Supplementary Materials for "EMER"

I. ADDITIONAL DETAILS IN EMER

A. Selection of Stimulus Videos

Following existing databases [1], [2], we employ the method of inducing spontaneous emotional states, by presenting participants with specific emotional stimulus materials to induce their short-term emotional states. Initially, we crawled 115 stimulus videos based on seven basic emotion categories: happiness, sadness, disgust, anger, fear, surprise, and neutral. 60 videos were sourced from the MAHNOB-HCI database [2], while the remaining 55 were collected by four emotion experts from video-sharing websites such as BiliBili and Youtube. To ensure diversity, the videos were selected from various movies, TV shows, and short videos. Then, the crawl stimulus videos were scored by four other emotion experts. Finally, we ranked the videos according to their average ratings and chose the top 4 videos from each emotion category as the final emotional stimulus materials. Ultimately, we selected 28 videos as the final stimulus materials, with each stimulus material lasting between 1 and 2 minutes. Table I displays the detailed content of the selected stimulus materials. Some selected stimulus materials are shown in Figure 1.

TABLE I
DETAILS OF THE SELECTED 28 STIMULUS MATERIALS.

Emotion category	Video source	Average rating	Length(s)
Anger	School Bullying	4.33	64
Anger	Better Days	4.33	87
Anger	The Flowers of War	4	70
Anger	Double-sided Adhesive	4	107
Disgust	Hannibal	4	68
Disgust	The Walking Dead	4	90
Disgust	Eats Worms	4	63
Disgust	Wilderness Survival	4.6	100
Fear	Dark Web	4	93
Fear	Pictured	4	97
Fear	Lights Out	4	117
Fear	Ju-on: The Grudge	4	120
Happiness	Detective Chinatown 3	4	90
Happiness	Mermaid	4.33	68
Happiness	iPartment	4	90
Happiness	The Glorious Era	4.33	63
Sadness	Aftershock	4	90
Sadness	Pack	4.33	90
Sadness	Mom's Waiting	4	80
Sadness	A Little Reunion	4	80
Surprise	Magic Show	4	108
Surprise	Hooks	4	63
Surprise	Dr. Strange	4	64
Surprise	America's Got Talent	4	84
Neutral	Scenic Video 1	4	65
Neutral	Scenic Video 2	4	65
Neutral	Scenic Video 3	4	65
Neutral	Scenic Video 4	4	65

B. Details of Experimental Procedure

We recruit 121 participants from different regions, professions, ages, and genders to participate in the experiment of data collection. There are 76 males and 45 females, with ages ranging from 18 to 40 years old. The 121 participants are required to watch the selected emotional stimulus materials in a lab setting to induce the production of short-term and spontaneous emotion states. Before the commencement of data collection, each participant is required to sign an informed consent form, following the GDPR [1] principles. The informed consent form for our data collection can be seen in the supplementary material.

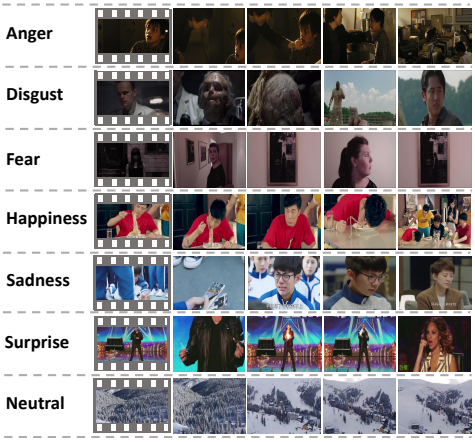


Fig. 1. Some examples of stimulus materials.

