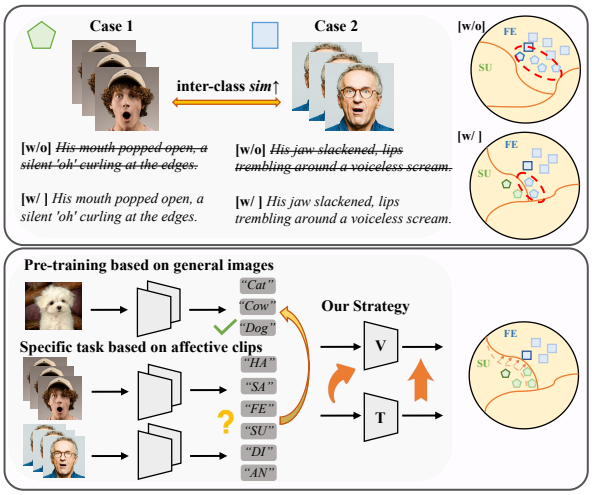

Fig. 2. Supervision from natural language can effectively alleviate the problem of difficulty in classifying expressions with inter-class similarities. Our method aims to bridge the gap between general image-based pre-training models and specialized recognition tasks based on dynamic expression sequences.

the hippocampus and medial temporal lobe integrate these experiences into conceptual knowledge, which is then accessed through the default mode network. As shown in the Fig. 1, these theories inspire us to explore the relationship between temporality and semantics, as well as between cues and knowledge, within the process of emotional modeling.

In response to actual task requirements, as illustrated in Fig. 2, subtle inter-class similarities and intra-class variations complicate DFER. Existing methods—whether supervised or self-supervised—typically rely on visual cues alone [11], overlooking the semantic depth offered by language. One-hot labels lack contextual information, limiting generalization and interpretability. In contrast, natural language supervision introduces richer, human-aligned guidance that helps disambiguate similar expressions. Vision-language models like CLIP [12] have been adapted to DFER via prompt tuning and fine-tuning [13], showing promising results.

Although CLIP exhibits strong generalization to novel concepts, its large-scale architecture and the scarcity of task-specific data render full-model fine-tuning impractical for downstream DFER tasks. While prior work in DFER has largely focused on adapting the visual encoder, findings from semantic segmentation [14] indicate that CLIP's text encoder contains rich semantic priors that remain underexploited in emotion recognition. Limited modal interaction, challenging knowledge transfer, and interpretability constraints aligned with human emotional cognition restrict the potential of visual-language models in dynamic emotion modeling.

To address the aforementioned challenges, we propose DuSE, a cognition-inspired dual-stream semantic enhancement framework for vision-based dynamic emotion modeling. Specifically, DuSE integrates a cross-modal prompt streaming to align textual emotion descriptions with fine-grained facial features, and a cross-domain knowledge streaming to transfer general visual knowledge to the facial expression domain. The algorithmic implementation of this "dual-mechanism" framework—which integrates preset expectations and knowledge through prompts—enables embodied cross-modal emotion perception.

Our main contributions are summarized as follows:

- Through cognitive affective analysis, we have revealed the gap between human multimodal dynamic emotion perception and unimodal DFER systems. Integrating cognitive theory with current applications, we propose the DuSE method.
- Inspired by the priming effect and knowledge integration process in cognitive theory, we design the Hierarchical Temporal Prompt Cluster (HTPC) to support cross-modal prompt streaming and we design the Latent Semantic Emotion Aggregator (LSEA) to support cross-domain knowledge streaming.
- We validate the effectiveness of our approach through extensive experiments on two challenging in-the-wild DFER benchmark datasets, demonstrating its superior performance over state-of-the-art methods.

II. RELATED WORKS

A. Cognitive Science and Neuroscience

Advances in cognitive science and neuroscience provide theoretical foundations for model design in artificial intelligence. In the human emotional generation process, priming effects and knowledge integration play pivotal roles. In cognitive science, priming effects refer to how external semantic or contextual cues alter the brain's processing of sensory stimuli. For instance, linguistic cues accelerate emotional categorization of ambiguous facial expressions. This process relies on the prefrontal cortex and hippocampus regulating visual pathways, enabling cross-modal semantic-sensory interaction [15]. Simultaneously, the brain does not process emotions in isolation but relies on long-term semantic memory and contextual knowledge to interpret sensory inputs. The Conceptual Act Theory [16] proposes that emotions are constructed through the integration of bodily signals with semantic knowledge via the hippocampus, default mode network, and prefrontal cortex, rather than being directly read. The hierarchical temporal processing mechanisms of the brain [17] and the information integration mechanisms between semantic memory and the prefrontal cortex [18] provide interpretability for the bionic design of neurotransmitters. Inspired by this, we designed an algorithmic implementation of a complementary "dual-mechanism" cognitive framework: achieving embodied cross-modal emotion perception through cue-based expectation setting and knowledge integration.

