

A New Multi-Domain Benchmark for Micro-Action Recognition and Detection

Yanbin Hao*, *Member, IEEE*, Pengyu Liu*, Xing Wei, Xun Yang, Dan Guo, *Senior Member, IEEE*, Meng Wang†, *Fellow, IEEE*

Abstract—Micro-actions are short-duration, low-amplitude subtle body movements at the whole-body level that can reveal latent intentions, involuntary reactions, and fine-grained affective changes. Our previous MA-52 benchmark has provided an important foundation for micro-action recognition, but it remains limited in scale, scene diversity, task coverage, and evaluation protocols. To advance micro-action analysis toward more realistic and comprehensive settings, we introduce MMA-82, a large-scale multi-domain extension of MA-52. MMA-82 expands the label space from 52 to 82 fine-grained micro-action categories and covers four distinct domains, including laboratory interviews, street interviews, psychiatric patient interviews, and emotion-rich television videos, resulting in 79,574 annotated instances from 454 subjects. Built upon MMA-82, we establish two core tasks: Micro-Action Recognition and Multi-label Micro-Action Detection. For recognition, we further define in-domain and cross-domain protocols, including few-shot and zero-shot settings, to evaluate model robustness, transferability, and generalization. Extensive experiments show that current methods still struggle with realistic micro-action understanding, especially under domain shift, long-tailed category distributions, and complex temporal localization. Beyond benchmarking, we investigate the relationship between micro-actions and emotion, showing that micro-actions are strongly associated with emotional states and provide complementary cues to facial micro-expressions for improved emotion recognition. These results demonstrate that MMA-82 serves as a comprehensive and challenging benchmark for realistic micro-action analysis and a valuable resource for human-centered AI. MMA-82 is available at <https://lpynow.github.io/MMA-82-AIM/>.

Index Terms—Multi-domain benchmark, micro-action recognition, micro-action detection, human behavior understanding

I. INTRODUCTION

HUMAN action recognition in computer vision has long been dominated by the study of macro-actions, namely voluntary and goal-directed activities characterized by large motion amplitudes and relatively long temporal durations, such as running, playing, or object manipulation. This research line has steered the development of large-scale benchmarks, such as UCF101 [1], Kinetics [2], and Something-Something [3], as well as increasingly powerful recognition models, including SlowFast [4], Video Swin [5] and GPT4Ego [6]. However, macro-actions mainly capture coarse and explicit aspects of human behavior, while subtle yet behaviorally informative

motion patterns, namely **micro-actions**, remain largely under-explored.

Micro-actions refer to short-duration, low-amplitude subtle body movements at the whole-body level that may convey important cues about human intention, emotional state, and spontaneous response. Analogous to facial micro-expressions, which reveal concealed affective states through fleeting facial changes [7], micro-actions reflect fine-grained motion dynamics that may disclose latent intentions, unconscious reactions, and incipient behavioral tendencies. Such signals are potentially valuable for a wide range of applications, including human-computer interaction, security screening, healthcare monitoring, and sports analytics.

Recognizing micro-actions is inherently challenging, which largely explains their scarcity in current research. First, micro-actions are extremely short in duration, often lasting only a few frames, making them difficult to capture reliably and easy to overlook during annotation and recognition. Second, their motion amplitudes are typically very small, so the relevant signals can be easily obscured by background clutter, camera noise, illumination variations, natural body fluctuations, and other irrelevant motions. Third, collecting micro-action data at scale is particularly challenging, since these behaviors are often spontaneous, weakly controlled, and highly context-dependent, making them difficult to elicit, record, segment, and annotate consistently. Nevertheless, studying micro-actions is of fundamental importance: humans can naturally infer intention and internal state from subtle body movements, and enabling machines to acquire such fine-grained perceptual ability is crucial for advancing embodied intelligence and human-machine interaction.

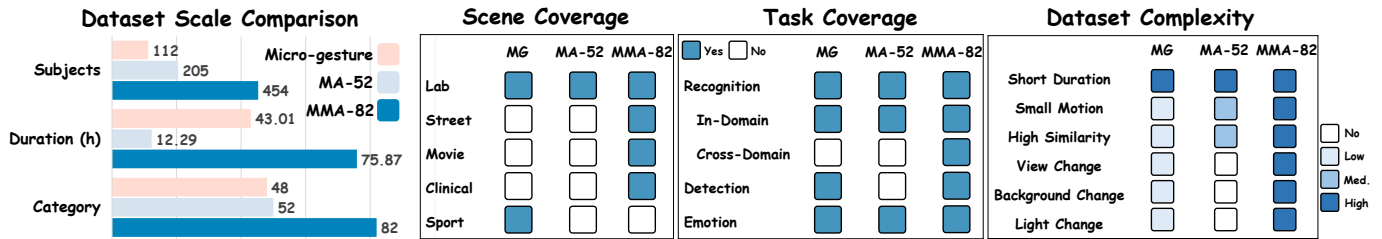
Our previous MA-52 [8] is the pioneering benchmark specifically introduced for micro-action recognition and marked an important early milestone in this emerging research area. It has already attracted growing community attention: to date, we have organized three editions of the *Micro-Action Analysis Grand Challenge* in ACM Multimedia 2024&2025&2026, involving more than 100 participating teams, and benchmark performance has improved substantially over time. Building upon MA-52, a series of follow-up studies, including MMAD [9], MA-Bench [10], Motion Matters [11], and PCAN [12], have further advanced this field by developing stronger models and extending the task scope from recognition to more challenging settings such as detection and QA-oriented tasks. Nevertheless, despite its pioneering role and growing impact, MA-52 still has limited capability to support large-scale and systematic study. Its data are predominantly collected in controlled laboratory-style environments

Yanbin Hao, Pengyu Liu, Xing Wei, Dan Guo and Meng Wang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China (e-mail: haoyanbin@hotmail.com; lpynow@gmail.com, xing.owners@gmail.com, guodan@hfut.edu.cn, eric.mengwang@gmail.com).

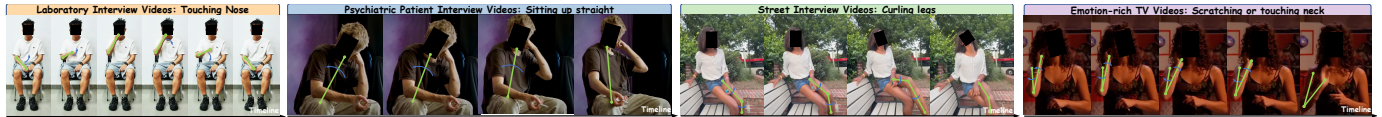
Xun Yang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China (e-mail: xyang21@ustc.edu.cn).

*: Co-first author.

†: Corresponding author.



(a) Comparison between existing micro-gesture/action datasets and our MMA-82. “MG” is the combination of iMiGUE and SMG datasets.



(b) Representative micro-action video samples from four sources in our MMA-82.

Fig. 1. Overview of the proposed MMA-82 benchmark.

with fixed viewpoints, stable illumination, clean backgrounds, and constrained body postures, which are far removed from the complexity and unpredictability of real-world scenarios. Moreover, its scale and annotation granularity remain limited in terms of videos, subjects, scenes, and action categories, while its task and protocol coverage is still relatively narrow, making it difficult to systematically assess model performance under realistic conditions such as cross-domain transfer, low-resource learning, and more challenging recognition settings.

Accordingly, our work is centered around the following four research questions (RQs):

- **RQ1:** *How can we build a realistic micro-action benchmark beyond controlled laboratory environments?*
- **RQ2:** *How can micro-action understanding be evaluated more comprehensively in realistic settings?*
- **RQ3:** *Are micro-actions associated with emotional states in videos?*
- **RQ4:** *Do micro-actions provide complementary affective cues beyond facial micro-expressions?*

These questions jointly target the central bottlenecks of current micro-action research, including realistic data collection, comprehensive evaluation, and the broader behavioral significance of subtle whole-body-level motion patterns.

To answer these questions, we introduce **MMA-82**, a comprehensive multi-domain benchmark for micro-action understanding. MMA-82 is designed to extend and improve MA-52 in a structured and one-to-one manner. *On the data side*, it expands the source domain from a single laboratory interview setting to four distinct domains, including laboratory interview videos, street interview videos, psychiatric patient interview videos, and emotion-rich television videos, thereby providing substantially richer variations in scene context, viewpoints, illumination, background clutter, recording conditions, and subject diversity. *On the benchmark side*, MMA-82 not only enlarges the label space from 52 to 82 categories, but also supports both *Micro-Action Recognition* and *Multi-label Micro-Action Detection*, together with *in-domain*, *cross-domain*, *few-shot*, and *zero-shot* evaluation settings. Beyond benchmark construction, MMA-82 further enables us to investigate the relationship between micro-actions and emotions, and to examine whether subtle body movements provide complementary affective cues beyond facial micro-expressions.

