MER 2025: When Affective Computing Meets Large Language Models

Zheng Lian
MAIS, Institute of Automation, Chinese Academy of Sciences
Beijing, China

Rui Liu
Inner Mongolia University
Hohhot, China

Kele Xu
National University of Defense Technology
Changsha, China

Bin Liu
MAIS, CASIA
Beijing, China

Xuefei Liu
Tianjin Normal University
Tianjin, China

Yazhou Zhang
Tianjin University
Tianjin, China

Xin Liu
Lappeenranta-Lahti University of Technology
Lappeenranta, Finland

Yong Li
Southeast University
Nanjing, China

Zebang Cheng
Shenzhen University
Shenzhen, China

Haolin Zuo
Inner Mongolia University
Hohhot, China

Ziyang Ma
Shanghai Jiaotong University
Shanghai, China

Xiaojiang Peng
Shenzhen Technology University
Shenzhen, China

Xie Chen
Shanghai Jiaotong University
Shanghai, China

Ya Li
Beijing University of Posts and Telecommunications
Beijing, China

Erik Cambria
Nanyang Technological University
Singapore

Guoying Zhao
University of Oulu
Oulu, Finland

Björn W. Schuller
Technical University of Munich
Munich, Germany

Jianhua Tao
Tsinghua University
Beijing, China

Abstract

MER2025 is the third year of our MER series of challenges. Previously, MER2023[1] focused on multi-label learning, noise robustness, and semi-supervised learning, while MER2024[2] introduced a new track dedicated to open-vocabulary emotion recognition. This year, MER2025 centers on the theme "When Affective Computing Meets Large Language Models (LLMs)". We aim to shift the paradigm from traditional categorical frameworks reliant on predefined emotion taxonomies to LLM-driven generative methods, offering innovative solutions for more accurate and reliable emotion understanding. The challenge contains four tracks: **MER-SEMI** focuses on fixed categorical emotion recognition enhanced by semi-supervised learning; **MER-FG** explores fine-grained emotions, expanding recognition from basic to nuanced emotional states; **MER-DES** incorporates multimodal cues (beyond emotion words) into predictions to enhance model interpretability; **MER-PR** reveals whether emotion prediction results can improve personality recognition performance. For the first three tracks, the baseline code is available at MERTools[3] and datasets can be accessed via Hugging Face[4]. For the last track, the dataset and baseline code are available on GitHub[5].

CCS Concepts

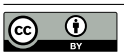
• **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

MER2025, basic and fine-grained emotion recognition, descriptive emotion understanding, emotion-enhanced personality recognition

ACM Reference Format:
Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2025. MER 2025: When Affective Computing Meets Large Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746027.3762007

[1] *http://merchallenge.cn/mer2023*
[2] *https://zeroqiaoba.github.io/MER2024-website*



MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762007

[3] *https://github.com/zeroQiaoba/MERTools*
[4] *https://huggingface.co/datasets/MERChallenge/MER2025*
[5] *https://github.com/cai-cong/MER25_personality*

1 Introduction

Human emotion expression is a complex process that typically involves multiple modalities, including facial expressions, vocal tones, body movements, gestures, and even physiological signals. This complexity has spurred the development of Multimodal Emotion Recognition (MER), a critical task that aims to integrate cross-modal cues to identify human emotions. Recently, MER research has evolved in two key directions: a) from coarse-grained [11, 16] to fine-grained emotion recognition [12], and b) from categorical approaches to descriptive emotion understanding [4, 10], aiming to enhance both prediction accuracy and interpretability. In line with these advancements, MER2025@ACM Multimedia introduces four tracks aligned with current research priorities.

Track 1. MER-SEMI. Recent studies have demonstrated that pre-training on large-scale unlabeled data, especially domain-matched data, can significantly enhance model performance [5, 16, 24]. This track provides a substantial collection of unlabeled samples from the same domain as the labeled data. Participants are encouraged to leverage semi-supervised learning techniques, such as masked auto-encoders or contrastive learning, to achieve better results.