During the data collection, participants were seated in a suitable position in the lab and their data were recorded by using a Tobbi Pro Fusion eye-tracking device [2] and a high-definition camera. Before the start of the experiment, each participant underwent a 2-minute eye tracking calibration to ensure the reliability of the collected data. Participants were then asked to watch 14 stimulus videos, of which two stimulus videos were randomly selected for each emotional category. The collection procedure for each data consists of the following three steps:

- The participants watch a neutral landscape video or rested with their eyes closed to calm their mood before watching the stimulus video.
- The participants watch a 1-2 minute stimulus video.
- After watching the video, participants complete a self-assessment questionnaire that includes discrete emotional categories, emotional intensities, valence, and arousal ratings.

Through this data collection procedure, we collected 1,623 data. After data cleaning and pre-processing, we obtained 1,303 multimodal emotional data as the final EMER database.

C. Details of the Eye Movement Subset

The detailed description of the eye movement subset is presented in Figure 2. Our eye movement subset contains more comprehensive and richer information, such as time stamp, gaze point coordinates, gaze direction, pupil diameter,

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

eye position, gaze time, and eye movement event type (sweep and gaze).

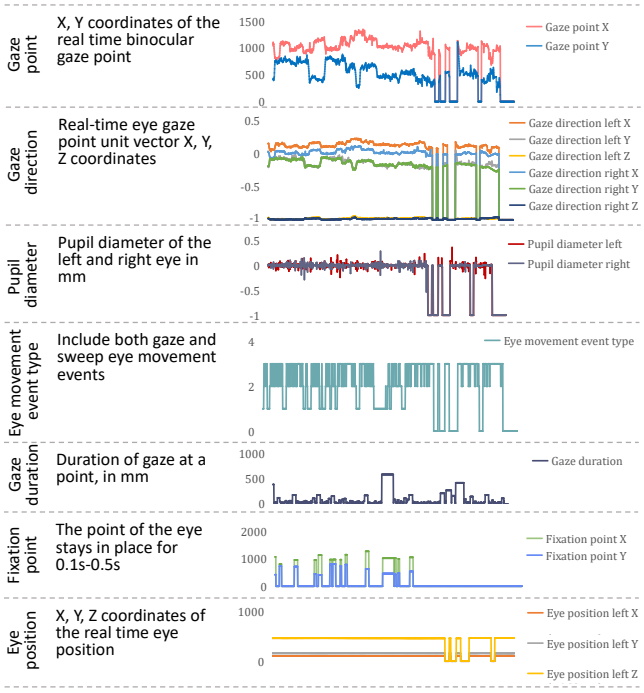


Fig. 2. Examples of eye movement data in EMER.

D. Alignment of Face Expressions, Eye Movements, and Eye fixation maps

We recruited four emotional experts to manually align and clip the collected facial expression videos, eye movement data, and eye fixation maps. We first selected only clips of facial expressions lasting longer than 12 seconds, along with their corresponding emotion-related eye movement data and eye fixation maps. Then, these experts confirmed the prominent clips by reviewing facial expression videos and assessing data quality using emotion-related eye fixation maps. They marked the beginning and end of the selected clips. As shown in Figure 3, the experts used the TobiiProLab software for data alignment and clipping, ensuring the synchronization of the collected data.

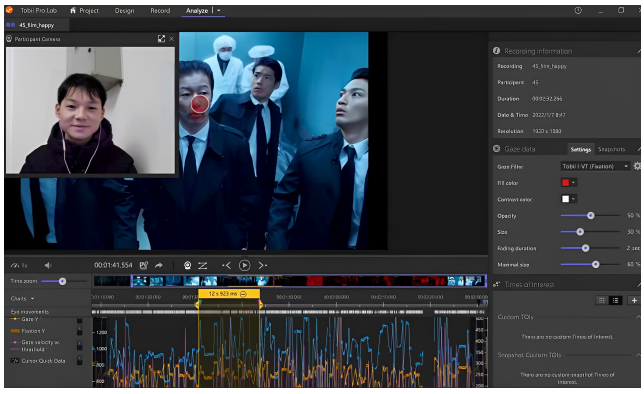


Fig. 3. Data alignment with the TobiiProLab software.

