Supplementary Material for MA-Bench: Towards Fine-grained Micro-Action Understanding

Kun Li[1], Jihao Gu[2], Fei Wang[3,4], Zhiliang Wu[5], Hehe Fan[5], Dan Guo[3,4*]
[1] CVLab, College of Information Technology, United Arab Emirates University [2] University College London
[3] Hefei University of Technology [4] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[5] CCAI, Zhejiang University

Overview

This supplementary material provides more details of the MA-Bench, including Micro-Action Benchmark Generation, Evaluation Metric, Supervised Fine-Tuning, Qualitative Results and Definition of Micro-Action label. These topics are organized as follows:

Contents

A. Micro-Action Benchmark Generation

A.1. Micro-Motion Tracker

As introduced in Section 1. Introduction of the main paper, we propose a micro-motion tracker to extract motion descriptors for each body part and build the micro-action captions. As shown in Figure S1, we illustrate the pipeline of micro-action benchmark generation. Specifically, we first use the CoTracker3 model [6] to extract dense optical flow with four directional components that capture detailed motion dynamics, and then utilize the YOLOv8x-Seg model [11] to segment human-centric regions. As shown in Figure S2, we provide more samples from the micro-motion tracker. It's obvious that the human-centric optical flow can reflect the motion of key body parts.

*Corresponding author.

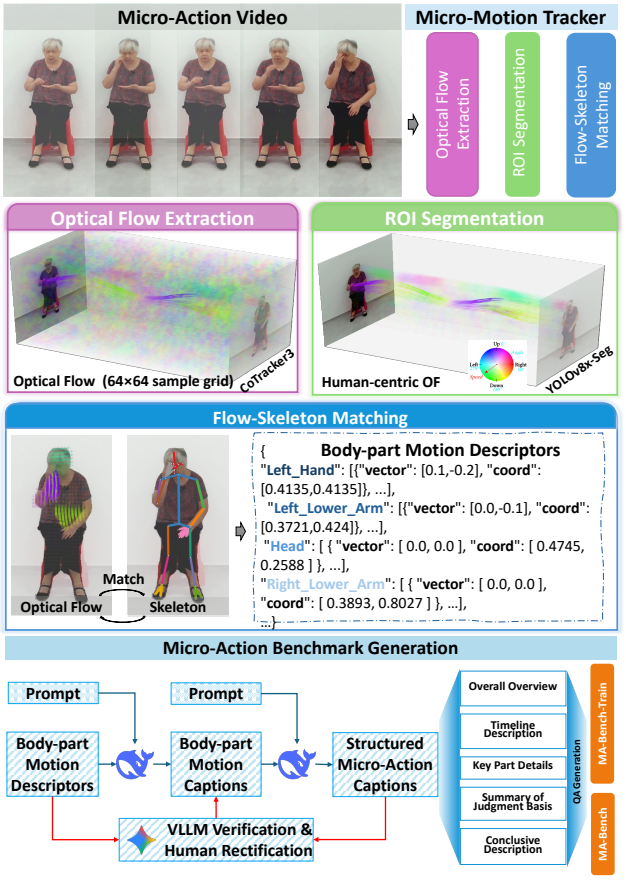

Figure S1. **The pipeline of the micro-action benchmark generation.**

As shown in Figure S2, we present additional examples produced by the proposed micro-motion tracker. The human-centric optical flow clearly highlights the local motion of key body parts and captures subtle motion patterns around the head, torso, and limbs. By matching these human-centric optical flows with the corresponding skeleton data, we can derive motion vectors and spatial coordinates for each body part, forming a structured, part-level description of the underlying dynamics.

