MMA-MTU: BENCHMARKING MULTI-TASK MICRO-ACTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have shown strong capabilities in video understanding, yet existing benchmarks largely focus on daily activities and general actions. Their ability to perceive and understand micro-actions, which are brief, low-amplitude, and spatially localized, remains insufficiently assessed. Beyond recognizing action categories, micro-action understanding requires reasoning about temporal relations, locating action boundaries, recovering relevant events, and describing how movements unfold. Our work addresses this evaluation gap through five key contributions: ***1) Multi-task benchmark.*** We introduce **MMA-MTU**, a benchmark that jointly evaluates Recognition, Temporal Reasoning, Grounding, Detection, and Description on shared video evidence. ***2) Large-scale, multi-domain data.*** MMA-MTU contains 4 domains, 11,180 untrimmed videos, and 122,845 queries covering 82 fine-grained categories, and 5 tasks with 12 subtasks. It provides dedicated training, validation, and test splits for instruction tuning and model learning. ***3) Paired queries and fine-grained descriptions.*** MMA-MTU includes 34,876 label-free queries paired with label-based counterparts, replacing action names with descriptions of observable movements. It also provides motion-evidence-guided event descriptions that capture how individual actions are performed. ***4) Extensive evaluation.*** We evaluate **36** MLLMs under the same protocol. The results expose a clear divide between different tasks: the best Temporal Reasoning accuracy reaches 66.62%, whereas the best Grounding average R@1 and Detection average mAP reach only 26.65% and 5.44%, respectively, showing that success on one ability does not reliably transfer to the others. Supervised fine-tuning significantly improves performance across all 5 tasks. ***5) Downstream exploration.*** A preliminary emotion-recognition study further explores how predicted micro-action information can support human emotion understanding.

1 INTRODUCTION

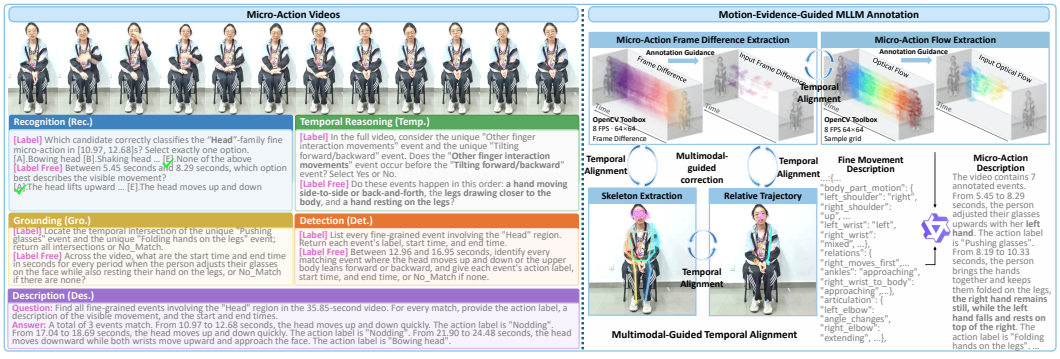


Figure 1: **Overview of MMA-MTU.** The benchmark evaluates Recognition, Temporal Reasoning, Grounding, Detection, and Description using label and, where applicable, label-free queries. Our Motion-Evidence-Guided MLLM Annotation pipeline combines frame differences, optical flow, skeletons, and relative trajectories to generate fine-grained ground-truth descriptions.

1.1 BACKGROUND AND IMPACT OF MICRO-ACTION UNDERSTANDING

Micro-actions (MAs) are brief and subtle body movements that provide fine-grained cues to human behavior Guo et al. (2024); Hao et al. (2026). They include head movements, hand gestures,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with representative benchmarks. N/R: Not Reported. The statistics refer to the test set

Benchmark	Videos	QAs	Untrim.	Annotation	Answer				★ Supported Task				
					Choice	Intervals	Events	Description	Recognition	Temporal	Grounding	Detection	Description
Daily Activities — Everyday human activities and interactions													
NExT-QA (Xiao et al., 2021)	1,000	8,564	✓	Manual	✓	✗	✗	✗	✗	✓	✗	✗	✗
AGQA (Grunde-McLaughlin et al., 2021)	1,814	669K	✓	Auto & Manual	✓	✗	✗	✗	✗	✓	✗	✗	✗
Charades-STA (Gao et al., 2017)	1,334	3,720	✓	Auto & Manual	✗	✓	✗	✗	✗	✗	✓	✗	✗
General Actions — Diverse video events across domains and temporal scales													
TempCompass (Liu et al., 2024b)	410	1,580	✗	Auto & Manual	✓	✗	✗	✓	✓	✓	✗	✗	✓
MVBench (Li et al., 2024b)	3,641	4,000	✓	Auto & Manual	✓	✗	✗	✗	✓	✓	✗	✗	✗
Video-MME (Fu et al., 2025)	900	2,700	✓	Manual	✓	✗	✗	✗	✗	✓	✗	✗	✗
MLVU (Zhou et al., 2025)	N/R	509	✓	Manual	✓	✗	✗	✓	✗	✓	✗	✗	✓
LongVideoBench (Wu et al., 2024b)	3,011	5,341	✓	Manual	✓	✗	✗	✗	✗	✓	✗	✗	✗
Micro-Actions — Short duration, low amplitude, and subtle movements													
MA-Bench (Li et al., 2026a)	1,000	12,000	✗	Auto & Manual	✓	✗	✗	✗	✓	✓	✗	✗	✗
★MMA-MTU (Ours)	**1,214**	**13,663**	✓	Auto & Manual	✓	✓	✓	✓	✓	✓	✓	✓	✓

leg actions, posture changes, body-part interactions, and human-object interactions, and can convey nonverbal information that complements facial expressions and speech Liu et al. (2021); Chen et al. (2023). Understanding MAs therefore can contribute to affective computing and human-computer interaction. Visually similar MAs may differ only in direction, trajectory, repetition, or the body parts involved, while background motion and repeated or overlapping events further complicate their interpretation in untrimmed videos. Recognizing only coarse interactions is therefore insufficient. As shown in Fig. 3, MLLMs confuse "Touching nose" with "Covering mouth," "Rubbing lip," or "Rubbing eyes", indicating that they may capture a broad hand-to-face interaction while failing to identify the precise contact location. Such distinctions can be behaviorally meaningful: facial self-touch locations and durations vary with cognitive and emotional demands Mueller et al. (2019), and facial self-touch has been associated with attention, working memory, and emotion regulation Grunwald et al. (2014). By emphasizing subtle motion within short temporal intervals, micro-action understanding (MAU) complements long-video understanding, which emphasizes integrating information across extended temporal contexts (Fu et al., 2025; Zhou et al., 2025). MAU thus provides a challenging setting for evaluating whether MLLMs can move beyond coarse-grained action semantics toward fine-grained spatiotemporal understanding of human behavior.

1.2 Current State of Micro-Action Understanding

Current MA research primarily focuses on recognition and detection. iMiGUE Liu et al. (2021) and MA-52 Guo et al. (2024) study recognition in pre-segmented clips, while SMG Chen et al. (2023) and MMAD Li et al. (2025) address temporal detection in untrimmed videos. MMA-82 Hao et al. (2026) extends both tasks across multiple domains, broadening the foundation for studying what micro-actions occur and when they occur. In practical systems, understanding human behavior requires connecting individual actions with their temporal context. A behavior-analysis system may need to locate a movement described by a user, identify other actions occurring around it, and determine their order or overlap. It may further need to describe the observed sequence and the relationships among these movements. Moreover, users may refer to "lifting one leg and placing it over the other" rather than the category name "crossing legs". These requirements connect recognition, grounding, detection, description, and temporal reasoning, and motivate evaluation with both label-based and label-free queries.

MLLMs offer an opportunity to evaluate these connected abilities within a single model through natural-language instructions. The challenge is to ground their responses in subtle visual changes and the temporal structure of micro-actions. MA-Bench Li et al. (2026a) extends evaluation beyond recognition to relational comprehension and interpretive reasoning within short clips. However, it does not address query-conditioned temporal grounding or multi-instance detection in untrimmed videos. Despite extensive benchmarking across video domains, a unified evaluation connecting all five micro-action understanding tasks through shared videos and event annotations remains lacking (Tab. 1). Such an evaluation should span multiple domains and compare label-based and label-free queries to assess whether models can identify the same movement from either its category name or a natural-language description.

1.3 MMA-MTU: A New Benchmark for Multi-Task Micro-Action Understanding

To address this evaluation gap, we introduce MMA-MTU, a benchmark for assessing MLLMs' ability to understand brief, low-amplitude, spatially localized human movements. Built on MMA-82-Det, it evaluates Recognition, Temporal Reasoning, Grounding, Detection, and Description. As illustrated in Figure 1, the five tasks share videos and event annotations while varying the query

conditions and required outputs, enabling systematic assessment of complementary micro-action understanding abilities. Our main contributions are summarized as follows:

- **Multi-task benchmark.** We construct 12 subtasks across 5 tasks using shared videos and fine-grained annotations. This design tests whether models can identify subtle movements, reason about their temporal relations, and accurately locate, recover, and describe the corresponding events.
- **Large-scale, multi-domain data.** MMA-MTU contains 11,180 untrimmed videos, 39,758 action instances, and 122,845 question-answer pairs spanning 82 micro-action categories. The videos cover laboratory, psychiatric, and street interviews, as well as television programs. Its cross-subject split provides 86,501 training pairs and 22,681 validation pairs for instruction tuning, alongside 13,663 test pairs from 1,214 videos for evaluation.
- **Paired queries and fine-grained descriptions.** We construct interval-aligned descriptions through a motion-evidence-guided MLLM pipeline that combines action labels and boundaries with four model motion cues. The annotations undergo automatic checks and human review. We further construct 34,876 label-free queries paired with label-based counterparts, supporting controlled comparison when labels are replaced by observable movement descriptions.
- **Extensive evaluation.** We evaluate 36 proprietary and open-source MLLMs in a zero-shot setting and examine the effects of recording domain, query form, frame sampling rate, and input resolution. The results identify temporal grounding and multi-instance detection as persistent challenges. Supervised fine-tuning Qwen3.5-4B on MMA-MTU improves performance on five tasks, demonstrating the value of its training data for model adaptation.
- **Downstream exploration.** We conduct a preliminary emotion recognition study to examine whether predicted micro-action information can support human emotion understanding. The results suggest that this information provides complementary behavioral cues, particularly when facial information is unavailable.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) connect visual encoders with large language models to support visual perception, instruction following, and language generation. Early models such as Flamingo, BLIP-2, and LLaVA established this design for image-language tasks (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). Later work extended MLLMs to video by introducing temporal visual tokens, video-language instruction tuning, and longer visual contexts (Zhang et al., 2023; Maaz et al., 2024; Lin et al., 2024; Bai et al., 2025b). These models can answer questions about video content and perform several tasks through natural-language instructions.

Video understanding still poses two related challenges for MLLMs. First, models often rely on static appearance and coarse scene cues rather than small changes across frames. MotionBench and TempCompass reveal persistent errors in motion direction, action order, repetition, and other temporal properties (Hong et al., 2025a; Liu et al., 2024b). MA-Bench reports a larger performance gap for micro-actions whose categories differ in body part, direction, or movement extent (Li et al., 2026a). Second, predicting temporal boundaries remains difficult. MLLMs must associate language with precise timestamps while preserving information from long or densely sampled videos. Results on E.T. Bench and CG-Bench show that strong video-level QA performance does not directly transfer to event grounding (Liu et al., 2024a; Chen et al., 2025). Micro-action understanding combines both difficulties: the model must distinguish a subtle movement and identify its short temporal extent.

2.2 Action-Centric MLLM Datasets and Benchmarks

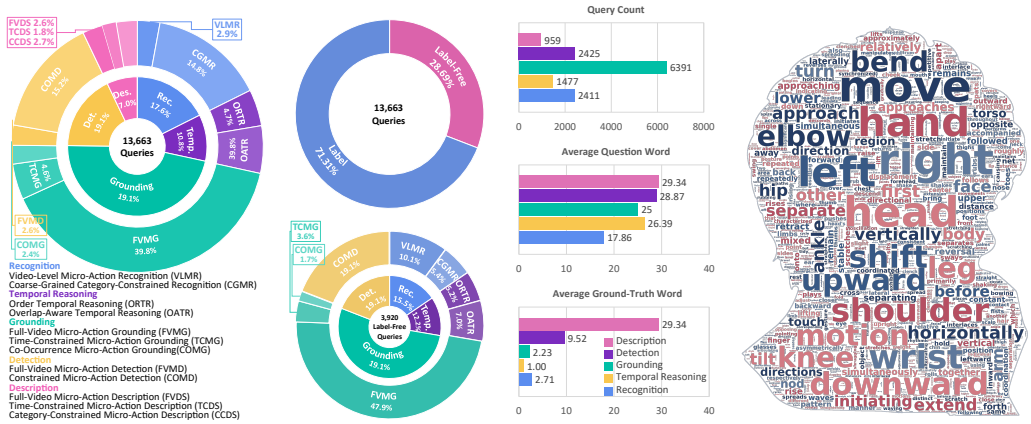
Action-related video benchmarks cover different parts of this problem. Recognition and motion-oriented benchmarks, including MA-Bench, MotionBench, and MVBench, test action categories or fine-grained motion properties through discrete questions (Li et al., 2026a; Hong et al., 2025a; Li et al., 2024b). NExT-QA, STAR, AGQA, and TempCompass focus on temporal order, causal relations, or compositional reasoning (Xiao et al., 2021; Wu et al., 2024a; Grunde-McLaughlin et al., 2021; Liu et al., 2024b). Temporal grounding benchmarks such as Charades-STA and QVHighlights (Gao et al., 2017; Lei et al., 2021) require models to return the interval associated with a

language query. CG-Bench and E.T. Bench extend grounding to long videos and event-level questions (Chen et al., 2025; Liu et al., 2024a). Detection benchmarks such as THUMOS14 require a set of action labels and temporal boundaries (Idrees et al., 2017). ActivityNet Captions and YouCook2 pair temporal intervals with natural-language event descriptions (Krishna et al., 2017; Zhou et al., 2018). General benchmarks, including Video-MME and MLVU, evaluate broader video perception and reasoning across multiple domains and durations (Fu et al., 2025; Zhou et al., 2025).

