Smile on the Face, Sadness in the Eyes: Bridging the Emotion Gap with a Multimodal Dataset of Eye and Facial Behaviors

Kejun Liu, Yuanyuan Liu*, *Member, IEEE,*, Lin Wei, Chang Tang, *Senior Member, IEEE,* Yibing Zhan, *Member, IEEE,* Zijing Chen, *Member, IEEE,* Zhe Chen, *Member, IEEE,*

***Abstract*—Emotion Recognition (ER) is the process of analyzing and identifying human emotions from sensing data. Currently, the field heavily relies on facial expression recognition (FER) because visual channel conveys rich emotional cues. However, facial expressions are often used as social tools rather than manifestations of genuine inner emotions. To understand and bridge this gap between FER and ER, we introduce eye behaviors as an important emotional cue and construct an Eye-behavior-aided Multimodal Emotion Recognition (EMER) dataset. To collect data with genuine emotions, spontaneous emotion induction paradigm is exploited with stimulus material, during which non-invasive eye behavior data, like eye movement sequences and eye fixation maps, is captured together with facial expression videos. To better illustrate the gap between ER and FER, multi-view emotion labels for mutimodal ER and FER are separately annotated. Furthermore, based on the new dataset, we design a simple yet effective Eye-behavior-aided MER Transformer (EMERT) that enhances ER by bridging the emotion gap. EMERT leverages modality-adversarial feature decoupling and a multitask Transformer to model eye behaviors as a strong complement to facial expressions. In the experiment, we introduce seven multimodal benchmark protocols for a variety of comprehensive evaluations of the EMER dataset. The results show that the EMERT outperforms other state-of-the-art multimodal methods by a great margin, revealing the importance of modeling eye behaviors for robust ER. To sum up, we provide a comprehensive analysis of the importance of eye behaviors in ER, advancing the study on addressing the gap between FER and ER for more robust ER performance. Our EMER dataset and the trained EMERT models will be publicly available at https://anonymous.4open.science/r/EMER-database.**

***Index Terms*—Multimodal emotion dataset, Emotion recognition, Facial expression recognition, Eye behaviors, Emotion gap.**

I. Introduction

EMOTION recognition (ER) aims to understand and identify human psycho-emotional states across diverse behaviors and contexts, playing a key role in human-computer interaction and cognitive science [1]. It also finds broad applications in multimedia scenarios and applications, such as video surveillance, intelligent education systems, affective healthcare, and personalized advertising [2]. Recent advances in Facial Expression Recognition (FER) have driven progress in ER [3]–[5], as facial expressions are widely regarded as strong indicators of emotional states. Most FER-based ER approaches rely on visual cues from images or videos [6], [7], supported by both static and dynamic FER datasets. Static datasets like SFEW [8] and JAFFE [9] consist of still facial images with emotion labels, while dynamic datasets such as AFEW 7.0 [10] and DFEW [11] provide temporally evolving facial expressions from videos. Using these datasets, many FER-based ER methods have been developed to classify Ekman's six basic emotions (*i.e.*, happiness, sadness, fear, surprise, disgust, and anger) [12]. However, relying solely on visual facial expressions may be insufficient, as such cues can be consciously masked or suppressed, leading to unreliable recognition in certain contexts [6].

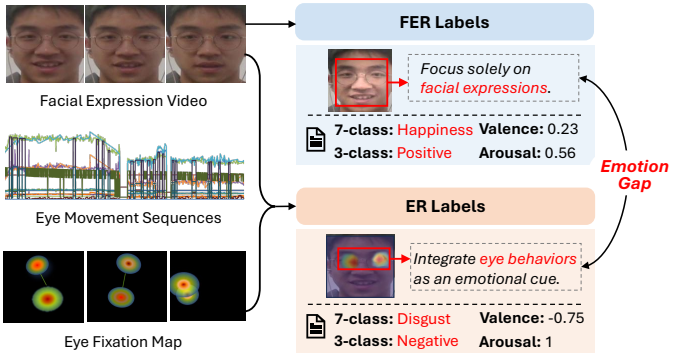

Fig. 1. An example from our EMER dataset. EMER comprises facial expression videos, eye movement sequences, and eye fixation maps, along with multi-view emotion annotations, including FER labels and ER labels, providing more comprehensive emotion analysis.

In the area of ER, existing literature has shown that relying solely on visual facial expression signals is inadequate due to the subjective and camouflaged nature of facial expressions [1], [13]. This limitation results in a significant '***emotion gap***' when applied to various scenarios. Here, the ***emotion gap*** refers to the disparity between facial expressions and genuine emotions of individuals [14]. As an intuitive example, when a person is concealing his or her sad feelings, he or she may put a big smile on the face as a natural response. In such scenarios, conventional FER methods would be misled by the smile expression and could not robustly recognize the true sadness emotion. To bridge this ***emotion gap*** between FER and ER, recent studies have explored the integration of additional physiological modalities, such as electroencephalogram (EEG) signals and eye behavior signals [15]–[17], which are more intuitively reflective of the true state of human emotions. These efforts have led to more robust and comprehensive multimodal ER (MER) systems. To achieve this, some physiological signals enhanced MER datasets have been developed, as shown in Table I. For example, the DEAP [16] collects facial expression videos and EEG signals from 32 participants, and MAHNOB-

Kejun Liu, Yuanyuan Liu, and Lin Wei are with the School of Computer Science, China University of Geosciences (Wuhan), China. E-mail: liukejun, liuyy, linw@cug.edu.cn.

Chang Tang is with the School of Software Engineering, Huazhong University of Science and Technology, China. E-mail: tangchang@cug.edu.cn.

Yibing Zhan is with the School of Computer Science, Wuhan University, China. E-mail: zybjy@mail.ustc.edu.cn.

Zijing Chen and Zhe Chen are with the School of Computing, Engineering and Mathematical Sciences, La Trobe University, Australia. They are also with the Cisco-La Trobe Centre for Artificial Intelligence and Internet of Things. E-mail: zijing.chen, zhe.chen@latrobe.edu.au.

*Corresponding author: Yuanyuan Liu

TABLE I
SUMMARY OF EXISTING POPULAR MULTIMODAL EMOTION DATASETS AND OUR PROPOSED EMER DATASET.

Dataset	#Part.	Data Quality	Non-invasive Sensor	Visual Facial Images	Eye Movement Data	Eye Fixation Maps	Both ER & FER Anno.	Emotion Gap Analysis
CMU-MOSI [18]	N/A	Noisy	✓	✓				
CMU-MOSEI [19]	N/A	Noisy	✓	✓				
IEMOCAP [20]	10	Clean	✓	✓				
eNTERFACE'05 [21]	42	Clean	✓	✓				
DECAF [22]	30	Clean		✓				
DEAP [23]	32	Clean		✓				
SEED-IV [17]	15	Clean			✓			
SEED-V [24]	20	Clean			✓			
MAHNOB-HCI [15]	27	Clean		✓	✓			
Our EMER	**121**	Clean	✓	✓	✓	✓	✓	✓

HCI dataset [15] contains eye movement data and facial images from 27 participants. Although these MER datasets help mitigate the gap between expression and emotion, most they rely on specialized and expensive sensing devices to capture high-quality EEG or eye signals. This makes these ***existing datasets collected on a small scale, with a relatively homogenous set of participants and labeling***. This limits the scalability and utility of such systems for real-world applications.

To address the limitations, in this paper, we construct an Eye-behavior-aided Multimodal Emotion Recognition dataset (**EMER**) featuring larger scale, diverse participants, and multi-view annotations. EMER captures rich emotional cues by integrating facial expression videos, eye movement sequences, and eye fixation maps, along with multi-view emotion annotations, including FER labels and ER labels. The inclusion of such eye behaviors is inspired by Hess *et al.* [25] and other psychological studies [26], which demonstrate that eye movements and fixation patterns serve as natural, intuitive responses to emotional states.

To construct the EMER dataset, we adopt a stimulus-induced spontaneous emotion elicitation protocol. Four emotion experts first curated 28 emotional video clips, which were then shown to 121 participants to induce short-term, spontaneous emotional responses. This process yielded 1,303 high-quality multimodal sequences, simultaneously capturing eye behaviors and facial expressions using a non-invasive Tobii Pro Fusion eye tracker[1] and a high-definition camera. Each sample in EMER is annotated with both emotion and facial expression labels using a combined labeling strategy to ensure accuracy and depth, enabling a detailed analysis of the ***emotion gap*** between FER and ER. To our knowledge, EMER is the first eye-behavior-aided multimodal dataset specifically designed for both ER and FER, offering unique insights into this gap. Fig.1 illustrates the multi-view annotations in EMER, and Table I compares EMER with existing MER datasets.