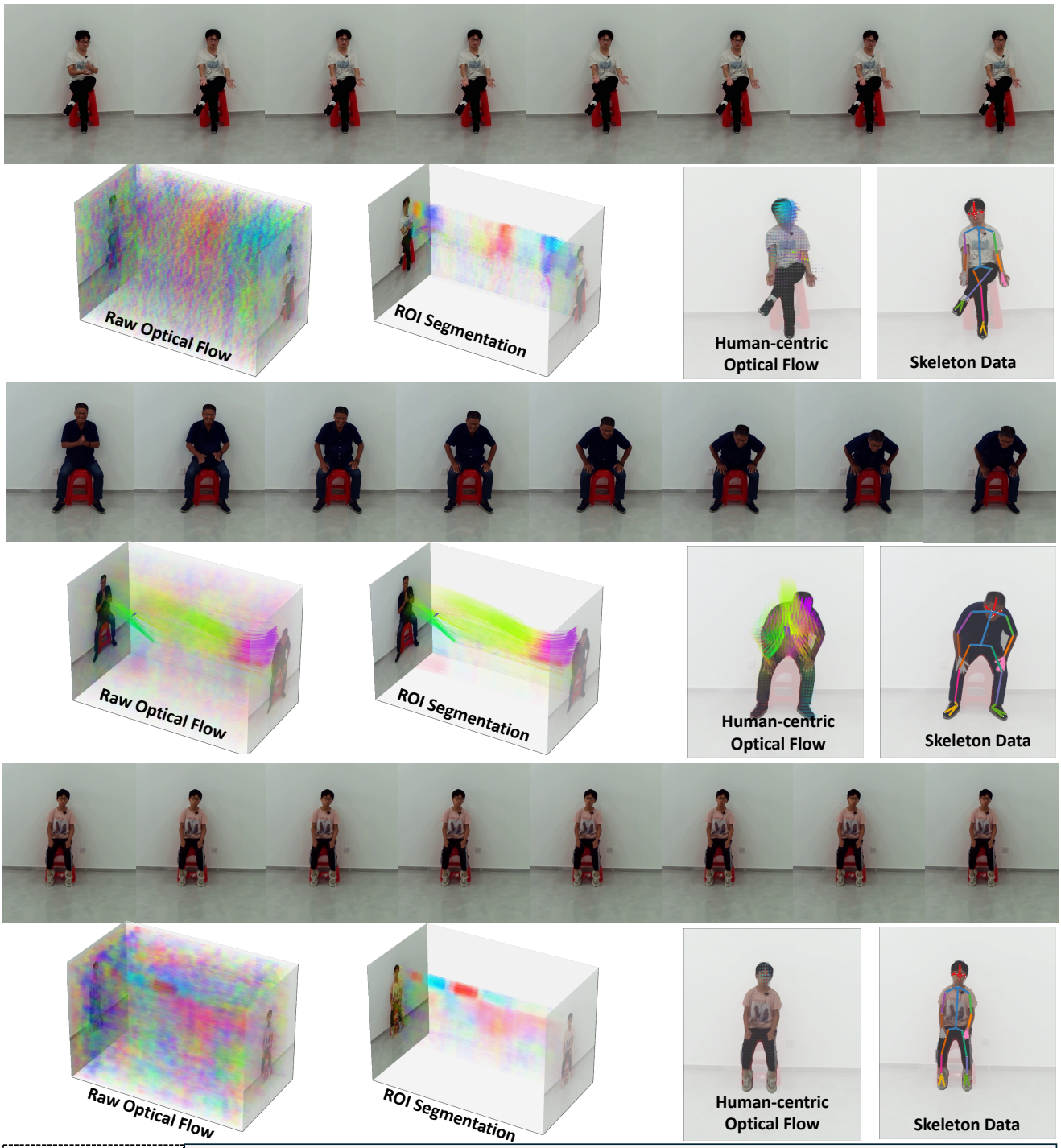

Figure S2. **Examples from micro-motion tracker.** We calculate the optical flow from four directions.