Our main contributions are summarized as follows:

- We present **MMA-82**, a comprehensive multi-domain benchmark for micro-action understanding. Compared with our prior version MA-52, MMA-82 substantially improves realism, scale, category richness, scene diversity, and overall complexity, providing a more rigorous testbed for subtle whole-body-level human action understanding.
- We establish two core benchmark tasks, namely *Micro-Action Recognition* and *Multi-label Micro-Action Detection*, together with *in-domain*, *cross-domain*, *few-shot*, and *zero-shot* evaluation protocols. These settings enable systematic assessment of model robustness, transferability, and generalization under realistic conditions.
- We conduct extensive experiments and analyses on MMA-82. The results reveal the intrinsic difficulty of realistic micro-action understanding, especially under domain shift, long-tailed category distributions, and challenging temporal localization scenarios.
- We investigate the behavioral significance of micro-actions and show that they are strongly associated with emotional states. We further demonstrate that micro-actions provide complementary affective cues to micro-expressions for emotion recognition.

II. RELATED WORK

In this section, we compare MMA-82 with existing human subtle motion datasets, including micro-expression, micro-gesture, and micro-action benchmarks. We also briefly review several representative fine-grained action datasets to position our work within the broader landscape of video action understanding. Fig. 1 and Table I summarize the differences between MMA-82 and representative datasets in terms of scale, scene coverage, task coverage, and dataset complexity.

In video action understanding, several datasets focus on relatively fine-grained human actions, including Something-Anything [3], NTU RGB+D 60 [21], NTU RGB+D 120 [22], EPIC-KITCHENS [23], Ego4D [24] and SAV [25]. Compared with more coarse-grained action datasets such as UCF101 [1] and Kinetics [2], these datasets place greater emphasis on fine-grained human-object interactions, procedural differences, and temporal reasoning. However, they are still mainly designed for explicit, long-duration, and semantically salient actions, rather than brief, subtle, and weakly controlled body motions.

In the broader area of subtle human behavior analysis, many existing datasets focus on facial expressions or upper-body

TABLE I
COMPARATIVE OVERVIEW OF RELEVANT DATASETS. “MA”, “MG” AND “ME” DENOTE MICRO-ACTION/GESTURE/EXPRESSION, RESPECTIVELY.

	Dataset	Venue	Resolution	Labels	Instances	Duration (h)	Frames	Subjects	Emotion Labels	Scene	Body Area	Task
MG	iMiGUE [13]	[CVPR'21]	1280×720	32	18,499	34.87	3.14M	72	2	Real	Up-body	Action Recognition
	SMG [14]	[ICCV'23]	1920×1080	17	3,712	8.14	0.82M	40	2	Lab	Whole body	Action Detection
ME	CAER [15]	[ICCV'19]	-	-	20,484	12.82	1.11M	-	7	TV videos	Whole body	Emotion Recognition
	SAMM [16]	[TAC'16]	2040×1088	-	159	-	-	32	8	Lab	Face	Emotion Recognition
	SMIC [17]	[FG'13]	640×480	-	164	-	-	16	3	Lab	Face	Emotion Recognition
	CASME [18]	[FG'13]	640×480	-	195	-	-	19	7	Lab	Face	Emotion Recognition
	CASME II [19]	[TAC'17]	640×480	-	247	-	-	26	5	Lab	Face	Emotion Recognition
	CAS(ME) ³ [20]	[TPAMI'22]	1280×720	-	1,030	8	8M	216	7	Lab	Face	Emotion Recognition
MA	MA-52 [8]	[TCSVT'24]	1080×1296	52	22,422	12.29	1.33M	205	5	Lab	Whole body	Action Recognition
	MMA-52 [9]	[ICCV'25]	1080×1296	52	19,782	18.67	2.01M	203	-	Lab	Whole body	Action Detection
	MMA-82 (ours)	-	-	82	79,574	75.87	8.05M	454	-	Mix	Whole body	Both
	-MMA-82-Rec	-	Mix	82	39,816	28.94	2.99M	454	8	Mix	Whole body	Action Recognition
	-MMA-82-Det	-	Mix	82	39,758	46.93	5.06M	434	-	Mix	Whole body	Action Detection

gestures. CASME II [26] and SMIC [17] are two representative micro-expression datasets. Micro-expressions are subtle and involuntary facial muscle movements that occur within an extremely short duration and often reveal genuine emotional states. Micro-gesture datasets, such as iMiGUE [13] and SMG [14], focus on subtle upper-body and hand movements that may reflect hidden emotions, suppressed intentions, or stress-related responses. These datasets demonstrate the practical value of subtle body cues for affective and stress-related analysis, but they mainly capture local facial, upper-body, or hand movements rather than subtle whole-body motions.

MA-52 [8] was the first benchmark dedicated to whole-body micro-actions, collected in a laboratory interview setting, with MMA-52 [9] further developed on top of it for the detection task. Building upon MA-52, our MMA-82 advances micro-action research toward a larger-scale, more diverse, and more realistic benchmark. Compared with MA-52, MMA-82 expands the number of action categories from 52 to 82, increases the numbers of videos and subjects, and broadens the data sources from a single laboratory interview scenario to four distinct domains. Moreover, MMA-82 establishes a more comprehensive evaluation framework, including two core benchmark tasks, micro-action recognition and multi-label micro-action detection, together with in-domain and cross-domain protocols, where the latter is further divided into few-shot and zero-shot settings. As a result, MMA-82 provides not only a substantial extension in scale and diversity but also a more comprehensive benchmark for advancing micro-action recognition and detection under realistic conditions.

III. THE MMA-82 BENCHMARK

MMA-82 is constructed through a multi-stage pipeline designed to extend prior micro-action resources toward richer category coverage, higher-quality temporal annotations, and more generalizable samples. As an extension of MA-52, MMA-82 aims to support more realistic and systematic micro-action understanding across diverse scenarios. Specifically, the construction process consists of the following stages: (A) defining and refining the extended label system, (B) collecting large-scale data from multiple sources, and (C) training annotators and establishing annotation guidelines. Based on this pipeline, we further build two benchmark datasets, namely MMA-82-Rec for micro-action recognition and MMA-82-Det for multi-label micro-action detection. The detailed annota-

tions and benchmark settings of MMA-82-Rec and MMA-82-Det will be introduced in the next section. Through its multi-domain data collection, extended 82-category taxonomy, and multi-stage annotation pipeline, MMA-82 directly answers **RQ1** by providing a realistic, large-scale, and richly annotated benchmark beyond controlled laboratory environments.

A. Label Definition and Extension

To build a more comprehensive micro-action taxonomy, we expand the 52 categories of MA-52 to 82 categories. The extended label set places greater emphasis on fine-grained whole-body movement patterns and is organized into seven coarse-grained groups: C1: Body, C2: Head, C3: Upper Limb, C4: Lower Limb, C5: Body-Hand, C6: Head-Hand, and C7: Leg-Hand, as shown in Fig. 2. Compared with the original label space, this refined taxonomy captures temporal dynamics and spatial motion patterns in greater detail, making it more suitable for describing the complexity of micro-actions in both constrained and natural scenes. For example, we additionally add ‘Thumbs up’, ‘Holding fingers’, ‘Folding arms’ for Upper Limb actions, and ‘Placing hands on legs with elbows outward’, ‘Rubbing knee’, ‘Folding hands on legs’ for Leg-Hand. These actions frequently appear in everyday nonverbal cues and carry specific meanings, which complement the micro-action structure of MA-52. This 82-category label system serves as the basis of MMA-82, supporting more precise modeling, evaluation, and analysis of micro-actions.

B. Dataset Collection

Capturing spontaneous micro-actions is inherently challenging, because they are typically subtle in intensity, short in duration, and difficult to annotate accurately. Existing collection strategies can generally be divided into three categories. The first is *prompted performance*, in which participants are instructed to act out predefined micro-actions, as done by iMiGUE [13]. This strategy is useful for collecting rare categories, but the resulting actions are often more standardized and exaggerated than naturally occurring behaviors. The second is *interview-based recording*, where subjects are recorded during face-to-face conversations or psychological interviews, allowing more spontaneous micro-actions to emerge, as done by MA-52 [8]. The third is *in-the-wild annotation*, where micro-actions are identified and annotated from street interviews, television videos, or other uncontrolled videos, as done by SMG [14].

Body (C1)	Head (C2)		Upper limb (C3)						
Swing body	Nodding	Tilting head	Pointing	Other symbolic gestures	Spire gestures	Rolling thumbs	Wrist movements	Folding arms	
Shaking body	Shaking head	Compound head movements	Clapping	Interlace fingers	Other finger interaction movements	Holding hands	Rubbing hands	Folding arms rubbing	
Change body direction	Bowing head	Semantic head movements	OK gestures	Holding fingers	Other hand movements	Clenching fists	Playing objects	Hands behind back	
Tilting forward/backward	Raising head	Swing hair	Thumbs up	Illustrative gestures	Retracting and stretching arms	Spreading hands	Pulling collar	Rubbing spire hands	
Sitting straightly	Turning Head		V gestures	Putting hands together	Rubbing hands with clenched	Shaking	Checking time	Scratching arms	
Shrugging			Waving						
Lower limb (C4)	Body-hand (C5)		Head-hand (C6)			Leg-hand (C7)			
Tiptoe	Closing legs	Scratching or touching shoulder	Scratching or touching forehead	Touching ears	Pushing glasses	Folding hands on legs	Touching legs		
Ankle movements	Spread legs	Scratching or touching neck	Scratching or touching cheek	Cradling chin	Covering mouth	Flattening hands on legs	Patting legs		
Stretching legs	Crossing legs	Scratching or touching back	Scratching or touching chin	Rubbing lip	Covering eyes	Scratching legs	Rubbing knee		
Retracting legs	Curling legs	Arms akimbo	Scratching or touching head	Touching nose	Playing or tidying hair	Placing hands on legs with elbows outward	Scratching feet		
Shaking legs		Scratching or touching chest or abdomen	Scratching or touching philtrum	Rubbing eyes					

Fig. 2. MMA-82 comprises 7 Body-level and 82 Action-level micro-actions, covering the majority of common micro-action categories.