In addition, based on the new EMER dataset, we design a simple yet effective Eye-behavior-aided MER Transformer (**EMERT**) method. Rather than the existing multimodal methods [27], [28], our EMERT applies adversarial learning and multi-task Transformer to help explicitly extract modality-complementary affective features, so that the gap between facial expression information and eye behavior information can be better modeled and bridged for more effective ER, providing a strong benchmark for future research.

To sum up, we summarize the key contributions as follows:

- We create EMER, a novel eye-behavior-aided multimodal ER dataset containing 1,303 spontaneous samples from 121 participants. EMER includes eye movement sequences, eye fixation maps, and facial expression videos, with both FER and ER labels to enable comprehensive emotion gap analysis. To our knowledge, EMER is the first dataset of its kind, offering a new direction for emotion gap research in ER.
- The EMER dataset introduces comprehensive annotation strategies for achieving multi-view emotion labels. We cover 3-class coarse ER and FER labels (namely positive, negative and neutral), 7-class fine ER and FER labels (namely happiness, sadness, fear, surprise, disgust, anger, and neutral), 2-dimensional continuous emotion ratings (valence and arousal), as well as facial expression intensity (0-3). All of the annotation information contributes to the explicit investigation of the emotion gap, aiming to delve into the details of how to improve ER with multimodal data.
- We devise a simple yet effective benchmarking method, EMERT, to achieve robust ER performance by explicitly and effectively bridging the emotion gap between facial expressions and eye behaviors. The EMERT has shown significant benefits in ER.
- We carried out a comprehensive evaluation of various multimodal methods on our EMER dataset, with seven benchmarking protocols. By addressing the gap between FER and ER, we can further demonstrate that both two tasks benefits from the emotional cues from multimodal data, highlighting the importance of explicitly analyzing the emotion gap for future research.

II. RELATED WORK

A. *Facial Expression-based Multimodal Emotion Datasets*

Currently, there are two main types of facial expression-based multimodal emotion datasets, namely in-the-wild collected emotion datasets [18] and lab-induced spontaneous emotion datasets [15], [23]. In-the-wild collected emotion datasets mainly contain facial expression data, audio, and text gathered from the web or social medias. These datasets often contain various sources of noise and can be challenging to annotate accurately, ultimately compromising their utility for complex applications in emotion recognition. For example, CMU-MOSI [18] consists of 2199 clips with video, audio, and text data, collected from YouTube and annotated with emotional scores in the range [-3,3]. Lab-induced spontaneous emotion datasets contain facial expressions and other physiological signals, such as EEG, Electrocardiography(ECG), Galvanic Skin Response(GSR), and so on. DEAP [23] contains 1280 multimodal samples from 32 participants, with annotations for valence, arousal, dominance, likability, and familiarity, along with facial videos and physiological signals (EEG, GSR, ECG). MAHNOB-HCI [15] includes 565 samples from

[1]https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion

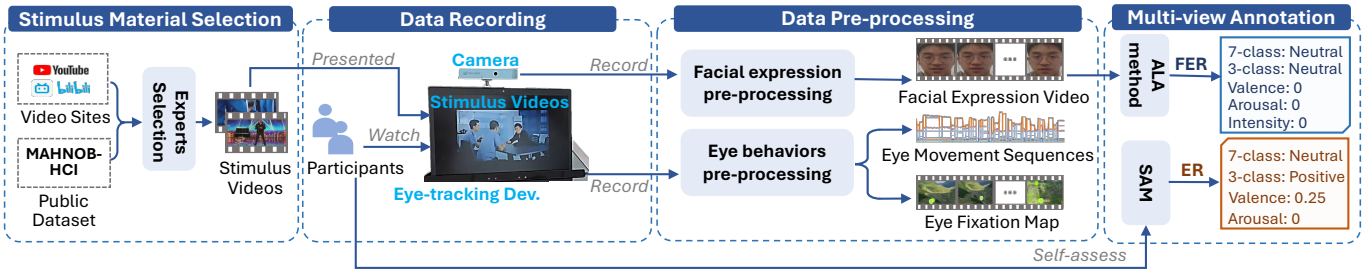

Fig. 2. The collection framework for our EMER dataset. The EMER dataset is multimodal, participant-rich, and multi-view annotation emotion dateset, providing a novel research direction in understanding the emotion gap between ER and FER.

27 participants, each with eye movements, EEG, physiological signals, video, audio, and labels for 9-class emotion, valence, arousal, dominance, and predictability. Despite progress, the former suffers from web-induced noise, while the latter is constrained by limited scale and participant diversity.

B. Facial Expression Recognition

The FER task aims to understand individual emotions from his/her visual facial expressions. Currently methods are divided into two main categories, namely static FER methods and dynamic FER methods [5], [29]. Static FER methods focus on recognizing facial expressions from static face images and have achieved significant achievements. Wang *et al.* [29] recognized facial expressions from low-quality images by introducing an effective self-healing network (SCN). By Contrary, dynamic FER methods explores spatio-temporal information from video sequences, obtaining more robust FER performance. Ma *et al.* [30] proposed the spatial-temporal Transformer to capture discriminative emotion features within each frame and model contextual relationships among frames. Liu *et al.* [5] proposed Expression snippet Transformer (EST) to decompose videos into expression snippets to enhance intra- and inter-snippet visual modeling capabilities, respectively. Although progress has been made in FER, these methods can only recognise information from facial expressions, which can be easily camouflaged in some scenes, and this can lead to recognition results that deviate from the true emotion.

C. Multimodal Emotion Recognition

Multimodal ER aims to predict human emotions from multiple modalities, such as video, audio, and physiological signals. Most existing methods mainly are divided into two categories, *i.e*, representation learning-based methods [31], [32] and multimodal fusion-based methods [33]. Representation learning-based methods focus on learning specific modality representations by considering the difference and consistency of different modalities, thus improving multimodal emotion recognition. For example, VAANET [32], which integrated spatial, channel-wise, and temporal attentions for audio-video emotion recognition. Multimodal fusion-based methods attempt to learn the interactive information between different modalities by designing complex fusion mechanisms. For example, MulT [34] used a set of Transformer encoders to capture both unimodal and cross-modal interactions. Kernel-based Extreme Learning Machine (ELM) [35] recognized video emotions by combining video content and EEG signals. Despite the progress, most methods did not consider the raw noises existing in the specific modality features, leading to sub-optimal results.

III. Proposed EMER Dataset

To gain deeper insights into the gap between FER and ER, we construct the Eye-behavior-aided Multimodal Emotion Recognition (EMER) dataset. As illustrated in Fig. 2, the construction pipeline comprises four stages: stimulus selection, data recording, data pre-processing, and multi-view annotation. Through this pipeline, we collects 1,303 spontaneous emotional sequences collected from 121 participants, covering 3 modalities: facial expression videos, eye movement sequences, and eye fixation maps. In addition, we provide both ER and FER labels using distinct annotation strategies, enabling a comprehensive analysis of the gap between FER and ER.

A. Stimulus Material Selection and Participants

Following the protocols of MAHNOB-HCI [15] and SEED [20], we adopt a stimulus material-induced paradigm to elicit spontaneous emotions. We first collect 115 candidate videos across seven basic emotion categories from public datasets and platforms such as BiliBili and YouTube. Four emotion experts then select the top 4 most emotionally evocative clips per category, yielding 28 final stimuli (1–2 minutes each). Some examples are shown in Fig.3., with details in the *supplementary material (Sec.I)*. We recruit 121 participants (76 males, 45 females; aged 18–40) from diverse backgrounds. In a controlled lab environment, each participant views the selected stimuli to evoke short-term, spontaneous emotional responses. All participants provided informed consent in accordance with GDPR[2], with the consent form included in the supplementary material.

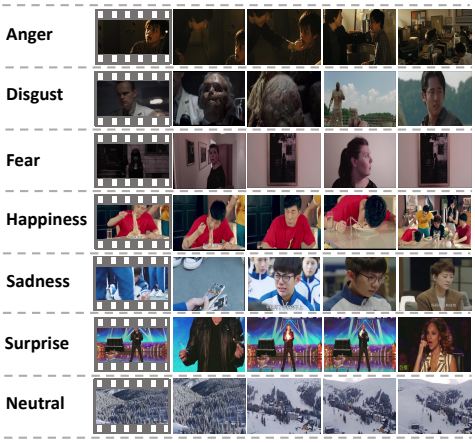

Fig. 3. Some examples of stimulus materials.