Track 2. MER-FG. Current frameworks primarily focus on basic emotions, often failing to capture the complexity and subtlety of human emotions. This track shifts the emphasis to fine-grained MER, enabling the prediction of a broader range of emotions. Following previous works [10, 12], participants are encouraged to leverage large language models (LLMs) for this purpose. Given that LLMs possess extensive vocabularies, they hold the potential to generate more diverse emotion categories beyond basic labels.

Track 3. MER-DES. The first two tracks primarily focus on emotion words, neglecting the integration of multimodal clues during the inference process. This omission results in prediction outcomes that lack interpretability. Moreover, emotion words struggle to fully capture the dynamic, diverse, and sometimes ambiguous nature of human emotions. This track seeks to leverage free-form, natural language descriptions to represent emotions [10, 17], offering greater flexibility to achieve more accurate emotion representations and enhance model interpretability.

Track 4. MER-PR. Personality and emotion are deeply intertwined in human behavior and social interactions, yet current research often treats them as separate tasks, neglecting their inherent correlations. This track seeks to investigate the interplay between emotion and personality, exploring whether emotion recognition can enhance the accuracy of personality predictions. Participants are encouraged to employ techniques such as multi-task learning to analyze the influence of emotion on personality prediction.

2 MER-SEMI

2.1 Dataset

MER-SEMI spans three consecutive MER challenges, aiming to enhance the performance of categorical emotion recognition algorithms through semi-supervised learning and unlabeled data. This year, we expanded the dataset by incorporating more labeled and unlabeled samples. Our raw data comes from two sources: a) conversational emotion datasets MC-EIU [18] and M3ED [26], with explicit approval from dataset owners; and b) 19 Chinese TV series sourced from publicly available platforms. Then, we follow the video segmentation and filtering process used in prior MER challenges [13, 14], ensuring each video contains predominantly one speaker with relatively complete speech content. Compared to MER2024 (120k samples), we increased the dataset size to 132k samples in MER2025, introducing richer topic diversity and more characters. Detailed dataset statistics are provided in Table 1.

Table 1: Dataset Statistics for MER-SEMI. In MER2025, we further expand the dataset, comprising 7,369 labeled samples and 124,802 unlabeled samples. Additionally, we select 2,026 samples from the unlabeled subset for evaluation.

	Train&Val	# of test samples (labeled/whole)	Total
MER-SEMI (2023) [13]	3,373	834/73,982	77,355
MER-SEMI (2024) [14]	5,030	1,169/115,595	120,625
MER-SEMI (2025)	7,369	2,026/124,802	132,171

For evaluation purposes, manually annotating 124k unlabeled data is impractical due to the substantial time and financial costs. Thus, we select a subset of the unlabeled data for performance assessments. To ensure label reliability, we recruit nine annotators and conduct a preliminary test to evaluate their alignment with emotion experts. In this process, two annotators are excluded based on insufficient agreement, leaving seven annotators for the labeling task. Then, we randomly sample 10k instances from the unlabeled subset and ask the annotators to choose the most likely label from eight categories: *neutral*, *happy*, *angry*, *sad*, *surprise*, *worry*, *others*, *unknown*. In this process, each annotator is assigned a portion of the 10k samples to reduce annotation time. Meanwhile, we ensure that each sample is labeled by at least five annotators. Samples that receive at least 80% agreements and whose majority vote is neither *others* nor *unknown* are included in the test set. This process yields 2,026 high-quality test samples, ensuring strong inter-annotator agreement and reliable ranking results.

2.2 Evaluation Metrics

This track aims to identify the most likely label from six candidate emotions: *neutral*, *happy*, *angry*, *sad*, *surprise*, and *worry*. For evaluation purposes, we use the same metrics as previous MER challenges [13, 14]: accuracy and weighted average F1-score (WAF). Given the inherent class imbalance in the dataset, we prioritize WAF as the primary metric for final ranking.