Body-part Motion Descriptor

```
{
  "Head": [
    {"vector": [0.02, -0.05], "coord": [0.537, 0.2693]},
    {"vector": [0.17, -0.05], "coord": [0.5372, 0.2688]},
    "..."
  ],
  "Left_Foot": [
    {"vector": [-0.12, -0.19], "coord": [0.607, 0.7922]},
    {"vector": [0.03, 0.01], "coord": [0.6038, 0.7933]},
    "..."
  ],
  "Left_Hand": [
    {"vector": [0.16, -0.81], "coord": [0.5808, 0.559]},
    {"vector": [0.02, 0.12], "coord": [0.5676, 0.547]},
    "..."
  ],
  "Left_Lower_Arm": [
    {"vector": [0.33, -0.24], "coord": [0.4844, 0.4977]},
    {"vector": [-0.17, -0.12], "coord": [0.4828, 0.4995]},
    "..."
  ],
  "Left_Lower_Leg": [
    {"vector": [-0.08, -0.4], "coord": [0.5667, 0.6691]},
    {"vector": [0.22, 0.13], "coord": [0.5667, 0.6696]},
    "..."
  ],
  "Left_Upper_Arm": [
    {"vector": [0.48, 0.0], "coord": [0.4294, 0.3874]},
    {"vector": [0.29, 0.18], "coord": [0.4316, 0.3874]},
    "..."
  ],
  "Left_Upper_Leg": [
    {"vector": [0.04, -0.35], "coord": [0.5044, 0.5477]},
    {"vector": [0.22, 0.13], "coord": [0.5667, 0.6696]},
    "..."
  ],
  "Right_Foot": [
    {"vector": [-0.06, -0.04], "coord": [0.5063, 0.8044]},
    {"vector": [0.01, 0.07], "coord": [0.5046, 0.8057]},
    "..."
  ],
  "Right_Hand": [
    {"vector": [0.02, -1.57], "coord": [0.5598, 0.528]},
    {"vector": [-0.1, -0.04], "coord": [0.5628, 0.5287]},
    "..."
  ],
  "Right_Lower_Arm": [
    {"vector": [-0.35, -1.29], "coord": [0.6263, 0.4754]},
    {"vector": [-0.27, -0.08], "coord": [0.6246, 0.4778]},
    "..."
  ],
  "Right_Lower_Leg": [
    {"vector": [-0.12, -0.19], "coord": [0.5117, 0.6951]},
    {"vector": [0.01, -0.03], "coord": [0.5044, 0.7034]},
    "..."
  ],
  "Right_Upper_Arm": [
    {"vector": [-0.27, -0.35], "coord": [0.6396, 0.3828]},
    {"vector": [-0.14, -0.09], "coord": [0.6346, 0.3795]},
    "..."
  ],
  "Right_Upper_Leg": [
    {"vector": [0.02, -1.0], "coord": [0.5462, 0.5649]},
    {"vector": [0.06, 0.05], "coord": [0.5412, 0.5728]},
    "..."
  ],
  "Torso": [
    {"vector": [0.14, -0.27], "coord": [0.5257, 0.4097]},
    {"vector": [0.07, -0.04], "coord": [0.5274, 0.4074]},
    "..."
  ]
}
```

Figure S3. Example of body-part motion descriptor.

A.2. Body-part Motion Descriptor

As shown in Fig S3, we give the example of body-part motion descriptors. This JSON data contains positional information for various body parts, with keypoint data for each part. Each body part's data includes vectors (`vector`) and coordinates (`coord`) at multiple time steps. The `vector` represents the motion change of the keypoint, while the `coord` indicates the keypoint's position in the 2D space. This motion descriptor is then input to LLMs to build micro-action captions.

Prompt for Body-part Motion Cap.

```
You will receive the following input:
**Motion data of a specific body part (in JSON format)**, including:
* 'part': the name of the body part (one of 14: Head /
Right_Upper_Arm / Right_Lower_Arm / Right_Hand / Left_Upper_Arm /
Left_Lower_Arm / Left_Hand / Torso / Right_Upper_Leg /
Right_Lower_Leg / Right_Foot / Left_Upper_Leg / Left_Lower_Leg /
Left_Foot);
* 'num_frames': total number of frames;
* 'fps': frame rate (frames per second);
* 'frames': a list of length 'num_frames', each containing:
  * 'coord': '[x_norm, y_norm]', the **normalized coordinates** of
this body part in the current frame (top-left = (0,0), bottom-right
(1,1); corresponding pixel position = '(x_norm900, y_norm1080)');
  * 'vector': '[vx, vy]', representing the **displacement vector from
 the previous frame to the current frame**, in units of pixels/frame.

    * 'vx > 0': moving right; 'vx < 0': moving left;
    * 'vy > 0': moving upward; 'vy < 0': moving downward.
**Background information:**
This data is extracted from a **professional psychological interview
experiment**.
Participants completed the SCL-90 (Symptom Checklist-90) and then
took part in a one-on-one open interview with a psychologist. The aim
 was to elicit **natural, spontaneous nonverbal behaviors** (e.g.,
micro-expressions, head movements, limb adjustments) during authentic
 communication.
Each participant remained seated throughout the session, recorded by
a high-resolution camera (9001080 px) from a full-body front view
allowing precise capture of subtle movements involving the head,
limbs, and inter-body coordination.
You should generate a **hierarchical (generalspecific) natural
language description** based on the data:
* **First sentence (General):**
  Summarize the overall **movement trajectory and magnitude** of the
body part throughout the video, and clearly indicate the **main
period(s) of motion** in seconds (time = frame_index / fps, rounded
to one decimal).
* **Subsequent sentences (Specific):**
  Describe the detailed dynamics during the main movement interval(s)
, including:
  * Changes in movement direction (e.g., moves upward to the right);
  * Magnitude of velocity (e.g., 0.8 px/frame);
  * Trends inferred from the sign and magnitude of 'vector' (e.g., '
vx' changing from positive to negative indicates a shift from
rightward to leftward motion).
**Requirements:**
* The description must be strictly **data-based**  avoid any
psychological or interpretive language (e.g., nervous, avoidant).
* Use intuitive directional terms (left / right / up / down / upper-
left / upper-right / lower-left / lower-right); note that 'vy > 0'
means **upward**.
* Express time in seconds, rounded to one decimal (e.g., 1.22.5
seconds), without referring to specific frame numbers.
* Although the data originates from a psychological interview,
describe only the **physical movement behavior**, not its emotional
meaning.
```