[2]https://gdpr-info.eu/

B. Data Recording

Using the selected emotion-stimulus videos, each participant is asked to watch different categories of stimulus videos in sequence, and his eye movement sequences, emotion-related eye fixation maps, and facial expression videos are recorded by a Tobbi Pro Fusion eye-tracking device and a high-definition camera, simultaneously. After viewing the stimulus videos, each participant completes an emotional self-assessment questionnaire for each stimulus video. Ultimately, we collected 1,623 original multimodal data samples from the 121 participants. *More details can be seen in our supplementary material (Sec.II).*

C. Data Pre-processing

To maintain the integrity and synchronization of the collected multimodal data, we meticulously align, trim, and filter the original data, obtaining a total of 1,303 high-quality multimodal emotional data samples processed for our EMER.

*1) **Facial expression pre-processing**:* Due to various emotion-irrelevant visual noise (*e.g.*, illumination, head poses, etc.) in raw facial data, we carry out a 2-step pre-processing. In the first step, we use illumination normalization [36] to remove lighting variations across different frames of a video. Then, we employed a state-of-the-art deep learning model, MTCNN [37], to extract facial landmarks and we then performed face alignment according to the landmarks to ensure consistency across all video frames.

*2) **Eye behaviors pre-processing**:* To align unsynchronised eye behavior signals, we employ a 3-step process to perform blink correction, sweep correction, and pupil correction[3] and align asynchronous eye movement sequences and eye fixation maps. Regarding blink correction, we first identify eye movement sequences with blink durations outside the range of 75 ms and 425 ms as invalid blinks [6], and then use linear interpolation to correct the invalid blinks for blink correction. In sweep correction, linear interpolation is also used to correct the eye sweep data. Lastly, following [6], we use the difference between the pupil diameters corresponding to the current and the previous timestamp for pupil correction in eye movement sequences. It is worth mentioning that, to make the eye fixation maps related to emotions, we remove invalid eye fixations according to invalid eye movements.

D. Multi-view Annotation

As introduced earlier, we provide multi-view emotion annotation with both ER and FER to help analyze the gap between emotions and facial expressions. To further clarify, the ER labels are based on participant- and organizer-inducing emotion annotations, while FER labels are assigned by experts through analysis of recorded facial videos. Here, we will explain label formats and annotation methods in detail.

*1) **Label Formats.**:* Each ER label contains three key aspects: (1) 3-class coarse ER labels, *i.e.*, positive, negative, and neutral; (2) 7-class fine ER labels, *i.e.*, happiness, sadness, fear, disgust, surprise, anger, and neutral; (3) valence and arousal ratings in the range [-1,1], where a higher valence score signifies a greater level of happiness, while a higher arousal score indicates a greater degree of excitement. Meanwhile, our EMER also offers four FER annotations (see Fig.1), which consist of 3-class coarse FER labels, 7-class fine FER labels, valence, and arousal ratings within the [-1, 1] range, and facial expression intensity within the [0,3] range. Although ER and FER labels are obtained separately, it is important to note that ER and FER labels come from the shared set of emotional categories, such as happiness, sadness, fear, surprise, disgust, anger, and neutral within 7-class fine labels, as well as positive, negative, and neutral within 3-class coarse labels.

In the rest of this paper, we use j to index each data sample, and we utilize e_j and f_j to represent the corresponding ER label and FER label, respectively.

*2) **Annotation Method**:* To achieve multi-view emotion annotation, we introduce two different annotation strategies to provide ER and FER labels, respectively, for comprehensive emotion analysis.

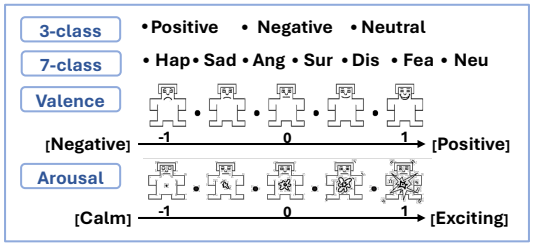

Fig. 4. The SAM self-assessment for the ER annotation.

ER annotation. Following the annotation methods in [15], [20], we annotate ER labels e_j. Using participant self-assessments via the Self-Assessment Manikin (SAM) [23], as shown in Fig.4. For each collected data sample, participants use the SAM to rate their emotional states, ensuring that the ER labels reflect both individual experiences and stimulus effects—providing rich, human-centered annotations for multimodal emotion data.

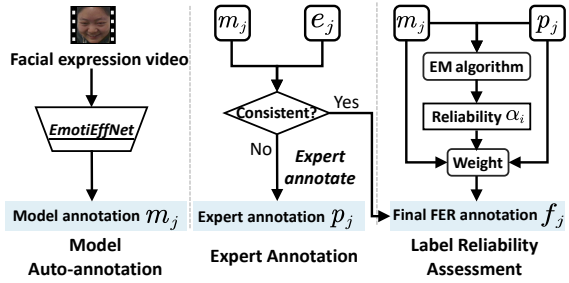

Fig. 5. The ALA pipeline for high-reliability FER annotation, including model auto-annotation, expert annotation, and annotation reliability assessment.

FER annotation. Manual FER annotation is often labor-intensive, time-consuming, and prone to subjectivity [19], [23]. To mitigate this issue, we adopt an Active Learning-based Annotation method (ALA) that combines deep model auto-annotation with manual expert annotation. As illustrated in Fig.5, ALA involves model auto-annotation, expert annotation, and label reliability assessment, resulting in efficient and high-quality FER labels.

Model auto-annotation represents the process of using machines to annotate facial videos. We employ EmotiEffNet

[3]Pupil correction: Pupil data alone may contain more noise, while pupil fluctuation is more capable of expressing emotion [6], pupil correction is performed to replace pupil data with pupil fluctuation.

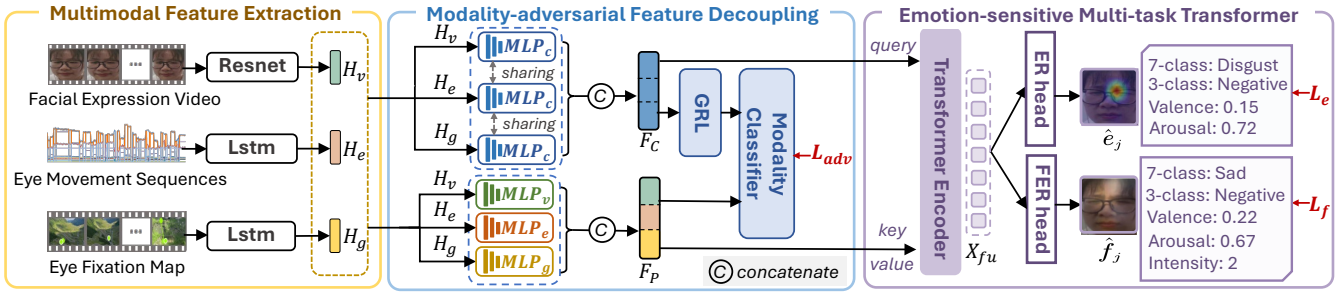

Fig. 6. The general framework for our EMERT method. The EMERT achieves robust ER performance by explicitly and effectively bridging the emotion gap between facial expressions and eye behaviors.

[38], a pre-trained deep neural network specialized in FER, to automatically annotate each collected data, resulting in the model-generated FER labels denoted m_j. This significantly expedites the annotation process, leading to substantial time and resource savings.

After model auto-annotation, there would be inherent prediction biases of machine models, and we enhance the machine-generated labels by incorporating expert annotations. Specifically, we first use ER labels to identify and filter inconsistent model-generated FER labels, and the remaining FER labels that are consistent with ER labels are directly stored as annotated FER labels. Due to the emotion gap, there could be many inconsistent labels. For these inconsistent labels, we enlist the expertise of four emotion specialists for re-annotation, ultimately yielding the ER label set termed as $T_j = \{m_j, p_j\}$, where p_j represents expert labels and m_j represents model-generated labels consistent with ER labels.

Using the FER label set T_j, we then employ the EM algorithm [39] to assess annotation reliability, enhancing the quality and reliability of FER labels. The EM algorithm comprises two key optimization steps: the E-step and the M-step. The E-step calculates the posterior probability for potential correct annotations, while the M-step optimizes the log-likelihood of each label by estimating each label reliability α_i within T_j. By iteratively cycling through E-step and M-step until convergence, we ultimately obtain high-quality FER labels f_j through a weighted voting process, which can be formulated as: $f_j = \sum_{i=1}^{5} \frac{\alpha_i}{\sum_{i=1}^{5} \alpha_i} \cdot t_j^i$, where $t_j^i \in T_j$.