E. FER Annotation Software and Format

To effectively annotate emotions, we developed a tool called AnnotationTool to generate and save annotation files. The main interface of the AnnotationTool is shown in Figure 4.

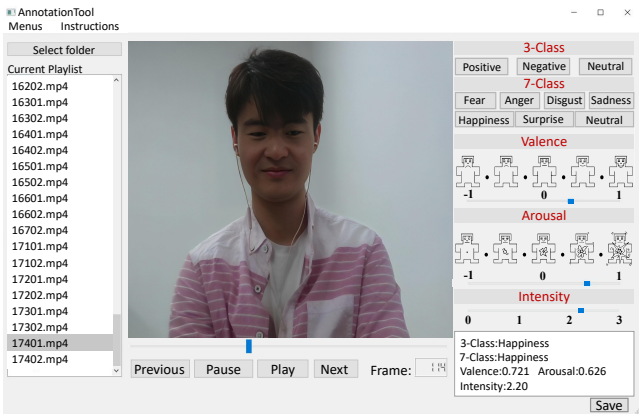


Fig. 4. The main interface of the AnnotationTool for FER annotation.

The Tabel II shows the annotation file format used in EMER. The "ID" represents the index of the facial expression clip, "7-class" represents the fine FER/ER labels annotated by the annotator for the clip, "3-class" represents the coarse FER/ER labels, the "Intensity" represents the intensity of the expression categories, the "Valence" represents the score of pleasure, and the "Arousal" represents the score of excitement.

TABLE II
EXAMPLES OF THE EMOTION ANNOTATION FORMAT IN EMER.

ID	7-class	3-class	Intensity	Valence	Arousal
03401	Happiness	Positive	2.96	1.000	0.982
42102	Sadness	Negative	2.28	-0.583	0.108
65202	Disgust	Negative	0.37	-0.750	0.221
53101	Anger	Negative	2.28	-1.000	0.706
01302	Fear	Negative	2.80	-0.750	0.535
57702	Surprise	Negative	2.58	-0.105	-0.250
09502	Neutral	Neutral	0.00	0.080	0.100

II. ETHICAL STATEMENT

Although this is a purely academic investigation, the potential sensitivity of facial information makes it necessary to clearly articulate the ethics involved. We claim that our EMER database is used for Academic Research Only and is compliant with GDPR [3] principles. We strictly adhered to ethical guidelines and principles to ensure the protection of subjects' privacy, the responsible handling of data, and in accordance with all applicable legal and ethical standards.

Informed Consent. All subjects are required to sign a written Informed Consent Form prior to participating in the data collection. The Informed Consent Form used in this study is shown in Figure 5. Participants were fully informed of the purpose, procedures, and potential risks of the study, and they were assured that their participation was voluntary, and they could withdraw from the study at any time without any repercussions.

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

Informed Consent

We are about to conduct data collection for the Eye-behavior-aided Multimodal Emotion Recognition (EMER) dataset and you are eligible for the study, so we would like to invite you to participate in the study. This informed consent will tell you about the purpose, procedures, benefits, risks of the study, so please read it carefully and make a careful decision about whether to participate in the study. When the researcher explains and discusses the informed consent form with you, you can ask questions at any time and have him/her explain to you what you do not understand. We promise that the data collected by this experiment will only be used for scientific research and not for commercial use, and we will take effective measures to strengthen the protection of your personal information, ensure the security of information, and never disclose your personal privacy.If you agree to participate, please sign the bottom of this consent.