2.3 Baseline Framework

A categorical model primarily relies on two key components: feature selection and model architecture. For feature selection, we evaluate the performance of both handcrafted and model-driven features. For model architecture, MERBench highlights that a simple attention mechanism, which maps different unimodal features to the same dimension and then fuses them using attention weights, can already achieve strong performance [16]. In contrast, more complex fusion architectures may lead to overfitting problems and are not suitable for MER, where labeled data is usually limited. For implementation details, please refer to MERBench and our baseline code.

Table 2: MER-SEMI baseline results (%). We also report five-fold cross-validation results on the Train&Val set. For multimodal results, Top1 means selecting the best-performing unimodal feature, while Top2 means selecting the two best-performing unimodal features.

Feature	Train&Val		MER-SEMI	
	WAF (↑)	ACC (↑)	WAF (↑)	ACC (↑)
Visual Modality				
ResNet-FER2013 [7]	57.83±0.20	58.71±0.34	52.16±0.25	52.90±0.22
DINOv2-large [20]	58.75±0.11	59.77±0.11	52.43±0.34	53.19±0.27
SENet-FER2013 [9]	57.77±0.29	58.76±0.22	53.69±0.08	54.90±0.24
MANet-RAFDB [27]	59.90±0.15	60.87±0.08	54.31±0.19	54.95±0.14
CLIP-base [22]	61.72±0.19	62.40±0.22	56.30±0.19	57.53±0.26
CLIP-large [22]	66.58±0.15	66.95±0.13	60.50±0.19	61.08±0.15
Acoustic Modality				
WavLM-base [3]	58.62±0.11	58.91±0.09	54.55±0.28	55.81±0.19
wav2vec 2.0-large [1]	67.64±0.23	67.63±0.20	61.66±0.27	62.76±0.26
wav2vec 2.0-base [1]	67.55±0.27	67.52±0.29	62.59±0.27	63.43±0.30
Whisper-large [23]	66.51±0.15	66.56±0.17	66.52±0.26	67.05±0.29
HUBERT-base [8]	72.36±0.09	72.41±0.08	68.13±0.26	69.05±0.20
HUBERT-large [8]	76.29±0.07	76.36±0.07	72.27±0.28	72.90±0.36
Lexical Modality				
MacBERT-base [6]	53.23±0.12	53.40±0.15	52.54±0.13	52.56±0.26
MacBERT-large [6]	53.68±0.13	53.83±0.13	52.75±0.14	52.91±0.29
BLOOM-7B [25]	54.03±0.10	54.13±0.15	53.32±0.32	53.08±0.30
RoBERTa-large [19]	53.80±0.11	54.01±0.09	53.66±0.45	53.55±0.42
RoBERTa-base [19]	53.05±0.08	53.30±0.10	53.78±0.23	53.71±0.21
Acoustic + Visual				
Top1	80.65±0.09	80.67±0.09	77.10±0.44	77.54±0.50
Top2	80.63±0.10	80.67±0.13	75.89±0.23	76.28±0.25
Acoustic + Lexical				
Top1	76.95±0.22	77.06±0.17	73.64±0.24	74.31±0.16
Top2	76.80±0.07	76.90±0.08	73.57±0.17	74.13±0.22
Visual + Lexical				
Top1	73.52±0.18	73.56±0.16	72.09±0.21	72.55±0.32
Top2	73.89±0.12	73.98±0.12	72.13±0.30	72.75±0.27
Acoustic + Visual + Lexical				
Top1	82.05±0.11	82.10±0.12	**78.63**±0.53	**78.77**±0.55
Top2	81.82±0.08	81.86±0.06	77.47±0.26	77.58±0.25

2.4 Baseline Results

Our baseline code is designed to automatically and randomly select hyperparameters. In practice, we execute each command 50 times, identify the optimal hyperparameter combination, and then run the code six times under this configuration to report the average results along with the standard deviation. Table 2 summarizes the baseline performance for MER-SEMI. To ensure reproducibility, we have included all feature extraction code and pretrained weights in the baseline code. From Table 2, a strong correlation emerges between the five-fold cross-validation results on the Train&Val set and the test set performance. This suggests that participants can use Train&Val results as a reliable indicator for model selection. Notably, trimodal fusion achieves the highest performance, underscoring the complementary value of each modality in emotion recognition. However, for multimodal results, Top2 does not yield better performance than Top1. This is because the fusion process may inadvertently introduce emotion-irrelevant features, potentially degrading overall performance.