Figure S4. Prompt for Body-part Motion Captions Generation.

A.3. Body-part Motion Captions

As shown in Figure S4, we provide the prompt for LLMs to generate Body-part Motion Captions. In Figure S5, we provide the example of Body-part Motion Captions.

B. Evaluation Metric of Open-ended QA

In MA-Bench, we define two open-ended tasks: ***Micro-Action Descriptive Understanding (MADU)*** and ***Micro-Action Reasoning and Explanation (MARE)***. We adopt an LLM-as-Judge protocol [13], using GPT-4o [5] as the evaluator. The detailed prompts for these two tasks are shown in Fig. S6 and Fig. S7, respectively.

C. Supervised Fine-Tuning

In experiments, we fine-tuned the QwenVL-8B model [12] on the proposed MA-Bench-Train set. Here, we provide the details of implementation. For data processing, the maximum sequence length (cutoff length) was set to 131,072.

Body-part Motion Captions

```
{
  "body_part_motion": {
    "Left_Upper_Leg": "The left upper leg shows small,
multidirectional movements with a total displacement magnitude of 8.5
 pixels, primarily active between 0.11.3 seconds. It initially moves
downward to the right at speeds up to 0.35 px/frame, then shifts to
upward rightward motion, followed by leftward movement, before
transitioning to smaller alternating motions...",
    "Right_Upper_Arm": "The right upper arm shows a general leftward
and downward movement trajectory with small overall displacement,
primarily active between 0.1 and 1.3 seconds. It initially moves
downward to the left at approximately 0.4 px/frame, then shifts to
upward-left motion with reduced velocity, followed by brief periods
of rightward motion...",
    "Right_Lower_Arm": "The right lower arm shows an overall downward
and leftward movement trajectory with a total displacement magnitude
of 16.6 pixels, primarily occurring during the 0.10.4 second interval
. The arm initially moves downward to the left with a velocity of
1.34 pixels/frame, then continues leftward with decreasing velocity
...",
    "Right_Lower_Leg": "The right lower leg exhibits small,
multidirectional movements throughout the 2.8-second recording, with
the main motion occurring between 0.1 and 0.3 seconds. The leg first
moves downward and leftward with velocities up to 0.23 px/frame, then
 reverses direction upward and rightward at 0.09 px/frame...",
    "Left_Foot": "The left foot shows small, multidirectional
movements throughout the 2.8-second recording, with the main motion
occurring between 0.10.9 seconds. The foot initially moves downward
and leftward at approximately 0.20.3 px/frame, then shifts to upward
and rightward motion...",
    "Right_Hand": "The right hand shows a small overall displacement
of 14.7 pixels, with the main movement occurring from 0.1 to 2.8
seconds. The hand initially moves downward with a velocity of 1.6
pixels per frame, then shifts to small rightward and leftward
adjustments...",
    "Left_Lower_Arm": "The left lower arm shows a small overall
displacement of about 9.6 pixels with the main movement period
occurring from 0.1 to 1.2 seconds. It first moves rightward and
downward at approximately 0.4 px/frame, then shifts leftward with
small vertical adjustments...",
    "Left_Hand": "The left hand exhibits moderate movement primarily
between 0.1 and 1.4 seconds, with an overall trajectory that shifts
slightly rightward and downward from its starting position. The hand
first moves downward to the right at 0.1 seconds with a velocity of
0.82 px/frame...",
    "Right_Foot": "The right foot exhibits minimal overall
displacement, with the main movement occurring between 0.9 and 2.7
seconds. It moves downward and slightly leftward at a velocity of
approximately 0.15 px/frame, then shifts upward and rightward with a
velocity of 0.11 px/frame...",
    "Left_Lower_Leg": "The left lower leg shows small overall
displacement with the main movement occurring between 0.1 and 0.9
seconds, characterized by brief directional changes and low
velocities generally below 0.4 pixels per frame. It moves downward
and leftward with a velocity of 0.4 pixels per frame...",
    "Left_Upper_Arm": "The left upper arm moved primarily rightward
and downward with small adjustments throughout the 2.8-second
recording, primarily active between 0.1 and 1.3 seconds. During
0.1-0.3 seconds, it moved rightward at 0.48 px/frame, then shifted to
 upper-right motion...",
    "Right_Upper_Leg": "The right upper leg exhibits moderate movement
 with a total displacement of 10.8 pixels, primarily occurring
between 0.10.9 seconds and 1.61.8 seconds. It moves downward and
slightly rightward with velocities reaching 1.0 px/frame downward,
then shifts upward to the right...",
    "Torso": "The torso exhibits a small overall displacement with
subtle rightward and slight upward movement, primarily active between
 0.10.9 seconds and 2.32.8 seconds. It initially moves downward to
the right at 0.3 px/frame, then shifts to upward rightward motion
...",
    "Head": "The head moves primarily downward and rightward with a
total displacement of approximately 64.5 pixels, with the main
movement period occurring from 0.4 to 2.7 seconds. The head first
moves downward at 3.4 px/frame, then shifts upward to the right at
1.8 px/frame..."
  }
}