To address the limitations of any single strategy, MMA-82 is constructed through a multi-source and multi-type data collection process. Our goal is to capture micro-actions across different environments, subjects, and recording conditions, and to improve both category diversity and real-world realism. To this end, we combine multiple collection strategies and carefully construct several complementary sub-datasets.

Laboratory interview videos: We reuse the interview videos from MA-52, which were collected through a specialized face-to-face psychological interview protocol based on participants’ SCL-90 test results. These videos provide a valuable foundation for capturing relatively spontaneous micro-actions in a semi-controlled environment. However, some rare micro-action categories are difficult to elicit naturally and are therefore severely underrepresented in the original dataset. To alleviate this long-tailed distribution, we first analyze the category distribution of MA-52 and identify a set of rare micro-actions. We then recruit 24 additional participants and ask them to perform these rare micro-actions according to their own understanding. Although this process is guided, it helps supplement categories that are difficult to capture in sufficient quantity through natural interviews alone. At the same time, allowing participants to interpret the actions in their own way preserves individual variation in performance style and motion characteristics. This collection stage is therefore used to intentionally enrich rare categories and improve category balance in the benchmark.

Psychiatric patient interview videos: This subset is collected from publicly available psychiatric patient interview videos on YouTube. The videos typically consist of topic-guided self-narratives from individuals with different psychiatric disorders, covering life experiences, daily challenges, emotional fluctuations, and disorder-related cognition. Compared with laboratory-style interviews, these videos provide more realistic behavioral expressions under less constrained conditions, and thus offer a valuable resource for studying subtle micro-actions in psychiatric contexts.

Street interview videos: This subset is mainly collected from unscripted street interview videos on YouTube, which are highly unstructured and recorded in open, uncontrolled environments. Interview settings, camera viewpoints, and recording conditions exhibit substantial randomness and variability. The subjects often come from socially marginalized or underrepresented groups, such as homeless individuals and transgender people, and the interviews typically focus on daily survival, social exclusion, personal hardship, and past trauma.

Such narratives are often accompanied by big individual differences, emotional fluctuations, and irregular speaking rhythms, which naturally induce frequent subtle body movements. At the same time, environmental noise, illumination changes, camera shake, and cluttered backgrounds make micro-action detection significantly more challenging. Therefore, this subset is particularly valuable for studying multi-label micro-action detection in realistic scenarios, and serves as an important complement to laboratory and semi-structured settings.

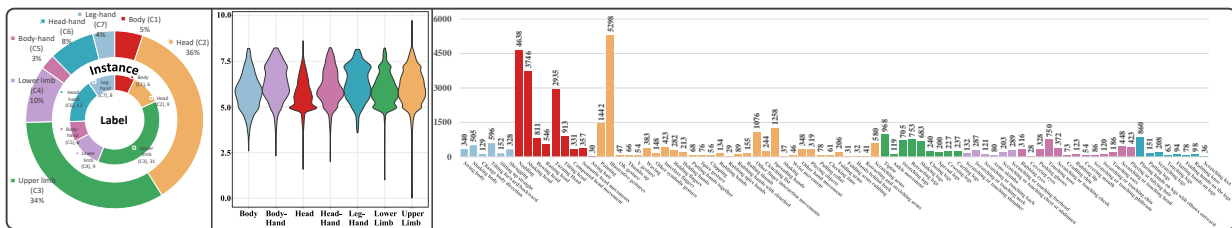
Emotion-rich television videos: This subset is derived from the CAER [15] dataset, one of the most widely used benchmarks for emotion recognition. CAER is mainly collected from television shows, documentaries, and other real-world videos, and contains rich emotional expressions in diverse natural scenes. Based on CAER, we select a subset of videos and further annotate them with fine-grained micro-actions, thereby establishing explicit links between emotional states and subtle body movements. This subset provides a useful resource for investigating how emotional changes are reflected through micro-actions and for studying the relationship between action and emotion in naturalistic settings.

C. Action Segment Annotation

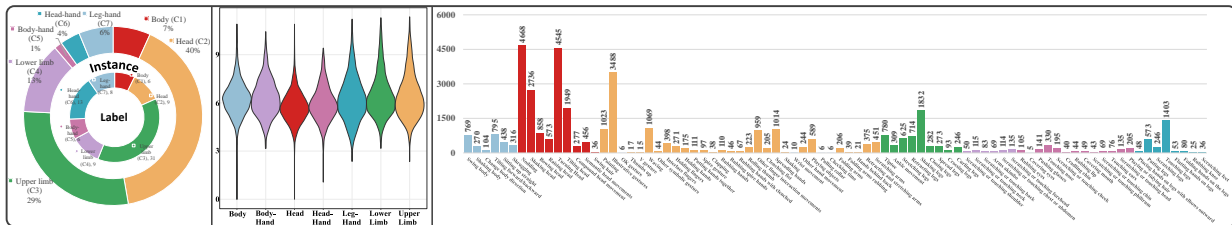
For each video, we use LabelU [27] to annotate every micro-action instance, including its category label and the corresponding start and end times in the temporal sequence. To ensure annotation quality and consistency, all annotators undergo standardized and long-term professional training before participating in the annotation process. Specifically, we provide a detailed annotation manual, which describes the category definitions, classification rules, and temporal boundary criteria for each micro-action class.

To reduce noise from extremely short clips, we impose a minimum duration threshold and require each annotated action segment to last longer than 0.2 seconds. Each video is annotated by three annotators. First, two annotators independently label all action instances in the video, including the action category and the corresponding temporal boundaries. Then, a third senior annotator manually reviews the results of the first two annotators, resolves category inconsistencies, and verifies the temporal annotations. This multi-stage process helps improve both label reliability and temporal consistency.

Because temporal boundary annotation is inherently subjective, different annotators may produce slightly different start and end times for the same action instance. To reconcile these discrepancies, we adopt the annotation merging strategy proposed in EPIC-KITCHENS [28]. Let the temporal annotation



(a) Distribution of the MMA-82-Rec dataset.



(b) Distribution of the MMA-82-Det dataset.

Fig. 3. Comprehensive statistics of the MMA-82 from multiple perspectives. (a) and (b) show the statistical information of MMA-82-Rec and MMA-82-Det across different domains. The left figure of each subplot represents the Body-Level category distribution of action instances, with most instances concentrated in “Head” and “Upper Limb.” The middle figure shows the distribution between Body-Level categories and frames, exhibiting varied temporal distributions; for ease of representation, the y-axis is in \log_2^{frames} . The right figure further provides the detailed instance distribution across all 82 fine-grained micro-action categories, showing a pronounced long-tailed pattern.

for the i -th action instance provided by the j -th annotator be denoted as $A_i^j = [t_{s,i}^{(j)}, t_{e,i}^{(j)}]$, where $t_{s,i}^{(j)}$ and $t_{e,i}^{(j)}$ represent the start and end times of the action, respectively.

To further quantify annotation agreement, we compute an agreement score based on the Intersection-over-Union (IoU) between temporal segments. For the j -th annotator, the inter-annotator agreement score is defined as:

$$s_i^{(j)} = \frac{1}{K_a} \sum_{k=1}^{K_a} \text{IoU}(A_i^{(j)}, A_i^{(k)}), \quad (1)$$

where K_a is the number of annotators ($K_a = 3$). The annotator with the maximum agreement is selected as

$$\hat{j} = \arg \max_j s_i^{(j)}. \quad (2)$$

We then identify the annotation that has the highest IoU overlap with the selected annotation:

$$\hat{k} = \arg \max_k \text{IoU}(A_i^{(\hat{j})}, A_i^{(k)}). \quad (3)$$

Finally, the ground-truth temporal segment A_i is determined as

$$A_i = \begin{cases} \text{union}(A_i^{(\hat{j})}, A_i^{(\hat{k})}), & \text{if } \text{IoU}(A_i^{(\hat{j})}, A_i^{(\hat{k})}) > 0.5, \\ A_i^{(\hat{k})}, & \text{otherwise.} \end{cases} \quad (4)$$

As demonstrated in [28], this strategy aggregates the annotations with the highest agreement by taking their union, thereby preventing the resulting temporal segment from being overly tight and improving the reliability of boundary annotation.

