MAC'26: The 3rd Micro-Action Analysis Grand Challenge

Dan Guo
Hefei University of Technology
Hefei, China
Institute of Artificial Intelligence (IAI), Hefei Comprehensive National Science Center
Hefei, China
guodan@hfut.edu.cn

Xiaobai Li
Zhejiang University
Hangzhou, China
University of Oulu
Oulu, Finland
xiaobai.li@oulu.fi

Kun Li
United Arab Emirates University
AL Ain, United Arab Emirates
kunli.hfut@gmail.com

Haoyu Chen
University of Oulu
Oulu, Finland
chen.haoyu@oulu.fi

Guoying Zhao
University of Oulu
Oulu, Finland
guoying.zhao@oulu.fi

Meng Wang
Hefei University of Technology
Hefei, China
Institute of Artificial Intelligence (IAI), Hefei Comprehensive National Science Center
Hefei, China
eric.mengwang@gmail.com

ABSTRACT

The **M**icro-**A**ction Analysis Grand **C**hallenge (**MAC 2026**) builds on MAC 2024 and MAC 2025, with a specific focus on fine-grained micro-action understanding. This edition targets the automatic analysis of subtle whole-body behaviors using computer vision and machine learning. Micro-actions are spontaneous, often fleeting body movements that reflect genuine emotions and intentions, yet are difficult to recognize and distinguish due to their subtle appearance. MAC 2026 comprises three tasks, Micro-Action Recognition (MAR), Multi-label Micro-Action Detection (MMAD), and a newly introduced task on Fine-grained Micro-Action Understanding (FMAU), based on the MA-52, MMA-52, and MABench datasets, which contain 52 micro-action categories and 8 annotated body parts. The goal is to promote innovative research on fine-grained whole-body micro-actions, establish strong benchmarks, and advance technologies for human behavior understanding and emotional state analysis.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Human Behavior Analysis, Micro-action Recognition, Multi-label Micro-Action Detection, Fine-grained Micro-Action Understanding

Conference'24, February 2024, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Dan Guo, Xiaobai Li, Kun Li, Haoyu Chen, Guoying Zhao, and Meng Wang. 2026. MAC'26: The 3rd Micro-Action Analysis Grand Challenge. In . ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recently, there has been a lot of interest in empowering intelligent machines with the capabilities for understanding human behaviors [28, 33]. Micro-action behavior analysis has become an important research topic due to its potential applications in human-to-human communication and human emotion state analysis. Micro-action offers insights into the feelings and intentions of individuals and is important for human-oriented applications, such as emotion recognition and psychological assessment. It can also benefit a wide range of human-centered applications, *e.g.*, police interrogation, in-vehicle scenarios, clinical diagnosis , depression analysis, and business negotiation.

In previous works, efforts have been made mainly on facial expressions [11, 16, 26, 39, 41], speech [12, 17, 23, 42], or expressive body gestures [1, 3] to interpret classical expressive emotions. As the computer vision and multimedia communities delve deeper into human emotion state analysis, facial micro-expression and body micro-gesture (mainly referring to upper-limb micro-movements) have received more attention in recent years [3, 6, 13, 25, 37], they all emphasize and explore the link between human micro-behavior and human emotions. Despite this, automatic human behavior analysis based on **human whole-body micro-actions (compared to micro-gestures, micro-actions include the micro-movements of head, body, gestures, and lower-limb)** remains largely unexplored due to the lack of suitable datasets; and the identification, differentiation, and understanding of micro-actions pose challenges, because they are subtle compared to normal action. Different from existing mainstream challenges (such as AVEC [31], EmotiW [10], MuSe [4] and FME [7]) for human-centered emotion analysis, we mainly focus on a fundamental study, the whole-body micro-action

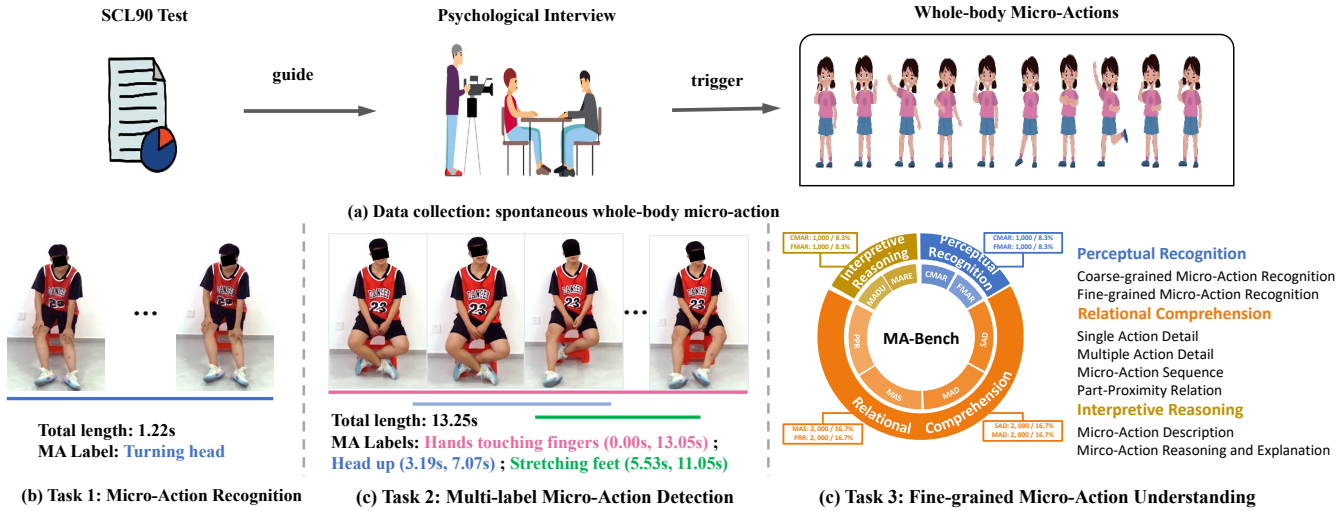


Figure 1: Overview of the 3rd Micro Action analysis grand Challenge (MAC): (a) data collection through professional psychological interviews, (b) Task 1: Micro-Action Recognition (MAR): single action occurrence; (c) Task 2: Multi-label Micro-Action Detection (MMAD): multi-action co-occurrence; and (d) Task 3: Fine-grained Micro-Action Understanding (FMAU). MAC 2026 focuses on the fine-grained micro-action understanding

based human behavior analysis, and provide a common platform and benchmark test sets for performance evaluation.

In MAC 2026, we present the 3rd challenge for whole-body micro-action analysis, as shown in Figure 1. We have fully annotated the MA-52 and MMA-52 datasets over the past two years, and both contain 52 micro-action categories along with seven body parts, covering a wide range of realistic micro-actions from 205 subjects, spanning children, youth, adults, and elderly participants. Building on these datasets, MAC 2026 consists of three tracks: Micro-action Recognition (MAR), Multi-label Micro-action Detection (MMAD), and Fine-grained Micro-Action Understanding (FMAU). The third track, Fine-grained Micro-Action Understanding, is newly introduced in MAC 2026 to push models toward more detailed, context-aware interpretation of subtle actions.

We present analyses of our novel datasets as well as evaluations of baseline approaches for all three tracks in MAC 2026. All collected datasets and baseline implementations will be publicly available on the MAC 2026 website[1]. Participants will submit their results on our MAC 2026 website, and the top 3 teams in each track are required to submit a paper in ACM MM format describing their results and the methods that produced them. These papers will undergo peer review by the technical program committee. We have carefully outlined the key schedule (data & code availability, evaluation set release, paper submission deadlines, *etc.*) on our MAC 2026 project website, which we are preparing to launch. **Sec. 6.2** also presents our planned Challenge schedule, which is consistent with the project website.