B. Dynamic Facial Expression Recognition

Early studies on facial expression recognition relied on handcrafted features in controlled environments [19]. With the advent of deep learning and large-scale DFER datasets, data-driven approaches have become mainstream. Unlike static FER, DFER requires modeling spatiotemporal dynamics to capture expressive variations over time. The growing

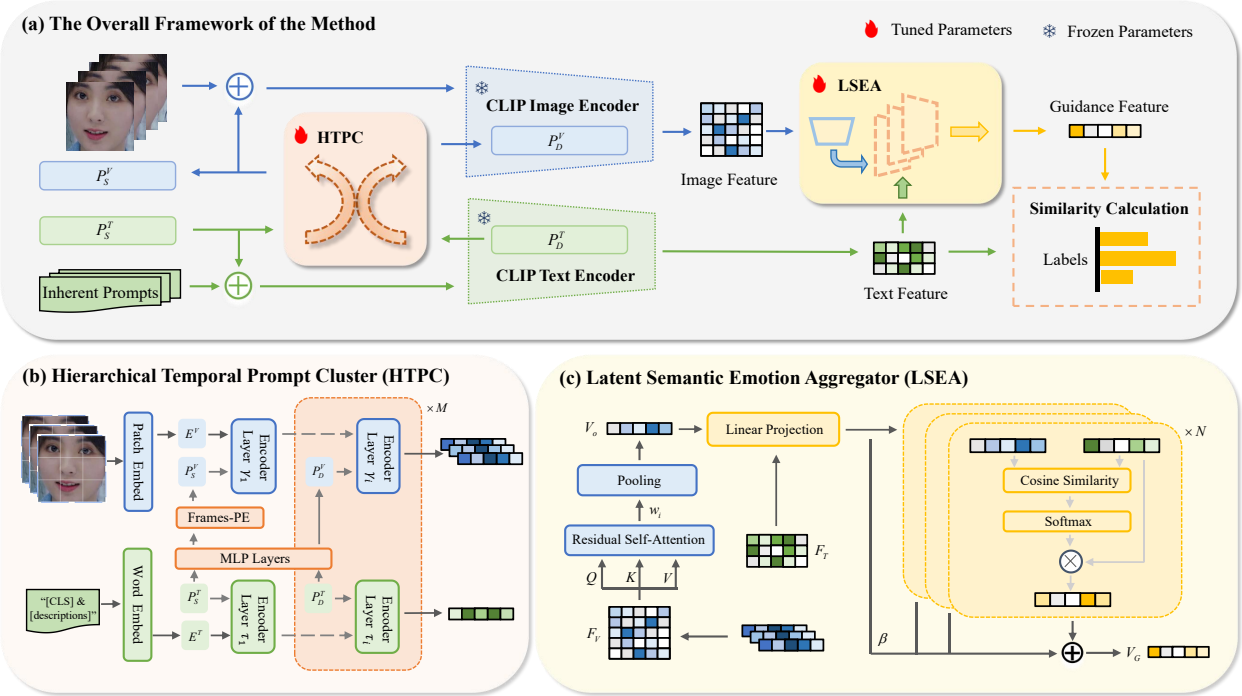

Fig. 3. **Overall architecture of DuSE**. (a) shows the overall methodological framework. (b) shows HTPC, which contributes to the cross-modal prompt streaming. (c) shows LSEA, which contributes to cross-domain knowledge streaming.

availability of in-the-wild datasets [20], [21] has established DFER as a distinct research task, prompting the development of specialized methods to address its unique challenges. Recent advances in deep learning have shifted FER from static image analysis to motion-aware frameworks. Models like C3D [22] effectively capture spatiotemporal features and long-range dependencies, essential for modeling dynamic expressions. Meanwhile, vision-language models such as CLIP [12] and CLIPER [13] enable robust semantic alignment across modalities. Unlike their approach, ours focuses more on mining pre-trained textual prior knowledge to augment dynamic sentiment modeling from a semantically guided perspective.

C. Dynamic Prompting with Cross-Domain Transfer

Prompt learning [23] has been extended from NLP to vision-language and vision-only models, enabling efficient adaptation to downstream tasks by learning task-specific prompts with minimal parameter updates. Recent text-based sentiment analysis leverages emotion-related prompts to reformulate classification as masked language modeling for data-efficient learning [24], while visual prompt tuning methods such as CoOp [25], VPT [26], and MaPLe [27] optimize learnable prompts to enhance cross-modal generalization in vision-language models like CLIP. However, these approaches mainly address coarse-grained object recognition and struggle with the fine-grained, temporally-evolving nature of DFER. To overcome this, we incorporate rich emotional textual prompts into facial imagery and introduce dynamic prompt tuning to capture temporal variations. Moreover, we enhance generalization via cross-domain knowledge transfer [28], which mitigates data scarcity and domain shifts (e.g., from controlled to in-the-wild settings through adversarial learning [29] and feature disentanglement [30]. While prior methods rely on static FER as intermediates or knowledge distillation [31], they offer limited gains and often suffer modal misalignment. In contrast, we propose a dynamic knowledge migration strategy that embeds emotional concepts into facial feature dynamics, aligning cross-modal semantics to improve real-world DFER performance on subtle and context-sensitive expressions.