These benchmarks provide strong evaluation protocols for their target tasks, but their annotations, videos, and output formats remain separate. Existing micro-action benchmarks mainly emphasize recognition and temporal reasoning. Therefore, existing benchmarks cannot evaluate whether a unified MLLM can comprehensively understand micro-actions in full videos. MMA-MTU addresses this gap by evaluating Recognition, Temporal Reasoning, Grounding, Detection, and Description under a shared micro-action taxonomy and shared temporal annotations.

3 MMA-MTU

3.1 Overview and Statistics of Benchmark



(a) (b) (c) (d)

Figure 2: **Statistics of the MMA-MTU.** (a) Distribution of queries across 5 tasks and their subtasks. (b) Distribution of label and label-free queries, together with the task and subtask composition of the label-free subset. (c) Distribution statistics of questions and answers. (d) Word Count Statistics.

MMA-MTU is built on MMA-82-Det, the untrimmed temporal detection subset of MMA-82 Hao et al. (2026) and preserves the distribution characteristics of the dataset. MMA-82-Det contains 11,180 videos ranging from 0.6 to 90 seconds in duration, and 39,758 action instances collected from laboratory interviews, street interviews, psychiatric interviews, and television programs. Its annotations span 82 fine-grained micro-action categories grouped into seven body-level classes. We retain the original train, validation, and test splits and construct five tasks from the same video annotations: Recognition, Temporal Reasoning, Grounding, Detection, and Description. The benchmark adopts a cross-subject split.

MMA-MTU contains 122,845 question-answer pairs. The training and validation sets comprise 7,839 and 2,127 videos, with 86,501 and 22,681 pairs, respectively. They are used for instruction tuning and model selection. The test set forms the evaluation benchmark, containing 1,214 videos and 13,663 pairs. Each test video lasts 14.73 seconds on average, 3.66 seconds on average per action instance, and is associated with 11.25 questions. Across all three splits, 34,876 questions are label-free, accounting for 28.39% of the dataset. The test set contains 3,920 label-free questions, or 28.69% of its queries. These questions describe observable movements without revealing their action labels.

Fig. 2 summarizes the query and annotation statistics of the MMA-MTU. Fig. 2(a) shows the distribution of all 13,663 queries across five tasks. Based on conditions such as temporal scope and action-category constraints, these tasks are further divided into 12 subtasks, covering different levels of micro-action understanding. Fig. 2(b) compares label and label-free queries. Label-free queries account for 28.69% of the MMA-MTU and span four tasks and eight subtasks. Fig. 2(c) shows statis-

tics on the number of queries, the number of words in each question, and the number of ground-truth words for different tasks in the Benchmark. Fig. 2(d) visualizes the vocabulary of the ground-truth descriptions. Compared with action labels alone, these descriptions contain finer-grained information about body parts, motion directions, and how actions are performed. Additional statistics are provided in Sec. B.5.

3.2 Task Definition

Human understanding of a video develops through a natural sequence of questions. We first notice ***What happened***. We then connect the observed actions and ask ***How are the actions related***. Once a particular action becomes relevant, we ask ***When did the action occur***. A fuller understanding requires identifying ***What actions occurred and when***. We may finally explain ***How did each action unfold***. Inspired by this process, MMA-MTU organizes Recognition, Temporal Reasoning, Grounding, Detection, and Description into a progressive chain of micro-action understanding. Questions may refer to actions by their labels or through label-free descriptions of visible movement. In addition, to evaluate the MLLM's true capabilities and avoid guessing answers, we provided a "None of the above" option for choice questions and a "No_Match" option for non-choice questions. We provide a detailed definition of the five tasks of the MMA-MTU. In Sec. B, we provide additional examples, as well as definitions and detailed examples for 12 subtasks.

① Recognition: What happened? The first step is to determine whether a micro-action is present and identify its category. Given a video segment or temporal interval, the model selects one fine-grained action from several candidates.

②Temporal Reasoning: How are the actions related? Recognizing individual actions does not explain how a video develops over time. Temporal Reasoning therefore examines relations among observed events, including order, overlap, and co-occurrence. The queries determine whether a specified relation holds between actions. These answers use a Yes/No format.

③Grounding: When did the action occur? After identifying an action of interest, an observer may look for where it appears on the timeline. Grounding follows this query-driven setting. Given an action label or a label-free movement description, the model returns every matching interval in the full video or a specified temporal window. In addition, we also examine the co-occurrence intervals of multiple actions.

④Detection: What actions occurred, and when? Grounding begins with a specified target action, but complete video understanding cannot assume that every relevant action is known in advance. Detection asks the model to discover all micro-actions within a full video or condition. For every event, it returns the fine-grained category, start time, and end time.

⑤Description: How did each action unfold? The final step turns the recovered event set into a description that people can understand. For each action, the model reports its category, temporal boundary, and a natural-language description of the visible movement. The description may capture active body parts, motion direction, left-right movement, and relations between body parts.

Together, the five tasks follow the way people move from noticing an action to forming a complete account of a video: recognizing individual events, relating them over time, locating a queried event, recovering the full timeline, and explaining how each action occurs.

3.3 Question-Answer Generation

MMA-MTU is built on MMA-82-Det Hao et al. (2026), which provides action- and body-level categories together with densely annotated temporal boundaries. MMA-82-Det was designed primarily for temporal action detection; its annotations do not directly support evaluating multi-task understanding or describing fine-grained micro-actions. We therefore extend these annotations: ① structured query generation, ② label-free query generation, and ③ motion-evidence-guided description annotation. The left side of Fig. 1 presents representative queries from the five tasks. In Sec. B.2, we provide more examples and a more detailed flow.

① **Structured QA Generation.** We first convert the MMA-82-Det annotations into standardized events, each containing a body-level label, an action-level label, and its start and end times. On top of these events, we implement five task-specific template generators. According to the task definition and conditions, each generator fills action labels, candidate-label sets, event relations,

temporal windows, and other conditions into natural-language templates. The answers are derived directly from the standardized events.

② **Label-Free QA Generation.** We construct label-free QA pairs from the ①. Here, "label-free" means that the predefined action-category name is not explicitly provided; the query may instead contain observable motion attributes expressed in natural language. We first select pairs whose question or answer contains an action-level label. After fixing all non-label content, we provide the label and its temporally aligned video clip to an MLLM, which describes the action through observable movement without revealing its name. The generated description replaces only the corresponding label in the original QA pairs. Each label-free QA is therefore paired with a label QA that shares the same video, conditions, structure, and target, differing only in how the action is expressed. This paired construction enables a controlled comparison between taxonomy-label queries and natural-language movement descriptions, while keeping the video, task conditions, query structure, and target unchanged.

③ **Motion-Evidence-Guided Description Annotation.** As shown on the right of Fig. 1, we design a motion-evidence-guided pipeline to construct the ground truth for the description task. For each annotated action instance, we extract RGB frames from its temporal interval and use the OpenCV toolbox Bradski et al. (2000) and YOLO v11 Khanam & Hussain (2024) to compute optical flow, frame differences, skeleton, and body-relative trajectories from the same interval. Instead of providing whole-body motion indiscriminately, we use the annotated action label to select the body regions relevant to that action. The resulting motion evidence is passed to an MLLM to generate a detailed instance-level description. For multi-instance outputs, the instance descriptions are combined to form the final ground truth.

④ **Quality Control.** The generated query-answer pairs first go through automatic validation. A validator checks their structure, answer options, and output format, while an MLLM reviewer examines the consistency between each query and its target output and assesses whether the language and event logic agree with the visual evidence. Each pair is then assigned to three human annotators for two rounds of review and correction. The annotators verify the action semantics, temporal information, and visual consistency. They also ensure that the queries are unambiguous, that choice options are mutually exclusive, and that every query has a clearly defined target output.

3.4 Evaluation Metrics

We use Accuracy for Recognition and Temporal Reasoning. For Grounding, we report interval-set $\mathrm{F1@0.5}$, $\mathrm{R@1}$ at tIoU thresholds from 0.1 to 0.9, and $\mathrm{mIoU@1}$. For Detection, we report label-aware event $\mathrm{F1@0.5}$, mAP at tIoU thresholds from 0.1 to 0.9, and the average mAP across these thresholds. For Description, we use CLIPScore, BLEU-4, CIDEr-D, BLEURT, and EMScore Papineni et al. (2002); Vedantam et al. (2015); Sellam et al. (2020); Hessel et al. (2021); Shi et al. (2022). We additionally report label accuracy and temporal IoU for the action labels and temporal boundaries contained in each description. Error Rate measures the percentage of queries that produce invalid outputs that do not follow the required response structure. In addition, we use Acc and $\mathrm{F1}$ to evaluate the model's accuracy on ground-truth data labeled as "No_Match". Label and label-free queries are evaluated using the same task-specific metrics. For their paired comparison, we report the scores under both conditions, Paired Primary, and the performance gap computed as the label-free score minus the label score. We report more metric details in Sec. C.

4 How Well Do MLLMs Understand Micro-Actions?

4.1 Experimental Setup

We evaluate 36 MLLMs, covering proprietary models, open-source large models, and open-source small models. The proprietary models include GPT 5.6 Luna Singh et al. (2025). The open-source models include DeepSeek v4.1-Flash Wu et al. (2024c), GLM-4.6V-Flash and GLM-5.3-Flash Hong et al. (2025b), Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct, Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, Qwen3.5-35B-A3B, Qwen3.6-35B-A3B, Qwen3.8-27B and Qwen3.8-Max Bai et al. (2025a), VideoLLaMA3-7B Zhang et al. (2025); Phi-3.5-Vision-Instruct and Phi-4-Multimodal-Instruct Abdin et al. (2024); InternVL3-8B Zhu et al. (2025); InternVL3.5-2B, InternVL3.5-4B, InternVL3.5-8B, InternVL3.5-14B, InternVL3.5-38B, InternVideo3-8B-Instruct, and InternVL3.5-GPT-20B-A4B Wang et al. (2025); Kimi-VL-A3B-Instruct Team et al. (2025); LLaVA-OV-1.5-4B-Instruct and LLaVA-OV-2-8B-Instruct Li et al. (2024a); MiniCPM-o-4.5 and

Table 2: The overall performance of **36** MLLMs on MMA-MTU, including proprietary, open-source large, and open-source small models.