IV. TASKS AND BASELINES

This section answers **RQ2** by establishing recognition and detection tasks under both standard and realistic generalization settings. Specifically, we introduce two benchmark tasks that together constitute the MMA-82 benchmark suite. The first task (Sec. IV-A) is *Micro-Action Recognition* (MAR), where each video clip is assigned a single micro-action label and its corresponding dataset is denoted as **MMA-82-Rec**. The

second task (Sec. IV-B) is *Multi-label Micro-Action Detection* (MAD), which is built upon **MMA-82-Det** dataset and aims to localize and classify multiple micro-action instances in untrimmed videos. Together, these tasks evaluate micro-action understanding at both semantic and temporal levels, encompassing category recognition and temporal localization.

A. Micro-Action Recognition

Task Definition. Given a video V , the MAR task aims to predict the category of the target micro-action. Following our annotation hierarchy, each micro-action can be described at two levels: the *body level* $C_{body} \in \{1, 2, \dots, N_{body}\}$ and the *action level* $C_{act} \in \{1, 2, \dots, N_{act}\}$, where N_{body} and N_{act} denote the numbers of body-level and action-level categories, respectively. This task requires the model to distinguish highly similar subtle motions under diverse recording conditions.

Evaluation Metrics. We adopt evaluation metrics widely used in action recognition to systematically assess model performance at different label levels. Specifically, we use accuracy and the F1 score as primary metrics. In particular, we report the accuracy of Top-1 (Acc@1) and Top-5 (Acc@5) at both the body level and the action level, together with macro-F1 and micro-F1 scores, following previous work on video classification [8], [29]. Since our dataset exhibits class imbalance, we also report mean class accuracy (MCA) following [30], [31]. MCA computes the accuracy for each class independently and then averages the results, thus treating all classes equally regardless of their sample sizes.

MMA-82-Rec Dataset. We select a total of 39,816 video clips from the four different sources of MMA-82, namely laboratory interview videos, psychiatric patient interview videos, street interview videos, and emotion-rich television videos, containing 27,509, 7,011, 3,925, and 1,371 video clips, respectively. Together, these clips constitute the recognition benchmark MMA-82-Rec. Fig. 3(a) illustrates the dataset statistics from three perspectives. The left panel shows the

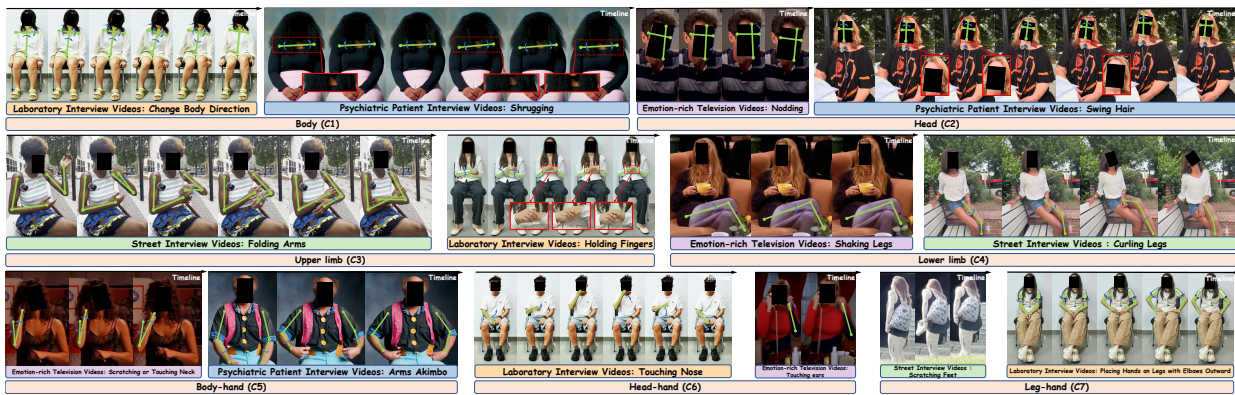


Fig. 4. Example video clips and annotations for the MMA-82-Rec dataset.

TABLE II
STATISTICS OF THE MMA-82-REC DATA SPLITS ACROSS DIFFERENT SOURCES.

Data Sources	Splits	Clips	Duration	Avg. Length	#Subj.
Lab Interview	Train	15,820	11.85h	2.70s	152
	Val	5,636	4.12h	2.63s	52
	Test	6,053	4.52h	2.69s	25
	Total	27,509	20.48h	2.68s	229
Psychiatric Interview	Train	4,203	3.19h	2.73s	8
	Val	1,398	1.13h	2.92s	6
	Test	1,410	1.40h	3.56s	5
	Total	7,011	5.72h	2.94s	19
Street Interview	Train	2,358	1.18h	1.80s	12
	Val	792	0.44h	2.00s	7
	Test	775	0.47h	2.20s	7
	Total	3,925	2.10h	1.92s	26
Emotion videos	Train	795	0.37h	1.70s	52
	Val	247	0.11h	1.58s	79
	Test	329	0.15h	1.68s	49
	Total	1,371	0.64h	1.67s	180
Total		39,816	28.94h	2.62s	454

distribution of body-level action instances, where *head* and *upper limb* account for more than two-thirds of the total. This is consistent with human behavioral patterns, as nonverbal information is often conveyed primarily through head and upper-limb movements. The middle panel presents the frame-length distribution of action instances in each body-level category, revealing subtle differences in temporal duration across categories. The right panel shows the action-level category distribution, which exhibits a pronounced long-tail pattern. Such imbalance increases the difficulty of learning robust representations and may lead to biased model performance. We further provide representative examples of MMA-82-Rec in Figure 4. The dataset covers a wide variety of videos with diverse sources, viewpoints, and backgrounds. In the visualization, green lines highlight the moving joints, while blue arrows indicate motion trends. As can be observed, micro-actions usually occur within very small spatial ranges, further underscoring the difficulty of the task.

For benchmarking, MMA-82-Rec is split into training, validation, and test sets with a ratio of 3:1:1, while preserving a similar action distribution across the three splits. To prevent identity leakage, *each subject is assigned exclusively to a single split*. Detailed dataset statistics and split information are reported in Table II. To comprehensively evaluate recognition performance under both seen and unseen conditions, we further define two evaluation settings: **In-Domain** and **Cross-Domain**.

1) In-Domain

This setting evaluates the model’s basic recognition ability when the training and test samples are drawn from the same domain and thus share similar scene characteristics. As baselines, we first adopt PoseConv3D (PoseC3D) [32], a widely used framework for skeleton-based action recognition. PoseC3D consists of two main stages: 2D skeleton sequences are first converted into a 3D heatmap volume, and then a 3D CNN is used to extract spatio-temporal features for action classification. We further include a pure RGB-based video recognition model, GC-TSM [33], which enhances TSM [34] with a family of efficient element-wise calibrators. More details about the baselines and training protocols are provided in the Supplementary. Table III reports the benchmark results of these two methods under the in-domain setting.

Several observations can be drawn from the results. First, on the overall MMA-82-Rec benchmark, the RGB-based GC-TSM achieves better action-level recognition performance than the skeleton-based PoseC3D, improving the test Top-1 Acc from 56.62 to 60.43, and MCA from 36.77 to 38.98. This indicates that appearance and contextual cues are helpful for recognizing subtle micro-actions. Second, the performance varies substantially across different sub-datasets. The laboratory interview subset is the easiest one, where both methods achieve their best results. In contrast, performance drops markedly on the psychiatric and street interview subsets, and becomes particularly poor on the emotion video subset. These results suggest that domain complexity, viewpoint variation, background clutter, and more spontaneous motion patterns introduce substantial challenges for micro-action recognition.

2) Cross-Domain

To further evaluate model robustness and transferability, we introduce a more challenging Cross-Domain setting, in which training and validation are conducted in one domain while testing is performed in a different domain. In our benchmark, the laboratory interview videos are used for training and validation, whereas the other three domains are separately used as test sets. Under this setting, we further define two evaluation protocols: **Zero-Shot** and **Few-Shot**. We adopt PoseC3D as the baseline model.

In the **Zero-Shot** setting, the model is trained only on the laboratory interview data, without accessing any samples from the target test domains. That is, the PoseC3D model trained on

the laboratory interview videos is directly applied to predict the micro-action category of each video clip in the other three domains. Since the target domains are entirely unseen during training, this setting evaluates the model’s direct cross-domain generalization capability. In the **Few-Shot** setting, the model is allowed to access a small number of labeled samples from the target domain for adaptation. Specifically, we follow a standard K -shot protocol with $K \in \{1, 5, 10\}$, where K labeled clips are randomly sampled from each class in the target domain for adaptation, and the remaining clips are used for evaluation. To reduce the effect of random sampling, each experiment is repeated over 20 independent runs, and the average performance is reported.

Table IV reports the results under the cross-domain setting. In particular, under the zero-shot protocol, we observe a substantial performance drop compared with the in-domain setting, indicating that direct transfer across domains remains highly challenging for micro-action recognition. Although performance improves in the few-shot setting, even with only one labeled sample per class, the gains gradually diminish as K increases from 1 to 5 and 10. This suggests that simply increasing the number of labeled target-domain samples provides only limited improvement for micro-action understanding in uncontrolled scenarios. These observations underline the need for more robust generalization and adaptation strategies for realistic micro-action recognition.