3 MER-FG

3.1 Dataset

Unlike MER-SEMI, which restricts the prediction scope to six candidates, MER-FG does not limit the label space, allowing predictions for any number and any emotion categories for each sample. In this track, we utilize two recently released datasets, OV-MERD [12] and MER-Caption+ [10], as training datasets. Their statistics are summarized in Table 3. Specifically, OV-MERD employs a *human-led, model-assisted* annotation strategy, where MLLMs first provide pre-extracted multimodal clues and then rely heavily on manual verification to ensure label quality. In contrast, MER-Caption+ adopts a *human-assisted, model-led* annotation strategy, leveraging human priors to guide description generation and sample filtering, ultimately achieving an automatic annotation process. Consequently, OV-MERD provides small-scale but high-quality labels, whereas MER-Caption+ offers large-scale labels that may contain some errors in emotion annotations.

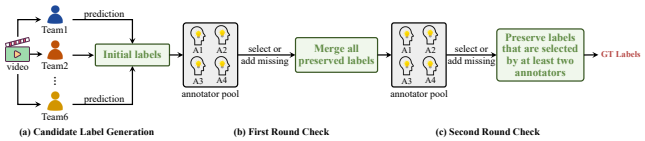


Figure 1: Annotation pipeline for MER-FG.

Table 3: Dataset statistics for MER-FG.

	Train&Val	# of test samples (labeled/whole)
OV-MERD [13]	332	1,200/124,802
MER-Caption+ [14]	31,327	

For evaluation purposes, we randomly selected 1,200 samples from the unlabeled set in MER-SEMI and engaged four annotators to label them. Figure 1 presents the annotation pipeline. Specifically, we first aggregated the Top-6 teams' submission results from the previous year's MER-OV track to generate initial labels. Each annotator was then tasked with either selecting labels they deemed correct or adding any additional labels they considered appropriate but not included in the provided candidate list. This approach enabled us to obtain richer emotion annotations for each sample compared to directly asking annotators to provide labels without candidate options. Our manual verification process consisted of two rounds. In the first round, we preserved all labels selected by different annotators to ensure comprehensiveness. In the second round, we retained only those labels confirmed by at least two annotators to ensure accuracy. Through this two-step verification process, each final label was validated at least three times, ensuring both comprehensiveness and accuracy of the annotation results.

3.2 Evaluation Metrics

This track does not restrict the number or category of labels during prediction. Therefore, traditional metrics like accuracy are not suitable for MER-FG. For evaluation purposes, we follow the previous work [12] and compute results in two stages.

3.2.1 Grouping. First, we apply a three-level grouping strategy to reduce the impact of synonyms:

M1. We normalize different forms of emotion words to their base form. For example, we map *happier* and *happiness* to *happy*. This function is denoted as $F(\cdot)$.

M2. We map synonyms to a unified label. For example, we map *happy* and *joyful* to *happy*. We call this function $G(\cdot)$.

M3. The emotion wheel provides natural hierarchical grouping, with basic emotions located in the innermost layer and more nuanced labels arranged in the outer layers [21]. First, we group the labels by their levels from the innermost to the outermost as L_w^1, L_w^2, and L_w^3. We then introduce two grouping functions, $W_1(\cdot)$ and $W_2(\cdot)$. For $W_1(\cdot)$, all labels are mapped to their corresponding labels in L_w^1. For $W_2(\cdot)$, all labels are mapped to corresponding labels in L_w^2. The above grouping functions can be summarized as:

$$L_1(\cdot) = W_1(G(F(\cdot))), \tag{1}$$

$$L_2(\cdot) = W_2(G(F(\cdot))). \tag{2}$$

In this paper, we employ five emotion wheels $\{w_i\}_{i=1}^5$, following the approach of previous works [12]. This process mitigates the influence of the choice of emotion wheels, thereby yielding more reliable evaluation results. Consequently, the aforementioned grouping function is dependent on the specific emotion wheel used. Therefore, we denote the final grouping functions as $L_{w_i}^1(\cdot)$ and $L_{w_i}^2(\cdot)$, where the former focuses on coarse-grained grouping, while the latter emphasizes more fine-grained grouping.