Model	Date	Rec. Acc.	Temp. Acc.	Grounding R@1 @.5	Grounding R@1 @.7	Grounding R@1 Avg.	Grounding mIoU	Detection mAP @.2	Detection mAP @.5	Detection mAP @.7	Detection mAP Avg.	Description CLIP	Description BL-4	Description CI-D	Description BLRT	Description EMS	Description Acc_L	Description tIoU	Err.↓
Proprietary MLLMs																			
GPT 5.6 Luna	2026-08	38.32	66.55	23.31	46.60	23.18	11.40	25.68	26.09	**12.41**	**10.06**	**4.73**	2.07	**5.44**	64.58	1.41	1.66	44.40	24.40	25.21	**34.84**	**0.00**
Open-source Large MLLMs																			
DeepSeek v4.1-Flash	2026-08	**38.82**	51.59	6.30	12.20	6.09	2.88	6.67	6.79	2.60	3.61	1.32	0.69	1.84	73.47	1.47	1.18	44.41	27.32	16.47	29.65	**0.00**
GLM-5.3-Flash	2026-08	29.49	63.78	23.80	**49.30**	24.18	11.04	**26.65**	**27.15**	8.24	7.27	3.39	1.36	3.83	72.60	0.83	1.62	41.11	27.28	24.32	34.10	**0.00**
Qwen3.8-Max	2026-08	37.99	**66.62**	**24.46**	46.60	**24.19**	**12.01**	26.16	26.46	10.54	8.20	4.39	**2.23**	4.59	**78.58**	1.70	**3.02**	**46.30**	29.28	**25.56**	34.78	**0.00**
Open-source Small MLLMs																			
Phi-3.5-Vision-Instruct	2024-08	0.58	3.11	5.04	14.31	3.98	1.30	6.64	7.32	0.23	0.35	0.03	0.00	0.13	0.41	0.00	0.00	0.25	0.15	0.06	0.04	49.45
VideoLLaMA3-7B	2025-01	26.01	56.40	0.02	0.05	0.02	0.02	0.03	0.03	0.19	0.53	0.27	0.08	0.28	2.41	0.00	0.00	1.38	0.91	0.42	0.92	8.97
Phi-4-Multimodal-Instruct	2025-02	26.75	44.01	6.71	21.69	6.86	3.35	11.00	12.14	0.26	0.27	0.22	0.03	0.18	60.97	0.40	0.61	32.83	22.76	4.21	16.14	4.94
InternVL3-8B	2025-04	30.82	50.51	11.51	32.02	11.82	5.51	16.48	17.75	0.97	2.19	0.66	0.07	1.11	65.09	0.12	0.34	39.41	24.96	11.70	21.92	0.64
Kimi-VL-A3B-Instruct	2025-04	11.36	30.06	4.16	12.84	4.21	2.19	6.38	6.85	0.58	0.50	0.12	0.02	0.20	9.81	0.00	0.09	6.04	3.67	2.10	2.78	34.64
Mistral-Small-3.2-24B	2025-06	31.11	52.06	13.59	35.69	13.67	7.05	18.68	19.78	1.41	1.87	0.48	0.20	0.88	77.40	**1.72**	1.49	**46.30**	28.89	10.18	26.31	2.40
InternVL3.5-14B	2025-08	34.63	47.46	8.08	22.93	8.17	4.27	11.61	12.42	0.97	1.46	0.45	0.31	0.72	68.80	0.55	0.88	40.03	26.05	9.74	21.64	1.65
InternVL3.5-2B	2025-08	19.66	16.52	3.15	11.43	3.19	1.57	5.71	6.47	0.18	0.17	0.03	0.00	0.07	9.99	0.00	0.00	4.01	3.86	0.69	3.95	61.35
InternVL3.5-38B	2025-08	28.91	30.94	0.31	1.00	0.30	0.14	0.46	0.51	0.05	0.43	0.01	0.00	0.12	19.14	0.01	0.19	12.39	7.30	2.42	6.59	4.87
InternVL3.5-4B	2025-08	1.82	10.22	1.31	3.45	1.32	0.63	1.78	1.91	0.44	0.76	0.19	0.01	0.33	45.59	0.09	0.51	24.63	17.28	3.83	15.30	36.16
InternVL3.5-8B	2025-08	12.94	25.39	6.34	18.73	6.28	2.87	9.23	9.98	0.82	1.04	0.22	0.01	0.49	56.17	0.12	0.52	33.69	21.57	10.01	17.41	15.51
InternVL3.5-GPT-20B-A4B	2025-08	36.46	44.01	4.65	11.93	4.68	2.30	6.19	6.67	0.70	1.08	0.27	0.15	0.53	65.96	0.21	0.42	36.75	24.99	9.42	22.02	2.47
LLaVA-OV-1.5-4B-Instruct	2025-09	18.04	52.88	0.63	2.25	0.66	0.14	1.07	1.22	0.11	0.11	0.00	0.00	0.05	20.81	0.00	0.00	13.12	7.95	2.06	7.75	1.27
Qwen3-VL-2B-Instruct	2025-10	6.80	13.88	8.03	22.74	8.33	4.26	11.65	12.30	0.65	1.11	0.28	0.10	0.51	8.82	0.00	0.08	4.97	3.33	0.89	1.96	27.83
Qwen3-VL-4B-Instruct	2025-10	31.81	47.94	13.78	37.25	13.98	6.89	19.66	21.37	1.74	2.32	1.44	0.17	1.20	65.49	0.34	0.59	36.30	24.55	7.66	14.91	2.57
GLM-4.6V-Flash	2025-12	30.28	54.91	12.80	30.71	12.93	5.28	15.70	16.24	1.01	1.50	0.34	0.27	0.73	39.15	0.07	0.62	25.95	14.87	9.01	21.22	24.98
Ministral-3-14B-Instruct	2025-12	35.05	44.82	5.44	10.88	5.36	2.91	6.08	6.25	0.77	1.31	0.40	0.05	0.62	21.65	0.55	0.17	15.69	8.16	1.60	11.44	0.12
Ministral-3-3B-Instruct	2025-12	19.20	47.19	5.88	14.97	6.01	3.02	7.87	8.27	1.25	0.88	0.30	0.13	0.47	15.57	0.04	0.13	11.05	5.92	0.78	8.12	3.05
Ministral-3-8B-Instruct	2025-12	35.63	47.33	6.73	15.72	6.70	3.32	8.35	8.83	1.83	2.18	0.51	0.06	0.99	64.67	1.15	0.68	41.66	23.93	3.88	18.82	0.93
MiniCPM-o-4.5	2026-02	22.02	52.13	3.89	11.23	3.74	1.77	5.65	6.21	2.08	1.70	0.48	0.01	0.74	56.72	0.36	0.62	34.23	21.28	11.60	20.79	4.72
Qwen3.5-27B	2026-02	34.59	65.47	8.83	16.47	8.75	4.95	9.47	9.55	3.51	3.12	1.84	1.53	2.07	33.95	0.57	0.83	22.18	12.82	8.62	19.30	0.12
Qwen3.5-35B-A3B	2026-02	33.64	56.13	1.88	3.43	1.88	0.99	2.01	2.05	0.60	0.72	0.20	0.14	0.33	69.49	0.69	0.67	39.36	26.34	15.17	27.67	0.34
Qwen3.5-4B	2026-02	30.65	64.12	3.65	6.89	3.74	1.94	3.93	4.01	2.14	2.70	1.43	0.81	1.48	52.25	0.26	0.55	26.63	19.68	7.72	16.38	0.79
Qwen3.5-9B	2026-02	28.87	61.88	17.23	40.54	17.40	8.68	21.62	22.56	4.86	5.33	2.83	1.74	2.94	72.58	0.93	0.74	41.59	27.47	13.71	24.02	1.19
Molmo2-4B	2026-03	0.00	0.41	4.81	12.86	4.85	2.54	6.72	7.13	0.24	0.70	0.41	0.18	0.41	18.24	0.00	0.22	11.22	6.98	1.50	5.77	76.67
MiniCPM-V-4.6	2026-04	6.35	13.00	3.39	10.48	3.16	1.44	5.29	5.91	0.21	0.19	0.12	0.00	0.10	38.31	0.04	0.35	25.05	14.38	4.84	8.70	42.61
Molmo2-8B	2026-04	18.08	64.05	1.87	4.57	1.89	0.94	2.35	2.40	0.80	1.09	0.35	0.03	0.55	2.50	0.00	0.01	1.57	0.98	0.49	1.60	7.43
Qwen3.6-35B-A3B	2026-04	37.04	56.26	17.93	35.61	17.27	8.64	19.71	20.31	6.83	4.86	2.27	1.27	2.63	77.80	0.85	0.87	43.32	29.32	21.83	31.26	1.87
InternVideo3-8B-Instruct	2026-05	0.91	26.88	4.59	14.00	4.56	2.32	7.14	7.91	0.03	0.37	0.00	0.00	0.21	59.47	0.57	0.71	35.20	22.38	4.13	16.63	23.84
LLaVA-OV-2-8B-Instruct	2026-05	35.59	55.92	3.25	5.64	3.30	1.86	3.37	3.37	3.88	3.06	0.82	0.26	1.29	5.77	0.00	0.03	3.57	2.21	0.31	4.18	7.53
VideoChat3-4B	2026-07	31.81	53.89	14.92	37.75	15.50	6.67	19.70	20.95	2.13	2.81	1.82	0.05	1.47	4.00	0.00	0.06	2.35	1.54	0.28	1.75	2.88
Qwen3.8-27B	2026-08	38.37	52.47	16.41	33.15	16.24	8.21	18.31	18.79	7.08	5.12	2.52	1.04	2.84	78.16	1.26	1.44	45.30	**29.33**	21.21	32.65	0.07

MiniCPM-V-4.6 Yu et al. (2026); Mistral-Small-3.2-24B, Ministral-3-3B-Instruct, Ministral-3-8B-Instruct, and Ministral-3-14B-Instruct (Liu et al., 2026); Molmo2-4B and Molmo2-8B (Clark et al., 2026); and VideoChat3-4B (Li et al., 2026b). These models span both dense and mixture-of-experts architectures and vary substantially in model scale and release date. Following common Video-LLM evaluation practice, we evaluate all models in a zero-shot setting. Each model receives either a video or frames sampled from the video, depending on its supported input interface, together with the corresponding query and output instructions. We evaluate Recognition, Temporal Reasoning, Grounding, Detection, and Description using the metrics defined in Sec. 3.4. For compact presentation, Tab. 2 reports Grounding R@1 and Detection mAP at tIoU thresholds of 0.2, 0.5, and 0.7, together with their averages over thresholds from 0.1 to 0.9.

4.2 From Recognition to Description: Overall Results

Tab. 2 reports the performance of 36 MLLMs across the five tasks and all 12 subtask results are reported in Sec. G. No model performs best throughout the complete understanding chain. DeepSeek v4.1-Flash achieves the highest Recognition accuracy, while Qwen3.8-Max leads Temporal Reasoning and Grounding (F1@0.5). GLM-5.3-Flash obtains the strongest top-1 localization results. GPT 5.6 Luna performs best on Detection. Description results are distributed across several models. A clear gap remains between recognizing or reasoning about actions and recovering their precise temporal evidence. Model scale also does not yield consistent improvements: several smaller models remain competitive with larger models on individual tasks, while models strong in Recognition or Temporal Reasoning may perform poorly on Grounding and Detection. These results indicate that the five tasks capture distinct aspects of micro-action understanding and cannot be represented reliably by performance on a single task.

4.3 Beyond Action Names: Label-Free Understanding

Label queries refer to an action using its annotated label. In practice, users may not know the exact label and instead describe what they see in everyday language. MMA-MTU includes 3,920 label-free queries paired with label queries from the same videos and semantic facts. For instance,

Table 3: Comparison between label (**L**) and label-free (**LF**) queries. Small values beside LF results denote the performance difference LF − L.

Model	Rec. Acc.		Temp. Acc.		Grounding F1@.5		Grounding R@1 @.5		Grounding R@1 Avg.		Grounding mIoU		Detection F1@.5		Detection mAP @.5		Detection mAP Avg.	
	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF
GLM-5.3-Flash	27.35	30.64 +3.29	61.72	65.90 +4.18	24.76	24.49 -0.27	**24.68**	24.34 -0.34	**27.56**	26.06 -1.51	**28.10**	26.52 -1.58	7.25	9.28 +2.03	4.50	3.97 -0.53	5.47	3.99 -1.48
Qwen3.8-Max	**36.74**	**38.06** +1.32	**68.41**	65.90 -2.51	**24.99**	**25.58** +0.59	24.25	**25.01** +0.76	26.57	**26.57** 0.00	26.91	**26.78** -0.13	**11.89**	**11.73** -0.16	**6.10**	**4.96** -1.14	**6.78**	**5.98** -0.80
Phi-4-Multimodal-Instruct	24.38	26.69 +2.31	43.10	49.58 +6.48	4.84	9.62 +4.78	4.89	9.68 +4.79	8.13	16.12 +7.98	8.96	17.88 +8.92	0.32	0.13 -0.19	1.68	0.13 -1.55	1.06	0.25 -0.81
InternVL3.5-14B	30.81	32.45 +1.64	46.44	50.42 +3.98	8.05	8.89 +0.84	8.00	8.86 +0.86	11.09	12.07 +0.98	11.89	12.95 +1.06	1.15	0.53 -0.62	2.10	0.37 -1.73	2.47	1.10 -1.37
Ministral-3-14B-Instruct	30.81	32.45 +1.64	42.26	42.26 0.00	5.12	6.25 +1.13	5.13	6.09 +0.96	5.60	6.46 +0.86	5.74	6.59 +0.85	0.81	1.07 +0.26	2.17	1.28 -0.89	1.99	1.58 -0.41
Qwen3.5-4B	27.84	29.98 +2.14	62.34	64.44 +2.10	4.30	2.85 -1.45	4.26	2.78 -1.48	4.28	2.93 -1.35	4.33	3.00 -1.33	1.91	1.88 -0.03	3.82	1.83 -1.99	3.66	1.71 -1.95
Molmo2-8B	19.28	9.39 -9.89	67.15	**66.11** -1.04	2.09	2.38 +0.29	2.11	2.35 +0.24	2.42	2.93 +0.51	2.44	3.00 +0.56	0.95	0.97 +0.02	0.77	0.77 0.00	1.41	1.04 -0.37
LLaVA-OV-1.5-4B	16.47	20.10 +3.63	52.30	50.84 -1.46	0.67	0.37 -0.30	0.67	0.38 -0.29	1.01	0.76 -0.26	1.15	0.89 -0.26	0.00	0.00 0.00	0.00	0.00 0.00	0.05	0.05 0.00
MiniCPM-o-4.5	25.86	21.42 -4.44	49.16	55.02 +5.86	5.96	1.39 -4.57	5.80	1.34 -4.46	8.18	1.66 -6.53	8.96	1.80 -7.16	1.74	1.57 -0.17	1.69	2.22 +0.53	2.08	1.98 -0.10
InternVideo3-8B-Instruct	0.33	0.00 -0.33	33.89	13.60 -20.29	5.31	0.79 -4.52	5.13	0.81 -4.32	8.75	1.40 -7.35	9.76	1.60 -8.16	0.10	0.00 -0.10	0.01	0.00 -0.01	0.43	0.26 -0.17
VideoChat3-4B	31.30	34.27 +2.97	48.12	55.02 +6.90	15.52	16.32 +0.80	15.91	16.63 +0.72	20.16	20.37 +0.22	21.49	21.45 -0.04	2.60	1.83 -0.77	1.44	1.15 -0.29	2.42	1.31 -1.11

Table 4: Effect of supervised fine-tuning on MMA-MTU. The Δ row reports SFT minus Base.

Model	Rec. Acc.	Temp. Acc.	Grounding R@1 @.2	Grounding R@1 @.5	Grounding R@1 @.7	Grounding R@1 Avg.	Grounding mIoU	Detection F1@.5	Detection mAP @.2	Detection mAP @.5	Detection mAP @.7	Detection mAP Avg.	Description CLIP	Description BL-4	Description CI-D	Description BLRT	Description EMS	Description Acc_L	Description tIoU	Err.↓
Qwen3.5-4B	30.65	64.12	3.65	6.89	3.74	1.94	3.93	4.01	2.14	2.70	1.43	**0.81**	1.48	52.25	0.26	0.55	26.63	19.68	7.72	16.38	0.79
Qwen3.5-4B-SFT	**59.98**	**86.46**	**27.97**	**53.30**	**29.00**	**13.39**	**30.27**	**30.49**	**13.03**	**4.83**	**1.95**	**0.81**	**2.41**	**60.36**	**32.55**	**62.20**	**59.56**	**23.60**	**38.51**	**48.12**	**0.13**
Δ	+29.33	+22.34	+24.32	+46.41	+25.26	+11.45	+26.34	+26.48	+10.89	+2.13	+0.52	0.00	+0.93	+8.11	+32.29	+61.65	+32.93	+3.92	+30.79	+31.74	-0.66

"Retracting legs" may be described as "pulling the legs back toward the body." Tab. 3 compares the two query forms. Eight of the eleven models improve on Recognition with label-free queries, and six improve on Temporal Reasoning. Six models also obtain higher Grounding performance. The added movement details appear to help models relate a query to the observed action, particularly when a short label alone does not clearly express how the action is performed. Detection shows a different trend: average mAP decreases for ten models and remains unchanged for one. Unlike Recognition and Grounding, Detection requires the model to recover all relevant events under a given constraint. A less precise description can blur that constraint and introduce irrelevant motion cues, leading to missed or incorrectly matched events.

5 More Analysis

5.1 Learning Micro-Actions with Supervised Fine-tuning

As introduced in Sec. 1.3, MMA-MTU provides dedicated training and validation sets for improving performance on MMA-MTU. We fine-tune Qwen3.5-4B with QLoRA Dettmers et al. (2023) while keeping its visual encoder frozen. The complete training configuration is provided in Sec. H. As shown in Tab. 4, supervised fine-tuning produces improvements across all five tasks, showing that MMA-MTU helps the model connect fine-grained action semantics with temporal evidence. Detection also improves, although its mAP indicates that recovering complete event sets with precise boundaries remains difficult.

5.2 What Do Models Actually See?

Fig. 3 provides a qualitative comparison on the Description task. Most models capture the coarse progression of the video and produce fluent descriptions of hand-arm or hand-head interactions. However, their predictions often remain at the level of broad body regions. For example, the ground truth identifies "Touching nose" from 6.80 to 9.35 seconds, whereas the models variously interpret the same movement as "Covering mouth", "Rubbing lip", or "Rubbing eyes". Similar confusion occurs among pointing, gestures, and other hand movements. Thus, observing that a hand approaches the head is relatively easy, but resolving the precise contact target and assigning the corresponding fine-grained action remains difficult. Supervised fine-tuning yields a more structured output, but does not eliminate fine-grained category and boundary errors. These results show that linguistic fluency alone does not imply an accurate account of what happens in the video: successful Description requires the model to jointly distinguish micro-actions, recover all relevant events, and align them with their temporal boundaries. More qualitative examples are provided in Sec. I.