B. Multi-label Micro-Action Detection

Task Definition. Multi-label Micro-Action Detection (MAD) aims to detect and classify all micro-action instances in an untrimmed video. Formally, given an untrimmed video \mathbf{V}_{utr} with a total number of T frames, the ground-truth action annotations are represented as

$$\Phi_g = \left\{ \phi_i = (t_s^i, t_e^i, c_i) \right\}_{i=1}^{N_g}, \quad (5)$$

where t_s^i and t_e^i denote the start and end times of the i -th micro-action instance, $c_i \in \mathcal{C}$ denotes its category label, \mathcal{C} is the set of all micro-action categories, and N_g is the total number of ground-truth instances. The goal of MAD is to predict a set of action proposals

$$\Phi_p = \left\{ \hat{\phi}_i = (\hat{t}_s^i, \hat{t}_e^i, \hat{c}_i, s_i) \right\}_{i=1}^{N_p}, \quad (6)$$

where \hat{t}_s^i and \hat{t}_e^i denote the predicted temporal boundaries, \hat{c}_i denotes the predicted category label, s_i denotes the confidence score, and N_p is the number of predicted action instances.

MMA-82-Det Dataset. We construct the MMA-82-Det dataset by cropping 11,180 videos from the original MMA-82

collection, with video durations ranging from 0.6 seconds to 1.5 minutes and a total of 39,758 action instances. Fig. 3(b) presents the statistical analysis of MMA-82-Det, while Table V summarizes its detailed split information. Similar to MMA-82-Rec, MMA-82-Det also exhibits a noticeably imbalanced distribution over different body-level and action-level categories, with a pronounced long-tail pattern at the fine-grained action level. Such imbalance increases the difficulty of learning robust temporal representations and may lead to biased detection performance toward frequent categories. In addition, the temporal lengths of action instances vary considerably, and multiple micro-actions often co-occur or appear in rapid succession within a single video, making precise temporal localization particularly challenging. On average, each video contains 3.56 micro-action instances, and each instance lasts approximately 3.66 seconds. We also visualize several representative detection examples in Fig. 5, which illustrate the diversity of recording conditions, action transitions, and co-occurring micro-actions in both controlled and real-world scenarios. These properties make MMA-82-Det a challenging benchmark for multi-label micro-action detection, requiring models to jointly address fine-grained category discrimination and accurate temporal localization.

Evaluation Metrics. We adopt Detection mean Average Precision (Detection-mAP) [9], [35] as the evaluation metric. Detection-mAP primarily measures the match between predicted action instances and ground-truth in localization and category recognition. It evaluates the model’s performance by calculating the temporal intersection over union (tIoU) between predicted and ground-truth segments. Specifically, we compute Detection-mAP across multiple tIoU thresholds and report the mAP for tIoU from 0.1 to 0.9 (in steps of 0.1). Considering the two-level micro-actions, we evaluate at both the body-level and action-level and report the average mAP.

Baselines. We use AdaTAD [36] as the baseline for MMA-82-Det. AdaTAD is a state-of-the-art end-to-end framework for temporal action detection that addresses the memory and computational constraints typically associated with scaling up TAD models. Unlike traditional feature-based methods that rely on pre-extracted backbone features, AdaTAD integrates a lightweight adapter directly into the frozen backbone. These adapters allow the framework to capture complex temporal dependencies across long video sequences while maintaining efficient memory usage.

Results. Table VI reports a comparison of our performance using AdaTAD [36] on the MMA-82-Det dataset with Video-

TABLE III
BASELINE RESULTS FOR THE MICRO-ACTION RECOGNITION UNDER THE IN-DOMAIN SETTING.

Sub-Dataset	Method	Action Level								Body Level											
		Top-1 Acc		Top-5 Acc		MCA		Macro F1		Micro F1		Top-1 Acc		Top-5 Acc		MCA		Macro F1		Micro F1	
		Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
MMA-82-Rec (All)	Skeleton	54.43	56.62	80.44	80.45	36.26	36.77	37.35	39.39	54.43	56.62	79.04	81.93	99.18	99.11	69.62	71.71	71.12	74.62	79.04	81.93
	RGB	57.59	60.43	83.01	86.14	37.39	38.98	35.64	39.56	57.59	60.43	79.18	83.17	98.79	98.95	67.24	69.48	67.77	71.98	79.18	83.17
Laboratory Interviews	Skeleton	57.81	64.84	84.28	87.81	39.82	44.88	41.77	47.03	57.81	64.84	81.74	86.73	99.59	99.80	75.19	79.78	76.16	81.87	81.74	86.73
	RGB	62.01	68.15	87.31	92.83	43.00	48.01	39.69	46.48	62.01	68.15	83.61	86.75	99.13	99.45	73.59	75.96	73.85	77.66	83.61	86.75
Psychiatric Interviews	Skeleton	50.64	41.63	75.46	68.79	27.33	15.74	24.93	17.65	50.64	41.63	76.97	77.73	98.71	99.01	57.29	47.46	52.54	51.33	76.97	77.73
	RGB	52.00	46.74	75.89	72.13	24.77	14.10	17.33	13.64	52.00	46.74	74.25	81.99	98.00	99.08	52.64	44.26	45.34	45.36	74.25	81.99
Street Interviews	Skeleton	44.82	40.77	68.18	70.71	19.55	19.49	21.20	20.70	44.82	40.77	69.95	71.35	97.60	98.84	45.44	43.39	49.09	49.52	69.95	71.35
	RGB	42.80	39.87	69.70	73.55	13.79	12.75	12.09	12.56	42.80	39.87	63.26	68.90	97.85	97.29	37.53	37.02	36.25	38.23	63.26	68.90
Emotion Videos	Skeleton	29.55	7.29	60.32	17.93	19.25	3.72	19.01	3.11	29.55	7.29	58.30	35.87	97.57	88.45	38.50	19.26	39.74	17.89	58.30	35.87
	RGB	35.63	25.53	67.61	52.89	21.83	12.20	22.19	11.05	35.63	25.53	57.09	55.93	98.38	93.01	31.04	33.20	29.69	31.01	57.09	55.93

TABLE IV
BASELINE RESULTS FOR THE MICRO-ACTION RECOGNITION UNDER THE CROSS-DOMAIN SETTINGS (ZERO-SHOT AND FEW-SHOT).

Source	Task	Action Level					Body Level				
		Top-1 Acc	Top-5 Acc	MCA	Macro F1	Micro F1	Top-1 Acc	Top-5 Acc	MCA	Macro F1	Micro F1
Psychiatric Interviews	Zero-Shot	27.30	50.28	14.25	7.96	27.30	68.58	96.74	49.40	42.43	68.58
	1-Shot	30.27	58.62	22.82	13.31	30.28	72.48	93.22	55.50	44.47	72.48
	5-Shot	30.12	58.60	22.72	13.29	30.12	72.60	93.28	55.43	44.41	72.60
	10-Shot	30.13	58.59	22.71	13.24	30.12	72.64	93.30	55.42	44.48	72.64
Street Interviews	Zero-Shot	20.65	44.90	10.65	6.91	20.65	53.16	92.65	38.68	35.12	53.16
	1-Shot	21.38	45.64	15.60	12.40	21.38	54.95	84.46	38.92	37.27	54.95
	5-Shot	21.58	46.10	16.17	13.28	21.58	64.91	95.61	40.33	17.00	64.91
	10-Shot	22.52	45.92	17.76	14.23	22.52	65.64	97.22	39.37	38.91	65.64
Emotion Videos	Zero-Shot	14.13	32.98	6.63	4.18	14.13	43.92	90.43	25.18	23.77	43.92
	1-Shot	17.26	39.77	14.00	11.62	17.26	41.43	79.84	27.11	25.13	41.43
	5-Shot	17.48	41.75	15.20	12.10	17.49	42.35	90.80	26.93	24.56	42.35
	10-Shot	17.70	42.53	15.43	12.26	17.70	42.03	92.30	26.95	24.39	42.03

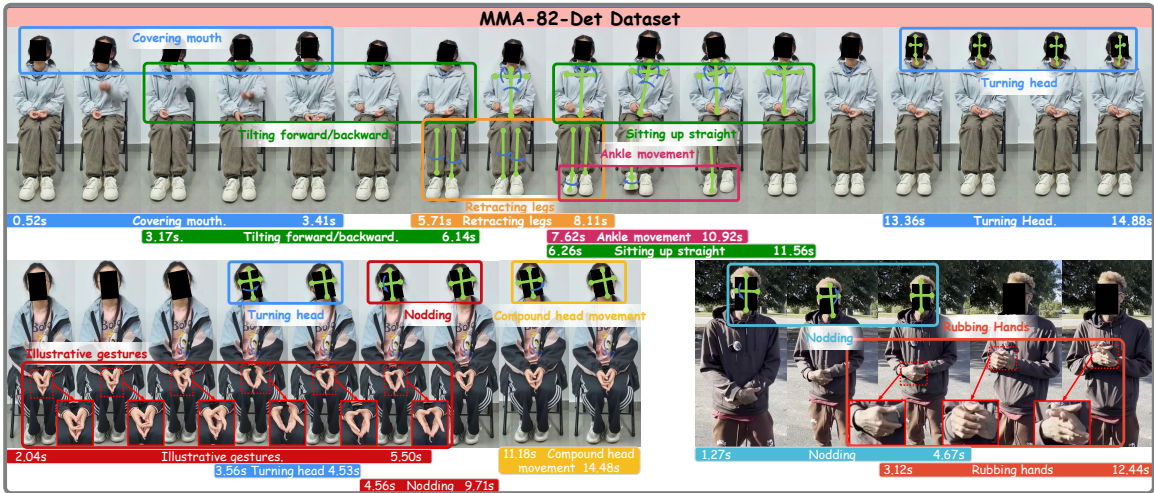


Fig. 5. Example video clips and annotations from the MMA-82-Det dataset.