In addition, considering the complexity and continuity of the challenge, we plan to organize a series of challenges over at least the next 3-5 years. The next effort is to extend the existing datasets by adding annotations of human emotional states, linking human micro-action with emotional expression for a deeper understanding of human mental states. The audio descriptions can also be released as cues to link micro-actions with emotions. We expect that a series of challenges will bring together researchers from all over the world to focus on this new research direction and find more valuable challenges to advance technologies in the deep psychological assessment and human emotion state analysis communities.

2 RELATED WORK

This section is a summary of the current state-of-the-art methods in automatic micro-action analysis with a focus on (i) micro-action behavior analysis and (ii) multi-label micro-action detection.

2.1 Micro-action Behavior Analysis

Currently, the research on human micro-behavior mainly focuses on micro-expressions [6, 7, 13, 37] and micro-gesture (mainly referring to upper-limb micro-movements) [2, 3, 18, 25]. The famous Facial Micro-Expression Grand Challenge (FME) [7] has continued to attract a large number of multimedia researchers in recent years. As for micro-gesture, Liu *et al.* [25] proposed a micro-gesture dataset named iMiGUE, which contained 32 micro-gesture labels and 5 body-grained labels. Chen *et al.* [3] proposed SMG dataset consisting of 3,692 spontaneous gesture clips was built through a story-telling game under the setting of positive and negative emotions. On the basis of iMiGUE and SMG, the MiGA competition organized at IJCAI 2023[2] and IJCAI 2024[3] takes further discussion for finer levels of human movement like micro-actions, as well as to explore the links between gestures and hidden emotions.

Macro-bodily behaviours [1, 29, 30] in social interaction have also been explored. The BBSI (bodily behaviors in social interaction)

[1]Competition pages: https://sites.google.com/view/micro-action

[2]1st MiGA competition: https://cv-ac.github.io/MiGA2023/

[3]2nd MiGA competition: https://cv-ac.github.io/MiGA2/

dataset [1] adopts a naturalistic multi-view group conversation way to study the influence of body language in the context of social phenomena such as leadership, rapport, or liking. BBSI discusses human social interaction with only 15 body macro movements such as "gesture", "adjusting clothing", "scratching". BBSI was further applied to the MultiMediate Grand Challenge [30] at ACM Multimedia 2023 to explore two key human social behaviour analysis tasks: engagement estimation and bodily behavior recognition in social interactions. These works reveal that **micro-action recognition and detection** is more challenging in terms of subtle detail capture.

In brief, previous works explored the micro-gestures of the upper limbs and head [28, 44] or coarse-grained body movements (with merely 15 categories) in social interactions [1, 30]. In contrast, in this challenge, unlike the aforementioned works and competitions, we collect and analyze the spontaneous whole-body micro-actions that include the head, hands, and upper and lower limbs. In summary, the advantages of the dataset we collected are as follows: 1) multiple types of micro-actions, *e.g.*, 52 categories *vs.* 15 of the BBSI dataset; 2) large-scale data, *e.g.*, 17,974 samples *vs.* 3,692 of the iMiGUE dataset; 3) diverse and large-scale participants, *e.g.*, 205 persons from children, youth, adult and elder *vs.* 32 youth athletes of the iMiGUE dataset, and 4) micro-actions are composed of body movements throughout the whole body. This facilitates the acquisition of rich leg and foot movements, such as "crossing legs" and "stretching feet". Micro-actions in the lower body contain rich information related to psychological and mental states, with the same significance comparable to those of facial expressions and gestures.

2.2 Multi-label Micro-Action Detection

There is almost no publicly available dataset specifically designed for academic use in multi-label micro-action detection. Considering the frequent co-occurrence of human micro-actions in real life, *i.e.*, the same micro-action may be repeated over time and different micro-actions may occur at the same time, multi-label micro-action detection (MMAD) is necessary for a deeper understanding of human bodily behavior. The MMAD task can be applied to fields such as video surveillance and behavioral analysis, which is of great significance. Therefore, we propose this challenge task. Herein, we review the approximate task and technology: multi-label normal action detection (localization). Charades[34] and MultiTHUMOS[43] are two commonly used multi-label action detection datasets, derived from indoor actions and sports videos on YouTube, respectively. In addition, two studies have contributed valuable insights to the field of multi-label action detection. Tirupattur *et al.* [36] introduced an attention-based Multi-label Action Dependency layer (MLAD) into their model, which demonstrated significant improvements in co-occurrence dependencies and temporal dependencies of actions. Tan *et al.* [35] presented an end-to-end action detection model (PointTAD) that leverages learnable query points for precise localization and differentiation of actions in multi-label videos. We apply these classic methods to the MMAD task, and the experimental results are presented in Section 5.2. The unique challenges of MMAD lie in the small movement amplitudes and brief durations of micro-actions in the case of multi-action concurrency, which calls for further research and innovation in the development of effective models tailored to address these intricacies.

Table 1: Definition of micro-actions in different parts of the whole body in datasets MA-52 and MMA-52. Micro-actions occurring in the lower-body are in the grey cells, and the rest are in the upper-body.

Part	Micro-action
Body (A)	Shaking body (A1); Turning around (A2); Sitting straightly (A3); Shrugging (A4); Rising up (A5);
Head (B)	Nodding (B1); Shaking head (B2); Turning head (B3); Tilting head (B4); Bowing head (B5); Head up (B6);
Upper limb (C)	Illustrative gestures (C1); Other finger movements (C2); Hands touching fingers (C3); Stretching arms (C4); Waving (C5); Scratching arms (C6); Spreading hands (C7); Playing objects (C8); Putting hands together (C9); Retracting arms (C10); Pointing oneself (C11); Clenching fist (C12); Rubbing hands (C13);
Lower limb (D)	Tiptoe (D1); Retracting feet (D2); Shaking legs (D3); Stretching feet (D4); Closing legs (D5); Spread legs (D6); Curling legs (D7); Crossing legs (D8);
Body -hand (E)	Scratching or touching chest (E1); Scratching or touching neck (E2); Scratching or touching back (E3); Arms akimbo (E4); Crossing arms (E5); Scratching or touching shoulder (E6);
Head -hand (F)	Touching nose (F1); Scratching or touching face (F2); Playing or tidying hair (F3); Scratching or touching hindbrain (F4); Pushing glasses (F5); Rubbing eyes (F6); Scratching or touching forehead (F7); Touching ears (F8); Covering mouth (F9); Covering face (F10);
Leg -hand (G)	Touching legs (G1); Patting legs (G2); Scratching legs (G3); Scratching feet (G4);

3 CHALLENGE DATASET

3.1 Dataset Construction

To support the Micro-Actions Analysis Challenge (MAC 2026), we have collected three spontaneous whole-body micro-action video datasets consisting of 52 action categories through psychological interviews [14]. Figure 1 illustrates the data collection process and provides sample annotations from the two new datasets, which correspond to the two tasks in MAC 2026.