6 Future Exploration: From Micro-Actions to Emotion

Micro-actions have long been regarded as behavioral cues for understanding human affect and intention Guo et al. (2024); Hao et al. (2026). However, most existing studies focus on recognizing and localizing micro-actions rather than examining their value for downstream behavioral understanding. As a preliminary exploration, we study whether the predictions produced on MMA-MTU can assist emotion recognition, particularly when facial information is unavailable for privacy protection.

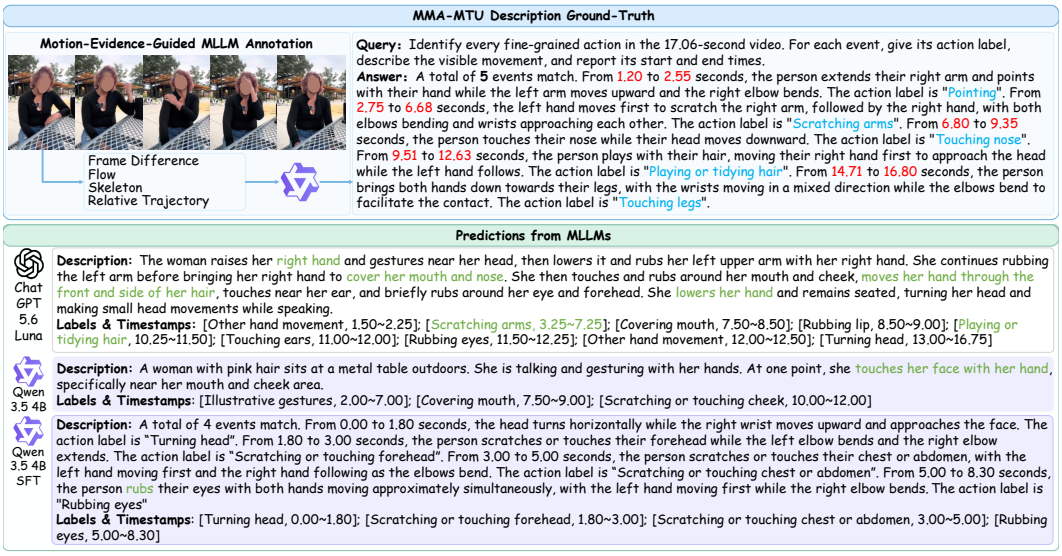


Figure 3: **Qualitative comparison on the Description task of MMA-MTU.** The top shows the ground-truth events and their descriptions, while the bottom compares predictions from different MLLMs.

We report experiments on 80 videos with annotations for eight emotion categories in MMA-MTU and evaluate 6 MLLMs under five input settings. Additional results are reported in Sec. J. "RGB (Mask)" uses RGB frames with the face masked, while "RGB" uses the complete frames. "Pre." uses only the predicted information in MMA-MTU. "RGB (Mask) + Pre." and "RGB + Pre." combine this prediction information with masked and complete RGB frames, respectively. The prediction information is model predictions rather than ground-truth annotations.

As shown in Figure 4, most models adding predicted information to complete RGB improve by 1.25 to 5.00 points. The effect is more consistent when facial information is removed. These preliminary results suggest that micro-actions provide emotion-related evidence beyond facial appearance. Their usefulness still depends on prediction quality, but the consistent gains under facial masking point to a promising role for micro-action understanding in privacy-sensitive emotion analysis.

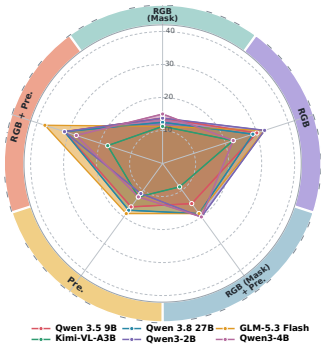


Figure 4: Emotion recognition accuracy.

7 Conclusion

We introduced MMA-MTU, a unified benchmark for micro-action Recognition, Temporal Reasoning, Grounding, Detection, and Description. It contains 11,180 untrimmed videos, 39,743 action instances, and 122,845 queries spanning 82 fine-grained categories. The five tasks share the same video evidence and temporal annotations, following the progression from identifying an action to relating, locating, recovering, and describing micro-actions. MMA-MTU further includes 34,876 label-free queries that describe observable movements without directly providing their action labels, together with fine-grained ground-truth descriptions. Our evaluation of 36 MLLMs reveals markedly different strengths across the five tasks. Temporal reasoning is substantially easier than recovering precise temporal evidence, while grounding and detection remain major bottlenecks. Performance also varies across recording domains and between label and label-free queries, showing that recognizing an action name does not necessarily imply robust understanding of how that action unfolds in a video. MMA-MTU provides a common training and evaluation framework for developing MLLMs that move beyond isolated action recognition toward complete micro-action understanding.

AI USE STATEMENT

In this work, generative AI tools were used to assist with translation, literature and information search, interpretation and presentation of experimental results, and manuscript editing. MLLMs were also incorporated into the benchmark-construction pipeline: Qwen3.5-27B generated observable-motion expressions for label-free queries, Qwen3.8-27B generated motion-evidence-guided event descriptions, and an MLLM reviewer supported annotation quality control. All generated annotations underwent automatic validation and human review and correction. We checked numerical and experimental claims against the original results and verified literature-related statements against their sources. The authors made the final decisions on all data, analyses, scientific claims, conclusions, and manuscript content and take full responsibility for the final work.

REPRODUCIBILITY STATEMENT

We have made substantial efforts to ensure that the construction, evaluation, and analysis of MMA-MTU are reproducible. The main paper describes the benchmark formulation, dataset construction, task definitions, evaluation protocol, model evaluation settings, and supervised fine-tuning experiments, while the supplementary material provides additional details required for reproduction. Specifically, the Appendix includes detailed procedures and statistics for QA generation, complete definitions and implementation details of the evaluation metrics, the prompts used for different tasks, additional analyses on input settings, NO-MATCH behavior, domain variation, and label-free queries, full results and definitions for all 12 subtasks, detailed supervised fine-tuning configurations, additional qualitative examples, further downstream exploration results, and a discussion of the limitations of the benchmark. Together, these materials document the data construction, inference, evaluation, and training procedures used throughout the paper. Following the completion of the double-blind review process, we will publicly release the MMA-MTU benchmark data and annotations in accordance with the redistribution and usage restrictions of the underlying MMA-82 data, with the goal of facilitating independent evaluation and future research.

ETHICS STATEMENT

MMA-MTU is constructed entirely from the videos and temporal annotations of MMA-82 and is intended for academic research on fine-grained micro-action and video understanding. We therefore follow the data collection, privacy, and release principles of the original dataset. For laboratory interview videos, informed consent was obtained from the recruited participants. For videos collected from publicly available online sources, including psychiatric interviews, street interviews, and emotion-rich television programs, the original dataset followed the corresponding platform policies and used the videos for research purposes. The source data were screened to exclude clearly identifiable minors, inappropriate content, and samples unsuitable for academic use.

Our benchmark annotations focus on observable human movements and their temporal structure. MMA-MTU does not collect additional private personal information or introduce identity-related, clinical, or psychological labels. The newly constructed label-free queries and motion descriptions describe only visually observable behavior, such as body parts involved, motion direction, temporal order, and interactions between body parts. Although we additionally explore whether predicted micro-action information can assist emotion recognition, this experiment is intended only as an exploratory study of the relationship between observable movement cues and affective understanding; it should not be interpreted as supporting clinical diagnosis, psychological profiling, or other high-stakes judgments about individuals.

We also recognize that advances in fine-grained human-behavior understanding have potential dual-use implications. Models capable of recognizing and temporally localizing subtle movements could potentially be applied to surveillance, behavioral profiling, or other privacy-sensitive settings. MMA-MTU is designed as a research benchmark for evaluating model capabilities, rather than for identifying individuals or making decisions about them. We encourage researchers to consider consent, privacy, context, and potential harm when transferring methods developed on this benchmark to real-world applications.

Finally, we respect the copyright and redistribution conditions of the source videos. When raw videos cannot be redistributed because of copyright or platform restrictions, we release benchmark annotations and source references rather than the restricted video content itself. By providing the benchmark construction, annotations, and evaluation protocol, we aim to support reproducible research while preserving the usage constraints of the underlying data.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Gary Bradski, Adrian Kaehler, et al. Opencv. *Dr. Dobb's journal of software tools*, 3(2):1–81, 2000.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

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

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. *arXiv preprint arXiv:2103.16002*, 2021.

Martin Grunwald, Thomas Weiss, Stephanie Mueller, and Lysann Rall. Eeg changes caused by spontaneous facial self-touch may represent emotion regulating processes and working memory maintenance. *brain research*, 1557:111–126, 2014.

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. doi: 10.1109/TCSVT.2024.3358415.

Yanbin Hao, Pengyu Liu, Xing Wei, Xun Yang, Dan Guo, and Meng Wang. A new multi-domain benchmark for micro-action recognition and detection. *arXiv preprint arXiv:2606.14096*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460, 2025a. doi: 10.1109/CVPR52734.2025.00791.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025b.

Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13225–13236. IEEE, 2025.

Kun Li, Jihao Gu, Fei Wang, Zhiliang Wu, Hehe Fan, and Dan Guo. Ma-bench: Towards fine-grained micro-action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20118–20128, June 2026a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024b.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10626–10637. IEEE, 2021.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Stephanie Margarete Mueller, Sven Martin, and Martin Grunwald. Self-touch: Contact durations and point of touch of spontaneous facial self-touches differ depending on cognitive and emotional load. *PloS one*, 14(3):e0213677, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7881–7892, 2020.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17908–17917. IEEE, 2022.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024a.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024b.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024c.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Overview of Appendix

This appendix provides more details of the MMA-MTU. The topics are organized as follows:

- B: Task definition, generation details, and additional dataset statistics.
- C: Evaluation metrics.
- D: Prompt details.
- E: More results: additional ablations, domain, and label-free results.
- F: Do MLLMs really look at micro-actions?
- G: Results for the 12 subtasks.
- H: Supervised fine-tuning configuration and training details.
- I: Additional qualitative examples of model predictions.
- J: Further exploration of micro-actions for emotion recognition.
- K: Limitations of MMA-MTU.
- L: Future work.

B More Details of MMA-MTU QAs Generation and Statistics

B.1 Subtasks Definition

In Sec.3.2, MMA-MTU divides the 5 tasks into 12 subtasks according to the query condition and temporal scope. Together, they follow the process of observing a video: recognizing the actions, reasoning about their temporal relations, locating particular events, recovering all relevant events, and describing how they occur. Representative queries and answers are shown below.

Recognition. Recognition asks the model to identify a micro-action through a single-choice query. Video-Level Micro-Action Recognition (**VLMR**) identifies the fine-grained action visible in a video or a given interval. Its candidates may include visually similar actions and "None of the above". Coarse-Grained Category-Constrained Recognition (**CGMR**) additionally specifies a body-level category, such as "Head" or "Upper Limb", and asks the model to distinguish among fine-grained actions within that category.

Recognition Examples

Query (VLMR). Between 0.0 and 2.43 seconds, which option best describes the visible movement?
Options: (A) fingers rub or brush a lip; (B) a hand adjusts, arranges, twirls, or handles the hair; (C) None of the above; (D) a hand contacts or scratches the philtrum; ✓ (E) fingers rub one or both eyes.

Query (CGMR). Recognize the "Head"-family event supplied at $[3.97, 5.50]$ s; select its correct fine-grained label.
Options: ✓ (A) Shaking head; (B) Bowing head; (C) None of the above; (D) Compound head movement; (E) Semantic head movements.

Temporal Reasoning. Temporal Reasoning examines how identified actions relate over time. Order Temporal Reasoning (**ORTR**) considers two actions and asks whether they follow a specified temporal order, such as whether one action occurs before the other. Overlap-Aware Temporal Reasoning (**OATR**) considers up to three actions and asks whether they temporally overlap or co-occur. Both subtasks use binary choices and may express the actions through either labels or label-free movement descriptions.

Temporal Reasoning Examples

Query (ORTR). Does the event when a hand contacts the nose happen before the event when fingers curl tightly into the palm?
Options: (A) Yes; ✓ (B) No.

Query (OATR). Do the periods when the head turns and when the arms extend overlap?
Options: ✓ (A) Yes; (B) No.

Grounding. Grounding takes an action label or label-free movement description as input and returns all temporal intervals that satisfy the query. Full-Video Micro-Action Grounding (**FVMG**) searches the complete video for every occurrence of one action. Time-Constrained Micro-Action Grounding (**TCMG**) restricts the search to a specified temporal window. Co-Occurrence Micro-Action Grounding (**COMG**) localizes the intervals in which two actions overlap for a specified minimum duration. A grounding output may contain one interval, multiple intervals, or "No_Match".

Grounding Examples

Query (FVMG). What are the start time and end time in seconds for every instance where the legs extend outward, or "No_Match" if there are none?
Target: $[[1.90, 2.53]]$ seconds.

Query (TCMG). Restrict attention to $[0.00, 4.39]$ s. Where does "Turning head" occur? Return all matching intervals, or "No_Match".
Target: $[[0.00, 1.67], [1.70, 2.64]]$ seconds.

Query (COMG). Find all intervals where the unique "Spread legs" and "Closing legs" events overlap for at least 0.10 seconds, or "No_Match".
Target: "No_Match".

Detection. Detection recovers a set of micro-action events rather than locating one queried action. Each detected event contains an action label, start time, and end time. Full-Video Micro-Action Detection (**FVMD**) detects all fine-grained actions in the complete video. Constrained Micro-Action Detection (**COMD**) limits detection to a specified temporal window, body-level category, or candidate-label set.

Detection Examples

Query (FVMD). Return all fine action events visible in the complete video as label/start/end records.
Target: Rubbing eyes: $[0.00, 2.43]$; Curling legs: $[0.00, 27.27]$; Rubbing hands: $[6.87, 22.63]$; Waving: $[9.43, 11.41]$.