III. METHOD

In DuSE design, we fully incorporate the complementary relationship between prompts and knowledge within the brain. The Hierarchical Temporal Prompt Cluster (HTPC) provides context-driven anticipatory expectations, simulating the process of modulating sensory processing sensitivity and the brain's hierarchical structure. The Latent Semantic Emotion Aggregator (LSEA) analogizes knowledge aggregation and emotional semantic processing, performing a posteriori construction to generate complete emotional concepts. DuSE's algorithmic implementation of a "dual-mechanism" neuroscience framework achieves embodied cross-modal emotion perception through pre-set expectations via prompts and knowledge-integrated construction.

A. Preliminaries

The overall architecture of DuSE is depicted in Fig. 3. Specifically, the overall framework requires input downsampled video frame sequences $\mathcal{V}$ and enriched and expanded multi-class text descriptions $\mathcal{T}$. For the text part we have adopted the tagging combined with salient sentiment category features natural language description composition. They will be integrated as $\mathcal{V}_{in} \in \mathbb{R}^{t\times\mathcal{C}\times\mathcal{H}\times\mathcal{W}}$ and $\mathcal{T}_{in} \in \mathbb{R}^{c}$ as inputs to CLIP image encoder $\mathcal{F}(\cdot)$ and text encoder $\mathcal{G}(\cdot)$ under the shallow and deep prompting action of the HTPC. Where t is the number of downsampled frames, $\mathcal{H}$, $\mathcal{W}$ and $\mathcal{C}$ are the information on pixel points of the image and c is the number of categories to be categorized. The subsequently obtained video features $\mathcal{F}_{\mathcal{V}} \in \mathbb{R}^{t\times d}$ and text features $\mathcal{F}_{\mathcal{T}} \in \mathbb{R}^{c\times d}$, where d is the encoder output dimension of CLIP. Video features and text features after passing through the LSEA module will output the temporally modeled and semantically guided video fusion feature $\mathcal{V}_g \in \mathbb{R}^{d}$. Subsequently $\mathcal{V}_g$ will be aligned with $\mathcal{F}_{\mathcal{T}}$ and the contrast learning loss will be computed. Where cls is the category corresponding to the correct label and i traverses all categories.

$$\mathcal{L} = -\log \frac{\exp\left(\mathcal{V}_g \cdot \mathcal{F}_{\mathcal{T}}\left(cls\right)\right)}{\sum_i \exp\left(\mathcal{V}_g \cdot F_{\mathcal{T}}\left(i\right)\right)} \tag{1}$$

B. Hierarchical Temporal Prompt Cluster

The prompt stream is not a single unit but a cluster of both shallow and deep prompts. The shallow prompts are designed for cross-modal interaction at the input stage, before the encoder, while the deep prompts are incorporated between layers of the encoder.

If we design n learnable tokens and $\mathcal{M}$ prompt streams, where $\mathcal{M}$ cannot exceed the intrinsic number of layers $\mathcal{K}$ of the encoder, then the shallow prompting corresponds to when $\mathcal{M} = 1$, and the rest of the cases can be classified as deep prompting. The following two formulas can briefly summarize the process of prompting on the text side. Where $i = 1, 2, ..., \mathcal{M}$ represents the serial number of the layer affected by the prompt flow and $j = \mathcal{M}+1, \mathcal{M}+2, ..., \mathcal{K}$ represents unaffected, both $\mathcal{P}_i^{\mathcal{T}} \in \mathbb{R}^{n\times d_{\mathcal{T}}}$ and $\mathcal{P}_j^{\mathcal{T}} \in \mathbb{R}^{n\times d_{\mathcal{T}}}$ are learnable tokens, while τ_i and τ_j are corresponding text encoder layers. $d_{\mathcal{T}}$ is the text encoder hidden layer feature dimension. The underscore "_" in the following formulas denotes the output of the corresponding dimension, which our algorithm does not consider.

$$\left[\mathcal{E}_i^{\mathcal{T}}, \_\right] = \tau_i\left(\left[\mathcal{E}_{i-1}^{\mathcal{T}}, \mathcal{P}_{i-1}^{\mathcal{T}}\right]\right) \tag{2}$$

$$\left[\mathcal{E}_j^{\mathcal{T}}\right] = \tau_j\left(\left[\mathcal{E}_{j-1}^{\mathcal{T}}\right]\right) \tag{3}$$

Correspondingly, the following two formulas can briefly summarize the process of prompting on the video side. Where both $P_i^V \in \mathbb{R}^{n\times d_V}$ and $P_j^V \in \mathbb{R}^{n\times d_V}$ are learnable tokens, while γ_i and γ_j are corresponding image encoder layers. d_V is the image encoder hidden layer feature dimension.

$$\left[\mathcal{E}_i^{\mathcal{V}}, \_\right] = \gamma_i\left(\left[\mathcal{E}_{i-1}^{\mathcal{V}}, \mathcal{P}_{i-1}^{\mathcal{V}}\right]\right) \tag{4}$$

$$\left[\mathcal{E}_j^{\mathcal{V}}\right] = \gamma_j\left(\left[\mathcal{E}_{j-1}^{\mathcal{V}}\right]\right) \tag{5}$$