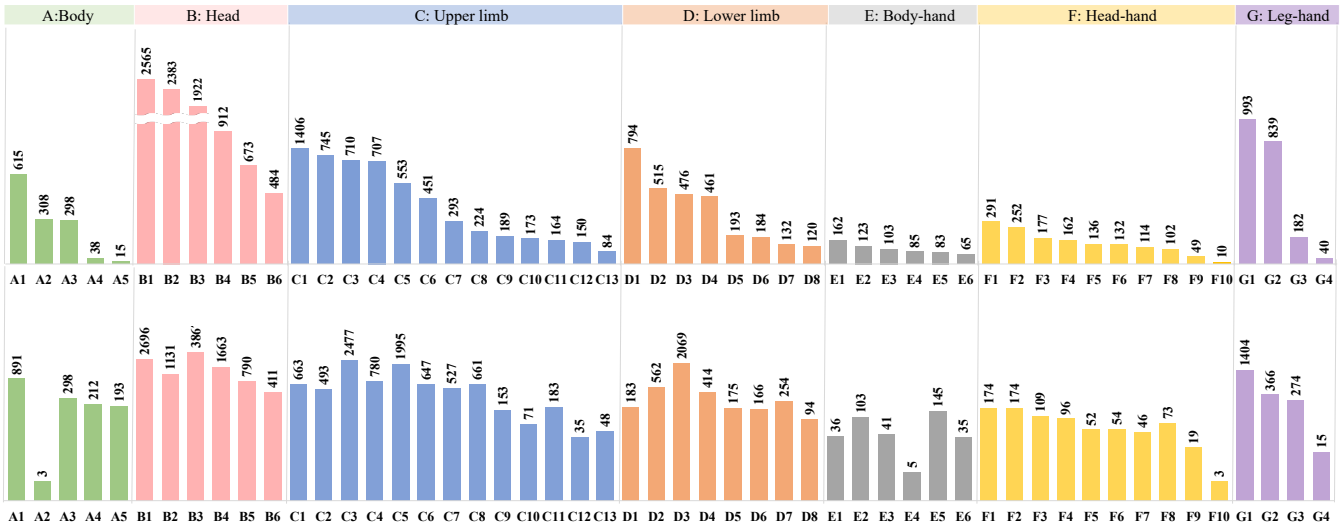
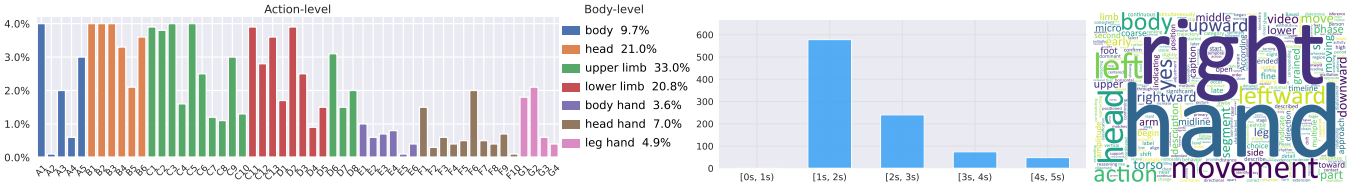


Figure 2: The overall distribution of the micro-action categories of 7 body parts in MAC 2026 datasets. (1) Above: MA-52 dataset. (2) Below: MMA-52 dataset (Log representation of histogram length). Detailed definition of micro-actions are introduced in Table 1.



(a) The distribution of micro-action categories. **(b) The duration of videos** **(c) Word Cloud Statistics.**

Figure 3: Data statistics of MA-Bench. Action category definitions are consistent with the Micro-Action-52 dataset [14].

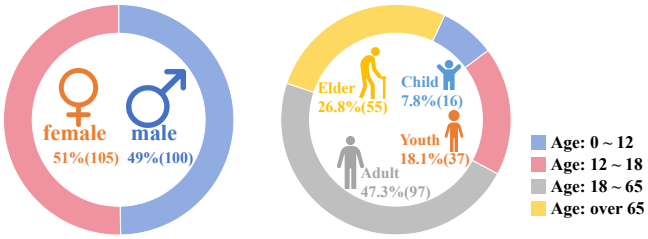
Data collection. We recruited 205 volunteers who were informed about the requirements for data collection and provided their signed informed consent for academic purposes. Professional face-to-face psychological interviews were conducted under the guidance of the Symptom Checklist 90 (SCL90) test, which is a comprehensive 90-item psychological assessment questionnaire [9]. The interviews were recorded using a high-resolution 1920×1080 camera to capture authentic and spontaneous micro-behaviors. It is important to note that the volunteers were not instructed to adopt specific gestures or postures but were encouraged to express their true spirit or state-of-mind naturally. To ensure participants' comfort and relaxation, they remained seated during the interview. The collected micro-actions are defined in Table 1 and categorized using a two-level labeling system: body-grained categories (body, head, upper limb, lower limb, head-hand, body-hand, and leg-hand) and action-grained categories (52 different micro-action categories). As discussed in Section 2, previous works mainly focused on collecting upper-limb movements, while our dataset includes several lower-limb movements such as "shaking legs", "crossing legs", "tiptoe", and "scratching feet".

Annotations of two datasets. We have created three datasets based on the collected interview videos in this challenge: the Micro-Action-52 dataset (MA-52) and the Multi-Label Micro-Action-52 dataset (MMA-52), designed for task MAR and task MMAD, and MA-Bench designed for fine-grained micro-actin understanding respectively.

The **MA-52** dataset consists of trimmed video samples with single micro-action (MA) annotations, as shown in Figure 1(b). The duration of these samples ranges from 1 second to 7 seconds, with an average duration of 1.9 seconds and a total span of 12.29 hours. During the annotation of the MA-52 dataset, our focus was on annotating the beginning and ending segments of a single micro-action occurrence, to avoid annotating multiple micro-action behaviors within the same time period. This approach ensures clarity and accuracy of the annotations, resulting in a consistent and reliable dataset.

The **MMA-52** dataset includes video samples of multiple micro-actions (MMA) with overlapping and intersecting occurrences, as shown in Figure 1 (c). The duration of these samples ranges from 5 seconds to 15 seconds, with an average duration of 10.61 seconds and a total duration of 30.44 hours.

The **MA-Bench** dataset consists of 1,000 carefully curated micro-action videos covering all 52 categories. The videos have an average duration of 2.12 seconds, with a maximum length of 5.01 seconds. Through the semi-automatic annotation pipeline, we construct 12,000 challenging question–answer (QA) pairs organized



(a) Gender Distribution **(b) Age Distribution**

Figure 4: Data distribution over the gender and age of respondents on the MA-52 dataset.

Task	Partition	Samples	Duration
MAR	Training	11,250	6.19h
	Validation	5,586	3.05h
	Test	1,138	0.71h
MMAD	Training	4,534	12.91h
	Validation	1,475	4.34h
	Test	492	1.42h

Table 2: Statistical information for the MA-52 and MMA-52 datasets in MAC.

in a three-tier hierarchical architecture encompassing eight fine-grained micro-action understanding tasks. The dataset statistics of MABench is illustrated in Fig. 3.

We believe that these datasets can serve as new benchmarks in micro-action emotion analysis, especially within the Chinese research community.

3.2 Dataset Split

MAC 2026 focuses on three tasks as shown in Figure 1. The tasks are detailed in Section 4. This part introduces and explains the data set labeling information as shown in Figures 1(b) and (c) and the partitioning of the data set.

3.2.1 MA-52 for Micro-Action Recognition.

Training data. The micro-action recognition (MAR) task dataset is derived from the MA-52 [14] dataset, which contains a total of 22,422 micro-action video samples. The training, validation, and test sets are distributed in a 2:1:1 ratio, resulting in 11,250, 5,586, and 5,586 samples, respectively. We also provide a text file containing annotation information with the following properties:

- `id`: Unique ID for each video.
- `label`: Presented as a number. Each number represents an action-level micro-action category and has a corresponding body-part label. The affiliation between body part labels and micro-action labels can be found in Table. 1.

Evaluation data. Taking into account factors such as competition time, computational resources, and server pressure, we randomly select 1/5 of the test data [14] from each micro-action category as the test set for the MAR task, resulting in a sample size of 1,138. Table **??** presents the dataset partitioning for the micro-action recognition (MAR) task, with sample sizes of 11,250, 5,586, and 1,138 for the training, validation, and test sets, respectively.