Query (COMD). Perform fine-event detection over the whole video, restricted to the "Head" coarse family.
Target: Nodding: $[2.85, 7.81]$; Shaking head: $[17.45, 19.10]$.

Description. Description reports the action category and temporal boundary together with a natural-language account of the visible movement. Full-Video Micro-Action Description (**FVDS**) identifies and describes all micro-actions in the complete video. Time-Constrained Micro-Action Description (**TCDS**) describes every event that overlaps a specified temporal window. Category-Constrained Micro-Action Description (**CCDS**) describes every event belonging to a specified body-level category or candidate-label set.

Description Examples

Query (FVDS). Find the complete set of fine-grained actions in the 9.98-second video. Give each event's action label, describe what the person visibly does, and include the start and end times.
Target: A total of 4 events match. From 0.74 to 1.64 seconds, the person waves with both hands moving vertically, with the left shoulder initiating the motion while the right elbow bends. The action label is "Waving". From 2.34 to 3.07 seconds, the person holds their hands together while moving their upper limbs horizontally. The action label is "Holding hands". From 3.07 to 4.54 seconds, the person pats their legs with both hands moving downward in approximately simultaneous motion. The action label is "Patting legs". From 6.37 to 7.80 seconds, the person lowers their body while moving their left hand first and then their right hand to touch their legs. The action label is "Touching legs".

Query (TCDS). Which fine-grained actions overlap $[0.72, 3.71]$ seconds? For each matching event, report the action label, the visible movement, and its start and end times.
Target: A total of 2 events match. From 0.76 to 2.26 seconds, the right hand extends and moves forward first, followed by the left hand, as both arms articulate to point. The action label is "Pointing". From 2.38 to 3.68 seconds, the hands approach each other and interlace their fingers. The action label is "Interlace fingers".

Query (CCDS). Find every action in the 16.68-second video whose label belongs to {"Closing legs", "Spread legs"}. For each event, give its action label, describe the visible movement, and report its start and end times.
Target: A total of 2 events match. From 1.39 to 4.13 seconds, the legs move apart in a spreading motion, with the right knee and ankle initiating the separation before the left leg follows. The action label is "Spread legs". From 11.66 to 14.00 seconds, the legs move inward from a wide stance, with the knees and ankles approaching each other as the feet come together. The action label is "Closing legs".

B.2 Details of Template QAs Generation

MMA-MTU constructs QA pairs from the fine-grained temporal annotations of MMA-82-Det Hao et al. (2026). We first normalize each annotation as an event containing a fine-grained action label, its body-level category, and its start and end times.

We implement a separate template generator for each of the five tasks. Given the task and its conditions, the corresponding generator fills predefined natural-language templates with action labels, candidate sets, temporal windows, or event relations. Recognition uses a five-way single-choice format. Each query contains four action candidates and one "None of the above" option, and their positions are randomly permuted to ensure randomness. For positive queries, the four action candidates contain the target label and three distractors; for negative queries, all four are distractors and "None of the above" is the target. The distractors are selected from visually similar actions, so recognizing the target requires distinguishing fine-grained motion rather than eliminating unrelated categories. For example, a query whose ground-truth action is "Nodding" may include "Shaking head" as a distractor. Temporal Reasoning constructs propositions from the temporal relations between two or three events. Grounding asks for every interval satisfying an action condition. Detection returns the complete set of events within a specified scope, while Description additionally requires a natural-language account of each detected movement. The target outputs are computed directly from the source annotations rather than generated by an MLLM. Label-free variants preserve the same task condition and target output but replace action labels with observable movement descriptions, as described in Sec. 3.3. The boxes below present one fillable template for each of the 12 subtasks. An underline denotes a field supplied during generation.

Recognition

VLMR. Which candidate correctly classifies the fine-grained action occurring between ____ and ____ seconds?
CGMR. Between ____ and ____ seconds, which option best identifies the fine-grained action in the "_____" category?
Options for both subtasks, in randomized order:
(A) _____ (B) _____ (C) _____ (D) _____ (E) _____
Randomly, one slot is filled with "None of the above".
Answer: ___

Temporal Reasoning

ORTR. Does the unique "_____" event occur before the unique "_____" event? Select Yes or No.
OATR. Do the periods of the unique "_____", "_____", and "_____" events overlap in time? Select Yes or No.
Answer: ____

Grounding

FVMG. Locate every occurrence of "_____" in the full video. Return all matching start and end times, or "No_Match" if none.
TCMG. Between ____ and ____ seconds, locate every occurrence of "_____", or return "No_Match".
COMG. Find every interval where the unique "_____" and "_____" events overlap for at least ___ seconds, or return "No_Match".
Answer: [[___, ___], . . .] or "No_Match"

Detection

FVMD. Identify every fine-grained event in the full video. Return each event's label, start time, and end time.
COMD. Within [___, ___] seconds, identify every event satisfying the "_____" category or candidate-label constraint. Return each event's label, start time, and end time, or "No_Match" if none.
Answer: {(_____, ___, ___), . . .} or "No_Match"

Description

FVDS. Which fine-grained actions occur across the ____-second video? For every event, provide its label, visible-movement description, and start and end times.
TCDS. Which fine-grained actions overlap [___, ___] seconds? Describe each visible movement and provide its label and temporal boundary.
CCDS. Which fine-grained actions satisfying the "_____" category or candidate-label constraint occur in the video? Describe each movement and provide its label and temporal boundary.
Answer: {(_____, _____, ___, ___), . . .}

B.3 Label-Free Action Description Generation

Label-free QAs are derived from the structured label QAs rather than generated independently. These descriptions retain only visually supported information. For every structured QA that contains an action-level label, we keep all non-label content fixed and replace the label. The corresponding observable evidence and task contract are then provided to Qwen 3.5 27B. The contract specifies the temporal scope, temporal relation, option and their order, and requested output fields. The MLLM produces a complete natural-language query in which each hidden label is expressed through observable movement. For instance, the label "Rubbing eyes" may be replaced by a phrase such as "fingers rub one or both eyes." The generated query is accepted only when it preserves the semantic fact of its label parent. Automatic validation checks for label leakage, unsupported motion details, altered numerical boundaries, malformed options, missing participants, and reversed temporal relations. Each accepted label-free QA therefore shares the same video evidence, constraints, and target output as its paired label QA, differing only in how the action is expressed.

B.4 Motion-Evidence-Guided Description Annotation

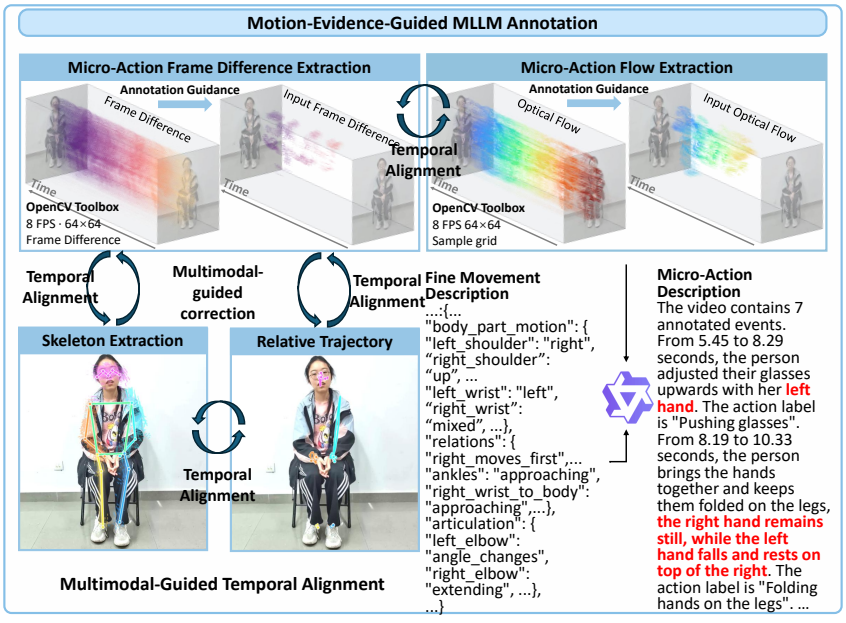


Figure 5: **Motion-Evidence-Guided MLLM Annotation Tool.** Each annotated event is converted into four timestamp-aligned motion views. Qwen3.8-27B uses the resulting 2×2 motion-only video to generate an instance-level micro-action description.

We introduce the **Motion-Evidence-Guided MLLM Annotation Tool**[1], an automated tool for constructing fine-grained ground-truth descriptions of micro-actions. The tool does not reclassify actions, discover temporal intervals, or determine the number of events. These elements are inherited directly from the verified action instances in MMA-82-Det and remain fixed throughout description generation. For a query containing multiple ground-truth events, we first decompose its event set into independent instances defined by the source video, action label, start time, and end time. The label identifies the annotated event and selects its relevant body parts, while the timestamps determine the segment used for motion extraction. Each unique event is processed once, and its description can be reused by every query referring to that event.

As shown in Fig. 5, OpenCV Bradski et al. (2000) computes optical flow and frame differences from the timestamp-aligned segment. YOLOv11 Khanam & Hussain (2024) localizes the target person and extracts the pose landmarks used to construct skeleton motion and body-relative trajectories. We retain only the body regions associated with the annotated action. For example, an upper-body action is represented using motion from the head, arms, and torso rather than unrelated leg movement. The extracted signals are temporally aligned and arranged as a 2×2 motion-only video containing frame differences, optical flow, skeleton motion, and relative trajectories. Qwen3.8-27B Bai et al. (2025a) receives no RGB frames or appearance information. These motion representations can contain estimation noise and temporal jitter. For example, unstable optical flow may make a "Nodding" instance appear to contain lateral head movement. We therefore use the trusted label as a semantic constraint to suppress motion patterns that clearly contradict the verified event, while retaining instance-specific details consistently supported by the motion evidence. The label does not supply unsupported laterality, direction, contact, or temporal order. Generated descriptions are checked against the source action semantics and extracted motion evidence, stored in an event bank, and assembled chronologically into the final references for single- and multi-event Description queries. The complete quality-control process is described in Sec. B.6.

Motion-Evidence-Guided Description Annotation Prompt

```
Write one concise ground-truth description of the annotated action in
this motion-only video.
Trusted canonical action label: "__________".
Annotation-backed semantic reference: __________
Core body parts or objects: __________
Describe the visible mechanics of the action instead of merely
repeating its label. Preserve the action's defining body-part,
contact, direction, or object relation. Add one concrete clip-specific
detail when it is clearly visible.
Ignore unrelated motion. If a motion candidate conflicts with the
trusted label, ignore that candidate. Mention left/right, temporal
order, speed, or magnitude only when the video supports it reliably.
Do not add emotion, intention, identity, cause, or scene information.
Do not mention timestamps, optical flow, skeletons, panels, keypoints,
statistics, annotations, datasets, or raw measurements. Timestamps are
inserted later by the program.
Return valid JSON only:
{"description": "ONE English sentence"}
```

B.5 More Statistics

We also report an interesting statistic. Fig. 6 compares the action frequency of ten subjects with the average frequency of each action in non-emotion videos. The results reveal clear subject-specific differences in the observed action distribution. For example, Subj.6 performs "Pointing" and "Illustrative gestures" more frequently than average, whereas "Turning head" occurs less frequently. Similar deviations appear across the remaining subjects, indicating that micro-action distributions vary considerably across individuals.

[1]The annotation tool will be released after the double-blind review period.

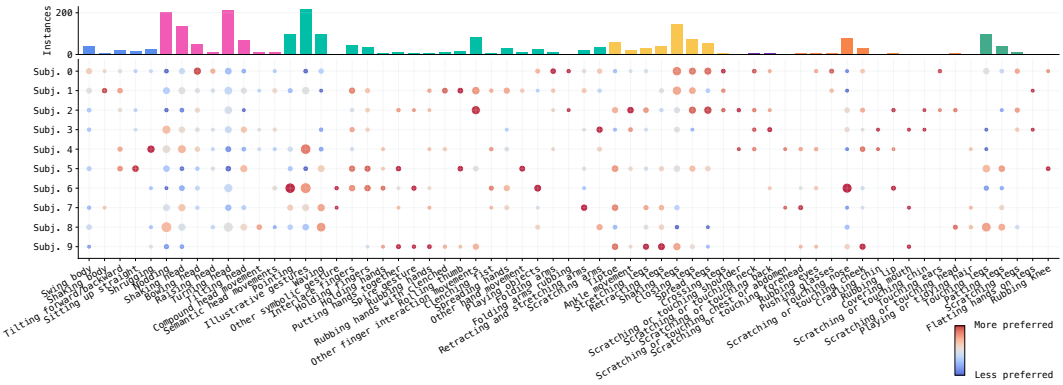


Figure 6: Subject-Action Associations in videos. The larger marker size represents how much a subject preferred or less preferred the action category; marker color represents the relative action preference of each subject.

B.6 Quality Control

All QA pairs undergo automatic validation against the labels and temporal boundaries inherited from MMA-82-Det. Task-specific checks cover option construction, temporal relations, event membership, output structure, and "No_Match" conditions. The 34,876 label-free queries receive an additional paired check for label leakage, altered constraints, unsupported motion details, and consistency with their label-based counterparts. Ground-truth descriptions are further validated at the event level against the corresponding annotated actions; descriptions containing unsupported or incorrect details are rejected, while unresolved cases are forwarded for manual review.

An MLLM reviewer (Qwen3.8-27B) then examines the semantic consistency among each query, its target output, the temporal annotations, and the corresponding video evidence. Human quality control is subsequently conducted by three experienced annotators with substantial experience in micro-action understanding and annotation: one has approximately four years of experience in this area, while the other two have approximately two years each. Before the formal review, the annotators jointly established and calibrated a unified set of review criteria using representative examples and disagreement cases.

Human verification is performed in two complete review rounds over approximately two months. In each round, every QA pair retained in the final benchmark is inspected by all three annotators, resulting in six human inspections per released QA pair. The annotators check and correct ambiguous language, inconsistent targets, incorrect temporal relations, label leakage, and descriptions unsupported by the video evidence. In the first round, 8.7% of the candidate samples are identified for correction, while only 1.2% require further revision in the second round. After the two-stage review process, 99.4% of the pre-release candidate samples satisfy all release criteria. The remaining unresolved or ambiguous cases are excluded before constructing the final benchmark, which contains 122,845 QA pairs.

C Details of Evaluation Setup

C.1 Details of Evaluation Metrics

We evaluate the five tasks using task-specific metrics. Label and label-free queries use the same evaluation protocol.