TABLE V

MMA-82-DET DATASET STATISTICS. “DURATION” DENOTES THE TOTAL VIDEO DURATION, WHILE “AVG. VIDEO”, “AVG. INSTA. DU” AND “AVG. INSTA.” DENOTE THE AVERAGE VIDEO DURATION, AVERAGE INSTANCE DURATION, AND AVERAGE NUMBER OF INSTANCES PER VIDEO, RESPECTIVELY.

Split	Videos	Instances	Duration	Avg. Video	Avg. Insta. Du.	Avg. Insta.
Training	7,839	28,240	33.05h	15.18s	3.62s	3.60
Validation	2,127	7,236	8.92h	15.10s	3.86s	3.40
Testing	1,214	4,282	4.97h	14.73s	3.59s	3.53
All	11,180	39,758	46.93h	15.11s	3.66s	3.56

TABLE VI

THE EXPERIMENTAL RESULTS ON THE MMA-82-DET DATASET.

Backbone	Action-level				Body-level				AVG
	@0.2	@0.5	@0.7	Avg	@0.2	@0.5	@0.7	Avg	
VideoMAE-S	20.88	12.72	5.56	12.09	48.18	28.78	13.91	25.44	18.77
VideoMAE-B	22.62	14.67	6.32	13.59	50.95	30.46	12.23	29.13	21.36
VideoMAE-L	22.74	15.60	7.68	14.98	55.48	33.01	14.06	31.83	23.41
VideoMAE-H	26.53	17.56	7.55	16.08	54.71	33.64	14.17	30.05	23.07

MAE [37] backbones of different scales. As the model scales up from VideoMAE-S to VideoMAE-L, the average mAP (AVG) continues to increase, rising from 18.77 to 23.41. This indicates that larger pre-trained models can provide more discriminative spatio-temporal representations for micro-action detection. Notably, compared to VideoMAE-L, VideoMAE-H shows only a small improvement at the action level and a slight drop at the body level, which may be due to increased optimization difficulty resulting from the larger model size.

V. EMOTION WITH MICRO-ACTIONS

Video-based emotion recognition has been widely studied in affective computing and human-centered AI. Existing methods mainly rely on facial affective cues, such as facial expressions and micro-expressions, together with holistic video context, including body posture, scene context, and temporal dynamics. However, the role of *micro-actions* remains largely unexplored. Compared with micro-expressions, micro-actions capture subtle whole-body-level movements that may reveal spontaneous reactions and fine-grained affective changes, and can provide complementary behavioral evidence when facial cues are occluded or insufficient. This section addresses **RQ3** and **RQ4** by examining whether micro-actions are associated with emotional states and whether they provide complementary affective cues beyond facial micro-expressions. To answer them, we conduct empirical analysis on the emotion-rich subset of MMA-82 and evaluate the contribution of micro-actions to video emotion recognition.

Emotion-Action Dataset. We build the Emotion-Action dataset based on CAER [15], which originally contains seven emotion categories: *anger*, *disgust*, *fear*, *happy*, *neutral*, *sad*, and *surprise*. We extend it in two aspects. First, we introduce an additional category, *melancholy*, to better capture subtle differences within negative emotions. In our definition, *melancholy* denotes a milder and more implicit affective state than *sad*. Second, we annotate the micro-actions in each video clip to explicitly link emotional states with subtle body movements.

We select 100 clips for each category, resulting in a total of 800 clips.

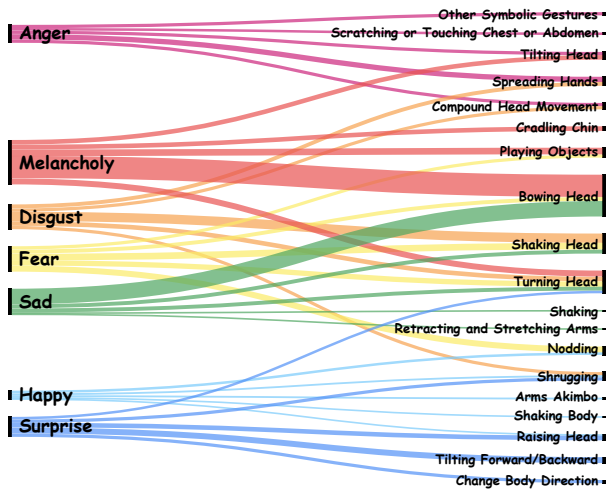


Fig. 6. Sankey-style visualization of the Top-5 micro-actions associated with each emotion category. Flow width denotes relative importance.

Exploration of Emotion-Action Relationships. To answer RQ3, we analyze the relationship between emotion categories and micro-actions using decision-tree-based feature importance and Sankey-style visualization. More details for the decision-tree construction are provided in the Appendix. Based on the results of the decision-tree, Fig. 6 visualizes the Top-5 micro-actions for each emotion. Two observations can be made. First, each emotion is strongly associated with several specific micro-actions, indicating that subtle whole-body movements provide meaningful cues for affective understanding. Second, similar emotions share overlapping micro-action patterns. For example, *sad* and *melancholy* both correlate strongly with *bowing head* and *turning head*, suggesting partially common behavioral patterns. At the same time, their differences are also informative: *sad* is more associated with *shaking* and *retracting and stretching arms*, whereas *melancholy* is more related to *playing objects* and *cradling chin*. This suggests that *sad* tends to involve more explicit negative bodily expressions, while *melancholy* is reflected through milder and more inward subtle movements.

We further validate these findings from an interdisciplinary perspective using *Laban Movement Analysis* (LMA) [38]¹. The high-importance micro-actions identified above show strong consistency with established movement psychology theory [39], [40]. For instance, positive emotions are often associated with upward, open, and rhythmic movements, which align well with actions such as *Nodding*, *Tilting Forward/Backward*, *Raising Head*, and *Shrugging*². In contrast, negative emotions are more related to downward or contractive movements (LMA descriptor of *Head-Drop*), such as *Bowing Head*. This agreement suggests that micro-actions are not only

¹It is a classical framework that connects human movement with psychological and emotional states.

²In our analysis, *Nodding* and *Tilting Forward/Backward* are closely related to positive affect and are well aligned with the LMA element of *Rhythmicity*. Similarly, *Raising Head* and *Shrugging* correspond to the *Up and Rise* element, while *Arms Akimbo* is related to the *Spread* element.

TABLE VII
PERFORMANCE COMPARISON OF FACIAL MICRO-EXPRESSION AND MICRO-ACTION BASED METHODS.

Task	Method	Top-1 Acc	F1
Micro-Expression Only	DeepFace [41]	22.86	17.54
Micro-Action Only	TSM [34]	32.38	31.86
Both	DeepFace + TSM	32.86	32.36

statistically correlated with emotions, but also psychologically meaningful.

Micro-Actions for Emotion Recognition. To answer RQ4, we compare facial micro-expression cues, micro-action cues, and their combination for video emotion recognition. Specifically, we evaluate three settings: (1) *Micro-Expression Only*, (2) *Micro-Action Only*, and (3) *Both*. The first uses only facial micro-expression cues as the baseline, the second uses only micro-action cues, and the third combines both to assess whether micro-actions provide complementary affective information beyond micro-expressions. Table VII shows that micro-actions alone already achieve meaningful emotion recognition performance, indicating a strong relationship between subtle body movements and emotional states. More importantly, combining micro-actions with micro-expressions further improves recognition over the facial baseline, demonstrating that micro-actions provide complementary behavioral evidence for affective analysis. It is worth noting that the micro-action cues used in this experiment are derived from ground-truth annotations rather than model predictions. This design is intended to verify the intrinsic value of micro-actions for emotion recognition and to examine whether they provide useful affective information when accurately identified.

VI. CONCLUSION

In this paper, we have presented **MMA-82**, a comprehensive multi-domain benchmark that extends MA-52 toward more realistic micro-action analysis. MMA-82 expands the taxonomy from 52 to 82 fine-grained categories and covers four distinct domains beyond controlled laboratory settings. Based on this benchmark, we established two core tasks, **MMA-82-Rec** and **MMA-82-Det**, together with in-domain, cross-domain, few-shot, and zero-shot evaluation protocols. Extensive experiments show that current methods still face significant challenges under domain shift, long-tailed distributions, and temporal localization. We further showed that micro-actions are strongly associated with emotional states and provide complementary cues to facial micro-expressions for emotion recognition. We hope MMA-82 will serve as a useful foundation for future research on subtle action understanding, temporal action detection, cross-domain generalization, and multimodal affective behavior analysis.