3.2.2 MMA-52 for Multi-label Micro-Action Detection.

Training data. The training and validation sets of the MMA-52 dataset consist of 6,515 and 1,974 samples for the multi-label micro-action detection (MMAD) task, respectively. We provide a JSON file in which each data sample is stored as a dictionary. The properties in the JSON file are as follows.

- `id`: Unique ID for each video.
- `subset`: Indicates whether the data belongs to the training, validation, or test.
- `duration`: The total duration of the entire video.

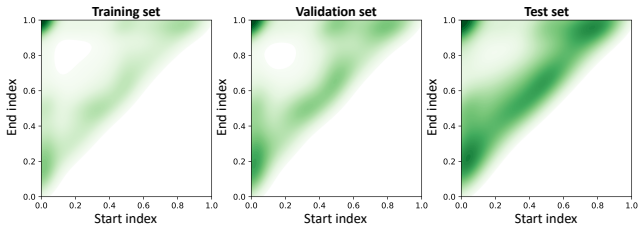


Figure 5: The distribution of micro-action instances of the MMA-52 dataset. The start and end indexes indicate the normalized starting and end times of micro-action instances. The deeper the color, the greater the density of the distributions.

- `actions`: It contains a list of multiple items, with each item consisting of three numbers. The first number represents the category of a micro-action, while the second and third numbers represent the start and end times of the action, respectively.

Evaluation data. Taking into account factors such as competition time, server computing power, and workload, we collected 492 data entries as the competition's test set. Despite the relatively small dataset size, we carefully considered factors such as the proportion of each action category to the total and label length and sought to balance the distribution of each category as much as possible. The test set segmentation file provides only video IDs.

3.2.3 MA-Bench for Fine-grained Micro-Action Understanding. Training data. To facilitate fine-grained micro-action understanding, we introduce **MA-Bench-Train**, a micro-action training corpus designed for adapting multimodal large language models (MLLMs) to subtle whole-body motions. MA-Bench-Train contains **20,510** short micro-action clips collected from **166** participants under a cross-subject setting (no participant overlap with the evaluation benchmark). Each clip is paired with a **structured micro-action caption**, describing the observed motion with explicit spatio-temporal details. Concretely, each caption is organized into five fields: (1) *Overall Overview*, (2) *Timeline Description*, (3) *Key Part Details*, (4) *Judgment Basis*, and (5) *Conclusive Summary*. These structured annotations provide fine-grained supervision for training models to perceive subtle motions and to articulate motion evidence.

Evaluation data. We further introduce **MA-Bench**, a benchmark for fine-grained micro-action understanding. MA-Bench comprises **1,000** curated micro-action videos (covering all 52 categories) with an average duration of **2.12**s (maximum **5.01**s), and a total of **12,000** question–answer pairs. The benchmark follows a **three-tier evaluation architecture** with **eight** tasks, ranging from perceptual recognition to relational comprehension and interpretive reasoning. For evaluation, we provide the video clips and questions; participants are required to predict the answers in the required formats (close-ended / open-ended) on the hidden test set.

3.3 Dataset Analysis

Figure 2 shows the distribution of the micro-actions in datasets MA-52 and MMA-52. To collect spontaneous micro-actions, we aimed to minimize interference with the respondents' bodily action expressions and capture as many varied and rich micro-actions as possible. In order to ensure the reliability and representativeness of the data, conscious efforts were made to achieve a balance in gender and age demographics, as depicted in Figure 4. The proportion of males to females among the respondents was close to 1:1, and there was a relatively even age distribution. However, some micro-actions such as "nodding" and "shaking head" were notably more common than others like "rubbing eyes" or "touching ears". This resulted in a long-tailed distribution where the frequency of sample occurrences varied significantly across different micro-action categories. It is challenging to avoid such data imbalance in wild data collection environments, and participants need to be aware of the impact of data distribution.

4 CHALLENGE DESCRIPTION

The aim of this challenge is to develop and benchmark models that are capable of human micro-action recognition (MAR), multi-label micro-action detection (MMAD), fine-grained micro-action understanding (FMAU). The MA-52 dataset is used for the MAR task, the MMA-52 dataset is used for the MMAD task, and MA-Bench is used for the FMAU task. These three datasets have the distinct advantage of the high-quality annotation of human whole-body micro-actions. We then describe task definitions for both challenge tasks.

4.1 Track 1: Micro-Action Recognition

Micro-Action Recognition (MAR) aims to recognize and distinguish subtle body actions that typically occur in a brief instant. The MAR task is similar to conventional action recognition, as it involves using video instances as input and requires precise and efficient algorithms. However, it is uniquely complex due to the presence of low-amplitude fluctuations in gestures and postures.

Micro-Action Annotations. We review the data annotation process here. The annotations for micro-action recognition have several specificities. In this task, we focus on single micro-action recognition. For example, a person consistently maintains the action of "curling legs". Each action sample is labeled with both a body-grained label and an action-grained label. The former is based on the region where the action occurs, including "body", "head", "upper limb", "lower limb", "body-hand", "head-hand" and "leg-hand". The latter labels are the sub-action categories under the body-grained labels, totaling 52 micro-action categories.

Task Definition. We consider a human micro-action video defined as $V = \{I_1, I_2, \ldots, I_T\}$, where T is the length of the video. Our goal is to classify the micro-action category contained in the video by selecting it from a set of pre-defined micro-action labels $\{1, \ldots, C\}$. We annotate the affiliation relationship between body parts $y_b \in [1, B]$ and micro-actions $y_c \in [1, C]$, where B and C represent the total number of body parts and micro-action categories, respectively.

4.2 Track 2: Multi-label Micro-Action Detection

Considering the co-occurrence of human micro-actions, *i.e.*, the same micro-action may be repeated in time and different micro-actions may occur at the same time, multi-label micro-action detection (MMAD) is necessary for a deeper understanding of human bodily behavior. **Multi-label Micro-Action Detection (MMAD)** refers to the task of identifying and localizing all micro-actions in a given uncut and densely annotated video, determining their corresponding start and end times, as well as their categories. This task takes an entire video as input and requires a model capable of accurately capturing both long-term and short-term action relationships to detect and locate multiple micro-actions. Designing a model for MMAD is more challenging due to the brief duration and small magnitude of micro-actions.

Multi-Label Micro-Action Annotations. The annotation process for multi-label micro-action is similar to that of micro-action recognition but with a focus on the time when the action undergoes a change and disregarding the duration of the action. For instance, for the action "stretching feet" we annotate only the duration during which the human foot transitions from a retracted state to an extended state, without annotating the duration during which the foot remains continuously extended. When referring to the 52 action categories in Micro-Action Recognition, we extend these action categories to the Multi-label Micro-Action Detection Task.

Task Definition. For a video sequence of length T, each time step t includes a ground-truth action label $y_{t,c} \in \{0, 1\}$, where $c \in \{1, \ldots, C\}$ denotes an action category. For each time step, the action detection model is tasked with predicting the category probability $\tilde{y}_{t,c} \in [0, 1]$.

4.3 Track 3: Fine-grained Micro-Action Understanding

Fine-grained Micro-Action Understanding aims to evaluate whether a model can *perceive*, *compare*, and *reason* about subtle whole-body micro-actions beyond label prediction. Following MA-Bench, we formulate this track as a video question answering problem.

Task Definition. Given a micro-action video $V = \{I_1, I_2, \ldots, I_T\}$ and a question q, the model is required to output an answer $\hat{a}$. Depending on the task type, $\hat{a}$ can be: (i) a single-choice option, (ii) a Yes/No decision, or (iii) an open-ended textual response.