In order to reflect the guiding role of textual semantics, the video-side learnable visual tokens are generated from their textual counterparts by the parameter-shared multi-layer perceptron $\mathcal{MLP}$. Where $\mathcal{W}$ is the parameter matrix, b is bias parameters, and since it is a generalized regression task, no activation function is used before the output layer.

$$\mathcal{P}_i^{\mathcal{V}} = \mathcal{MLP}(\mathcal{P}_i^{\mathcal{T}}) = \mathrm{ReLU}(\mathcal{W}\cdot\mathcal{P}_i^{\mathcal{T}} + b) \tag{6}$$

In particular, due to the nature of the CLIP architecture as it applies to images, we would like to add dynamic prompts between frames. Therefore, for the first layer of prompts, after mapping the multilayer perceptron, a sinusoidal position encoding is performed in the frame-level dimension t, and then the position encoding vectors are summed up in the dimension of the number of learnable tokens n for the broadcast mechanism.

With the HTPC, it is also possible to obtain $\mathcal{F}_{\mathcal{V}}$ and $\mathcal{F}_{\mathcal{T}}$ as described before. Shallow and deep prompts will be injected in a layered manner according to the hierarchy. For prompt depth, we define three strategies: $Shallow$ prompting strategy means affecting only the input layer, $Normal$ prompting strategy means affecting one-third of the encoder layers and $Deep$ prompting strategy means affecting two-thirds of the encoder layers. Since the performance met expectations and the large size of the CLIP ViT-L/14 model, $Deep$ prompting strategy was not applied to it. This is consistent with the subsequent ablation experiments.

$$\mathcal{F}_{\mathcal{V}} = \mathcal{F}\left(\mathcal{E}^{\mathcal{V}};\ \mathcal{P}_s^{\mathcal{V}}@1,\ P_d^{V}@2..\mathcal{M}\right) \tag{7}$$

$$\mathcal{F}_{\mathcal{T}} = \mathcal{G}\left(\mathcal{E}^{\mathcal{V}};\ \mathcal{P}_s^{\mathcal{T}}@1,\ \mathcal{P}_d^{\mathcal{T}}@2..\mathcal{M}\right) \tag{8}$$

C. Latent Semantic Emotion Aggregator

To effectively transfer the knowledge learned by CLIP to the expression recognition domain, we propose a text-guided knowledge transfer module to reduce the domain gap. This module leverages textual knowledge to guide visual feature learning, acting as a bridge that connects facial expression images with knowledge from the natural domain.

Since $\mathcal{F}_{\mathcal{V}} \in \mathbb{R}^{t\times d}$ is a multi-frame feature, we model its multi-frame fusion using a spatio-temporal split-attention mechanism. Since spatial information has already been taken into account in the CLIP image encoder, only time series modeling is performed here. This step is mainly realized based on the self-attention mechanism, where $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$ represent the Query, Key, and Value matrices, respectively, and d_k is the dimensionality factor for appropriate scaling. Subsequently passing it through a linear layer and applying temporal attention pooling to aggregate frame-level representations into a single fused feature $\mathcal{V}_o \in \mathbb{R}^{d}$. Where the weights w_i are learned via via a learned scoring function

followed by softmax over time. w_i denotes the self-attention weight of the i-th frame.

$$\text{Attention }(\mathcal{Q},\mathcal{K},\mathcal{V}) = \text{softmax}\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}}\right)\mathcal{V} \tag{9}$$

$$\mathcal{V}_m = \text{Linear}(\text{Attention}(\mathcal{F}_\mathcal{V},\mathcal{F}_\mathcal{V},\mathcal{F}_\mathcal{V})) \in \mathbb{R}^{t\times d} \tag{10}$$

$$\mathcal{V}_o = \sum_{i=1}^{t} w_i \cdot \mathcal{V}_m^{(i)}, \quad \text{with} \sum w_i = 1 \tag{11}$$

Text features and video features are taken as inputs and measure their similarity by calculating the cosine of the vectors. Then, Softmax function is applied to normalize the similarity and construct the semantic vector $\mathcal{T}_o$ that is most similar to the image side, which is weighted and summed with the original features to achieve the latent feature embedding and obtain the text-guided image feature $\mathcal{V}_g$. In the specific implementation process, we introduce hyperparameter $\mathcal{N}$ to the generation process of $\mathcal{T}_o$, using the design of multiple heads. Specifically, we utilize a linear layer to map $\mathcal{V}_o$ and $\mathcal{F}_\mathcal{T}$ to $\mathcal{N}$ heads of the same dimension to obtain $\tilde{\mathcal{V}}_o^{(i)}$ and $\tilde{\mathcal{F}}_\mathcal{T}^{(i)}$, and perform the operations described earlier on these $\mathcal{N}$ pairs of features and finally average them to obtain $\mathcal{V}_g$.