3.2.2 Metrics. For each sample, the number of labels is variable. Let the dataset consist of N samples. For sample x_i, the true labels are denoted as $\mathbf{Y}_i$ and the predicted labels are denoted as $\hat{\mathbf{Y}}_i$. For a grouping function $M \in \{L_{w_i}^1, L_{w_i}^2\}$, we define the following evaluation metrics:

$$\text{Precision}_\text{S}^M = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|M(\mathbf{Y}_i) \cap M(\hat{\mathbf{Y}}_i)\right|}{\left|M(\hat{\mathbf{Y}}_i)\right|}, \tag{3}$$

$$\text{Recall}_\text{S}^M = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|M(\mathbf{Y}_i) \cap M(\hat{\mathbf{Y}}_i)\right|}{\left|M(\mathbf{Y}_i)\right|}, \tag{4}$$

$$\text{F}_\text{S}^M = 2 \times \frac{\text{Precision}_\text{S}^M \times \text{Recall}_\text{S}^M}{\text{Precision}_\text{S}^M + \text{Recall}_\text{S}^M}. \tag{5}$$

We define the following scores S_1 and S_2 based on different grouping functions, and use their average results for the final ranking:

$$S_1 = \text{Avg}[\text{F}_\text{S}^M], M(\cdot) \in L_{w_i}^1(\cdot), \tag{6}$$

$$S_2 = \text{Avg}[\text{F}_\text{S}^M], M(\cdot) \in L_{w_i}^2(\cdot). \tag{7}$$

3.3 Baseline Framework

3.3.1 Zero-shot Baselines. The primary objective of MER-FG is to generate appropriate emotion labels for a given sample without being constrained by a predefined emotion taxonomy. Consequently, LLM-driven baselines are well-suited for this task, as they have extensive vocabularies that enable the generation of fine-grained emotion labels. Given that emotions are often expressed through multimodal cues, we primarily select multimodal LLMs (MLLMs) as our baselines, including representative frameworks such as SALMONN and Chat-UniVi. To extract emotion-related clues, we employ the following prompt: *As an expert in the field of emotions, please focus on the facial expressions, body movements, environment, acoustic information, subtitle content, etc., in the video to discern clues related to the emotions of the individual. Please provide a detailed description and ultimately predict the emotional state of the individual in the video.*

Next, we use Qwen2.5 to extract emotion labels from the above descriptions using the following prompt: *Please assume the role of an expert in the field of emotions. We provide clues that may be related to the emotions of the characters. Based on the provided clues, please identify the emotional states of the main characters. Please separate different emotional categories with commas and output only the clearly identifiable emotional categories in a list format. If none are identified, please output an empty list.*

Finally, we obtain the emotion prediction results for MER-FG. This process does not involve any training and operates in a zero-shot setup. For MLLMs, we use their 7B parameter weights by default. All models are implemented in PyTorch, and all inference processes are conducted on an A100 GPU. Additionally, we leverage the vLLM toolkit to accelerate the inference process.

3.3.2 AffectGPT. We also evaluate the performance of an emotion-specific MLLM, AffectGPT [10]. This framework employs a pre-fusion mechanism to enhance multimodal integration. During our experiments, we train the model on two datasets: OV-MERD and MER-Caption+. All pretrained models are available in our official baseline code repository. For detailed implementation instructions, please refer to the baseline code.