E. Metadata in EMER

Following the construction process, EMER contains a collection of 1,303 videos, totaling 390,900 frames. In EMER, we provide three distinct emotional signal subsets, namely the facial expression subset, the eye movement subset, and the emotion-related eye fixation subset, each of which is carefully designed to capture different aspects of emotions. Examples of the metadata in EMER are presented in Fig. 1, with additional samples provided in our *supplementary material (Sec.III)*.

The facial expression subset comprises 1,303 videos and 390,900 frames. Each video has a duration of 1 to 2 minutes. **The eye movement subset** contains 1.91 million timestamp samples, offering a more comprehensive and richer set of information, such as time stamps, gaze point coordinates, gaze direction, pupil diameter, eye position, gaze time, and eye movement event type (sweep and gaze). **The eye fixation subset** is a massive collection of 7.50GB in size that contains 390,900 frames, each depicting a heat map of the video content, location, and trajectory of the participant's attention when emotional events occurred.

Moreover, each emotional data in EMER possesses dual labels for FER and ER, respectively, including 3-class FER and ER labels, 7-class FER and ER labels, 2-dimensional continuous emotion ratings, and facial expression intensity. Fig.7 reports the distributions of the rich ER and FER labels in our EMER. We also provide all *manual* and *automatic* annotations for each data in EMER.

In summary, EMER offers larger scale (1,303 samples from 121 participants), richer multi-view annotations (ER/FER, valence/arousal, and expression intensity), and non-invasive acquisition of synchronized facial and eye behaviors, making it a valuable benchmark for advancing research on FER and ER.

IV. Proposed EMERT Method

Leveraging the EMER dataset, we design the Eye-behavior-aided MER Transformer (**EMERT**), incorporates Modality-Adversarial Feature Decoupling (MAFD) and an Emotion-sensitive Multi-task Transformer (EMT) to learn modality-complementary affective features, effectively bridging the emotional gap between facial expressions and eye behaviors for robust ER. As illustrated in Fig.6, EMERT first employs a Multimodal Feature Extraction (MFE) module to extract unimodal features. Then, the MAFD applies a gradient reversal layer (GRL) with adversarial loss to decouple **emotion-generic features**, which are invariant to the emotion gap across various modalities, from **emotion-unique features**, which preserve their emotion discrepancies across modalities. Finally, the EMT further uses the emotion-generic features as query to guide the fusion of emotion-unique features, enabling robust, emotion-sensitive representation learning for enhanced ER performance.

MFE: Given multimodal data as input, we employ a pre-trained Resnet [40] to extract the expression features $H_v \in \mathbb{R}^{T_v \times S}$, and use an LSTM [41] to extract the eye movement features $H_e \in \mathbb{R}^{T_e \times S}$ and eye fixation features $H_g \in \mathbb{R}^{T_g \times S}$, respectively. T_v, T_e and T_g are the lengths of these three feature sequences, and S is the dimension of feature vector.

MAFD: With the multimodal features, we employ the MAFD module to separate the emotion-generic features F_C, invariant to the emotion gap across modalities, from the emotion-unique features F_P retaining modality-specific emotional cues. To achieve this, we first apply an emotion-generic feature extractor (*i.e.*, MLP_c) and three emotion-unique feature extractors (*i.e.*, MLP_v, MLP_e, MLP_g), each

implemented as a two-layer MLP. To ensure modality emotion invariance in F_C, we introduce a gradient reversal layer (GRL) after the MLP_c, and attach a modality classifier D that distinguishes feature origins.

By adversarially training D and MLP_c, we encourage the F_C to be indistinguishable across modalities, while F_P retains discriminative modality-specific cues. The process can be formulated as,

$$\min_{\theta_D} \max_{\theta_{MLP_c}} \mathcal{L}_{adv} = -\frac{1}{N_b} \sum_{j=0}^{N_b} f_j/e_j \cdot \log F_D\left(F_C/F_P; \theta_D\right), \tag{1}$$

where N_b is the number of training samples, f_j/e_j represents the FER or ER labels, and θ_{MLP_c} and θ_D are the parameters of the MLP_c and D, respectively. This adversarial setup ensures effective decoupling of emotion-generic and emotion-unique features, helping bridge the modality-induced emotion gap in ER.

EMT: Existing multimodal Transformer methods often uses queries from a single modality, which can overemphasize modality-specific cues and amplify the emotion gap between facial expressions and eye behaviors, leading to suboptimal fusion. To address this, EMT first adopts the decoupled emotion-generic features F_C as query q, and the emotion-unique features F_P as key k and value v, yielding more modality-complementary affective features. We follow the typical formulation of the Transformer structure as $Trans(\cdot)$ as:

$$X_{fu} = Trans\left(q = F_C, k/v = F_P\right). \tag{2}$$

With the modality-complementary affective features X_{fu}, we apply two multi-task prediction heads, namely the FER head and ER head, to predict the FER result $\hat{f}_j$ and the ER result $\hat{e_j}$, respectively. Each prediction head possesses a similar structure, comprising a 2-layer MLP. Formally, the objectives for the two prediction heads are written as:

$$L_e = \begin{cases} CE(\hat{e_j}, e_j), & \text{for classification} \\ Huber(\hat{e_j}, e_j), & \text{for regression} \end{cases} \tag{3}$$

$$L_f = \begin{cases} CE(\hat{f}_j, f_j), & \text{for classification} \\ Huber(\hat{f}_j, f_j), & \text{for regression} \end{cases} \tag{4}$$

For discrete emotion classification, we introduce the multi-class cross-entropy loss $CE()$; for continuous emotion regression, we employ the huber loss $Huber()$.

Overall Objective: The total objective function $\mathcal{L}$ of EMERT is the the summation of the above-mentioned three learning objectives, which can be writen as: $\mathcal{L} = \alpha\mathcal{L}_{adv} + \beta(\mathcal{L}_e + \mathcal{L}_f)$. Empirically, $\alpha = 0.3$ and $\beta = 0.1$ are hyper-parameters to balance the multi-task learning process.

V. EXPERIMENTS

A. Experimental Setup

*1) **Evaluation Protocols**:* To evaluate methods on the EMER dataset, we conducted both classification and regression protocols for both ER and FER tasks. *For the classification tasks*, consistent with the previous research, we chose three widely-used classification validation metrics, namely unweighted average recall (UAR), weighted average recall (WAR), and F-score (F1), to estimate our model. Larger values are preferred for all of these indicators. *For the regression tasks*, we also chose three widely-used regression validation metrics, like in other papers, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). All of these regression metrics are as small as possible.

In addition, following existing evaluation protocols [11], we employed a 5-fold cross-validation approach for these benchmarks on EMER. We utilized 1,043 data from the EMER dataset for training, and the remaining 260 data for testing.

*2) **Implementation Details**:* In this paper, we used the PyTorch framework to implement all models on our EMER dataset. We set the batch size of all models to 16. All models were trained on an NVIDIA GeForce RTX 3090 with an initial learning rate as 0.0001. Cosine decay was used to decrease the learning rate during training. In the training and testing phases, each multimodal data in EMER is initially sampled with 8 frames evenly extracted from the facial expression video, 32 frames from the eye movement sequence, and 32 frames from the emotion-related eye fixation map. These frames are then input into the proposed EMERT model as well as other benchmark models to acquire emotion representations for predicting the final results.

B. Benchmarking ER & FER on EMER Dataset

We conducted extensive benchmarks for ER and FER, respectively, on the EMER dataset. For ER, we evaluate classification (3-class and 7-class) and valence/arousal regression. For FER, we further include intensity regression in addition to classification (3-class and 7-class) and valence/arousal regression. In each task, we compared our EMERT with various cutting-edge methods including TMC [42], TAILOR [43], MulT [34], LMF [44], ResNet_LSTM [40], Self_MM [46], MMA-DFER [2], and so on. The results show that EMERT consistently outperforms these methods across all evaluation settings, demonstrating its effectiveness in handling eye behaviors information and mitigating the emotion gap.

*1) **3-class ER Classification**:* Following [18], we benchmarked EMERT against state-of-the-art methods on 3-class ER, as shown in Table II. EMERT achieved the highest accuracy of 59.28% WAR, 52.62% UAR, and 55.71% F1 on ER, improving over Self_MM [46] by 5.2%, 9.73%, and 8.23%, respectively.