1. Why is this study being conducted?
   Emotion recognition (ER) aims to identify human emotions from given data, and it predominantly relies on facial expression recognition (FER) results using facial images. However, we argue that FER can yield misleading results when facial expressions do not necessarily align with true emotions. To mitigate this issue, we introduce eye behaviors as an important context for ER, resulting in the creation of a new Eye-behavioraided Multimodal Emotion Recognition (EMER) dataset.
2. What is included in this study?
   (1) Watching emotionally stimulating videos under the guidance of staff to collect appropriate emotional data;
   (2) Fill out the self-assessment questionnaire after watching the video.
3. What are the risks of participating in this study?
   (1) Risks during data collection: some of the stimulus videos may cause physiological discomfort on a small scale due to different individual tolerances;
   (2) Risks of dataset disclosure: the psychological strain on individuals and families when information about a dataset is made public;
   (3) Unknown risks: There may be risks that are not currently known.
4. What are the benefits of participating in this study?
   Your participation will help advance the field of artificial intelligence and make a contribution to scientific research.
5. Is it mandatory to participate and complete this study?
   Your participation in the study is completely voluntary, and there will be no negative consequences for you if you withdraw from the study. Even after you have agreed to participate, you can change your mind at any time and tell the investigator to withdraw from the study.

Researcher's Signature Date

Subject's signature Date

Fig. 5. The Informed Consent Form.

Privacy and Anonymity. To protect the privacy and confidentiality of the participants, all identifiable information, such as names and addresses, were removed from the database. Where necessary, unique identifiers were used to replace personally identifiable information.

Data Processing Procedures. All data collection, storage, and processing procedures have been implemented, and compliance with the best practices of international data management and data protection is ensured. We will take strict security measures to protect the security of the data.

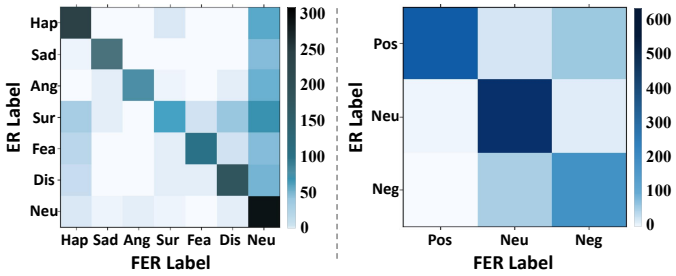
Data Sharing and Storage. The data collected throughout the study will be securely stored and made available to other researchers for Academic Purposes Only. Access to the database will be provided through a controlled and approved process to ensure that researchers adhere to ethical and data protection standards.

By taking these measures into account, our research aims to uphold the highest ethical standards and responsibly contribute to the academic community.

III. Qualitative Analysis

A. Deep Analysis of Emotion Gap between ER and FER

To illustrate the gap between ER and FER, Figure 6 displays the distribution discrepancy between ER labels and FER labels in our EMER dataset. In this context, Figure 6(a) represents the 7-class label confusion matrix, while Figure 6(b) represents the 3-class label confusion matrix. Apart from the diagonal of the matrix (representing consistent labels), each cell in the matrix reflects the disparity in distribution between the categories of the ER and FER labels. In the figure, we can find that many data with "Surprise" ER labels are annotated as "Neutral" in our FER labels.



(a) 7-class distribution discrepancy (b) 3-class distribution discrepancy

Fig. 6. The distribution gaps between the ER and FER labels in EMER. (a) The confusion matrix of 7-class ER and FER labels, (b) the confusion matrix of 3-class ER and FER labels. The horizontal axis represents the FER labels, and the vertical axis represents the ER labels. With the exception of the diagonal, each cell in the matrix indicates the variation in distribution between the FER and ER labeling categories. Hap, Sad, Ang, Sur, Fea, Dis, Neu, Neg, Pos are the abbreviations of the corresponding labels.

In addition, to facilitate a better understanding of the '*emotion gap*' between ER and FER, we provide two intuitive examples from our database. As depicted in Figure 7, when the expressions are the same, for instance, happiness (or positive), different emotion categories like happiness and sadness (or positive and negative) exhibit significant differences in their eye behavior. This includes variances in what they focus on, how frequently they blink, and how they regulate the size of their pupils (see the red rectangles and green spots in the figure). Accordingly, we can explicitly depict the 'emotion gap' by analyzing the discrepancy between facial expressions and eye behavior signals.