```

Figure S5. Example of Body-part Motion Captions

Prompt for MMDU Evaluation

```
In the **Micro-Action Descriptive Understanding (MMDU)** task, we
evaluate consistency between the model-generated description and the
ground-truth from **semantic**, **spatial**, and **temporal**
perspectives. For each level (**L1**, **L2**, **L3**), assign an **
integer score from 0 to 5**, where **0** means completely incorrect
or missing and **5** means fully accurate and consistent.

* **L1. Behavioral Semantic Alignment (05)**
  Evaluate whether the model accurately reproduces the **core
behavioral semantics** of the action, including:

  * dominant body parts,
  * primary action type (micro-motion label),
  * leadfollow relationships between parts,
  * motion tendency (continuous, repetitive, or single).

* **L2. Spatial-Topological Fidelity (05)**
  Evaluate the correctness of **spatial directionality** and **
positional relations** in the description, including:

  * directions (left/right, up/down, front/back, far/near),
  * relations w.r.t. body midline or torso (e.g., toward/away from
torso, crossing midline).

* **L3. Temporal-Structural Coherence (05)**
  Evaluate whether the **temporal structure and rhythm** of the
action are preserved, including:

  * clear action phases (initiation, transition, stabilization),
  * correct temporal order of sub-motions,
  * consistent rhythmic pattern (continuous, intermittent, repetitive
).

For each response, output three scores: `L1_score`, `L2_score`, and `
L3_score`, each in [0, 5].
```

Figure S6. Prompt for Micro-Action Descriptive Understanding (MADU) Evaluation.

Prompt for MARE Evaluation

```
In the **Micro-Action Reasoning and Explanation (MARE)** task, we
evaluate model reliability along three dimensions: **L1**, **L2**,
and **L3**. For each dimension, assign an **integer score from 0 to
5**, where **0** means completely incorrect or missing and **5**
means fully correct and consistent.

* **L1. Coarse-Grained Label Accuracy (05)**
  Evaluate whether the model correctly identifies the **primary body
parts** involved in the action, captures their **hierarchical
relationships** (e.g., which part is primary/secondary in contact),
and **avoids ambiguous or mixed body-part descriptions**.

* **L2. Fine-Grained Label Accuracy (05)**
  Evaluate whether the model correctly recognizes the **micro-action
category**, including **motion tendency** (single / continuous /
reciprocal), **contact attributes**, **directionality**, and **object
 participation**, and whether it **distinguishes similar action types
** without semantic confusion.

* **L3. Causal Reasoning Consistency (05)**
  Evaluate whether the model forms a coherent **
observationreasoningconclusion** chain, using observable cues (body
parts, direction, motion trend, contact, temporal pattern) to support
 its labels, **without internal logical contradictions**.

For each response, output three scores: `L1_score`, `L2_score`, and `
L3_score`, each in [0, 5].
```

Figure S7. Prompt for Micro-Action Reasoning and Explanation (MARE) Evaluation.