$$\tilde{\mathcal{T}}_o^{(i)} = \text{softmax}\left(\tilde{\mathcal{V}}_o^{(i)} \cdot \tilde{\mathcal{F}}_\mathcal{T}^{(i)}\right) \cdot \tilde{\mathcal{F}}_\mathcal{T}^{(i)} \tag{12}$$

$$\mathcal{V}_g = \frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} \left(\beta \tilde{\mathcal{V}}_o^{(i)} + (1-\beta)\,\tilde{\mathcal{T}}_o^{(i)}\right) \tag{13}$$

The visual features provide low-level structural information for dynamic representation, while the semantic vectors contain high-level affective semantics from category-wide textual guidance. The weighting factor β can balance the weights of the original visual information and the semantically guided information, and introduce enhancement through the semantic attention mechanism while preserving the visual details. The semantic bootstrapping mechanism in LSEA performs all-category soft bootstrapping via a multi-head semantic attention mechanism, which adaptively extracts and fuses the most relevant semantic vectors from all category texts for each visual feature. β is used to control the influence of semantic guidance while attenuating potential noise from inter-class similarity.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: Our study evaluates the effectiveness of DuSE on two in-the-wild, video-based DFER datasets: DFEW [20] and FERV39k [21]. We use 5-fold cross-validation on DFEW to ensure thorough performance evaluation, and follow the predefined splits of FERV39k to remain consistent with its official protocol. This approach ensures a fair and rigorous comparison across different datasets and experimental settings. The results demonstrate DuSE's robustness and accuracy, particularly in real-world conditions.

TABLE I

COMPARATIVE RESULTS (%). OUR PROPOSED DUSE PERFORMS WELL ON BOTH DATASETS FOR 7-CLASS CLASSIFICATION. BEST RESULTS ARE IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED.

Method	Publication	DFEW		FERV39k	
		UAR	WAR	UAR	WAR
C3D [22]	CVPR'15	42.74	53.54	22.68	31.69
P3D [32]	ICCV'17	43.97	54.47	23.20	33.39
I3D-RGB [33]	ICCV'17	43.40	54.27	30.17	38.78
3D ResNet18 [34]	CVPR'18	46.52	58.27	26.67	37.57
EC-STFL [20]	MM'20	45.35	56.51	-	-
Former-DFER [35]	MM'21	53.69	65.70	37.20	46.85
NR-DFERNet [36]	arXiv'22	54.21	68.19	33.99	45.97
DPCNet [37]	MM'22	57.11	66.32	-	-
EST [38]	PR'23	53.94	65.85	-	-
LOGO-Former [39]	ICASSP'23	54.21	66.98	38.22	48.13
GCA-IAL [40]	AAAI'23	55.71	69.24	35.82	48.54
MSCM [41]	PR'23	58.49	70.16	-	-
M3DFEL [42]	CVPR'23	56.10	69.25	35.94	47.67
AEN [43]	CVPRW'23	56.66	69.37	38.18	47.88
DFER-CLIP [44]	BMVC'23	59.61	71.25	41.27	51.65
MAE-DFER [45]	MM'23	63.41	74.43	43.12	52.07
EmoCLIP [46]	FG'24	58.04	62.12	31.41	36.18
SW-FSCL [47]	C&C'24	57.25	70.81	36.83	49.87
CLIPER [13]	ICME'24	57.56	70.84	41.23	51.34
CDGT [48]	NN'24	59.16	70.07	41.34	50.80
LSGTnet [49]	ASC'24	61.33	72.34	41.30	51.31
UMBEnet [5]	MM'24	<u>64.55</u>	73.93	**44.01**	<u>52.10</u>
DuSE(ours)	-	**64.88**	**75.36**	<u>43.39</u>	**53.05**

2) Implementation details: For the visual input, all fixed 16-frame sequences in our DuSE experiments followed the sampling strategy described in related works and were resized to 224×224 pixels. To mitigate overfitting, we employed several data augmentation techniques, including random resized cropping, horizontal flipping, random rotation, and color jittering. The text section was designed with fixed prompt words that combine emotion categories with descriptions of subtle facial expression changes, paired with the learnable tokens mentioned earlier. These textual prompts provided semantic guidance for the model during training.

All experiments were conducted in a high-performance computing environment equipped with 4 NVIDIA GeForce RTX 3090 GPUs. During training, we used the Adam optimizer with an initial learning rate of 0.001, and employed small-batch training with a batch size of 16. For the hyperparameters, we set the HTPC parameter n to 4, the LSEA parameter $\mathcal{N}$ to 4 and β to 0.7. To improve computational efficiency, we adopted automatic mixed-precision training, using half-precision floating-point numbers where applicable. This strategy accelerated the training process and reduced GPU memory usage, allowing for faster processing and more efficient scaling. Specifically, we conducted deployment and dynamic emotion recognition tests on an actual robotic head. DuSE is implemented as a pre-information sensing model for large multimodal large language models such as Qwen and LLaMA. The detected dynamic emotions are fed into the large model as part of the prompt. The entire framework can be deployed on the robot's head, utilizing the head camera to capture video data.