For Recognition and Temporal Reasoning, we use Accuracy over the predicted option. Let N be the number of queries and y_i and $\hat{y}_i$ denote the target and predicted options. Accuracy is computed as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i). \tag{1}$$

For temporal evaluation, the temporal intersection over union between a predicted interval $p = [s_p, e_p]$ and a ground-truth interval $g = [s_g, e_g]$ is:

$$\mathrm{tIoU}(p, g) = \frac{\max(0, \min(e_p, e_g) - \max(s_p, s_g))}{\max(e_p, e_g) - \min(s_p, s_g)}. \tag{2}$$

For Grounding, we report interval-set $\mathrm{F1@0.5}$, $\mathrm{R@1}$, and $\mathrm{mIoU@1}$, following common video temporal grounding evaluation (Zhang et al., 2026). Predicted and ground-truth intervals are matched one-to-one at a tIoU threshold of 0.5 to compute precision, recall, and F1. For $\mathrm{R@1}$ and $\mathrm{mIoU@1}$, predictions with confidence scores are ranked by confidence; otherwise, their original output order is retained. The first interval is used as the top-1 prediction. We report $\mathrm{R@1}$ at tIoU thresholds from 0.1 to 0.9 and their average, together with the mean top-1 tIoU.

For Detection, we report label-aware event $\mathrm{F1@0.5}$ and mAP, following temporal action detection protocols (Caba Heilbron et al., 2015; Idrees et al., 2017). Event $\mathrm{F1@0.5}$ is computed on queries containing ground-truth events. For mAP, predictions are ranked by confidence within each action category; predictions without confidence scores retain their output order. We report mAP at tIoU thresholds from 0.1 to 0.9 and the average across these thresholds. Unlike Event F1, mAP includes negative queries, and events predicted for such queries are counted as false positives.

We separately report "No_Match" Accuracy for Grounding and Detection. Let Q^- denote the set of queries with no matching ground-truth interval or event:

$$\mathrm{Acc}_{\mathrm{NM}} = \frac{1}{|Q^-|} \sum_{q \in Q^-} \mathbb{I}\Big(\hat{Y}_q = \varnothing\Big). \tag{3}$$

Any predicted interval or event is counted as an incorrect response. Negative queries are excluded from Grounding $\mathrm{F1@0.5}$, $\mathrm{R@1}$, $\mathrm{mIoU@1}$, and Detection Event $\mathrm{F1@0.5}$, but remain part of Detection mAP.

For Description, we evaluate the generated text using BLEU-4, CIDEr-D, BLEURT, CLIPScore, and EMScore (Papineni et al., 2002; Vedantam et al., 2015; Sellam et al., 2020; Hessel et al., 2021; Shi et al., 2022). We additionally report action-label accuracy (Acc_L) and temporal IoU for the structured events contained in the output. Let P and G denote the predicted and ground-truth event sets, respectively, where each event contains an action label and a temporal interval.

Acc_L ignores temporal boundaries and matches the two multisets of action labels:

$$\mathrm{Acc}_L = \frac{\sum_c \min(n_P(c), n_G(c))}{\max(|P|, |G|)}, \tag{4}$$

where $n_P(c)$ and $n_G(c)$ are the numbers of predicted and ground-truth events with category c. A missed or additional event increases the denominator without increasing the number of matches. Repeated predictions of one label cannot receive more matches than the number of corresponding ground-truth events, and an incorrect label receives no credit.

Temporal IoU is computed independently of the action labels. We construct the pairwise tIoU matrix between predicted and ground-truth intervals and use Hungarian assignment to find the maximum-weight one-to-one matching:

$$\mathrm{tIoU}_{\mathrm{Des}} = \frac{\max_{\pi \in \Pi(P,G)} \sum_{(i,j) \in \pi} \mathrm{tIoU}(P_i, G_j)}{\max(|P|, |G|)}, \tag{5}$$

where $\Pi(P, G)$ is the set of one-to-one matchings between P and G. Each event can be matched at most once; unmatched missed or additional events contribute zero through the denominator. For example, when the ground truth contains two events but the model predicts only one, the maximum score is 0.5, even if the predicted interval exactly matches one ground-truth event.

These two metrics provide complementary diagnostics. Acc_L measures event labels and counts, whereas $\mathrm{tIoU}_{\mathrm{Des}}$ measures temporal coverage without considering labels. A prediction may therefore obtain a high temporal score with incorrect labels, or recover the correct labels with inaccurate boundaries. We interpret them jointly rather than treating temporal IoU alone as complete event correctness.

Error Rate measures the percentage of queries whose final output is invalid under the required response structure. This includes malformed outputs and violations of the task-specific schema:

$$\mathrm{ErrorRate} = \frac{N_{\mathrm{invalid}}}{N_{\mathrm{queries}}}. \tag{6}$$

In addition to "No_Match" Accuracy, we report Overall "No_Match" F1 across Grounding and Detection. We treat "No_Match" as the positive class and compute precision and recall over both positive and negative queries:

$$\mathrm{P}_{\mathrm{NM}} = \frac{\mathrm{TP}_{\mathrm{NM}}}{\mathrm{TP}_{\mathrm{NM}} + \mathrm{FP}_{\mathrm{NM}}}, \qquad \mathrm{R}_{\mathrm{NM}} = \frac{\mathrm{TP}_{\mathrm{NM}}}{\mathrm{TP}_{\mathrm{NM}} + \mathrm{FN}_{\mathrm{NM}}}, \tag{7}$$

$$\mathrm{F1}_{\mathrm{NM}} = \frac{2\mathrm{P}_{\mathrm{NM}}\mathrm{R}_{\mathrm{NM}}}{\mathrm{P}_{\mathrm{NM}} + \mathrm{R}_{\mathrm{NM}}}. \tag{8}$$

Here, a true positive is a negative query correctly answered with "No_Match". A false negative occurs when the model predicts an interval or event for a negative query, while a false positive occurs when it returns "No_Match" for a query containing a valid ground-truth interval or event. Task-specific $\mathrm{Acc}_{\mathrm{NM}}$ measures rejection accuracy on negative queries, whereas Overall $\mathrm{F1}_{\mathrm{NM}}$ also penalizes models that indiscriminately reject actions that are present.

C.2 Retry Policy

Each query is first evaluated with a single inference attempt. We set the maximum generation length to 768 tokens for Description and 384 tokens for the other four tasks. The generated response is then checked by a task-specific validator. The validator verifies that the output can be parsed, follows the required field structure, contains valid temporal boundaries, and uses action labels from the predefined category set. When validation fails, the model receives the detected error together with a correction instruction and may regenerate its response up to two times. A query therefore has at most three inference attempts. The retry prompt is provided in Sec. D. It identifies the violated output constraint but does not reveal the correct option, action label, temporal boundary, or any other ground-truth answer. If neither retry produces a valid output, the final prediction is marked invalid, counted as an incorrect answer by the corresponding task metric, and included in Error Rate.

We record two terminal protocol-failure categories. Output-format errors: include unparsable responses, missing or malformed fields, invalid field types, and illegal temporal values. Non-standard action categories: occur when a structured prediction contains a label outside the 82-category action set. A response that satisfies the output protocol but predicts the wrong option, action, interval, or description is not counted as a protocol error; it is evaluated normally by Accuracy, F1, mAP, temporal IoU, or the Description metrics. Error Rate thus measures compliance with the inference interface after correction, while the task metrics measure the semantic quality of valid predictions.

D Prompt Details

All models are evaluated with the same task-specific prompts. Each input consists of the video, its duration, the task and subtask names, the query, and an explicit output schema. The system message instructs the model to use visible video evidence and return a single JSON object without additional explanation. This common structure keeps the inference interface consistent across models while allowing each task to retain its required output format.

Recognition and Temporal Reasoning share a single-choice interface and require the model to copy one option ID from the supplied candidates. Grounding returns all matching temporal intervals, whereas Detection returns action labels, temporal boundaries, and confidence scores. Both tasks support "No_Match" when no event satisfies the query and instruct the model to preserve complete continuous events rather than fragmenting them into fixed-length segments. Description jointly produces a natural-language account and a structured event list containing action labels and temporal boundaries. Detection and Description receive the complete set of 82 fine-grained labels at inference time to constrain their structured outputs.

Recognition Prompt

```
System: You are a precise benchmark answerer. Output JSON only.
User: <video>
You are answering one MMA-MTU micro-action benchmark query.
Use only visible video evidence. Return exactly one JSON object, with
no explanation or Markdown.
Required schema: a JSON object with option_ids (array of selected
option-ID strings); copy each selected ID exactly from the provided
options.
Video duration: {DURATION} seconds.
Task: Recognition / {SUBTASK}.
Question:
{QUESTION}
Options:
A. {OPTION_A}
B. {OPTION_B}
...
Keep option IDs exactly as shown.
```

Temporal Reasoning Prompt

```
System: You are a precise benchmark answerer. Output JSON only.
User: <video>
You are answering one MMA-MTU micro-action benchmark query.
Use only visible video evidence. Return exactly one JSON object, with
no explanation or Markdown.
Required schema: a JSON object with option_ids (array of selected
option-ID strings); copy each selected ID exactly from the provided
options.
Video duration: {DURATION} seconds.
Task: Temporal Reasoning / {SUBTASK}.
Question:
{QUESTION}
Options:
A. {OPTION_A}
B. {OPTION_B}
...
Keep option IDs exactly as shown.
```

Grounding Prompt

```
System: You are a precise benchmark answerer. Output JSON only.
User: <video>
You are answering one MMA-MTU micro-action benchmark query.
Use only visible video evidence. Return exactly one JSON object, with
no explanation or Markdown.
Required schema: either a JSON object containing only no_match
(boolean true), or a JSON object with segments (array of objects); each
segment object must contain start_time (number in seconds) and end_time
(number in seconds).
Video duration: {DURATION} seconds.
Task: Grounding / {SUBTASK}.
Question:
{QUESTION}
If no matching event is visible, return only a JSON object with
no_match set to boolean true; an empty result array is also accepted
as a no-match prediction.
Report each continuous action event exactly once, using its full start
and end times.
Do not divide an event into frames, one-second intervals, or other
fixed-length segments.
```

```
Merge adjacent intervals when they belong to the same continuous
action.
```

Detection Prompt

```
System: You are a precise benchmark answerer. Output JSON only.
User: <video>
You are answering one MMA-MTU micro-action benchmark query.
Use only visible video evidence. Return exactly one JSON object, with
no explanation or Markdown.
Required schema: either a JSON object containing only no_match
(boolean true), or a JSON object with detections (array of objects);
each detection object must contain label (string), start_time (number
in seconds), end_time (number in seconds), and score (number from zero
through one).
Video duration: {DURATION} seconds.
Task: Detection / {SUBTASK}.
Question:
{QUESTION}
If no matching event is visible, return only a JSON object with
no_match set to boolean true; an empty result array is also accepted
as a no-match prediction.
Report each continuous action event exactly once, using its full start
and end times.
Do not divide an event into frames, one-second intervals, or other
fixed-length segments.
Merge adjacent intervals when they belong to the same continuous
action.
Every detection label must exactly match one official fine action label
from the 82-class MMA82 ontology.
Do not output a synonym, paraphrase, generic body-motion description,
or wording copied from the question as a label.
Return only events that satisfy the question.
Allowed fine action labels:
{82_CLASS_FINE_LABEL_LIST}
Include a score in [0,1] for every detection and sort detections by
descending score.
```

Description Prompt

```
System: You are a precise benchmark answerer. Output JSON only.
User: <video>
You are answering one MMA-MTU micro-action benchmark query.
Use only visible video evidence. Return exactly one JSON object, with
no explanation or Markdown.
Required schema: either a JSON object containing only no_match
(boolean true), or a JSON object with description (string) and events
(array of objects); each event object must contain label (string),
start_time (number in seconds), and end_time (number in seconds).
Video duration: {DURATION} seconds.
Task: {UNDERLYING_TASK} / {DESCRIPTION_SUBTASK}.
Question:
{QUESTION}
Write a complete, fluent paragraph in the description field.
Name and describe every matching event requested by the question, and
give its start and end time.
Do not include events outside the requested time window, label set,
body region, or other scope.
Describe only observable body-part and motion cues.
If no matching event is visible, return only a JSON object with
no_match set to boolean true; an empty result array is also accepted
as a no-match prediction.
```

```
Report each continuous action event exactly once, using its full start
and end times.
Do not divide an event into frames, one-second intervals, or other
fixed-length segments.
Merge adjacent intervals when they belong to the same continuous
action.
Every label in the structured events array must exactly match one
official fine action label from the 82-class MMA82 ontology.
Allowed fine action labels:
{82_CLASS_FINE_LABEL_LIST}
```

Correction Prompt for Invalid Outputs

```
Your previous response was rejected for this reason: {ERROR_REASON}.
Correct the response now. Re-evaluate the video and question, obey the
required schema and ontology exactly, and return only one JSON object.
```

E MORE RESULTS

E.1 MORE EXPERIMENTAL SETUP

We use the same benchmark queries and evaluation protocol for all models while retaining each model's supported visual interface. Models differ in whether they accept videos or frame sequences, how many frames their context permits, and which spatial resolution their visual encoder expects. These properties form part of the model's video-understanding capability. We therefore do not restrict every model to the lowest frame count or resolution supported by all models. A model with native video input receives the complete video through its official processor. For frame-based models without a prescribed sampling rate, we sample RGB frames chronologically at 4 FPS over the complete duration. When a model does not specify an input resolution, frames are resized with their aspect ratio preserved and padded with black borders to 336×336. Models with an official resolution use that resolution and their corresponding visual processor.

Under this protocol, Qwen3-VL and Qwen3.5/3.6/3.8 Bai et al. (2025a), together with LLaVA-OneVision-1.5 Li et al. (2024a), generally use 336×336 inputs sampled at 4 FPS. LLaVA-OneVision-2 receives the original video path through its official 448-pixel processor while retaining the frame count determined at 4 FPS. InternVL3 Zhu et al. (2025), InternVL3.5 and InternVideo3 Wang et al. (2025) first cover the complete duration at 4 FPS and then uniformly retain a model-dependent number of frames: up to 38 for InternVL3, 44 for the 2B/4B/8B InternVL3.5 models, and 128 for the larger InternVL3.5 variants. Phi-3.5 and Phi-4 Abdin et al. (2024) uniformly retain 16 frames over the complete duration.