REFERENCES

- [1] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.
- [2] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [3] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, “The” something something” video database for learning and evaluating visual common sense,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

- [4] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.
- [5] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.
- [6] G. Dai, X. Shu, W. Wu, R. Yan, and J. Zhang, "Gpt4ego: unleashing the potential of pre-trained models for zero-shot egocentric action recognition," *IEEE Transactions on Multimedia*, vol. 27, pp. 401–413, 2024.
- [7] S. Mao, X. Li, F. Zhang, X. Peng, and Y. Yang, "Facial action units as a joint dataset training bridge for facial expression recognition," *IEEE Transactions on Multimedia*, vol. 27, pp. 3331–3342, 2025.
- [8] D. Guo, K. Li, B. Hu, Y. Zhang, and M. Wang, "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6238–6252, 2024.
- [9] K. Li, P. Liu, D. Guo, F. Wang, Z. Wu, H. Fan, and M. Wang, "Mmad: Multi-label micro-action detection in videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 225–13 236.
- [10] K. Li, J. Gu, F. Wang, Z. Wu, H. Fan, and D. Guo, "Ma-bench: Towards fine-grained micro-action understanding," *arXiv preprint arXiv:2603.26586*, 2026.
- [11] J. Gu, K. Li, F. Wang, Y. Wei, Z. Wu, H. Fan, and M. Wang, "Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 5461–5470.
- [12] K. Li, D. Guo, G. Chen, C. Fan, J. Xu, Z. Wu, H. Fan, and M. Wang, "Prototypical calibrating ambiguous samples for micro-action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 4815–4823.
- [13] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, "imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 631–10 642.
- [14] H. Chen, H. Shi, X. Liu, X. Li, and G. Zhao, "Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1346–1366, 2023.
- [15] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 143–10 152.
- [16] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Sammm: A spontaneous micro-facial movement dataset," *IEEE transactions on affective computing*, vol. 9, no. 1, pp. 116–129, 2016.
- [17] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.
- [18] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–7.
- [19] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas (me)²: a database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 424–436, 2017.
- [20] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 2782–2800, 2022.
- [21] A. Shahrroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
- [22] J. Liu, A. Shahrroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.
- [23] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4125–4141, 2021.
- [24] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 995–19 012.
- [25] Z. Tan, C. Gao, A. Qin, R. Chen, T. Song, F. Yang, and D. Meng, "Towards student actions in classroom scenes: New dataset and baseline," *IEEE Transactions on Multimedia*, 2025.
- [26] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS one*, vol. 9, no. 1, p. e86041, 2014.
- [27] C. He, W. Li, Z. Jin, C. Xu, B. Wang, and D. Lin, "Opendatalab: Empowering general artificial intelligence with open datasets," *arXiv preprint arXiv:2407.13773*, 2024.
- [28] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4125–4141, 2020.
- [29] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1192–1200.
- [30] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [31] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.
- [32] H. Duan, J. Wang, K. Chen, and D. Lin, "Pyskl: Towards good practices for skeleton action recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7351–7354.
- [33] Y. Hao, H. Zhang, C.-W. Ngo, and X. He, "Group contextualization for video recognition," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2022, pp. 928–938.
- [34] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7083–7093.
- [35] J. Tan, X. Zhao, X. Shi, B. Kang, and L. Wang, "Pointad: Multi-label temporal action detection with learnable query points," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 268–15 280, 2022.
- [36] S. Liu, C.-L. Zhang, C. Zhao, and B. Ghanem, "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*, 2024, pp. 18 591–18 601.
- [37] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.
- [38] R. Laban and L. Ullmann, "The mastery of movement." 1971.
- [39] T. Shafir, R. P. Tsachor, and K. B. Welch, "Emotion regulation through movement: unique sets of movement characteristics are associated with and enhance basic emotions," *Frontiers in psychology*, vol. 6, p. 2030, 2016.
- [40] A. Melzer, T. Shafir, and R. P. Tsachor, "How do we recognize emotion from movement? specific motor components contribute to the recognition of each emotion," *Frontiers in psychology*, vol. 10, p. 392097, 2019.
- [41] S. I. Serengil and A. Ozpinar, "Boosted lightface: A hybrid dnn and gbm model for boosted facial recognition," *Gazi University Journal of Science*, vol. 39, no. 1, pp. 452–466, 2026. [Online]. Available: <https://dergipark.org.tr/en/pub/gujs/article/1794891>
- [42] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in neural information processing systems*, vol. 30, 2017.
- [43] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 492–510.
- [44] H. Alwassel, F. C. Heilbron, V. Escorcia, and B. Ghanem, "Diagnosing error in temporal action detectors," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 256–272.
- [45] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

APPENDIX A
 BASELINE IMPLEMENTATION DETAILS

In this appendix, we provide the implementation details of all baselines evaluated on the MMA-82 benchmark.

A. Implementation Details of PoseConv3D

We use PoseConv3D [32] as the baseline for micro-action recognition based on skeleton. It uses a 3D convolutional neural network to process skeleton heatmap sequences. The model backbone is based on ResNet3D, with 17 input channels. It consists of three stages, containing 4, 6, and 3 layers respectively, and uses dilated convolutions in the second and third stages to expand the temporal receptive field. The spatial domain is downsampled with a stride of 2. The final features are fed into a 512-channel I3D [2] classification head, which outputs classification results for 82 classes. In the data pipeline, 32 frames are sampled uniformly from the raw skeleton sequence at a resolution of 64x64. The training process uses the SGD [42] optimizer with an initial learning rate of 0.2, a weight decay of 3e-4, and 60 training epochs.

B. Implementation Details of GC-TSM

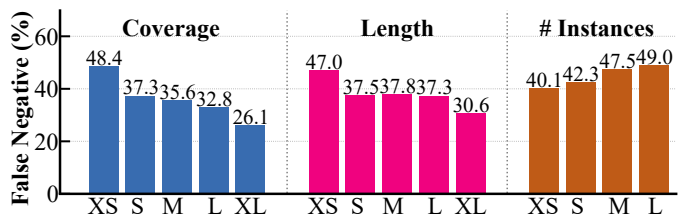
For RGB input, we use GC-TSM [33] as the baseline. GC-TSM embeds a lightweight group contextualization module into the TSM backbone network, enabling it to capture more spatio-temporal information critical to recognizing subtle actions while maintaining high computational efficiency. Specifically, we use ResNet-50 as the backbone and select 8 frames from each video as input. Additionally, SGD is used as the optimizer, with an initial learning rate of 0.01 and a weight decay of 1e-4. Decay is applied at the 20th and 40th epochs, with a total of 50 training epochs and a batch size of 16.

C. Implementation Details of AdaTAD

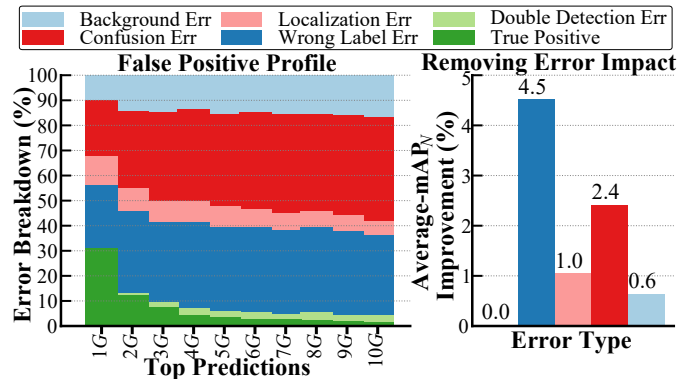
For the micro-action detection task, we use the end-to-end action detection framework AdaTAD on the MMA-82-Det dataset. Its backbone uses a vision transformer with an adapter, with VideoMAE pre-trained on Kinetics-400 serving as the base weights. The model uses sliding windows for sampling, with a window size setting of 256. After center-cropping, the data is resized to 160x160 and fed into the network. AdaTAD uses AdamW for optimization, with the initial learning rates for both the backbone and the adapter set to 1e-4, and the model trained for 20 epochs. During the inference stage, Soft-NMS is used for post-processing, with the NMS threshold parameter set to 0.5. It retains the top 200 highest-confidence bounding boxes and uses a voting strategy to further refine the predicted temporal boundaries.

APPENDIX B
 MORE RESULTS ANALYSIS ON MAD

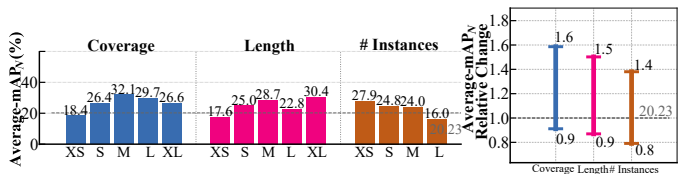
Following TAD conventions [36], [43] and micro-action detection task specifications [9], we used detection analysis tools [44] to evaluate the model’s performance on VideoMAE-H at the action level.