TABLE II

COMPARATIVE RESULTS (%) ACROSS DIFFERENT METHODS ON VARIOUS EMOTION CATEGORIES IN DFEW. BEST RESULTS ARE IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED.

Method	Publication	Happy	Sad	Neutral	Angry	Surprise	Disgust	Fear	UAR	WAR
C3D [22]	CVPR'15	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
I3D-RGB [33]	CVPR'17	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27
P3D [32]	ICCV'17	74.85	43.40	54.18	60.42	50.99	0.69	23.28	43.97	54.47
3D ResNet18 [34]	CVPR'18	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27
EC-STFL [20]	MM'20	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
Former-DFER [35]	MM'21	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
GCA-IAL [40]	AAAI'23	87.95	67.21	70.10	76.06	62.22	0.00	26.44	55.71	69.24
SW-FSCL [47]	C&C'24	88.35	68.52	70.98	78.17	64.25	1.42	28.66	57.25	70.81
LSGTnet [49]	ASC'24	90.67	71.70	70.48	76.71	**65.01**	14.48	40.24	61.33	72.34
DuSE (Ours)	-	**92.89**	**81.05**	**72.76**	**78.51**	62.69	**20.69**	**45.57**	**64.88**	**75.36**

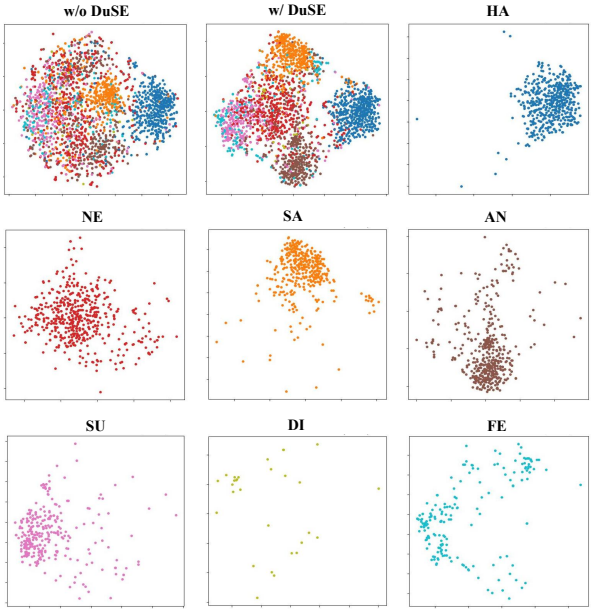

Fig. 4. Global and category-specific t-SNE visualization of DuSE on DFEW-fold5. The clustering results show that DuSE has a significant enhancement effect.

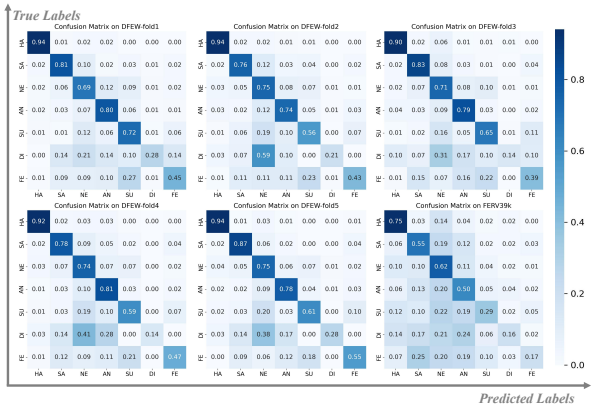

Fig. 5. Confusion matrices of the DuSE on DFEW and FERV39k.

B. Evaluation Metrics

To evaluate model performance, we adopt two key metrics: Weighted Average Recall (WAR) and Unweighted Average Recall (UAR). WAR computes the average recall weighted by class sample size, making it suitable for imbalanced datasets where certain classes dominate. In contrast, UAR treats each class equally by averaging recall across all classes, regardless of their frequency, and is more appropriate for balanced datasets. Together, WAR and UAR provide a comprehensive assessment of model effectiveness and are widely used in the DFER field.

C. Comparative Experiments

In the context of our approach, experiments were conducted on two established DFER datasets, DFEW and FERV39k, with video data as the primary input. As shown in Table I, the DuSE method performs well on both datasets. Its consistently strong performance in terms of WAR and UAR metrics highlights the effectiveness of the method in achieving optimal results across multiple datasets. Compared to other CLIP-based methods, our approach demonstrates the advantage of cross-modal information interaction without the need for fine-tuning the encoder. Even when compared to methods such as MAE-DFER [45] and HiCMAE [50], which are based on pre-training on a large amount of data, our method has some superiority. Table II demonstrates a comparison with previous work on the performance of DFEW in categorizing individual classes, where our method surpasses recent state-of-the-art methods and achieves the best results in essentially all classes.

Figures 4 and 5 show the the t-SNE visualization during DFEW training and confusion matrix for the best DuSE results. It can be seen that the CLIP pre-trained model's own natural knowledge reserve allows for simple clustering of regular expression classes such as happy and sad, and the model's ability to perceive emotions such as anger and surprise is gradually enhanced as the training process proceeds and knowledge of the emotion domain is transferred.