3.4 Baseline Results

Table 4 presents all baseline results. In Table 4, AffectGPT outperforms the zero-shot baselines, achieving higher scores in both coarse-grained scores S_1 and fine-grained scores S_2. Additionally, the model trained on MER-Caption+ demonstrates superior performance compared to the one trained on OV-MERD. These findings indicate that although OV-MERD has high-quality labels, its limited dataset size is insufficient to support effective training. In contrast, MER-Caption+ offers a larger number of samples with relatively accurate labels, resulting in better performance on MER-FG.

Table 4: MER-FG baseline results.

Model	Metrics		
	S_1 (↑)	S_2 (↑)	Avg (↑)
Otter	14.64	04.46	09.55
Video-LLaVA	27.40	12.18	19.79
Qwen-Audio	28.22	16.27	22.25
VideoChat2	34.07	17.78	25.92
Video-ChatGPT	35.29	19.77	27.53
LLaMA-VID	40.89	21.60	31.25
SALMONN	41.33	22.50	31.92
Chat-UniVi	43.33	23.90	33.62
VideoChat	43.48	24.30	33.89
mPLUG-Owl	46.28	27.32	36.80
AffectGPT(OV-MERD) [10]	47.81	30.16	38.98
AffectGPT(MER-Caption+) [10]	**57.36**	**36.35**	**46.86**

4 MER-DES

4.1 Dataset

ER-SEMI and MER-FG focus on emotion word prediction, whereas MER-DES extends beyond emotion words by integrating multimodal clues to enhance the interpretability of each emotion label. Since OV-MERD and MER-Caption+, which are used in MER-FG, also provide emotion descriptions, they are employed as the training set for MER-DES.

4.2 Evaluation Metrics

MER-DES leverages multimodal cues to enhance the interpretability of each emotion label. Given that interpretability is inherently a relatively subjective metric, we will invite two members from each team to assist with the scoring process. The evaluation will be conducted along two dimensions. First, we will evaluate whether the submission results include interpretable clues. This criterion distinguishes MER-DES from MER-SEMI and MER-FG by ensuring that participants do not merely provide emotion labels without supporting evidence. Second, we will assess the quality of the emotion descriptions from two perspectives: 1) whether the provided emotion clues are present in the video; and 2) whether the emotion and its associated clues are logically connected.

Initially, we plan to provide real descriptions and calculate similarity scores between real and predicted descriptions, as done in EMER [17]. However, human emotional expression is complex and difficult to fully capture all emotion-related visual and acoustic cues. Therefore, we shift our focus to pairwise ranking, as we do in EMER-Ranker [15]. During the ranking process, each team is required to participate in the labeling process. We will exclude annotators with low inter-annotator agreement compared to others to ensure annotation quality. Additionally, we will mask the sample names and require participants to submit results in English while using their translated Chinese versions for ranking. This approach prevents participants from identifying their own submissions and deliberately assigning inflated scores. The final rankings will be determined based on these manual annotations. Please refer to EMER-Ranker [15] for more details.

4.3 Baseline Framework

4.3.1 Zero-shot Baselines. You can also use MLLMs as baselines. Specifically, you can prompt the MLLMs with the following instructions: *As an expert in the field of emotions, please focus on the facial expressions, body movements, environment, acoustic information, subtitle content, etc., in the video to discern clues related to the emotions of the individual. Please provide a detailed description and ultimately predict the emotional state of the individual in the video.*

4.3.2 AffectGPT. You can also fine-tune AffectGPT using the emotion descriptions from the OV-MERD and MER-Caption+ datasets. **The official GitHub repository already includes baseline code and pretrained weights for this track.**

5 MER-PR

5.1 Dataset

This track aims to explore whether emotion prediction results can enhance the performance of personality recognition. For this purpose, we utilize the MDPE dataset [2], which provides annotations for both emotion and personality traits. The dataset and baseline code are available on GitHub[6]. Table 5 provides dataset statistics.

For personality traits, we used a Big Five personality questionnaire consisting of 60 items. Responses were used to derive scores for five personality dimensions (*openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*), with each dimension represented as a floating-point value ranging from 0 to 1. The dataset includes 233 participants, all native Chinese speakers from diverse backgrounds.