*2) **7-class ER Classification**:* In this setting, we compared our EMERT method with other typical and popular methods for 7-class ER. Table II reports the comparison results of these benchmarks on EMER. EMERT achieved the highest accuracy of 59.28% WAR, 52.62% UAR, and 55.71% F1 on ER, improving over Self_MM [46] by 5.2%, 9.73%, and 8.23%, respectively. It indicates that EMERT can obtain more modality-complementary affective features for robust ER performance.

*3) **ER Valence/Arousal Regression**:* Table III reports the regression results on valence/arousal for ER. Unlike classification, regression tasks favor lower metric values. Regarding the performance on valence and arousal MSE and MAE, we

TABLE II
COMPARISON RESULTS OF 3-CLASS/7-CLASS ER AND FER ON EMER, RESPECTIVELY.

Models	3-class						7-class					
	ER			FER			ER			FER		
	WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)
TMC [42]	54.83	37.52	45.87	54.56	39.84	46.96	31.28	27.20	25.20	49.64	31.11	42.66
TAILOR [43]	52.98	34.70	47.58	63.91	43.61	53.12	33.39	24.85	28.40	48.9	30.55	37.12
MulT [34]	57.33	48.03	53.53	64.94	52.18	61.48	31.96	27.45	28.40	48.83	32.82	42.78
LMF [44]	58.78	49.36	54.80	66.90	56.55	63.64	33.07	27.12	27.81	49.34	32.05	42.53
ResNet_LSTM [41]	57.78	49.06	55.08	67.24	56.50	63.52	31.18	27.64	27.48	47.83	32.79	41.38
ResNet_Transformer [40]	56.75	49.63	55.70	61.01	46.56	56.23	29.21	23.49	25.98	47.33	31.10	40.43
C3D_Transformer [45]	52.51	42.44	53.12	59.67	45.88	55.01	31.66	20.49	26.85	42.50	21.54	29.33
Self_MM [46]	54.08	42.89	47.48	61.51	42.34	50.69	32.47	27.55	28.19	49.58	32.05	42.77
MMIM [47]	55.94	46.93	53.13	68.05	55.99	63.57	31.71	27.81	28.43	48.61	31.77	42.50
TMT [48]	58.72	50.75	55.02	67.14	55.93	63.69	33.65	27.68	30.26	50.28	32.21	42.35
MMA-DFER [2]	52.92	33.33	37.35	59.43	33.31	44.55	30.65	27.73	30.42	48.54	32.81	42.68
NORM-TR [49]	59.13	49.28	53.04	66.92	56.36	62.68	33.72	28.09	28.81	50.80	32.63	43.32
EMERT	**59.28**	**52.62**	**55.71**	**68.10**	**56.91**	**63.73**	**33.92**	**28.17**	**30.38**	**51.18**	**33.04**	**43.33**

TABLE III
COMPARISON RESULTS OF VALENCE AND AROUSAL REGRESSION FOR ER AND FER, RESPECTIVELY.

Models	ER						FER					
	Arousal			Valence			Arousal			Valence		
	MAE (↓)	MSE (↓)	RMSE (↓)	MAE (↓)	MSE (↓)	RMSE (↓)	MAE (↓)	MSE (↓)	RMSE (↓)	MAE (↓)	MSE (↓)	RMSE (↓)
TMC [42]	0.369	0.264	0.468	0.438	0.304	0.540	0.226	0.079	0.276	0.375	0.218	0.460
TAILOR [43]	0.373	0.223	0.465	0.514	0.390	0.619	0.246	0.089	0.294	0.291	0.231	0.474
MulT [34]	0.399	0.263	0.503	0.481	0.344	0.573	0.236	0.085	0.287	0.338	0.180	0.417
LMF [44]	0.368	0.219	0.457	0.440	0.283	0.526	0.228	0.080	0.278	0.286	0.128	0.359
ResNet_LSTM [41]	0.383	0.241	0.482	0.454	0.303	0.543	0.229	0.082	0.281	0.303	0.138	0.366
ResNet_Transformer [40]	0.388	0.241	0.481	0.450	0.293	0.537	0.234	0.086	0.288	0.299	0.136	0.362
C3D_Transformer [45]	0.376	0.224	0.465	0.497	0.333	0.571	0.239	0.087	0.288	0.342	0.179	0.416
Self_MM [46]	0.374	0.228	0.470	0.444	0.302	0.542	0.244	0.092	0.297	0.351	0.189	0.426
MMIM [47]	0.379	0.230	0.471	0.443	0.295	0.535	0.228	0.084	0.285	0.290	0.131	0.356
TMT [48]	0.376	0.227	0.465	0.456	0.313	0.551	0.231	0.084	0.284	0.301	0.141	0.371
MMA-DFER [2]	0.366	0.235	0.493	0.440	0.291	0.529	0.226	0.084	0.273	0.292	0.128	0.364
NORM-TR [49]	0.367	0.221	0.461	0.437	0.283	0.521	0.224	0.079	0.278	0.289	0.129	0.360
EMERT	**0.365**	**0.217**	**0.456**	**0.433**	**0.279**	**0.519**	**0.223**	**0.078**	**0.266**	**0.286**	**0.127**	**0.351**

observed that our EMERT obtained more precise regression results with smaller errors. Moreover, EMERT achieves 0.9% and 1.1% lower MAE, and 1.1% and 2.3% lower MSE than Self_MM [46], showing its fine-grained emotion perception.

*4) **3-class FER Classification**:* We compared EMERT with state-of-the-art methods for 3-class FER on EMER dataset, and the results are shown in the Table II. EMERT achieved the significant gains over Self_MM [46]: 6.59% WAR, 14.57% UAR, and 13.04% F1, demonstrating the effectiveness of integrating eye behavior cues for FER.

*5) **7-class FER Classification**:* The performance results of the cutting-edge methods and our proposed EMERT method for 7-class FER are shown in Table II. EMERT achieves the highest WAR (51.18%), UAR (33.04%), and F1 (43.33%), indicating a stronger ability to capture facial expressive behaviors from multi-source signals.

*6) **FER Valence/Arousal Regression**:* Table III provides the results of our EMERT and other cutting-edge methods for FER valence/arousal regression on EMER. Compared with TMT [48], EMERT achieves up to 0.8% and 1.5% improvement on MAE, 0.6% and 1.4% on MSE, and 1.8% and 2.0% on RMSE, confirming its robustness under continuous affective dimensions.

*7) **FER Intensity Regression**:* To the best of our knowledge, we are the first to propose benchmarks for FER intensity regression. Table IV shows that EMERT outperforms all Transformer-based methods, achieved the gains over MMA-DFER [2]: 2.6% MAE, 2.2% MSE, and 0.2% RMSE, demonstrating the effectiveness of integrating eye behavior cues for FER.

TABLE IV
COMPARISON OF FACIAL EXPRESSION INTENSITY REGRESSION FOR FER.

Models	Metric		
	MAE (↓)	MSE (↓)	RMSE (↓)
ResNet_Transformer [45]	0.701	0.707	0.829
MulT [34]	0.804	1.035	0.999
TMT [48]	0.668	0.690	0.818
MMA-DFER [2]	0.686	0.695	0.811
NORM-TR [49]	0.666	0.685	0.823
EMERT	**0.660**	**0.673**	**0.809**

C. Deepen Understanding of the EMER Dataset

*1) **Analysis of ER labels and FER labels**:* To illustrate the gap between ER and FER, Fig.7 displays the distribution discrepancy between ER labels and FER labels in our EMER dataset. As shown, the "Neutral" category is significantly more dominant in FER labels, while high-intensity emotions such as "Surprise" are notably underrepresented. This suggests that FER tends to capture surface-level facial cues, whereas ER reflects individuals' subjective emotional experiences. Such differences indicate a clear discrepancy in emotional perception between the two tasks.