Videos were sampled at 3.0 frames per second (FPS). The total video pixels were dynamically managed, with a minimum of $16 \times 32 \times 32$ (`video_min_pixels`) and a maximum of $1280 \times 32 \times 32$ (`video_max_pixels`). The LoRA [4] configuration was set with a rank (r) of 8, an alpha (α) of 16, and a dropout rate of 0. LoRA modules were applied to the visual encoder and the final two layers of the text decoder. Training was conducted using the `adamw_torch` [10] optimizer with `bf16` (BFloat16) mixed precision. We utilized a cosine learning rate scheduler with a peak learning rate of 5.0×10^{-5}. The model was trained for 1.0 epoch with no warmup steps. We used a per-device batch size of 2 with 8 gradient accumulation steps, resulting in an effective batch size of 16. Gradient clipping was applied with a maximum gradient norm of 1.0.

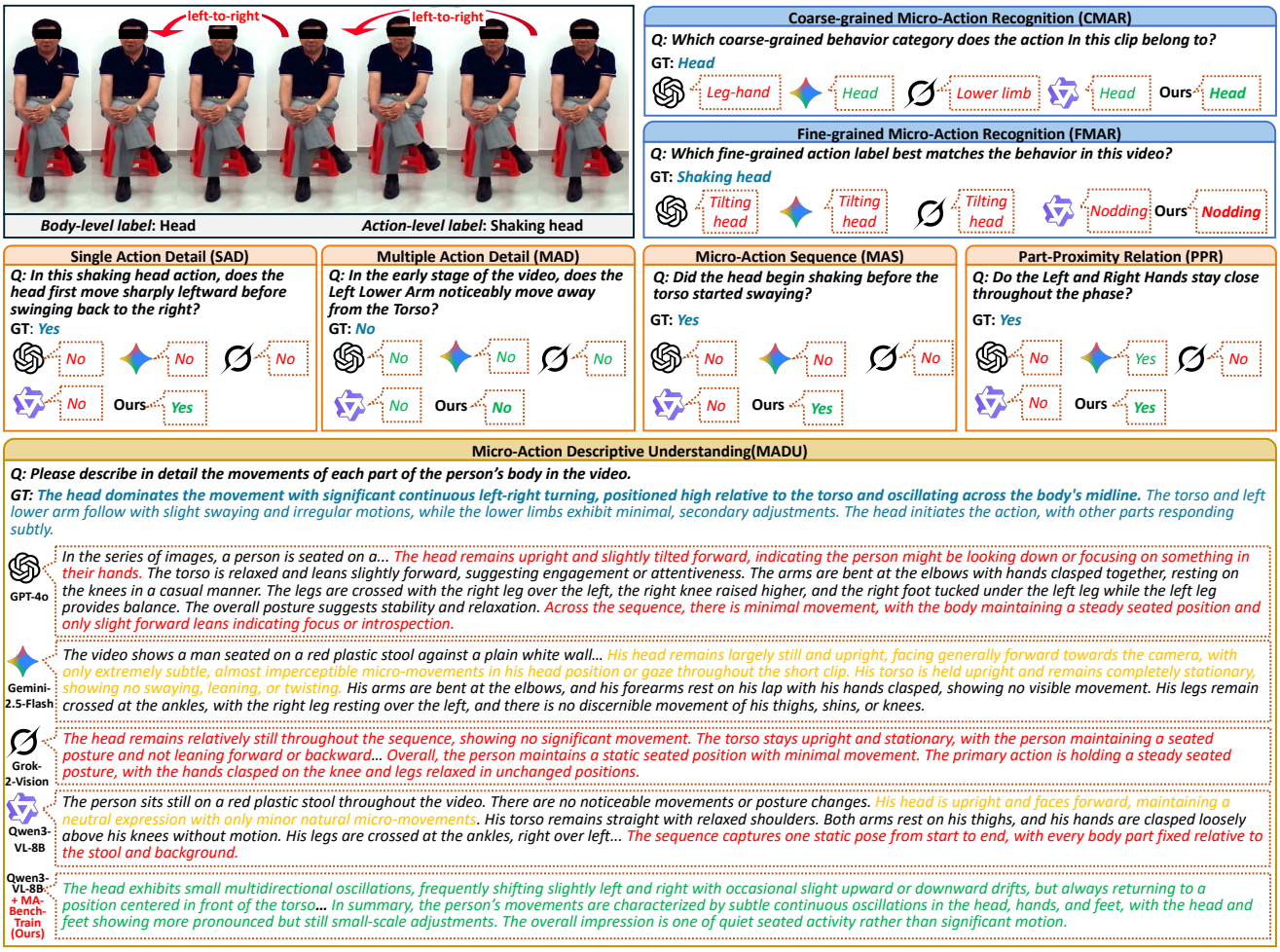

Figure S8. **Qualitative example from the MA-Bench.**Green font denotes correct predictions, red indicates completely incorrect ones, yellow marks partial errors, black shows results with no impact, and purple highlights faulty reasoning chains.

D. Qualitative Results

In the main paper, we have illustrated the qualitative results of the ***Micro-Action Reasoning and Explanation (MARE)*** task in MA-Bench. Here, we provide the qualitative results of the other tasks of MA-Bench in Figure S8. *First*, we can see that current models perform well in recognizing coarse-grained micro-actions, such as identifying dominant body parts (*e.g.