D. Ablation Study

In this work, we conduct ablation experiments on the DFEW and FERV39k datasets. This section focuses on intra-module and inter-module ablation experiments. The assessment metrics are consistent with the previous experiment.

Table III shows the results of the inter-module ablation experiments. Specifically, HTPC is replaced with the uni-

TABLE III

COMPARATIVE RESULTS (%) OF THE INTER-MODULE ABLATION EXPERIMENTS.

HTPC	LSEA	DFEW		FERV39k	
		UAR	WAR	UAR	WAR
×	×	55.64	66.80	33.74	46.26
×	✓	58.86	70.28	35.77	47.85
✓	×	60.27	71.83	36.53	48.80
✓	✓	**64.88**	**75.36**	**43.39**	**53.05**

TABLE IV

COMPARATIVE RESULTS (%) OF THE INTRA-MODULE ABLATION EXPERIMENTS OF HTPC (CROSS-MODAL PROMPT STREAMING).

Pre-trained model (parameters-frozen)	Strategy	DFEW		FERV39k	
		UAR	WAR	UAR	WAR
CLIP ViT-B/32	*Shallow*	49.95	61.35	33.47	44.92
	Normal	53.22	64.70	35.02	46.87
	Deep	57.75	68.76	36.93	48.54
CLIP ViT-B/16	*Shallow*	55.27	65.97	35.14	46.13
	Normal	56.84	68.15	37.32	48.05
	Deep	59.86	71.94	38.95	50.03
CLIP ViT-L/14	*Shallow*	57.03	68.98	37.28	47.94
	Normal	**64.88**	**75.36**	**43.39**	**53.05**

modal prompt tuning method CoOp, and LSEA is replaced with average pooling after ablation. The experimental results demonstrate that both modules are effective, with the introduction of each module individually significantly improving the model's performance. This highlights the importance of both the prompt and knowledge streams. Table IV presents the intra-module ablation experiments in HTPC. We perform experiments on the number of learnable tokens and the prompt depth for each of the three CLIP pretrained models with varying specifications. The results demonstrate that our approach achieves improvements across baseline models of varying scales, with increasingly pronounced effects as the influence on the visual encoder intensifies. Table V shows the impact of hyperparameters on the experiment in LSEA. We mainly conducted ablation experiments on the number of heads $\mathcal{N}$ and fusion weight β. The results indicate that too few attention heads weaken the ability to capture semantic relationships across multiple categories, while too many may lead to overfitting or excessive learning of irrelevant features. The fusion parameter balances visual features with semantically enhanced features. An excessively high value weakens semantic guidance, causing the model to confuse fine-grained emotions, while an excessively low value results in overly strong semantic information and introduces noise from non-target categories.

V. CONCLUSION

This paper analyzes the gap between human dynamic emotion perception and existing DFER methods through cognitive affect theory. Inspired by the priming effect and knowledge integration in affect cognitive theory, we propose the Dual-Stream Semantic Enhancement (DuSE) framework. This framework integrates emotional concepts into

TABLE V

COMPARATIVE RESULTS (%) OF THE HYPERPARAMETER ABLATION EXPERIMENTS OF LSEA (CROSS-DOMAIN KNOWLEDGE STREAMING).

Hyperparameter	Value	DFEW		FERV39k	
		UAR	WAR	UAR	WAR
$\mathcal{N}$	2	64.37	74.92	43.16	52.71
	4	**64.88**	**75.36**	**43.39**	**53.05**
	6	64.21	74.83	42.94	52.08
β	0.3	61.26	71.82	38.55	48.73
	0.5	62.01	73.56	40.98	50.09
	0.7	**64.88**	**75.36**	**43.39**	**53.05**
	0.9	63.38	74.33	42.47	51.74

facial appearance and leverages semantic information to transfer knowledge from general scenes to the data-scarce domain of facial expressions. This algorithmic implementation of a dual-mechanism cognitive science and neuroscience framework achieves embodied cross-modal emotion perception through prompting predefined expectations and knowledge integration. Extensive experiments on DFEW and FERV39k datasets validate our approach's effectiveness. The lightweight model framework serves as a deployment-friendly pre-processing emotion perception module for multimodal large language models. We will continue advancing research on agent emotion perception and expression, hoping this work provides valuable insights for other researchers.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No.62576109, 62072112, 62406075), National Key Research and Development Program of China (2023YFC3604802), Shanghai Key Technology R&D Program (Grant No. 25511107200).

REFERENCES

[1] R. Adolphs, "Neural systems for recognizing emotion," *Current opinion in neurobiology*, vol. 12, no. 2, pp. 169–177, 2002.

[2] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5g," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2399–2406, 2017.

[3] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.

[4] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Information Fusion*, vol. 99, p. 101847, 2023.

[5] X. Mai, J. Lin, H. Wang, Z. Tao, Y. Wang, S. Yan, X. Tong, J. Yu, B. Wang, Z. Zhou *et al.*, "All rivers run into the sea: Unified modality brain-inspired emotional central mechanism," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 632–641.