*2) **Analysis of Benefits from Eye Behaviors for ER/FER**:* To explore the effect of eye behaviors in EMER for both ER

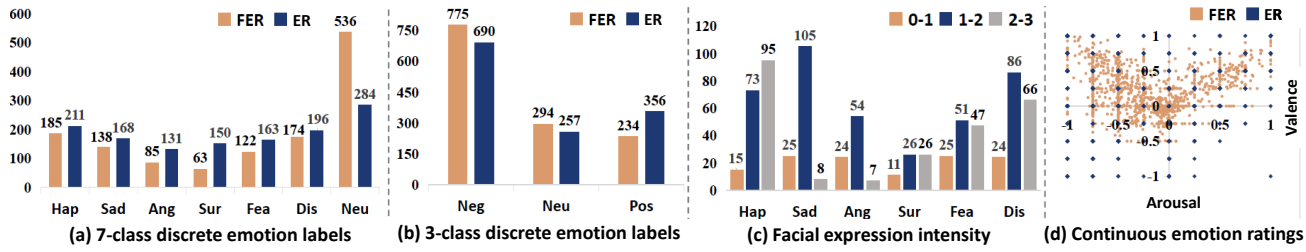

Fig. 7. The distribution differences between ER and FER labels in EMER. Hap, Sad, Ang, Sur, Fea, Dis, Neu, Neg, Pos are the abbreviations of the labels.

and FER tasks, we used our proposed EMERT to perform separate emotion and faical expression classification with different modality settings on EMER, as shown in Table V. The results show that using all modalities, including eye movement sequences, eye fixation maps and facial expression videos, yielded the most superior performance, underscoring that eye behaviors effectively enhance both ER and FER performance. In addition, we observed that adding the eye movement sequences led to the highest improvement (see the underline results in the table), verifying that the eye behaviors are the effective complement of facial expressions for robust ER. We also investigate the correlations between eye behaviors and facial expressions for ER and FER tasks by using three distinct correlation coefficients: Pearson Coefficient, Spearman Coefficient, and Kendall Coefficient [50].

TABLE V
ABLATION STUDY ON 7-CLASS ER AND FER TASKS. F, E, G DENOTE FACIAL EXPRESSIONS, EYE MOVEMENTS, AND FIXATION MAPS, RESPECTIVELY.

Modality			ER task			FER task		
F	E	G	WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)
✓			30.21	26.40	29.32	46.95	27.21	34.98
	✓		31.51	26.12	27.61	40.80	18.99	24.92
		✓	28.20	27.20	24.10	41.98	22.74	28.72
✓	✓		32.89	27.03	29.89	49.35	32.25	41.98
✓		✓	32.81	27.63	29.85	47.29	27.75	35.09
	✓	✓	31.88	27.64	29.01	43.32	24.71	31.22
✓	✓	✓	**33.92**	**28.17**	**30.38**	**51.18**	**33.04**	**43.33**

As depicted in Fig.8, the majority of correlation coefficients associated with the ER task exhibit higher values than the FER task. This observation indicates that eye movement data serves as a valuable complement to ER. Meanwhile, the correlation coefficients pertaining to facial expressions also demonstrate that eye movement data effectively contributes to the understanding of FER. By combining the information, we can achieve a more comprehensive analysis of both emotions and facial expressions. This complementarity not only deepens our understanding of the '***emotion gap***' but also holds promise for enhancing the performance of emotion analysis systems in various domains.

*3) **Effects of Different Annotation Methods**:* To assess the efficacy of our ALA annotation methods for FER, we employed Cronbach's Alpha [19] to evaluate the consistency of different annotation approaches, as presented in Table VI. The results reveal that model auto-annotation exhibits the lowest reliability, primarily due to dataset bias. The manual expert annotation, commonly used in many current datasets [19], is susceptible to subjective individual differences, such as identity and profession, resulting in an average low Cronbach's Alpha of 0.799. The combination of model auto-annotation and expert annotation can enhance the quality and consistency of emotion labels (as seen in the third row of Table VI). Ultimately, our ALA approach achieves the most consistent and reliable FER labels, significantly surpassing other methods, owing to the incorporation of annotation reliability assessment.

TABLE VI
ANNOTATION CONSISTENCY EVALUATION ON DIFFERENT ANNOTATION APPROACHES. THE BEST RESULTS ARE IN BOLD.

Methods	Label Type	Cronbach's Alpha
Model auto-annotation	Discrete emotion category	0.731
	Valence rating	0.789
	Arousal rating	0.679
Expert annotation	Discrete emotion category	0.784
	Valence rating	0.829
	Arousal rating	0.786
Expert annotation + Model auto-annotation	Discrete emotion category	0.852
	Valence rating	0.863
	Arousal rating	0.847
Proposed ALA	Discrete emotion category	**0.978**
	Valence rating	**0.927**
	Arousal rating	**0.982**

*4) **Effects of Different Eye Movement features**:* To further investigate the contribution of different Eye Movement information, we conducted an ablation study on EMER by progressively isolating individual features. Specifically, we evaluated models using only gaze point, gaze time, and pupil diameter, and compared them with our complete eye movement modeling design. As shown in Table VII, the results reveal that different features contribute unequally to ER and FER. For ER, gaze point and pupil diameter provide more emotion-related information, while gaze time yields relatively weaker performance. For FER, pupil diameter achieves the most significant improvements, followed by gaze time, with gaze point being the weakest. Importantly, when integrating all three types of features, our complete design consistently outperforms single-feature settings on both ER and FER tasks. These findings demonstrate that eye movement information not only carries meaningful affective cues but also provides complementary signals when different features are combined. This validates the necessity of modeling multiple eye movement features jointly, rather than relying on a single aspect, to enhance the robustness and accuracy of multimodal emotion recognition.

D. Deepen Understanding of the EMERT Method

We conduct extensive ablation studies to further validate our method, with feature distribution visualizations across tasks provided in the *supplementary material (Sec.ĨII)*.

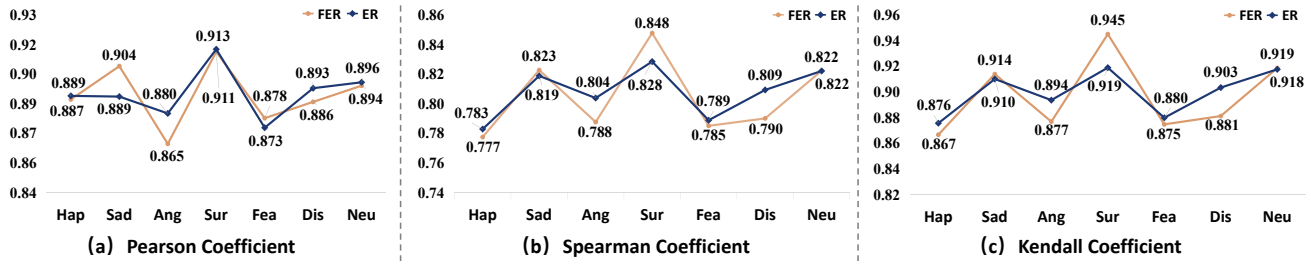

Fig. 8. The correlation analysis of eye behavior data for each category in ER and FER by three different correlation coefficients, *i.e.*, (a) Pearson's coefficient, (b) Spearman's coefficient, and (c) Kendall's coefficient. The orange line illustrates the correlation of eye movement data for 7-class FER annotations, while the blue line signifies the correlation of eye movement data for 7-class ER annotations. In both cases, the closer the value is to 1, the stronger the correlation.

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT EYE MOVEMENT FEATURES ON THE 7-CLASS ER AND FER TASKS. THE BEST RESULTS ARE IN BOLD.

Type	ER			FER		
	WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)
Gaze point	31.96	27.17	28.47	46.18	30.64	36.53
Gaze time	31.13	27.05	25.21	47.57	30.48	38.06
Pupil diameter	32.38	26.85	27.42	50.46	31.27	41.34
Ours	**33.92**	**28.17**	**30.38**	**51.18**	**33.04**	**43.33**

*1) **Analysis of Attention Gap between ER and FER**:* In addition to the label differences, we can further carry out deeper interpretation with our EMER dataset. In particular, by training an EMERT on our dataset, we can visualize the attention maps obtained by the FER and ER heads in EMERT, respectively, as illustrated in Fig.9. According to the figure, it can be seen that the ER head pays more attention to detailed areas such as the corners of the eyes, mouth, and nose, while the FER head pays more attention to the global areas, such as the entire eyes and mouth.

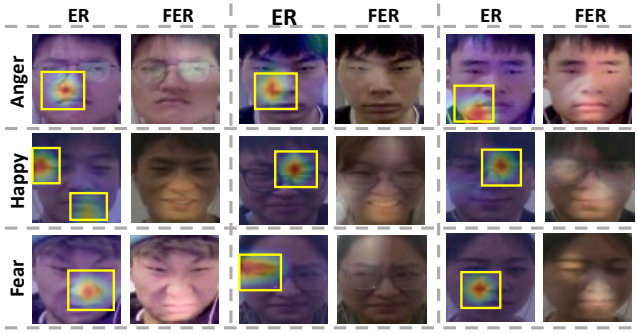

Fig. 9. Attention maps from the ER head and FER head in EMERT. For ER, the model pays more attention to detailed areas such as the corners of the eyes, mouth, and nose, while for FER, the model pays more attention to the global areas, such as the entire eyes and mouth.