Three-tier Task Taxonomy. The evaluation contains eight tasks grouped into three tiers:

Tier-1: Perceptual Recognition. (1) **Coarse-grained Micro-Action Recognition (CMAR)**: recognize the *body-grained* label (7

body parts). (2) **Fine-grained Micro-Action Recognition (FMAR)**: recognize the *action-grained* label (52 micro-action categories).

Tier-2: Relational Comprehension. These tasks evaluate whether models can understand fine-grained relations and dependencies among micro-actions; questions are organized in a Yes/No format. (3) **Single Action Detail (SAD)**: query fine-grained motion attributes (e.g., direction, amplitude) of a single micro-action. (4) **Multiple Action Detail (MAD)**: query details of multiple micro-actions, including co-occurrence and inter-action comparison. (5) **Micro-Action Sequence (MAS)**: query temporal ordering relations between micro-actions. (6) **Part-Proximity Relation (PPR)**: query spatial proximity relations among body parts during micro-actions.

Tier-3: Interpretive Reasoning. (7) **Micro-Action Descriptive Understanding (MADU)**: generate a detailed description of the micro-action video with motion evidence. (8) **Micro-Action Reasoning and Explanation (MARE)**: provide an explanatory reasoning process and (optionally) infer the most plausible coarse- and fine-grained micro-action labels.

4.4 Evaluation Metric

4.4.1 Micro-Action Recognition Task.
Participants are required to submit a ".csv" formatted file, which will be evaluated on the unpublished test set. Considering the diversity of micro-actions involving body parts and micro-action categories, we adopt multiple evaluation metrics for this task. **Acc-Top1**, **Acc-Top5** accuracy of micro-action categories are used to ensure the accurate recognition of micro-actions. To address the long-tail problem of data, we adopt the **F1 score** to evaluate model performance. Herein, $\mathbf{F1}_{macro}$ calculates the unweighted mean for each action category, independent of the sample size for the specific category. $\mathbf{F1}_{micro}$ treats all samples equally and is less influenced by any category with a predominant number of micro-actions. We consider **F1 score** with both "micro" and "macro" calculations. Furthermore, since we provide labels with different annotation granularities (body-grained and action-grained), we measure the prediction results from these two aspects. Additionally, we use $\mathbf{F1}_{mean}$ as the performance metric, and its formula is as follows:

$$\mathrm{F1}_{mean} = \frac{\mathrm{F1}^{b}_{macro}+\mathrm{F1}^{b}_{micro}+\mathrm{F1}^{c}_{macro}+\mathrm{F1}^{c}_{micro}}{4}, \tag{1}$$

where "b" and "c" denote the label granularity of body-part and micro-action categories, respectively.

4.4.2 Multi-label Mirco-Action Detection Task.
Participants are required to submit a ".txt" formatted file, which will be evaluated on the unpublished test set. Similar to typical multi-label normal action detection tasks [5, 35, 36], we use **frame-based mean Average Precision (mAP)** as the evaluation metric for the MMAD task. Additionally, to assess the effectiveness of the model in modeling relationships between different actions, we introduce **action dependency metrics** for model evaluation. Specifically, the calculation formula for **mAP** is as follows:

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i, \tag{2}$$

Method	Acc-Top1 (Action)	Acc-Top5 (Action)	$\mathbf{F1}_{mean}$
VSwinT-T [27]	**58.88**	89.72	64.45
VSwinT-S [27]	**58.88**	89.63	**65.64**
VSwinT-B [27]	58.17	89.89	63.55
VSwinT-L [27]	58.35	**90.25**	63.94

Table 3: Baseline results for Micro-Action Recognition under different variants of Video Swin Transformer [27].

where C represents the total number of categories, AP_i denotes the frame based average precision for the i-th category.

4.4.3 Fine-grained Micro-Action Understanding Task.
Participants are required to submit predictions for all questions in MA-Bench. We adopt different metrics for closed-ended and open-ended questions.

Closed-ended evaluation. For CMAR, FMAR, SAD, MAD, MAS, and PPR, we report **Accuracy** (%), i.e., the percentage of correctly predicted answers. We further compute a **Closed-AVG** score as the mean accuracy over the six closed-ended tasks.

Open-ended evaluation. For MADU and MARE, we adopt an **LLM-as-Judge** protocol to automatically score free-form responses. Specifically, each response is evaluated on a **0–5** scale under three difficulty levels (**L1–L3**). We report per-level scores (MADU-L1/L2/L3 and MARE-L1/L2/L3), and compute:

$$\text{Open-AVG} = \frac{1}{6}\sum_{i=1}^{3}\left(\mathrm{MADU}_{L_i} + \mathrm{MARE}_{L_i}\right). \tag{3}$$

5 BASELINES & RESULTS

5.1 Micro-Action Recognition

Baseline. Video Swin Transformer (VSwinT) [27], which is considered a high-performing and versatile method for video-based action recognition tasks, can serve as a strong comparative baseline for the participants. We use it as the baseline for Micro-Action Recognition (MAR) task. VSwinT [27] introduces an inductive bias of locality in video Transformers, resulting in a better speed-accuracy trade-off compared to previous approaches that compute self-attention globally, even with spatio-temporal factorization. We believe that such a baseline not only provides a high-performance starting point but also inspires participants to develop more innovative solutions.

Results. As shown in Table 3, both the Tiny and Small versions (VSwinT-T & VSwinT-S) of the VSwinT benchmark achieve 58.88% on the micro-action accuracy Acc-Top1 metric, while the Acc-Top5 result is highest on the Large version at 90.25%. The evaluation metrics for the combined evaluation of body-part and micro-action labels are highest on the Small version at 65.64%. It is important to note that in the experiments described above, the body-part prediction results are derived from the affiliation of micro-action prediction results for convenience. We encourage the participants to generate the prediction results of body-parts and micro-actions separately.

Analysis & Main challenges. The primary challenges of the MAR task are as follows: **1) Subtle Changes in Movement.** Micro-actions refer to subtle and rapid muscle movements that occur

across various body parts, such as the head, hands, arms, legs, and feet. Identifying and distinguishing between these subtle variations in movement is a complex task. **2) Approximate inter-class differences.** Categorizing samples from the same body part under distinct actions becomes complicated due to visual similarities. For example, it can be difficult to differentiate between "looking up" and "nodding" in a video clip. The former involves elevating the head, while the latter is characterized by repeated upward and downward head motions. **3) Data imbalance.** As discussed in Section 3.3, the problem of long-tailed distribution of data persists due to the diversity of micro-actions and the difficulty in ignoring individual differences.

5.2 Multi-label Micro-Action Detection

Baselines. To facilitate participants in the MMAD, we have released the baseline code of MS-TCT [5], PointTAD [35], and AdaTAD [24], which are utilized in multi-label normal action localization tasks. MS-TCT designed a multi-scale temporal encoder to capture short- and long-term information. PointTAD proposes an end-to-end query-based action detector that utilizes learnable query points to detect important frames for participating action instances, capturing significant features of actions. AdaTAD employs a parameter-efficient tuning mechanism by integrating temporal-informative Adapter modules into the backbone, aiming to reduce training resource consumption while enhancing model performance.

Results. As shown in Table 5, we report the results of three baselines for the MMAD task, among which AdaTAD-B achieves relatively high performance on Detection-mAP. Compared to the MAR task results presented in Table 3, the detection accuracy of the three baselines is significantly lower on the MMAD task, indicating that the MMAD task poses greater challenges. For the three baselines, MS-TCT models both long-term and short-term action dependencies. PointTAD leverages flexible learnable query points and achieves action localization and distinction through end-to-end training. AdaTAD uses VideoMAE as the backbone and introduces parameter-efficient tuning to enhance the model's performance on the MMAD task while reducing resource consumption. Therefore, we encourage participants to actively explore more reliable and efficient methods to address the challenging MMAD task.