(a) False Negative Analysis



(b) False Positive Analysis



(c) Sensitivity Analysis

Fig. 7. Error analysis for the MAD task.

1) False Negative Profiling: As shown in Figure 7(a), we report the model’s false negatives (FN) across three dimensions: coverage, length, and instances. The baseline achieves a low FN in most dimensions. As coverage and length increase, the FN decreases; however, as the number of instances increases, the FN rises.

2) False Positive Profiling: As shown in Figure 7(b), in the left figure, we analyze the false positives at tIoU=0.5. These errors fall into five major categories, with “Confusion Error” and “Wrong Label Error” accounting for the majority. Furthermore, in the right figure, we break down the impact of each error type on the average mAP. Removing “Wrong Label Error” brings the most significant performance improvement (+4.5%), followed by “Confusion Error” (+2.4%). Although action localization is crucial, distinguishing between similar actions and filtering out background noise are also critical tasks that cannot be overlooked.

3) Sensitivity Profiling: To evaluate the model’s robustness, we analyze the performance sensitivity across different scales, as shown in Figure 7(c). The model performs stably under varying coverage levels, with performance maximizing at 32.1% when coverage is set to M. Performance tends to improve as action duration increases, but decreases as action density increases, which is consistent with real-world patterns.



Fig. 8. Video clip examples from the Emotion-rich television video collection.

APPENDIX C

MORE ANALYSIS FOR EMOTION WITH MICRO-ACTIONS

A. Examples of Dataset

As shown in Figure 8, we visualize some example videos from the Emotion-Action Dataset. Each video generally contains one or more action instances, which together reflect non-spontaneous human emotions. For example, when a person is sad, they tend to show sinking and contracting movements such as “Retracting and stretching arms” and “Folding arms,” which is consistent with LMA theory [38]. When a person is happy, they exhibit rising, light-weight, rhythmic movements such as “Shrugging” and “Shaking body.”

B. Details of Micro-Actions for Emotion Recognition

To verify the contribution of micro-actions to emotion recognition, we use DeepFace and TSM to recognize emotions in facial expressions and micro-actions, respectively. The results are reported in Table VII. Specifically, we first use DeepFace to recognize emotions in cropped facial video frames and retain the logit vectors l_{face} after removing the final Softmax layer. Similarly, for micro-action video frames, TSM predicts emotions by capturing the spatiotemporal interactions of micro-actions, and we similarly retain the logit vectors l_{mas} . Subsequently, rather than concatenating features directly in high dimensions, we adopt a late fusion strategy, which mitigates issues of feature dimensional mismatch and redundant noise. We concatenate these two independent logits to obtain a concatenated logits l_{fusion} . This vector is then fed into a lightweight MLP for fine-tuning, resulting in an optimized model.

C. Details of Emotion-Action Tree

Due to the wide variation in the frequency of natural micro-actions among subjects, using absolute action counts directly would result in emotion-related key actions being overrepresented among high-frequency micro-actions. Therefore, we treat each video clip as a document and each action as a word, and process the data using Term Frequency–Inverse Document Frequency (TF-IDF) [45]. First, for each video containing multiple annotated micro-actions, we first compute a TF-IDF [45] feature vector and use it as the input to the decision tree. To better distinguish subtle and visually similar emotions (*i.e.*, *melancholy* and *sad*), we adopt a *One-Versus-Neutral* (OVN) strategy instead of the conventional *One-Versus-Rest* (OVR) setting. This design provides a more stable

and interpretable way to analyze the relationship between emotions and micro-actions.

For each OVN tree, we use the CART to construct it, with the choice of splitting points based on Gini impurity. To measure the global contribution of micro-action features within the tree, we calculate the feature importance for each micro-action, which is determined by the weighted sum of the Gini impurity values obtained at the nodes where the micro-action feature serves as a splitting node. We use Figure 9 to present the decision process of the decision tree for emotion discrimination, while highlighting the Top-5 micro-actions most strongly associated with each emotion category.

Based on the learned trees, we compute the importance of each micro-action for distinguishing a target emotion. Since feature importance reflects only contribution magnitude rather than association direction, we further define the following conditional probability ratio:

$$R = \frac{P(MA | E_t)}{P(MA | \bar{E}_t)}, \quad (7)$$

where E_t denotes the target emotion and \bar{E}_t denotes all non-target emotions. When $R \geq 1$, the micro-action MA is regarded as positively correlated with the target emotion.

To further verify whether the micro-actions selected by the OVN decision tree described above retain strong emotional representational capabilities, we use these filtered features as input to an MLP for emotion classification. The pseudocode for the entire algorithm is shown in Algorithm 1.

We conduct three sets of experiments to validate its effectiveness. First, we input all TF-IDF values as features into the MLP. Second, we retain the Top-5 MAs, set the others to 0, recalculate the TF-IDF, and input it into the MLP. Finally, we set all Top-5 MAs to 0, recalculate the TF-IDF, and input it into the MLP. We report the validation results in Table VIII.

TABLE VIII
RESULTS OF TOP-5 MICRO-ACTION AND EMOTION

No.	Experience	Acc	Delta	F1
1	Base Results	0.271	0	0.277
2	no Top-5 MAs only	0.186	-0.086	0.147
3	Top-5 MAs only	0.379	+0.107	0.350

This indicates that using only non-Top-5 MAs significantly reduces the accuracy of emotion recognition, whereas considering only the Top-5 MAs significantly improves it.

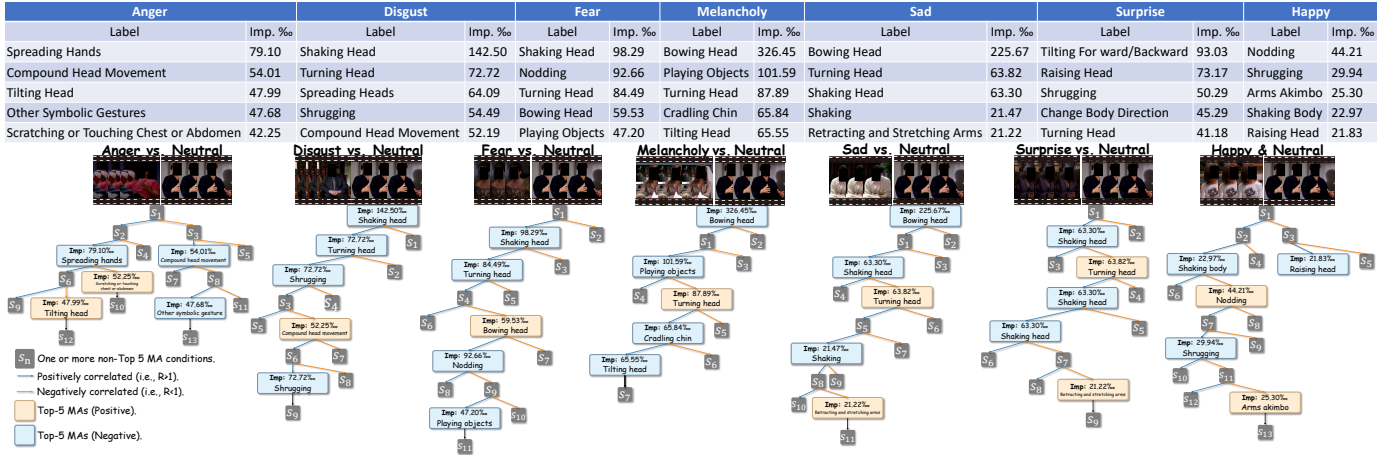


Fig. 9. Decision-tree-based visualization of emotion discrimination and the Top-5 micro-actions most strongly associated with each emotion category. For each emotion, the figure shows the decision process under the One-Versus-Neutral (OVN) setting and highlights the five micro-actions with the highest importance scores.

Algorithm 1 MLP-based Emotion Recognition

Require: TF-IDF feature matrix $X \in \mathbb{R}^{n \times m}$
Require: Emotion labels $y \in \{1, \dots, 7\}$
Require: Selected MA set M
Ensure: Predicted labels \hat{y}

- 1: Normalize MA frequency features
- 2: Construct TF-IDF representation X
- 3: Split X and y into training, validation, and test sets
- 4: Initialize MLP parameters θ
- 5: **for** $epoch = 1$ to E **do**
- 6: **for** each mini-batch (x_b, y_b) **do**
- 7: $h_b \leftarrow \text{ReLU}(W_1 x_b + b_1)$
- 8: $h_b \leftarrow \text{Dropout}(h_b)$
- 9: $z_b \leftarrow W_2 h_b + b_2$
- 10: $L \leftarrow \text{CrossEntropy}(z_b, y_b)$
- 11: Update θ using Adam optimizer
- 12: **end for**
- 13: Evaluate validation performance
- 14: Update best model if F1 score improves
- 15: **end for**
- 16: $\hat{y} \leftarrow$ Predict test labels using the best model
- 17: **return** \hat{y}

This demonstrates that the actions we identified are strongly correlated with the corresponding emotions and can effectively enhance the performance of emotion recognition.