*2) **Analysis of Whether FER and ER Reinforces Each Other**:* To investigate the potential benefits of ER and FER to each other, Table VIII compared their performance in single-task and multi-task settings on the 3-class ER and FER tasks, respectively. Specifically, we first conducted the experiments on the FER task with and without the ER head integrated into our proposed EMERT model (see Fig. 2), respectively. Following this, we proceeded to evaluate the ER task with and without the FER head. Our results revealed that the inclusion of the FER head in the ER task significantly enhances performance, resulting in a 5.94% increase in the UAR metric. Additionally, the integration of the ER head also improves FER results by 0.37% in WAR. These experimental results highlight the mutually reinforcing effects of ER and FER, indicating that our proposed EMER and the corresponding EMERT method help to understand the emotion gap between emotions and facial expressions, ultimately improving state-of-the-art performance.

TABLE VIII
INTER-AUGMENTATION OF FER AND ER IN EMERT.

Type	Task head		Metric		
	FER	ER	WAR (↑)	UAR (↑)	F1 (↑)
Single-task FER	✓	X	67.73	**58.31**	63.18
Multi-task FER	✓	✓	**68.10**	56.91	**63.73**
Single-task ER	X	✓	56.01	46.68	50.85
Multi-task ER	✓	✓	**59.28**	**52.62**	**55.71**

*3) **Evaluation on the Another SIMS Dataset**:* To validate EMERT's effectiveness, we followed [46] and evaluated our EMERT on the SIMS dataset using Acc-2/3/5. SIMS [51] provides video, audio, and text with multimodal and unimodal annotations. We utilize 1,824 samples for training and 457 samples for testing. As shown in Table IX, EMERT outperforms the other Transformer-based method, *i.e.*, MulT [34], with a relative increase of 1.86% (Acc-2), 3.34% (Acc-3), and 23.04% (Acc-5), highlighting its effectiveness and generalization.

TABLE IX
COMPARISON RESULTS OF MULTIMODAL EMOTION RECOGNITION ON SIMS. THE BEST RESULTS ARE IN BOLD.

Models	Acc-2 (↑)	Acc-3 (↑)	Acc-5 (↑)
MulT [34]	78.84	67.13	38.24
TFN [52]	**82.06**	66.16	39.74
MFN [53]	78.26	65.79	41.19
Self-MM [46]	78.99	66.52	44.20
EMERT	80.31	**69.37**	**47.05**

*4) **Effects of Different Modules in EMERT**:* To assess each module's impact, we conduct ablation studies on the 7-class ER and FER tasks using the EMER dataset (Table X). The baseline method, a Transformer-based multimodal fusion approach [45], directly combines facial expression features from the pre-trained ResNet [40] with eye movement and eye fixation features from the pre-trained LSTM [41]. Firstly, with the addition of the MAFD module, we observe consistent improvements compared to the baseline. Specifically, there is a relative increase of 2.31% in ER and 0.50% in FER for WAR. This demonstrates that modality-adversarial decoupling effectively reduces non-emotional interference and helps the model learn clearer modality-invariant emotion representations. Secondly, the EMT module further enhances performance with a relative increase of 8.19% in ER and 1.76% in FER for F1, indicating its capability to exploit complementary supervision from ER and FER labels. By jointly modeling both tasks, EMT alleviates task-specific bias and strengthens the cross-task generalization of the learned features. Ultimately, our full EMERT model, combining MAFD and EMT, achieves the best performance across all metrics, confirming that these modules are complementary and together contribute to both ER and FER tasks.

TABLE X
MODULE ABLATION STUDY ON THE 7-CLASS MER AND FER, RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Module			ER			FER		
Baseline	MAFD	EMT	WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)
✓			31.18	27.64	27.48	47.83	32.79	41.38
✓	✓		31.90	27.85	29.70	48.07	32.85	42.80
✓		✓	32.93	28.00	29.73	48.61	32.91	42.11
✓	✓	✓	**33.92**	**28.17**	**30.38**	**51.18**	**33.04**	**43.33**

*5) **Robustness of EMERT against Noisy Data**:* To evaluate robustness, we injected Gaussian noise with varying variances into each modality for the 3-class emotion recognition task. As shown in Table XI. Notably, existing methods such as MMIM [47] and Self-MM [46] exhibit a significant performance drop (over 5%) under noise. In contrast, EMERT shows only a 2.29% decrease in WAR when the variance is set to 0.01, with a slight increase in F1. These results demonstrate that EMERT is more resilient to modality-specific noise compared to prior approaches.

*6) **Effects of Different hyperparameter in EMERT**:* To investigate the sensitivity of EMERT to hyperparameter

TABLE XI
ROBUSTNESS COMPARISON OF DIFFERENT METHODS FOR 3-CLASS ER ON EMER. THE BEST RESULTS ARE IN BOLD.

Methods	Variance (Gaussian) setting		
	0.01(WAR/F1)	0.05(WAR/F1)	0.1(WAR/F1)
MMIM [47]	46.23/45.02	45.13/40.05	42.82/36.01
Self-MM [46]	42.98/42.24	42.27/42.16	37.20/40.45
EMERT	**56.99/57.01**	**55.42/57.30**	**56.33/56.23**

settings, we conduct experiments on the 7-class ER and FER tasks with varying values of the adversarial loss weights α and β. As shown in Table XII, the performance exhibits notable variation under different configurations. When $\alpha = 0.3$ and $\beta = 0.1$, EMERT achieves the best results, with WAR/UAR/F1 reaching 33.92/28.17/30.38 for ER and 51.18/33.04/43.33 for FER, respectively. This suggests that a moderate adversarial loss weight provides an effective balance: it is strong enough to guide modality-adversarial decoupling, but not so dominant as to destabilize training. In contrast, excessively small weights (e.g., $\alpha = 0.1$ or $\beta = 0.01$) reduce the effectiveness of the decoupling mechanism, while overly large weights (e.g., $\alpha = 0.5$ or $\beta = 1$) can overwhelm the supervised learning objective, leading to degraded performance. These findings confirm that EMERT is relatively robust to hyperparameter variations within a reasonable range, and that carefully tuning α and β further enhances the balance between adversarial feature decoupling and multi-task learning, thereby improving overall performance.

TABLE XII
HYPERPARAMETRIC ANALYSIS EXPERIMENT ON THE 7-CLASS ER AND FER, RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

α	β	ER			FER		
		WAR (↑)	UAR (↑)	F1 (↑)	WAR (↑)	UAR (↑)	F1 (↑)
0.1	0.01	30.19	26.31	26.51	45.31	29.83	38.13
	0.1	31.69	27.44	27.28	46.64	32.26	40.43
	1	31.2	26.57	26.36	48.28	32.13	40.47
0.3	0.01	28.67	24.39	21.57	44.66	26.38	34.31
	0.1	**33.92**	**28.17**	**30.38**	**51.18**	**33.04**	**43.33**
	1	33.09	27.36	29.73	51.15	32.8	40.71
0.5	0.01	25.14	19.59	16.68	41.79	19.58	26.75
	0.1	26.19	20.63	18.56	43.71	23.57	30.42
	1	29.55	23.77	21.79	47.84	28.93	36.93

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we explored the gap between Facial Expression Recognition (FER) and Emotion Recognition (ER) by introducing eye behaviors as a crucial emotional cue. To support this, we constructed the Eye-behavior-aided Multimodal Emotion Recognition (EMER) dataset, which contains 1,303 spontaneous emotional samples from 121 participants. Notably, EMER provides multi-view emotion labels for both ER and FER, enabling a comprehensive analysis to elucidate the gap between them. Building upon EMER, we propose the Eye-behavior-aided MER Transformer (EMERT), a simple yet effective model that integrates modality-adversarial feature decoupling and multitask learning to effectively fuse eye behaviors and facial expressions. Extensive experiments across seven benchmark protocols demonstrate that EMERT significantly outperforms state-of-the-art multimodal methods, highlighting the importance of modeling eye behaviors for robust and complementary emotion understanding. This paper provides a complete framework—from dataset to model—for advancing multimodal emotion recognition. Our findings offer insights into the gap between FER and ER, and emphasize the value of eye behaviors in enhancing emotional perception. In the future, we plan to extend EMER with more diverse and ecologically valid scenarios and release it to encourage further research in emotion-related tasks. Moreover, the proposed EMERT framework shows strong potential for real-world applications such as human–computer interaction, mental health monitoring, and emotionally intelligent virtual agents. We also aim to integrate large multimodal models to further investigate the relationship between FER and ER. These directions will not only enrich the academic value of this work but also enhance its practical impact.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