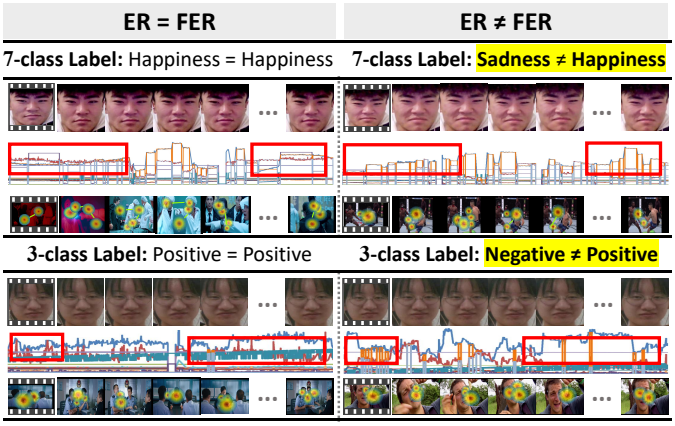


Fig. 7. A typical example of the emotion gap exists in the disparity between eye behaviors and facial expressions.

B. Visual Emotion Representation on EMER and SIMS

In addition, we added the visualization of emotion representations on EMER and SIMS in Figure 8 using 3D-TSNE [3]. Obviously, we observe that the emotion representations learned by our EMERT can be significantly separated according to the different emotion categories. This indicates that our EMER dataset enables the proposed method to learn modality-sensitive high-level features by effectively mitigating the emotion gap between facial expressions and eye behaviors, thus achieving more robust emotion recognition performance.

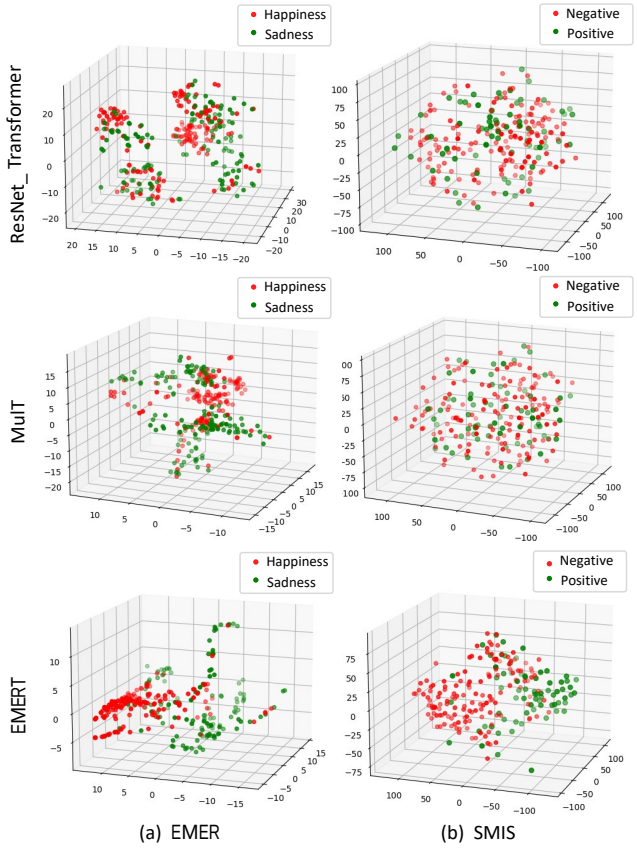


Fig. 8. The emotion representations from EMERT, MulT [4] and ResNet_Transformer [5], [6] on our EMER database and SIMS database, respectively.

REFERENCES

[1] 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.

[2] 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.

[3] Laurens, V. D. Maaten, and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.

[4] 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.

[5] 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.

[6] 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.