API models follow their native visual interfaces. Qwen3.8-Max and GLM-5.3-Flash Hong et al. (2025b) receive the complete video directly. Qwen3.8-Max transcodes inputs only when they exceed 20 MiB and limits the longest spatial side to 1920 pixels without changing the duration. GPT-5.6-Luna Singh et al. (2025) receives timestamped JPEG frames at 4 FPS, preserves the source resolution for non-4K videos, and resizes 4K videos to 1080p. DeepSeek v4.1-Flash Wu et al. (2024c) also receives timestamped 4-FPS frames, with inputs above 720p resized proportionally to fit within 1280×720. Frame order is preserved in all standard evaluations.

Every query is processed as an independent, stateless sample. Queries from the same video may reuse frames or cached visual features, but they do not share conversation history, query text, or previous predictions. Local models use deterministic decoding. The default maximum output length is 384 tokens for Recognition, Temporal Reasoning, Grounding, and Detection, and 768 tokens for Description. Qwen3.8-Max uses temperature 0.1 and top-p 0.8; GLM-5.3-Flash uses temperature 1.0, top-p 0.95, and low reasoning effort. GPT-5.6-Luna and DeepSeek v4.1-Flash use no additional reasoning.

Table 5: Ablation study on Qwen3.5-9B of frame sampling rate on micro-action understanding.

FPS	Rec.	Temp.	Grounding						Detection					Description						
	Acc.	Acc.	R@1 F1@.5	R@1 @.2	R@1 @.5	R@1 @.7	R@1 Avg.	mIoU	mAP F1@.5	mAP @.2	mAP @.5	mAP @.7	mAP Avg.	CLIP	BL-4	CI-D	BLRT	EMS	Acc_L	tIoU
4	28.87	61.88	**17.23**	**40.54**	17.40	8.68	**21.62**	**22.56**	4.86	**5.33**	**2.83**	**1.74**	**2.94**	72.58	**0.93**	0.74	41.59	27.47	13.71	24.02
6	29.82	**62.29**	17.11	39.12	**17.46**	**8.71**	21.02	21.89	5.14	4.48	2.11	1.54	2.42	72.33	0.91	0.73	**41.71**	27.33	**14.18**	24.48
8	**30.57**	61.34	16.87	37.54	16.88	8.16	20.29	21.10	5.58	3.15	1.32	0.48	1.60	72.32	0.88	0.81	41.38	27.32	13.77	24.67
12	30.53	61.21	15.66	35.11	15.61	8.08	18.84	19.59	**5.82**	4.10	2.01	0.81	2.30	72.42	0.81	**0.83**	41.00	27.26	12.64	25.20
16	29.08	61.21	15.47	33.57	15.39	8.11	18.21	18.91	5.79	4.10	1.80	0.70	2.16	**73.09**	0.92	0.66	41.28	**27.54**	13.85	**25.24**

Table 6: Ablation study on Qwen3.5-9B of input frame size on micro-action understanding.

Frame Size	Rec.	Temp.	Grounding						Detection					Description						
	Acc.	Acc.	R@1 F1@.5	R@1 @.2	R@1 @.5	R@1 @.7	R@1 Avg.	mIoU	mAP F1@.5	mAP @.2	mAP @.5	mAP @.7	mAP Avg.	CLIP	BL-4	CI-D	BLRT	EMS	Acc_L	tIoU
224 × 224	28.41	**62.29**	17.47	**41.86**	17.84	8.88	**22.37**	**23.36**	4.55	5.16	2.78	**2.19**	**3.04**	71.88	**1.12**	**0.92**	41.20	27.13	13.53	23.10
336 × 336	28.87	61.88	17.23	40.54	17.40	8.68	21.62	22.56	4.86	**5.33**	**2.83**	1.74	2.94	72.58	0.93	0.74	41.59	27.47	13.71	24.02
384 × 384	28.58	61.61	17.45	**41.86**	17.74	**9.10**	22.29	23.23	5.73	3.82	1.71	0.98	2.03	**73.12**	0.83	0.87	**41.69**	**27.66**	14.61	24.38
448 × 448	29.53	61.07	**17.50**	41.61	**17.90**	8.55	22.07	22.97	**6.46**	4.39	2.41	1.54	2.50	72.31	0.85	0.68	41.00	27.31	14.62	**24.44**
512 × 512	**29.57**	61.00	17.42	40.56	17.51	8.89	21.66	22.52	6.10	4.87	2.21	1.86	2.72	72.92	0.94	0.63	41.51	27.53	**15.19**	24.26

E.2 Can Seeing More Lead to Seeing Better?

We study the effects of temporal and spatial resolution on micro-action understanding. For temporal resolution, we sample videos at rates from 4 to 16 fps. For spatial resolution, each video is resized to fit within a square input of 224 to 512 pixels, with the remaining area padded in black to preserve its aspect ratio. Tab. 5 and 6 show that increasing either resolution does not yield consistent gains. Recognition benefits from moderately denser frames and larger inputs, while Grounding and Detection mAP are strongest at lower frame rates and resolutions. Description favors denser sampling and intermediate spatial resolution. This difference likely arises because larger and denser inputs retain more appearance and motion details but also introduce redundant frames and background information, which can obscure the short temporal evidence needed for precise localization.

E.3 Full Domain Results

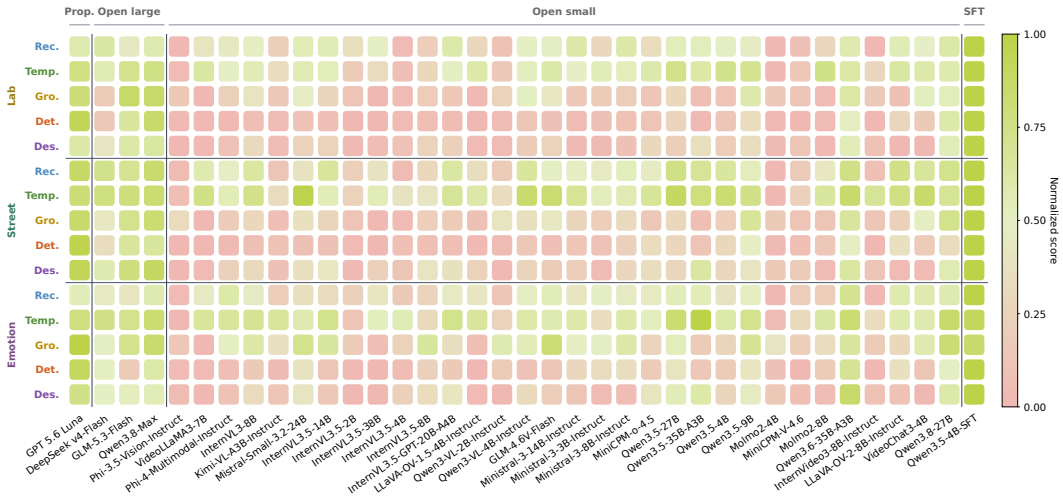


Figure 7: **Primary-metric performance across domains.** We select Accuracy for Recognition and Temporal Reasoning, F1@0.5 for Grounding and Detection, and label accuracy for Description, and normalize each domain–task row across models for visualization.

Different scenarios can substantially affect micro-action understanding. MMA-82-Det contains videos from laboratory interviews, psychiatric interviews, street interviews, and television programs with emotion-oriented content. We combine laboratory and psychiatric interviews into the Lab do-

Table 7: Results across three video domains: Lab, Street, and Emotion.