REFERENCES

[1] Y. Liu, L. Li, Y. Tu, B. Zhang, Z.-J. Zha, and Q. Huang, "Dynamic strategy prompt reasoning for emotional support conversation," *IEEE Trans. Multimedia*, vol. 27, pp. 108–119, 2025.
[2] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2024, pp. 4673–4682.
[3] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 6676–6688, 2024.
[4] Y. Li, M. Wang, M. Gong, Y. Lu, and L. Liu, "Fer-former: Multimodal transformer for facial expression recognition," *IEEE Trans. Multimedia*, vol. 27, pp. 2412–2422, 2025.
[5] 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.
[6] "Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources," *Information Fusion*, vol. 77, pp. 107–117, 2022.
[7] Z. Zhang, X. Tian, Y. Zhang, K. Guo, and X. Xu, "Label-guided dynamic spatial-temporal fusion for video-based facial expression recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 10 503–10 513, 2024.
[8] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proc. 23th ACM Int. Conf. Multimedia*, 2015, p. 423–426.
[9] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Automat. Face Gesture Recognit.*, 1998, pp. 200–205.
[10] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proc. 19th ACM Int. Conf. Multimedia*, 2017, p. 524–528.
[11] 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 *Proc. 28th ACM Int. Conf. Multimedia*, 2020, p. 2881–2889.
[12] P. Ekman, "Facial expression and emotion." *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.
[13] T. Sun, Y. Wei, J. Ni, Z. Liu, X. Song, Y. Wang, and L. Nie, "Muti-modal emotion recognition via hierarchical knowledge distillation," *IEEE Trans. Multimedia*, vol. 26, pp. 9036–9046, 2024.
[14] D. Heaven, "Why faces don't always tell the truth about feelings," *Nature*, vol. 578, pp. 502–504, 02 2020.
[15] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans Affect Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[16] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Trans Affect Comput.*, vol. 3, no. 1, pp. 18–31, 2012.
[17] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, 2019.
[18] V. Skaramagkas, E. Ktistakis, D. Manousos, E. Kazantzaki, N. S. IEEE Trans Affect Comput.hos, E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "esee-d: Emotional state estimation based on eye-tracking dataset," *Brain Sciences*, vol. 13, no. 4, 2023.
[19] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018, pp. 2236–2246.
[20] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans Auton Ment Dev*, vol. 7, no. 3, pp. 162–175, 2015.
[21] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface' 05 audio-visual emotion database," in *Proc. Int. Conf. Data Eng. Workshops*, 2006, pp. 8–8.
[22] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *IEEE Trans Affect Comput.*, vol. 6, no. 3, pp. 209–222, 2015.
[23] W. Liu, J. Qiu, W. Zheng, and B. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, 2022.
[24] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, 2021.
[25] E. H. Hess and J. M. Polt, "Pupil size as related to interest value of visual stimuli," *Science*, vol. 132, pp. 349 – 350, 1960.
[26] Q. Yang, Y. Li, C. Li, H. Wang, S. Yan, L. Wei, W. Dai, J. Zou, H. Xiong, and P. Frossard, "Svgc-ava: 360-degree video saliency prediction with spherical vector-based graph convolution and audio-visual attention," *IEEE Trans. Multimedia*, vol. 26, pp. 3061–3076, 2024.
[27] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Trans. Multimedia*, vol. 26, pp. 776–788, 2024.
[28] D. Chen, K. Pan, G. Dai, G. Wang, Y. Zhuang, S. Tang, and M. Xu, "Improving vision anomaly detection with the guidance of language modality," *IEEE Trans. Multimedia*, vol. 27, pp. 1410–1419, 2025.
[29] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6896–6905.
[30] F. Ma, B. Sun, and S. Li, "Spatio-temporal transformer for dynamic facial expression recognition in the wild," *arXiv preprint arXiv:2205.04749*, 2022.
[31] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1642–1651.
[32] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 303–311.
[33] "A multimodal fusion emotion recognition method based on multitask learning and attention mechanism," *Neurocomputing*, vol. 556, p. 126649, 2023.
[34] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, vol. 2019, 2019, p. 6558.
[35] L. Duan, H. Ge, Z. Yang, and J. Chen, "Multimodal fusion using kernel-based elm for video emotion recognition," in *Proceedings of ELM-2015 Volume 1: Theory, Algorithms and Applications (I)*. Springer, 2016, pp. 371–381.
[36] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new image contrast enhancement algorithm using exposure fusion framework," in *CAIP*. Springer, 2017, pp. 36–46.
[37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[38] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans Affect Comput.*, vol. 13, no. 4, pp. 2132–2143, 2022.
[39] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2852–2861.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[41] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
[42] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 2551–2566, 2022.
[43] Y. Zhang, M. Chen, J. Shen, and C. Wang, "Tailor versatile multi-modal learning for multi-label emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 8, 2022, pp. 9100–9108.
[44] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018, pp. 2247–2256.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst*, vol. 30, 2017.
[46] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, 2021, pp. 10 790–10 797.
[47] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *EMNLP*, 2021, pp. 9180–9192.
[48] G. Yin, Y. Liu, T. Liu, H. Zhang, F. Fang, C. Tang, and L. Jiang, "Token-disentangling mutual transformer for multimodal emotion recognition," *Eng. Appl. Artif. Intell*, vol. 133, p. 108348, 2024.
[49] Y. Liu, H. Zhang, Y. Zhan, Z. Chen, G. Yin, L. Wei, and Z. Chen, "Sample-cohesive pose-aware contrastive facial representation learning," *Int. J. Comput. Vis*, vol. 133, no. 5, pp. 3020–3040, 2025.
[50] E. I. Obilor and E. C. Amadi, "Test for significance of pearson's correlation coefficient," *IJIMSEP*, vol. 6, no. 1, pp. 11–23, 2018.
[51] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *ACL*. ACL, 2020, pp. 3718–3727.
[52] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1103–1114.
[53] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, S. A. McIlraith and K. Q. Weinberger, Eds., 2018, pp. 5634–5641.

Kejun Liu received the B.S. degree in Network Engineering from Henan University in 2022; she is currently working toward the Ph.D. degree at China University of Geosciences in Wuhan, China. Her research interests include affective computing.

Yuanyuan Liu received the PhD degree from Central China Normal University. She is currently an associate professor at the China University of Geosciences (Wuhan). She was also a visiting scholar in Nanyang Techonological University, Singapore. Her research interests include computer vision and multimodal analysis. She has published various top conferences and journals, such as CVPR, ACM MM, IEEE T-VCG, PR, IEEE TGRS, CIKM, INS, NC, IEEE FGR and so on.

Lin Wei received the B.S. and M.S. degrees in software engineering from the China University of Geosciences, Wuhan, China, in 2022 and 2025, respectively. Her research focus on affective computing.

Chang Tang (Senior Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2016. He joined the AMRL Laboratory, University of Wollongong, Wollongong, NSW, Australia, from September 2014 and September 2015. He has published various peer-reviewed articles, including those in highly regarded journals and conferences, such as TPAMI, TMM, TGRS, ICCV, CVPR, AAAI, ACM MM. His research interests include machine learning and computer vision.

Yibing Zhan received the bachelor's and doctor's degrees from the School of Information Science and Technology, University of Science and Technology of China, Hefei, China, in 2012 and 2018, respectively. From 2018 to 2020, he was an Associate Researcher with the School of Computer Science, Hangzhou Dianzi University, Hangzhou, China. Served as an algorithm scientist at JD Explore Academy from 2021 to 2025. He is currently employed at the School of Computer Science of Wuhan University. He has authored or coauthored many scientific papers in top conferences and journals such as NeurIPS, CVPR, ACM MM, ICCV. His research interests include graph generation, foundation model, and graph neural networks.

Zijing Chen is a Lecturer in the Department of Computer Science and Information Technology at La Trobe University, Australia, affiliated with the Cisco-La Trobe Centre for Artificial Intelligence and IoT. She received her Ph.D. from the University of Technology Sydney, Australia, in 2019. Her research interests include video processing, machine learning, computer vision, and artificial intelligence, with publications in high-quality journals and conferences.

Zhe Chen is a Lecturer in the Department of Computer Science and Information Technology at La Trobe University, Australia, affiliated with the Cisco-La Trobe Centre for Artificial Intelligence and IoT and the Australian Centre for AI in Medical Innovation. He received his Ph.D. from the University of Sydney in 2019. His research focuses on computer vision and artificial intelligence, with applications in healthcare, robotics, and related domains. His work is regularly published in top-tier venues such as CVPR and IJCV and has accumulated over 5,000 citations to date. Dr. Chen serves as a reviewer for leading journals and conferences in the field.