Analysis & Main challenges. The main challenges faced by the MMAD task include: **1) Subtle Changes in Movement**. **2) Dense Action Labeling**: Every frame of the video may contain multiple action boundaries that require precise detection and labeling. **3) Data imbalance as discussed in Section 3.3**. **4) Difficulties in Simultaneously Capturing Long-Term and Short-Term Temporal Relationships**: The distribution of micro-actions in the real world can be challenging, as multiple actions may occur over different terms, and background information about the actions may be limited. Understanding the long-term and short-term dependencies between micro-actions is crucial for the model to make accurate predictions. For example, actions like "stretching feet" and "retracting feet" may exhibit strong temporal relationships, requiring effective time modeling techniques to detect densely labeled actions. **5) A Significant Challenge: Capturing Action Co-occurrence Relationships**: Micro-actions not only have strong relationships over different terms but may also exhibit strong action co-occurrence relationships among different body-grained categories of micro-actions within the same term. For instance, hand movements may often co-occur with facial expressions, so considering the co-occurrence relationships between actions could have a positive impact on prediction performance.

5.3 Fine-grained Micro-Action Understanding

Baselines. We provide a strong MLLM baseline for Track 3 using an open-source vision–language model, *e.g.*, Qwen3-VL-8B. Following MA-Bench, we extract visual inputs by uniformly sampling **8 frames** from each video, corresponding to approximately **3 fps**, to preserve fine-grained temporal cues while maintaining computational efficiency. The sampled frames are then fed into the model, which is prompted to answer questions in the required output formats.

In addition, to enable participants to explore model adaptation, we release MA-Bench-Train as training data. We report a fine-tuned baseline by adapting the backbone on MA-Bench-Train with parameter-efficient tuning methods (e.g., LoRA), which improves both perceptual recognition and interpretive reasoning on subtle micro-actions.

Results. Table 4 summarizes baseline performance. Overall, fine-tuning on MA-Bench-Train brings clear improvements on both Closed-AVG accuracy and Open-AVG reasoning scores, demonstrating the effectiveness of structured micro-action captions for enhancing fine-grained micro-action understanding.

Analysis & Main challenges. Track 3 poses additional challenges beyond Track 1/2: **1) Fine-grained motion evidence grounding.** Models must capture subtle temporal changes and associate them with explicit body parts and motion attributes. **2) Relational reasoning under micro-action co-occurrence.** Understanding temporal ordering (MAS) and spatial proximity relations (PPR) is difficult due to small motion amplitudes and short durations. **3) Interpretable open-ended reasoning.** For MADU/MARE, models need to generate faithful descriptions and reasoning steps, which requires aligning vision perception with structured language generation.

6 CHALLENGE SCHEDULE & ORGANIZERS

6.1 Challenge History

The 1st Micro-Action Analysis Grand Challenge (MAC 2024)[4] [15] is hosted at ACM MM 2024, Melbourne, Australia, and the 2nd Micro-Action Analysis Grand Challenge (MAC 2025)[5]

In 2024, the achievements of this challenge are summarized as follows:

(1) **Large participants.** This challenge attracted over **43 teams** from **37 research institutions worldwide**, including prestigious organizations such as the Max Planck Institute, Tsinghua University, Peking University, Zhejiang University, Shanghai Jiaotong University, the Institute of Automation at the Chinese Academy of Sciences, China Mobile, and

[4]MAC 2024 pages: https://sites.google.com/view/micro-action2024
[5]MAC 2025 pages: https://sites.google.com/view/micro-action2025

Methods	Closed-ended							Open-ended						
	CMAR Acc.	FMAR Acc.	SAD Acc.	MAD Acc.	MAS Acc.	PPR Acc.	AVG	MADU L1	MADU L2	MADU L3	MARE L1	MARE L2	MARE L3	AVG
Qwen3-VL-8B (zero-shot)	32.23	37.11	56.10	53.69	47.94	54.74	46.97	0.50	0.51	0.47	1.18	1.00	0.84	0.75
Qwen3-VL-8B + MA-Bench-Train (one-shot)	47.90	32.60	60.30	59.65	50.00	53.60	50.68	1.50	1.67	1.54	1.98	1.78	1.67	1.69

Table 4: Baseline results for Track 3 Fine-grained Micro-Action Understanding on MA-Bench. Closed-ended tasks are evaluated by Accuracy (%), while open-ended tasks are evaluated by LLM-as-Judge scores (0–5). Open-AVG averages six level scores across MADU and MARE.

Method	Detection-mAP			
	@0.2	@0.5	@0.7	AVG
MS-TCT [5]	5.72	3.91	2.16	3.51
PointTAD [35]	9.46	3.79	1.02	4.51
AdaTAD-B [24]	28.31	19.25	9.69	17.64

Table 5: Baseline results for Multi-label Micro-Action Detection.

Unisound AI Technology Co., Ltd., highlighting the event's global appeal and significance in both academia and industry.

(2) **Broader audience.** This challenge garnered significant attention, with more than **30 scholars** attending in person, actively engaging with the presentations, exchanging insights, and fostering discussions on cutting-edge advancements in the field. Their participation not only underscored the event's impact and relevance but also facilitated meaningful academic and industry collaborations, further amplifying the challenge's influence within the research community.

(3) **Significant improvements.** The champion of Track 1 (Micro-Action Recognition) is from the University of Science and Technology of China, which developed an ensemble-based method [19] built upon InterVideoV2 [40]. By incorporating hybrid loss functions, their approach improved the **top-1 action-level accuracy from 58.88% to 70.83%**. The champion of Track 2 (Micro-Action Detection) was a collaborative team from the University of Science and Technology of China and Unisound AI. They introduced a temporal information aggregation method [45] based on VideoMAE [38], which increased **the mean average precision (mAP) from 17.64% to 27.27%**. Although these methods have achieved notable progress, the inherent challenges of micro-action analysis, such as subtle motion variations and complex temporal dependencies, indicate that there is still significant room for further research and improvement in this field.

If this challenge continues to be approved, we firmly believe it will have an even greater impact on the research community. With strong global participation, active scholarly engagement, and significant technical advancements, it has already proven its value in fostering innovation. Future editions can further drive progress in micro-action analysis, inspiring new methodologies and breakthroughs for both academia and industry.

6.2 Challenge Schedule

When setting up the Challenge schedule, we take into account the participants' schedules, the complexity of the challenge tasks, and the rationalization of the timeline.

- **April 30, 2026.** Data, baseline paper, and code are available.
- **July 1, 2026.** Evaluation datasets release.
- **July 8, 2026.** Results submission deadline.
- **July 15, 2026.** Paper submission deadline.
- **July 30, 2026.** Paper acceptance notification.
- **August 6, 2026.** Deadline for camera-ready papers.

Practically, we will work closely with the ACM Multimedia Conference organizers to make reasonable adjustments to the above schedule, if needed, based on the actual progress of the Challenge. Additional participant notes are provided in the "Evaluation" section of the MAC website[6]. If participants have any questions, they are welcome to contact us via the email address listed on the website.

6.3 Challenge Organizers

MAC 2026 is organized by senior researchers specializing in human affective computing and multimedia computing. The organizers are made up of experienced members from around the world who have organized several challenges in the past few years. All organizers will be responsible for coordinating, publicizing, reviewing, and judging the challenge submissions, as described in the proposal. For example, Prof. Dan Guo has led initiatives such as an ACM TOMM special issue and the 1st micro-action analysis grand challenge at ACM MM 2024 and the 2nd micro-action analysis grand challenge at ACM MM 2025. Prof. Xiaobai Li has successfully organized Challenge RePSS [20], FME [7], MEGC'23 [8] MEGC'24 [32] in the multimedia community. Prof. Guoying Zhao has successfully organized MuSe 2021 [4], FME [7], MRAC'23 [21], MER 2023 [22]. Prof. Meng Wang has chaired lots of related topics and has been highly influential in the field of multimedia computing and affective computing.