*, "Head") in ***Coarse-Grained Micro-Action Recognition (CMAR)***. However, they struggle with fine-grained actions, such as distinguishing between similar motions (*i.e.*, ***Fine-Grained Micro-Action Recognition***), where partial errors are observed. *Second*, for the relation comprehension tasks, we can see that our model outperforms previous work. This improvement is primarily due to our model's ability to better understand and capture the complex relationships between different body parts and their motion patterns. *Finally*, in ***Micro-Action Descriptive Understanding (MADU)***, our method outperforms traditional models by better capturing the dominant body part and its motion sequence, though it still faces challenges in fine-grained action details and temporal relations.

E. The Definition of Micro-Action Label

In micro-action understanding [1–3, 7–9], each video contains both `Body-level` and `Action-level` categories. `Body-level` category denotes the body part of micro-action occurring, including ["A: Body", "B: Head", "C: Upper limb", "D: Lower limb", "E: Body-hand", "F: Head-hand", and "G: Leg-hand".] `Action-level` category denotes the exact name of micro-actions. Taking the body-level category "A: Body" as an example, there are 5 action-level categories: "A1: Shaking body", "A2: Turning around", "A3: Sitting straightly", "A4: Shrugging" and "A5: Rising up". Body-level labels and action-level labels are naturally hierarchical structures. The detailed descriptions and label ID of each micro-action are from the MA-52 dataset [2], and are provided in Table S1.