For emotion labels, each participant watched 16 emotion-induction videos, comprising two videos designed to elicit each of the eight target emotions: *sadness*, *happiness*, *relaxation*, *surprise*, *fear*, *disgust*, *anger*, and *neutral*. After watching, participants described their emotional responses and completed a self-report emotion scale to quantify the intensity of each emotion on a 1–5 Likert scale (1 = no emotion, 5 = strongest emotion). This process collected 16 samples per participant, resulting in a total of 3,728 samples.

Table 5: Dataset statistics for MER-PR.

	Train	Val	Test	Total
Subjects	153	40	40	233
# Samples	2,448	640	640	3,728

5.2 Evaluation Metric

This track focuses on personality recognition. In the dataset, personality traits are quantified using five floating-point values. For evaluation, we employ the Root Mean Square Error (RMSE) metric to measure the discrepancy between true and predicted scores.

5.3 Baseline Framework

We adopt the same baseline as in Section 2.3. For unimodal features, we use fully connected layers to extract hidden representations and predict personality scores. For multimodal features, we employ feature concatenation for fusion, and the resulting fused features are used for personality prediction.

5.4 Baseline Results

5.4.1 Unimodal Results. We establish a unimodal benchmark for personality trait prediction across visual, acoustic, and textual modalities (see Table 6). For the visual modality, the ViT feature achieves competitive performance, particularly in predicting *conscientiousness*. Among acoustic features, Wav2Vec2-base yields the highest average results. The textual modality outperforms all other unimodal features, with Baichuan-13B achieving the best average scores on the test set (0.156), excelling in *extraversion* (0.157), *agreeableness* (0.214), and *neuroticism* (0.157). Meanwhile, although textual cues are the most informative, visual and acoustic signals provide complementary trait-specific insights. For instance, ViT is particularly effective for *conscientiousness*, whereas Wav2Vec2-base performs well for *agreeableness*.

5.4.2 Multimodal Results. In Table 7, we present the multimodal fusion results based on several best-performing unimodal features. While multimodal fusion does not consistently outperform individual modalities in terms of average test scores, it yields improvements for specific personality traits. This underscores the potential advantages of tailoring fusion strategies to individual traits.

6 Conclusions

This year's MER2025 focuses on the theme "When Affective Computing Meets Large Language Models" and contains four distinct

[6] *https://github.com/cai-cong/MER25_personality*

Table 6: Unimodal results for MER-PR.

Feature	Val						Test					
	O(↓)	C (↓)	E (↓)	A (↓)	N (↓)	Avg (↓)	O (↓)	C (↓)	E (↓)	A (↓)	N(↓)	Avg (↓)
VIT	0.179	**0.159**	0.174	0.231	0.183	0.185	**0.145**	0.106	0.178	0.217	**0.156**	0.160
ClipVIT-B16	0.171	0.169	0.177	0.225	0.187	0.185	0.158	0.108	0.180	0.220	0.165	0.166
ClipVIT-L14	0.172	0.168	0.180	0.230	**0.180**	0.186	0.173	0.114	0.176	0.229	0.166	0.172
HUBERT-base	0.173	0.162	0.176	0.228	0.184	0.185	0.160	**0.098**	0.183	0.218	0.167	0.165
HUBERT-large	0.200	0.171	0.194	0.240	0.200	0.201	0.189	0.150	0.210	0.228	0.174	0.190
Wav2vec2-base	0.169	0.160	**0.172**	**0.221**	0.182	**0.180**	0.159	0.092	0.177	0.215	0.158	0.160
Wav2vec2-large	0.198	0.168	0.190	0.237	0.200	0.198	0.190	0.146	0.205	0.223	0.173	0.187
WavLM-base	0.178	0.164	0.179	0.227	0.184	0.186	0.156	0.101	0.180	0.219	0.162	0.164
WavLM-large	0.191	0.174	0.191	0.246	0.206	0.201	0.200	0.147	0.209	0.251	0.205	0.202
Sentence-BERT	0.181	0.181	0.176	0.234	0.181	0.190	0.159	0.104	0.181	0.219	0.163	0.165
ChatGLM2-6B	0.188	0.180	0.195	0.248	0.202	0.203	0.182	0.133	0.200	0.240	0.180	0.187
Baichuan-13B	**0.167**	0.160	**0.172**	0.229	**0.180**	0.182	0.156	**0.098**	**0.157**	**0.214**	0.157	**0.156**

Table 7: Multimodal results for MER-PR.