- **Dan Guo** (Senior Member, IEEE) (https://scholar.google.com/citations?user=DsEONuMAAAAJ) is a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. Her research interests include visual sentiment analysis, micro-action analysis, and body language understanding. She has published 43 papers including top-tier journals and conferences such as

[6]MAC pages: https://sites.google.com/view/micro-action

TPAMI, TIP, TMM, CVPR, ACM MM, AAAI, IJCAI, ACM SIGIR, ECCV. She regularly serves as a PC Member for top-tier conferences and prestigious journals in multimedia and artificial intelligence, like ACM Multimedia, IJCAI, AAAI, CVPR, and ECCV. **Organization experiences:** She also serves as an SPC Member for IJCAI 2021, and Area Chair for ACM MM 2024. She has organized special issue of "Deep Learning for Robust Human Body Language Understanding" on ACM TOMM, and the special issue of "Trustworthy Cross-Modal Reasoning for Video-Language Understanding" on the CVIU journal.

- **Xiaobai Li** (Senior Member, IEEE) (https://scholar.google.com/citations?user=JTFfexYAAAAJ) is a ZJU100 professor at the College of Computer Science and Technology, Zhejiang University, China. She also works as an adjunct professor at the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland. She received her Ph.D. from CMVS in 2017 and was funded by the Academy of Finland postdoc position in 2019. She worked as a tenure track assistant professor at CMVS from 2020 to 2022. Her research focuses on analyzing subtle information from facial videos, including micro-expression analysis, remote physiological signals measurement, and related applications. **Organization experiences:** Dr. Xiaobai Li has organized multiple competitions and other academic activities, *e.g.*, the 1st RePSS with CVPR 2020, the 2nd RePSS with ICCV 2021, the 3rd RePSS with IJCAI 2024, the 3rd facial micro-expressions grand challenge (MEGC) with FG 2020, the 4th to the 7th MEGC with ACM MM 2021 to ACM MM 2024.
- **Kun Li** (Member, IEEE) (https://scholar.google.com/citations?user=UQ_bInoAAAAJ) is a Postdoctoral Fellow at the United Arab Emirates University, United Arab Emirates. He received his Ph.D. degree from Hefei University of Technology, China, in 2023. He regularly serves as a PC Member for top-tier conferences and prestigious journals in multimedia and artificial intelligence, like ICLR, CVPR, ECCV, ACM Multimedia, and AAAI. He won 9 times championships in top-tier conference challenges, such as Micro-Gesture Recognition in MiGA at IJCAI 2023, 2024 and 2025. **Organization experiences:** He serves as Area Chair for IJCNN 2025, and as a Program Committee member for CVPR, ECCV, ICCV, ICLR, and ACM Multimedia. He also organized the 1st Micro-Action Grand Challenge at ACM Multimedia 2024 as the lead Data Chair, the 2nd Micro-Action Grand Challenge at ACM Multimedia 2025 as the organizer, and MultiMediate 2025 as the Program Committee.
- **Haoyu Chen** (Member, IEEE) (https://scholar.google.com/citations?user=QgbraMIAAAAJ) is a Tenure-track Assistant Professor at CMVS, University of Oulu, Finland. He received his Ph.D. from CMVS, University of Oulu in 2022, and was a Postdoctoral Researcher with University of Oulu, CMVS. His current research interests include Emotion AI and human behaviour understanding. His publications are including top-tier journals and conferences, such as NeurIPS, AAAI, CVPR, ICCV, TIP, IJCV, etc. He won the 'The second prize of IEEE Finland Jt. Chapter SP/CAS Best Paper Award' in 2022. He won the 2nd Place on Action Recognition Track of ECCV 2020 VIPriors Challenges. **Organization experiences:** Assist. Prof. Haoyu Chen has organized the 1st MiGA workshop & challenge on IJCAI 2023, serving as the leading data chair. He also organized the 2nd and 3rd MiGA workshop & challenge on IJCAI 2024 and IJCAI 2025, and 1st to 2nd MAC on ACM MM 2024 and ACM MM 2025.
- **Guoying Zhao** (IEEE, IAPR, AAIA Fellow, Member of Academia Europaea) (https://scholar.google.com/citations?user=hzywrFMAAAAJ) is currently an Academy Professor with the Academy of Finland and (tenured) full professor with the CMVS, University of Oulu, Finland. She received a Ph.D. degree in computer science from the Chinese Academy of Sciences in 2005 and then joined the University of Oulu as a senior researcher. Her research interests include image and video descriptors, facial expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics. **Organization experiences:** Prof. Zhao is general co-chair of ICIPMC 2022, ICBEA 2020, 2019, panel chair for FG 2023, program co-chair for ACM International Conference on Multimodal Interaction (ICMI 2021), co-publicity chair for FG2018 and Late Breaking Results Co-Chairs of ICMI 2019. She has co-organized 20 international workshops/special sessions with CVPR, ICCV, ECCV, IJCAI, *etc.* regarding topics covering computer vision, machine learning, affective computing, and psychology. For example: "Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop" with ACM Multimedia 2021; "CEFRL: Compact and Efficient Feature Representation and Learning in Computer Vision" with ICCV 2017.
- **Meng Wang** (IEEE, IAPR Fellow) (https://scholar.google.com/citations?user=rHagaaIAAAAJ) is a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. He received his Ph.D. from University of Science and Technology of China in 2008. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He is an Associate Editor of IEEE TPAMI, IEEE TKDE, IEEE TCSVT, and the IEEE TNNLS. He published 600+ papers (G-scholar H-index 98, citation 37000+) in journals and conferences. He was a recipient of the ACM SIGMM Rising Star Award 2014. He received multiple Best Paper Awards, such as the Best Paper Award from ACM MM 2009, ACM MM 2010, MMM 2010, and ICIMCS 2012, the Best Deom Award from ACM MM 2012, and the Best Student Paper Award from ICDM 2014. **Organization experiences:** Prof. Wang has organized several academic conferences, such as the General Co-Chair of ICMR 2021, PCM 2018 and MMM 2013, and the Program Co-Chair of ICIMCS 2013.

All organizers will collaborate with ACM Multimedia Conference Organizers to promote this Grand Challenge and encourage researcher participation.

7 CONCLUSION & FUTURE CHALLENGES

We introduce MAC 2026, the third edition of our grand challenge on human whole-body micro-action recognition and multi-label action detection under well-defined conditions, and present baseline approaches for each track. Building on the progress from previous years, MAC 2026 further emphasizes fine-grained micro-action analysis driven by multimodal large models (MLLMs), aiming to exploit their powerful cross-modal representation and reasoning capabilities. To foster continued development in this direction, we are collecting and annotating new emotion labels and audio descriptions based on the interview videos. With these extended resources, we aim to more deeply investigate the relationship between micro-actions and human emotional states, and to leverage audio information as a key cue connecting them. Looking ahead, we will continue to release more human-centered tasks in future editions of the challenge, such as micro-action-based emotion recognition and joint audio–video understanding for both micro-action recognition and emotion recognition, ultimately advancing the study of micro-actions in relation to intrinsic emotions and personality.

A COMMITMENT

We commit to establishing and maintaining a dedicated website for our Grand Challenge for a minimum of three years. This site will host essential information, including challenge goals, participation guidelines, datasets, and task details. Regular updates will ensure access to the latest information and resources, fostering community engagement and collaboration.