ID	Categories	Descriptions
		Body-level label: Body (A)
A1	Shaking body	The movement of the upper body from side to side
A2	Turning around	The movement that causes changes in the facing angle
A3	Sitting straightly	The dynamic process of straightening the back, without involving the maintenance of an upright sitting
A4	Shrugging	Shrugging or moving the shoulders back and forth, indicating movements involving only the shoulders
A5	Rising up	The preparatory movements before standing up from a seated position
		Body-level label: Head (B)
B1	Nodding	The movement of continuous up-and-down head
B2	Shaking head	The movement of continuous turn of the head left and right
B3	Turning head	Turn the head backward or sideways, typically as a single motion
B4	Tilting head	Tilt the head toward the shoulder, distinguishing between tilting and turning
B5	Bowing head	Lower the head slowly or abruptly, typically as a single motion
B6	Head up	Raise the head slowly or abruptly, typically as a single motion
		Body-level label: Upper limb (C)
C1	Illustrative gestures	The illustrative gestures made with either the left arm or the right arm
C2	Other finger movements	Excluding finger movements as defined within the upper limb
C3	Hands touching fingers	The movement of changing from having the fingers of both hands uncrossed to interlaced
C4	Stretching arms	The movement of raising or extending the arm forward or sideways
C5	Waving	The movement of waving one's hand or arm, often accompanied by shaking one's head
C6	Scratching arms	The movement of scratching the other arm with one hand
C7	Spreading hands	The movement of holding one's hand open with the palm facing up
C8	Playing objects	The movement of playing with or adjusting objects in hand.
C9	Putting hands together	The movement of bringing both hands closer together until palms together
C10	Retracting arms	The movement of extending and retracting the arm
C11	Pointing oneself	The movement of the hand that indicates direction and self-reference
C12	Clenching fist	The movement of bending the fingers into a fist, including partial clenching of the fingers
C13	Rubbing hands	The movement of touching and rubbing the hands together
		Body-level label: Lower limb (D)
D1	Tiptoe	Place the foot sideways or rest the heel on a stool with the toe touching the ground
D2	Retracting feet	Retract the foot towards the body. If both feet retract into a tiptoe position, it is considered tiptoe
D3	Shaking legs	The movement of shaking the leg up and down, either continuously or briefly
D4	Stretching feet	The movement of extending the foot forward or sideways, often accompanied by leaning forward or turning
D5	Closing legs	The movement of moving the legs from apart to together
D6	Spread legs	The movement of increasing the distance between the knees
D7	Curling legs	The movement of sitting with one leg crossed over the other
D8	Crossing legs	The movement of legs gradually turning from crossing to crossing
		Body-level label: Body-hand (E)
E1	Scratching or touching chest	The movement of bringing the hand close to and touching the front of the upper body
E2	Scratching or touching neck	The movement of bringing the hand close to and touching the neck area
E3	Scratching or touching back	The movement of bringing the hand close to and touching the back of the upper body
E4	Arms akimbo	The movement of placing one or both hands on the hips
E5	Crossing arms	The movement of crossing one's arms over one's chest
E6	Scratching or touching shoulder	The movement of bringing the hand close to or touching the shoulder area
		Body-level label: Head-hand (F)
F1	Touching nose	The movement of bringing the hand close to and touching the nose area
F2	Scratching or touching face	The movement of touching the cheek area, excluding the nose, forehead, and chin
F3	Playing or tidying hair	Hand movements that adjust hair on the sides or top of the head, including women arranging their hanging hair
F4	Scratching or touching hindbrain	Touch the back of the head or the hair on the back of the head
F5	Pushing glasses	The movement of bringing the hand close to and touching the glasses
F6	Rubbing eyes	The movement of bringing the hand close to and touching the area around the eyes
F7	Scratching or touching forehead	The movement of scratching or touching the forehead
F8	Touching ears	The movement of bringing the hand close to and touching the ear, or touching and rubbing the ear
F9	Covering mouth	The movement of covering the mouth with the hand
F10	Covering face	The movement of covering the face region, excluding the mouth, with one or both hands
		Body-level label: Leg-hand (G)
G1	Touching legs	The movement of touching and rubbing the leg with one hand
G2	Patting legs	Move one hand close to and touch the leg in a dynamic process
G3	Scratching legs	Scratch the calf area with the hand, excluding the ankle
G4	Scratching feet	The movement of scratching the foot with the hand, excluding the legs

Table S1. The descriptions of each micro-action in the Micro-Action-52 dataset [2].

References

[1] Jihao Gu, Kun Li, Fei Wang, Yanyan Wei, Zhiliang Wu, Hehe Fan, and Meng Wang. Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition. In *Proceedings of the 33rd ACM International Con-*

ference on Multimedia, pages 5461–5470, 2025. 5

[2] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6238–6252, 2024. 5, 6

[3] Dan Guo, Xiaobai Li, Kun Li, Haoyu Chen, Jingjing Hu, Guoying Zhao, Yi Yang, and Meng Wang. Mac 2024: Micro-action analysis grand challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11304–11305, 2024. 5

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[6] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025. 1

[7] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. Prototypical calibrating ambiguous samples for micro-action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4815–4823, 2025. 5

[8] Kun Li, Dan Guo, Xiaobai Li, Haoyu Chen, Pengyu Liu, Fei Wang, Jingjing Hu, Guoying Zhao, and Meng Wang. Mac 2025: The 2nd micro-action analysis grand challenge. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 14216–14221, 2025.

[9] Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13225–13236, 2025. 5

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[11] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems*, pages 1–6, 2024. 1

[12] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[13] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 3