[6] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[7] D. Gündem, J. Potočnik, F.-L. De Winter, A. El Kaddouri, D. Stam, R. Peeters, L. Emsell, S. Sunaert, L. Van Oudenhove, M. Vandenbulcke *et al.*, "The neurobiological basis of affect is consistent with psychological construction theory and shares a common neural basis across emotional categories," *Communications Biology*, vol. 5, no. 1, p. 1354, 2022.

[8] A. B. Gerdes, M. J. Wieser, and G. W. Alpers, "Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains," *Frontiers in psychology*, vol. 5, p. 1351, 2014.

[9] M. Golesorkhi, J. Gomez-Pilar, F. Zilio, N. Berberian, A. Wolff, M. C. Yagoub, and G. Northoff, "The brain and its time: intrinsic neural timescales are key for input processing," *Communications biology*, vol. 4, no. 1, p. 970, 2021.

[10] M. C. Camacho, E. Deshpande, and M. T. Perino, "The cognitive–affective social processing and emotion regulation (casper) model," *Neuropsychopharmacology*, pp. 1–17, 2025.

[11] Y. Wang, S. Yan, Y. Liu, W. Song, J. Liu, Y. Chang, X. Mai, X. Hu, W. Zhang, and Z. Gan, "A survey on facial expression recognition of static and dynamic emotions," *arXiv preprint arXiv:2408.15777*, 2024.

[12] 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*. PmLR, 2021, pp. 8748–8763.

[13] 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)*. IEEE, 2024, pp. 1–6.

[14] Y. Zhang, M.-H. Guo, M. Wang, and S.-M. Hu, "Exploring regional clues in clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3270–3280.

[15] J. A. Bargh and T. L. Chartrand, "The unbearable automaticity of being." *American psychologist*, vol. 54, no. 7, p. 462, 1999.

[16] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[17] U. Hasson, J. Chen, and C. J. Honey, "Hierarchical process memory: memory as an integral component of information processing," *Trends in cognitive sciences*, vol. 19, no. 6, pp. 304–313, 2015.

[18] J. R. Binder and R. H. Desai, "The neurobiology of semantic memory," *Trends in cognitive sciences*, vol. 15, no. 11, pp. 527–536, 2011.

[19] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[20] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2881–2889.

[21] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 922–20 931.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[23] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[24] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, p. 100059, 2024.

[25] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[26] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European conference on computer vision*. Springer, 2022, pp. 709–727.

[27] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 113–19 122.

[28] S. A. Serrano, J. Martinez-Carranza, and L. E. Sucar, "Knowledge transfer for cross-domain reinforcement learning: a systematic review," *IEEE Access*, 2024.

[29] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, 2019.

[30] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1642–1651.

[31] H. Zhou, S. Huang, F. Zhang, and C. Xu, "Ceprompt: Cross-modal emotion-aware prompting for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[32] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[33] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[35] Z. Zhao and Q. Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 1553–1561.

[36] H. Li, M. Sui, Z. Zhu *et al.*, "Nr-dfernet: Noise-robust network for dynamic facial expression recognition," *arXiv preprint arXiv:2206.04975*, 2022.

[37] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge, and W. Zhang, "Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 101–110.

[38] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, and Y. Zhan, "Expression snippet transformer for robust video-based facial expression recognition," *Pattern Recognition*, vol. 138, p. 109368, 2023.

[39] F. Ma, B. Sun, and S. Li, "Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[40] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 1, 2023, pp. 67–75.

[41] T. Li, K.-L. Chan, and T. Tjahjadi, "Multi-scale correlation module for video-based facial expression recognition in the wild," *Pattern Recognition*, vol. 142, p. 109691, 2023.

[42] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 958–17 968.

[43] B. Lee, H. Shin, B. Ku, and H. Ko, "Frame level emotion guided dynamic facial expression recognition with emotion grouping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5681–5691.

[44] Z. Zhao and I. Patras, "Prompting visual-language models for dynamic facial expression recognition," in *BMVC*, 2023.

[45] L. Sun, Z. Lian, B. Liu, and J. Tao, "Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6110–6121.

[46] N. M. Foteinopoulou and I. Patras, "Emoclip: A vision-language method for zero-shot video facial expression recognition," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024, pp. 1–10.

[47] S. Yan, Y. Wang, X. Mai, Q. Zhao, W. Song, J. Huang, Z. Tao, H. Wang, S. Gao, and W. Zhang, "Empower smart cities with sampling-wise dynamic facial expression recognition via frame-sequence contrastive learning," *Computer Communications*, vol. 216, pp. 130–139, 2024.

[48] D. Chen, G. Wen, H. Li, P. Yang, C. Chen, and B. Wang, "Cdgt: Constructing diverse graph transformers for emotion recognition from facial videos," *Neural Networks*, vol. 179, p. 106573, 2024.

[49] L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "A joint local spatial and global temporal cnn-transformer for dynamic facial expression recognition," *Applied Soft Computing*, vol. 161, p. 111680, 2024.

[50] L. Sun, Z. Lian, B. Liu, and J. Tao, "Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *Information Fusion*, vol. 108, p. 102382, 2024.