ACKNOWLEDGMENTS

We would like to thank the ethics committee of the Institute of Artificial Intelligence (IAI), Hefei Comprehensive National Science Center, for supervising the collection and usage of the dataset. As well, we would like to thank all the participants interviewed for the dataset. Prior to data collection, we informed each participant of the requirement for data collection and obtained their signed consent for academic research purposes.

REFERENCES

[1] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and Francois Bremond. 2022. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the ACM International Conference on Multimedia*. 70–79.

[2] Ardhendu Behera, Zachary Wharton, Yonghuai Liu, Morteza Ghahremani, Swagat Kumar, and Nik Bessis. 2023. Regional attention network (ran) for head pose and fine-grained gesture recognition. *IEEE Transactions on Affective Computing* 14, 1 (2023), 549–562.

[3] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. 2023. SMG: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision* 131, 6 (2023), 1346–1366.

[4] Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Lukas Stappen, Eva-Maria Meßner, et al. 2023. The muse 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*. 1–10.

[5] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S Ryoo, and François Brémond. 2022. MS-TCT: multi-scale temporal convtransformer for action detection. (2022), 20041–20051.

[6] Adrian K Davison, Jingting Li, Moi Hoon Yap, John See, Wen-Huang Cheng, Xiaobai Li, Xiaopeng Hong, and Su-Jing Wang. 2023. FME'23: 3rd Facial Micro-Expression Workshop. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9736–9738.

[7] Adrian K Davison, Jingting Li, Moi Hoon Yap, John See, Wen-Huang Cheng, Xiaobai Li, Xiaopeng Hong, and Su-Jing Wang. 2023. FME'23: 3rd Facial Micro-Expression Workshop. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9736–9738.

[8] Adrian K Davison, Jingting Li, Moi Hoon Yap, John See, Wen-Huang Cheng, Xiaobai Li, Xiaopeng Hong, and Su-Jing Wang. 2023. MEGC2023: ACM Multimedia 2023 ME Grand Challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9625–9629.

[9] Leonard R Derogatis, RS Lipman, and L Covi. 2004. SCL 90. *GROUP* 1 (2004), 4.

[10] Abhinav Dhall, Monisha Singh, Roland Goecke, Tom Gedeon, Donghuo Zeng, Yanan Wang, and Kazushi Ikeda. 2023. Emotiw 2023: Emotion recognition in the wild challenge. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 746–749.

[11] Muhterem Dindar, Sanna Järvelä, Sara Ahola, Xiaohua Huang, and Guoying Zhao. 2020. Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1390–1400.

[12] Kexin Feng and Theodora Chaspari. 2021. Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing* 14 (2021), 1627–1633.

[13] Weijia Feng, Manlu Xu, Yuanxu Chen, Xiaofeng Wang, Jia Guo, Lei Dai, Nan Wang, Xinyu Zuo, and Xiaoxue Li. 2023. Nonlinear Deep Subspace Network for Micro-expression Recognition. In *Proceedings of the 3rd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis*. 1–8.

[14] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking Micro-action Recognition: Dataset, Methods, and Applications. *IEEE Transactions on Circuits and Systems for Video Technology* 1 (2024), 1–15.

[15] Dan Guo, Xiaobai Li, Kun Li, Haoyu Chen, Jingjing Hu, Guoying Zhao, Yi Yang, and Meng Wang. 2024. MAC 2024: Micro-Action Analysis Grand Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11304–11305.

[16] Saurabh Hinduja, Shaun Canavan, and Lijun Yin. 2020. Recognizing perceived emotions from facial expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*. 236–240.

[17] Shreyah Iyer, Cornelius Glackin, Nigel Cannings, Vito Veneziano, and Yi Sun. 2022. A comparison between convolutional and transformer architectures for speech emotion recognition. In *2022 International Joint Conference on Neural Networks*. 1–8.

[18] Okan Köpüklü, Thomas Ledwon, Yao Rong, Neslihan Kose, and Gerhard Rigoll. 2020. Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 77–84.

[19] Qiankun Li, Xiaolong Huang, Huabao Chen, Feng He, Qiupu Chen, and Zengfu Wang. 2024. Advancing micro-action recognition with multi-auxiliary heads and hybrid loss optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11313–11319.

[20] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. 2020. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 314–315.

[21] Zheng Lian, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2023. MRAC'23: 1st International Workshop on Multimodal and Responsible Affective Computing. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*. ACM, 9713–9714. doi:10.1145/3581783.3610936

[22] 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, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*. ACM, 9610–9614. doi:10.1145/3581783.3612836

[23] Na Liu, Yuan Zong, Baofeng Zhang, Li Liu, Jie Chen, Guoying Zhao, and Junchao Zhu. 2018. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. 5144–5148.

[24] Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. 2024. End-to-end temporal action detection with 1b parameters across 1000 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18591–18601.

[25] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10631–10642.

[26] Yang Liu, Xingming Zhang, Janne Kauttonen, and Guoying Zhao. 2024. Uncertain Facial Expression Recognition via Multi-task Assisted Correction. *IEEE*

Transactions on Multimedia 26 (2024), 2531–2543.

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3202–3211.

[28] Yang Mi and Song Wang. 2019. Recognizing micro actions in videos: learning motion details via segment-level temporal pyramid. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1036–1041.

[29] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, et al. 2024. MultiMediate'24: Multi-Domain Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11377–11382.

[30] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, et al. 2023. MultiMediate'23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9640–9645.

[31] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 3–12.

[32] John See, Jingting Li, Adrian K Davison, Gen Bing Liong, Moi Hoon Yap, Wen-Huang Cheng, Xiaobai Li, Xiaopeng Hong, and Su-Jing Wang. 2024. MEGC2024: ACM Multimedia 2024 Facial Micro-Expression Grand Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11482–11483.

[33] Henglin Shi, Wei Peng, Haoyu Chen, Xin Liu, and Guoying Zhao. 2022. Multiscale 3D-shift graph convolution network for emotion recognition from human actions. *IEEE Intelligent Systems* 37, 4 (2022), 103–110.

[34] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. (2016), 510–526.

[35] Jing Tan, Xiaotong Zhao, Xintian Shi, Bin Kang, and Limin Wang. 2022. Pointtad: Multi-label temporal action detection with learnable query points. *Advances in Neural Information Processing Systems* 35 (2022), 15268–15280.

[36] Praveen Tirupattur, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. Modeling multi-label action dependencies for temporal action localization. (2021), 1460–1470.

[37] Tuomas Varanka, Yante Li, Wei Peng, and Guoying Zhao. 2023. Data Leakage and Evaluation Issues in Micro-Expression Analysis. *IEEE Transactions on Affective Computing* (2023).

[38] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14549–14560.

[39] Shangfei Wang, Zhuangqiang Zheng, Shi Yin, Jiajia Yang, and Qiang Ji. 2019. A novel dynamic model capturing spatial and temporal patterns for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2019), 2082–2095.

[40] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*. 396–416.

[41] Yi Wu, Shangfei Wang, and Yanan Chang. 2023. Patch-Aware Representation Learning for Facial Expression Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6143–6151.

[42] Jiaxin Ye, Yujie Wei, Xin-Cheng Wen, Chenglong Ma, Zhizhong Huang, Kunhong Liu, and Hongming Shan. 2023. Emo-DNA: Emotion Decoupling and Alignment Learning for Cross-Corpus Speech Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5956–5965.

[43] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. 2018. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision* 126 (2018), 375–389.

[44] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. 2016. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2629–2638.

[45] Jun Yu, Mohan Jing, Guopeng Zhao, Keda Lu, Yifan Wang, Feng Zhao, Jiaqing Sun, Qingsong Liu, and Jiaen Liang. 2024. End-to-end Spatio-Temporal Information Aggregation For Micro-Action Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11306–11312.