Features			Val						Test					
V	A	T	O(↓)	C (↓)	E (↓)	A (↓)	N (↓)	Avg (↓)	O (↓)	C (↓)	E (↓)	A (↓)	N(↓)	Avg (↓)
VIT	W2V	–	0.171	**0.160**	0.175	**0.223**	0.185	0.183	0.159	0.095	0.177	0.215	**0.158**	0.160
–	W2V	BAI	0.170	**0.160**	**0.171**	**0.223**	0.184	**0.182**	0.159	0.096	0.175	0.215	0.160	0.161
VIT	–	BAI	**0.167**	0.161	0.172	0.228	**0.182**	**0.182**	**0.154**	0.097	0.175	0.216	**0.158**	0.160
VIT	W2V	BAI	0.172	**0.160**	0.175	0.224	0.187	0.184	0.156	**0.093**	**0.171**	**0.212**	0.160	**0.158**

tracks. This paper presents the datasets, baselines, evaluation metrics, and experimental results for all tracks. For MER-SEMI, we evaluate various unimodal features and multimodal fusion approaches, providing a strong baseline for categorical emotion recognition under a fixed taxonomy. For MER-FG and MER-DES, we employ LLM-driven baselines to generate fine-grained emotions with corresponding multimodal clues, ensuring accurate and interpretable emotion understanding. For MER-PR, we test diverse unimodal features for personality detection. We hope all participants enjoy this year's challenge! Your continued support and engagement make this challenge truly meaningful.

7 Acknowledgments

This workshop is supported by the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2024311) and the National Natural Science Foundation of China (62201572, 62322120, 61831022, 62276259, U21B2010).

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the Advances in Neural Information Processing Systems*. 12449–12460.

[2] Cong Cai, Shan Liang, Xuefei Liu, Kang Zhu, Zhengqi Wen, Jianhua Tao, Heng Xie, Jizhou Cui, Yiming Ma, Zhenhua Cheng, Hanzhe Xu, Ruibo Fu, Bin Liu, and Yongwei Li. 2024. Mdpe: A multimodal deception dataset with personality and emotional characteristics. *arXiv preprint arXiv:2407.12274* (2024).

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[4] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems* 37 (2024), 110805–110853.

[5] Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. 2023. Semi-supervised multimodal emotion recognition with expression mae. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9436–9440.

[6] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 657–668.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.

[8] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[9] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[10] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. 2025. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *Forty-second International Conference on Machine Learning*.

[11] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.

[12] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. 2025. OV-MER: Towards Open-Vocabulary Multimodal Emotion Recognition. In *Forty-second International Conference on Machine Learning*.

[13] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9610–9614.

[14] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2024. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*. 41–48.

[15] Zheng Lian, Licai Sun, Haoyu Chen, Zebang Cheng, Fan Zhang, Ziyu Jia, Ziyang Ma, Fei Ma, Xiaojiang Peng, and Jianhua Tao. 2025. DMER-Ranker: Learning to Rank Emotion Descriptions in the Absence of Ground Truth. *arXiv preprint arXiv:2507.04278* (2025).

[16] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2024. MERBench: A Unified Evaluation Benchmark for Multimodal Emotion Recognition. *arXiv preprint arXiv:2401.03429* (2024).

[17] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable Multimodal Emotion Reasoning. *arXiv preprint arXiv:2306.15401* (2023).

[18] Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. 2024. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751* (2024).

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[21] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1 (1980).

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 28492–28518.

[24] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6110–6121.

[25] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[26] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5699–5710.

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