Model	Domain	Rec.	Temp.	Grounding						Detection					Description						
				R@1	R@1	R@1	R@1			mAP	mAP	mAP	mAP								
		Acc.	Acc.	F1@.5	@.2	@.5	@.7	Avg.	mIoU	F1@.5	@.2	@.5	@.7	Avg.	CLIP	BL-4	CI-D	BLRT	EMS	Acc_L	tIoU
Proprietary MLLMs																					
GPT 5.6 Luna	Lab	34.61	65.93	22.00	44.40	22.05	10.66	24.41	24.79	11.67	9.23	4.80	2.28	5.29	65.95	1.41	1.60	44.34	24.80	24.69	34.28
	Street	54.15	70.77	23.23	51.04	21.84	9.33	26.15	26.90	14.40	15.11	6.70	2.66	7.87	57.39	1.45	2.05	44.56	22.11	28.01	37.22
	Emotion	34.69	73.17	**50.44**	**77.69**	**50.38**	**33.08**	**50.17**	**50.02**	22.63	**33.23**	**21.02**	8.22	**19.34**	56.98	0.00	2.95	45.43	23.18	27.84	40.49
Open-source Large MLLMs																					
DeepSeek v4.1-Flash	Lab	37.55	49.31	4.63	8.59	4.48	2.18	4.83	4.91	1.74	4.05	1.71	0.92	2.20	75.22	1.46	1.09	44.04	27.91	16.26	29.03
	Street	45.20	68.46	11.40	28.47	10.80	3.44	13.44	13.89	5.15	4.98	1.33	0.42	2.51	64.76	1.22	1.94	45.91	24.25	17.30	32.02
	Emotion	28.57	70.73	24.41	35.00	24.23	15.38	23.08	22.96	12.85	17.73	9.18	7.66	10.16	61.96	2.53	2.23	48.38	24.11	18.77	36.93
GLM-5.3-Flash	Lab	26.26	63.09	23.81	48.26	24.20	10.86	26.28	26.77	8.04	7.29	3.78	1.77	4.14	74.56	0.83	1.65	41.24	27.97	24.60	33.94
	Street	43.23	69.23	19.67	50.55	19.88	7.48	24.51	25.17	9.70	11.95	5.63	1.42	5.85	61.75	0.79	1.66	40.37	23.31	23.51	34.55
	Emotion	26.53	68.29	36.67	66.54	37.31	25.77	40.98	41.23	4.35	16.92	4.78	0.00	6.94	64.82	0.75	2.33	40.81	25.38	19.75	36.96
Qwen3.8-Max	Lab	34.61	65.70	23.87	44.70	23.52	11.52	25.18	25.47	10.59	7.05	4.06	2.39	4.18	**81.19**	1.68	2.99	46.26	30.21	25.59	34.66
	Street	52.18	72.31	22.42	51.53	22.33	9.69	26.49	26.93	9.64	10.58	4.66	1.21	5.40	65.32	1.90	3.27	46.79	24.44	26.42	35.45
	Emotion	36.73	78.05	42.88	70.00	43.85	29.23	45.30	45.14	15.22	22.87	13.25	5.41	12.52	62.70	1.52	4.86	45.23	24.04	20.68	35.01
Open-source Small MLLMs																					
Phi-3.5-Vision-Instruct	Lab	0.16	2.68	4.34	12.88	3.54	1.24	6.00	6.66	0.25	0.35	0.04	0.00	0.15	0.39	0.00	0.00	0.24	0.15	0.03	0.04
	Street	2.18	6.92	8.88	21.47	6.50	1.35	9.58	10.09	0.06	0.03	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Emotion	2.04	4.88	7.31	21.15	5.00	2.31	10.56	11.92	0.87	1.83	1.09	1.09	1.06	2.67	0.00	0.08	1.83	0.99	1.48	0.36
VideoLLaMA3-7B	Lab	23.58	55.28	0.02	0.04	0.02	0.02	0.03	0.03	0.22	0.71	0.55	0.13	0.45	2.15	0.00	0.00	1.18	0.82	0.37	0.93
	Street	35.81	65.38	0.00	0.12	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.01	4.09	0.00	0.02	2.65	1.53	0.79	1.11
	Emotion	28.57	63.41	0.00	0.00	0.00	0.00	0.00	0.00	0.72	1.38	0.22	0.00	0.39	2.18	0.00	0.00	1.47	0.95	0.00	0.00
Phi-4-Multimodal-Instruct	Lab	25.84	42.88	6.08	19.66	6.21	2.97	10.10	11.20	0.21	0.20	0.14	0.09	0.17	61.40	0.41	0.57	32.67	22.97	3.81	15.81
	Street	29.26	49.23	4.59	23.80	4.91	1.35	10.36	11.98	0.31	0.05	0.01	0.01	0.04	62.62	0.33	0.90	35.67	23.04	6.90	18.55
	Emotion	38.78	63.41	26.15	56.54	26.15	17.31	31.41	31.88	1.45	1.72	0.87	0.00	0.85	40.36	0.00	0.69	24.31	15.16	3.70	14.84
InternVL3-8B	Lab	28.78	48.62	11.14	30.33	11.35	5.23	15.73	17.00	0.86	2.35	0.59	0.15	1.18	67.06	0.12	0.32	39.77	25.70	12.06	21.20
	Street	39.52	64.62	7.56	34.48	8.71	2.09	15.31	16.81	1.04	0.41	0.11	0.02	0.46	54.21	0.09	0.09	37.12	20.75	9.30	24.99
	Emotion	28.57	65.85	31.47	58.85	31.15	21.92	35.60	36.09	3.99	7.15	2.03	0.00	3.04	56.89	0.00	2.29	39.24	22.39	12.35	28.95
Kimi-VL-A3B-Instruct	Lab	12.34	29.02	3.88	12.60	3.99	1.94	6.18	6.66	0.38	0.44	0.08	0.06	0.18	8.05	0.00	0.03	4.72	3.00	1.93	2.22
	Street	7.21	30.00	1.53	10.43	1.96	0.61	4.22	4.62	1.50	0.87	0.36	0.00	0.39	18.52	0.02	0.37	13.00	7.02	2.58	5.83
	Emotion	12.24	63.41	18.13	25.38	15.77	12.31	17.31	17.84	0.43	0.77	0.09	0.00	0.35	21.73	0.00	0.56	13.19	8.32	4.94	5.45
Mistral-Small-3.2-24B	Lab	32.98	48.62	13.05	34.46	13.18	6.85	18.15	19.29	1.17	1.89	0.50	0.23	0.93	79.66	1.59	1.46	45.98	29.71	9.92	25.96
	Street	24.24	85.38	9.86	32.52	9.45	1.72	14.29	15.35	1.32	0.49	0.02	0.00	0.37	66.13	1.91	1.10	48.15	24.65	10.74	27.67
	Emotion	22.45	56.10	36.31	70.77	36.92	27.69	43.33	43.74	9.78	7.24	3.93	1.03	3.94	62.58	1.64	4.36	47.26	24.23	15.12	30.33
InternVL3.5-14B	Lab	33.46	46.63	7.02	19.43	7.12	3.75	10.00	10.75	0.86	1.72	0.33	0.17	0.79	70.98	0.55	0.87	39.47	26.79	9.39	19.91
	Street	41.05	49.23	7.06	35.58	7.24	1.60	14.89	16.27	1.22	1.60	0.47	0.04	0.83	57.10	0.58	0.89	43.32	21.95	12.91	30.99
	Emotion	20.41	68.29	32.91	54.62	32.69	23.46	34.19	34.54	2.83	3.35	1.43	1.43	1.95	58.25	0.00	2.03	41.45	23.05	5.56	29.61
InternVL3.5-2B	Lab	21.06	15.93	2.92	11.11	2.97	1.53	5.56	6.37	0.11	0.12	0.03	0.00	0.05	8.98	0.00	0.00	3.43	3.45	0.74	3.49
	Street	14.41	23.08	2.87	11.66	2.70	0.86	5.26	5.91	0.47	0.09	0.02	0.00	0.03	11.26	0.00	0.00	4.97	4.40	0.50	4.90
	Emotion	14.29	14.63	8.74	17.31	9.23	4.62	10.13	10.24	0.33	0.35	0.05	0.00	0.12	33.99	0.00	0.00	17.10	13.51	0.00	13.23
InternVL3.5-38B	Lab	30.15	28.48	0.16	0.47	0.15	0.06	0.23	0.25	0.04	0.44	0.01	0.01	0.12	16.18	0.00	0.10	9.91	6.15	1.84	4.89
	Street	24.02	49.23	0.28	2.82	0.12	0.00	1.05	1.29	0.08	0.05	0.05	0.00	0.07	35.30	0.35	0.51	25.83	13.55	6.64	14.83
	Emotion	26.53	51.22	3.49	6.15	3.85	2.31	3.42	3.36	0.00	0.40	0.00	0.00	0.09	31.93	0.00	1.90	23.94	12.60	0.00	18.83
InternVL3.5-4B	Lab	1.47	6.36	0.88	2.32	0.88	0.38	1.21	1.32	0.34	0.56	0.01	0.00	0.20	49.70	0.10	0.54	26.36	18.82	3.71	16.48
	Street	2.40	35.38	0.70	5.28	0.86	0.37	2.07	2.28	0.83	0.26	0.06	0.02	0.12	17.59	0.00	0.20	11.66	6.71	3.64	6.32
	Emotion	10.20	53.66	11.88	20.77	11.54	6.54	12.48	12.70	1.09	2.56	0.87	0.00	1.20	53.78	0.00	2.82	33.56	20.78	8.46	22.04
InternVL3.5-8B	Lab	11.82	24.12	5.16	15.68	5.16	2.20	7.70	8.44	0.67	0.94	0.08	0.02	0.45	56.81	0.11	0.48	33.17	21.78	9.57	17.00
	Street	17.25	36.15	5.49	26.87	5.28	1.35	11.14	12.05	0.80	0.70	0.03	0.00	0.37	53.25	0.20	0.78	36.92	20.48	12.31	18.88
	Emotion	16.33	31.71	33.14	55.38	32.31	21.15	34.61	34.88	5.62	5.34	2.20	0.16	2.63	50.81	0.00	2.38	34.24	20.34	12.59	22.73
InternVL3.5-GPT-20B-A4B	Lab	35.98	41.65	4.02	9.34	4.03	2.11	5.08	5.50	0.58	1.26	0.59	0.35	0.80	67.43	0.20	0.42	36.36	25.53	8.92	21.35
	Street	39.30	60.00	4.72	22.58	4.91	0.74	9.30	10.33	0.83	1.68	0.03	0.00	0.74	56.83	0.25	0.38	38.02	21.58	11.27	24.34
	Emotion	28.57	68.29	17.43	31.54	17.31	11.15	18.97	19.17	3.88	2.17	0.87	0.33	1.12	64.60	0.00	1.38	42.42	24.93	15.74	31.04
LLaVA-OV-1.5-4B-Instruct	Lab	17.02	52.37	0.29	0.94	0.30	0.08	0.51	0.62	0.09	0.09	0.02	0.00	0.04	16.35	0.00	0.00	9.87	6.19	1.28	5.60
	Street	21.40	54.62	2.15	8.47	2.33	0.49	3.79	4.18	0.21	0.18	0.00	0.00	0.09	44.66	0.00	0.01	29.99	17.13	7.21	18.16
	Emotion	26.53	63.41	2.69	9.62	2.69	0.38	4.02	4.37	0.00	0.97	0.00	0.00	0.95	42.76	0.00	0.00	31.54	17.45	1.23	23.13
Qwen3-VL-2B-Instruct	Lab	5.83	11.79	6.59	19.98	6.91	3.56	10.17	10.80	0.56	1.17	0.38	0.17	0.60	9.68	0.00	0.09	5.34	3.65	0.99	2.22
	Street	10.70	30.77	10.70	29.69	11.17	4.29	14.44	15.16	0.54	1.10	0.02	0.01	0.59	3.71	0.00	0.03	2.58	1.42	0.44	0.62
	Emotion	8.16	26.83	29.22	57.31	28.46	18.46	33.25	33.87	4.31	5.30	5.01	0.08	2.93	7.14	0.00	0.02	4.83	2.74	0.00	0.45
Qwen3-VL-4B-Instruct	Lab	29.83	45.02	13.75	36.64	13.93	6.99	19.52	21.22	1.65	3.14	2.00	0.33	1.67	67.59	0.33	0.57	36.51	25.29	7.49	14.05
	Street	40.17	73.85	9.56	37.55	10.31	2.94	17.38	19.47	2.07	0.99	0.39	0.35	0.59	53.49	0.00	0.93	34.74	20.21	8.02	19.40
	Emotion	30.61	58.54	27.60	48.85	26.54	17.31	29.53	30.49	2.32	1.16	0.31	0.00	0.47	59.05	0.76	0.59	37.20	22.78	11.11	19.63
GLM-4.6V-Flash	Lab	30.30	53.06	11.60	29.00	11.67	4.88	14.69	15.28	1.01	1.41	0.51	0.45	0.80	40.11	0.07	0.72	26.11	15.15	9.60	21.25
	Street	29.69	73.08	11.59	30.43	12.02	3.80	14.71	15.18	0.75	0.89	0.19	0.08	0.46	33.06	0.03	0.51	24.03	12.79	5.56	20.02
	Emotion	34.69	56.10	41.15	66.54	41.54	18.08	39.44	39.03	2.90	6.43	5.22	0.87	3.58	38.89	0.18	0.38	30.25	15.99	7.41	26.00
Ministral-3-14B-Instruct	Lab	36.61	42.80	4.17	7.93	4.16	2.35	4.57	4.68	0.34	1.24	0.35	0.21	0.63	15.49	0.15	0.17	10.55	5.78	0.83	8.21
	Street	29.48	60.77	7.17	22.09	6.87	2.45	10.18	10.69	1.84	0.88	0.10	0.05	0.43	54.26	1.54	0.32	42.47	20.60	5.90	27.65
	Emotion	26.53	58.54	25.77	36.15	25.00	15.77	24.10	24.40	7.07	4.23	2.71	0.00	2.12	53.29	1.21	0.00	44.07	21.04	4.44	32.21
Ministral-3-3B-Instruct	Lab	16.81	47.55	5.02	12.26	5.12	2.64	6.62	6.98	1.00	1.15	0.52	0.24	0.66	16.33	0.04	0.16	11.28	6.19	0.72	8.44
	Street	27.95	47.69	5.57	22.70	5.77	0.98	9.79	10.60	1.40	0.59	0.21	0.04	0.35	10.46	0.00	0.02	8.77	4.04	1.03	5.50
	Emotion	30.61	34.15	24.42	46.15	25.00	17.31	27.35	27.49	8.15	14.70	6.85	0.43	5.75	16.57	0.00	0.00	14.71	6.68	1.23	10.77
Ministral-3-8B-Instruct	Lab	36.50	46.71	5.30	11.56	5.31	2.58	6.33	6.73	1.30	2.12	0.44	0.09	0.97	65.68	1.21	0.72	40.71	24.17	3.23	16.46
	Street	32.75	50.00	8.31	28.22	8.10	1.72	12.53	13.49	3.35	0.75	0.14	0.02	0.38	59.51	0.81	0.24	46.68	22.67	8.48	30.91
	Emotion	28.57	58.54	30.94	61.54	30.77	23.46	36.50	36.95	8.48	9.70	7.29	1.35	5.36	58.68	0.82	2.93	46.44	22.68	1.85	32.68
MiniCPM-o-4.5	Lab	20.22	51.45	3.48	10.09	3.43	1.56	5.09	5.64	1.38	1.33	0.14	0.02	0.50	57.35	0.35	0.64	33.52	21.39	11.68	20.18
	Street	28.82	59.23	3.70	14.60	2.82	0.74	6.09	6.74	4.96	1.87	0.32	0.01	0.69	53.82	0.28	0.70	38.45	20.77	10.39	24.07
	Emotion	28.57	51.22	12.93	23.85	13.08	9.23	15.56	16.32	4.49	6.38	2.10	0.00	3.15	51.19	0.66	1.43	35.49	20.29	14.81	23.72
Qwen3.5-27B	Lab	31.72	63.78	8.08	14.44	7.95	4.39	8.39	8.47	3.17	3.27	1.93	1.57	2.16	34.15	0.51	0.66	21.21	12.83	8.30	18.70
	Street	46.51	78.46	7.88	22.82	8.10	3.68	10.92	11.22	4.18	1.39	0.19	0.08	0.65	31.49	0.88	1.72	26.02	12.07	8.11	20.25
	Emotion	34.69	78.05	27.08	38.08	27.31	20.38	26.97	26.34	9.86	18.88	11.43	9.89	14.13	39.36	1.24	2.88	33.20	16.10	20.37	32.61

Table 8: Results across three video domains: Lab, Street, and Emotion (continued).

Model	Domain	Rec.	Temp.	Grounding						Detection					Description						
				R@1	R@1	R@1	R@1			mAP	mAP	mAP	mAP								
		Acc.	Acc.	F1@.5	@.2	@.5	@.7	Avg.	mIoU	F1@.5	@.2	@.5	@.7	Avg.	CLIP	BL-4	CI-D	BLRT	EMS	Acc_L	tIoU
Open-source Small MLLMs																					
Qwen3.5-35B-A3B	Lab	31.62	53.75	1.57	2.58	1.54	0.87	1.59	1.62	0.30	0.80	0.47	0.47	0.47	70.87	0.69	0.62	38.72	26.78	14.35	26.43
	Street	41.92	68.46	1.76	6.63	2.09	0.25	2.89	3.07	1.56	1.25	0.28	0.05	0.55	61.67	0.73	0.93	42.98	23.72	18.77	33.84
	Emotion	34.69	**92.68**	8.72	10.77	8.08	5.77	7.69	7.67	3.55	2.34	0.87	0.00	1.11	64.78	0.00	1.53	41.60	25.61	22.84	35.98
Qwen3.5-4B	Lab	28.52	63.48	2.92	5.03	2.97	1.62	2.99	3.04	1.61	2.64	1.47	1.02	1.51	52.00	0.25	0.50	25.92	19.59	7.60	15.47
	Street	40.17	72.31	5.46	14.85	5.77	1.96	7.04	7.31	4.29	1.39	0.41	0.06	0.90	53.85	0.24	0.75	30.50	20.13	8.13	20.85
	Emotion	24.49	58.54	12.95	20.00	13.08	8.46	13.42	13.44	4.28	4.09	2.66	0.43	2.42	52.40	0.00	2.32	29.87	20.20	9.48	22.48
Qwen3.5-9B	Lab	27.78	61.33	16.18	38.26	16.48	8.36	20.55	21.49	4.13	5.32	3.16	2.03	3.13	75.58	0.99	0.77	41.69	28.55	13.85	23.91
	Street	32.97	65.38	18.25	46.87	17.91	6.01	22.48	23.48	7.46	5.31	1.13	0.23	2.29	56.51	0.56	0.23	40.67	21.46	12.13	23.53
	Emotion	32.65	68.29	35.28	67.31	34.62	23.46	40.81	41.57	10.47	4.89	1.87	0.42	2.57	58.09	0.00	4.38	43.09	23.24	16.98	29.74
Molmo2-4B	Lab	0.00	0.00	4.09	10.68	4.14	2.18	5.72	6.10	0.17	0.36	0.18	0.02	0.22	18.81	0.00	0.18	11.19	7.16	0.87	5.15
	Street	0.00	2.31	4.41	18.28	4.42	0.74	7.55	8.15	0.29	0.62	0.00	0.00	0.17	14.69	0.01	0.32	11.06	5.75	4.60	8.58
	Emotion	0.00	7.32	20.77	40.38	20.77	15.38	24.66	24.84	2.17	2.44	1.96	0.87	1.77	17.61	0.00	1.19	12.94	7.14	5.86	11.15
MiniCPM-V-4.6	Lab	5.04	11.56	3.14	10.07	2.97	1.34	5.07	5.68	0.09	0.48	0.41	0.01	0.26	37.65	0.05	0.30	24.12	14.15	4.89	7.79
	Street	11.14	21.54	2.95	9.45	2.58	0.86	4.73	5.59	0.78	0.04	0.02	0.00	0.02	42.20	0.00	0.69	29.80	15.67	5.11	13.78
	Emotion	12.24	31.71	9.92	21.92	8.85	5.38	11.41	11.71	0.00	0.43	0.00	0.00	0.20	39.53	0.00	0.29	30.60	15.24	2.16	12.03
Molmo2-8B	Lab	16.18	64.78	0.90	2.50	0.90	0.43	1.25	1.28	0.42	1.05	0.38	0.01	0.55	1.66	0.00	0.00	0.96	0.63	0.37	0.89
	Street	26.64	57.69	3.24	12.02	3.44	1.23	5.26	5.43	2.07	1.18	0.18	0.02	0.63	5.69	0.00	0.05	4.07	2.27	0.95	5.08
	Emotion	12.24	60.98	17.37	23.46	17.31	10.38	15.81	15.82	4.35	4.67	1.41	0.54	2.17	12.70	0.00	0.00	8.02	5.19	1.85	6.48
Qwen3.6-35B-A3B	Lab	34.61	53.75	17.21	34.55	16.51	8.36	19.10	19.75	6.36	5.23	2.93	1.50	3.02	80.20	0.88	0.92	43.32	30.19	21.90	30.61
	Street	46.29	74.62	17.78	38.40	16.93	6.87	19.56	20.17	7.49	4.61	2.71	0.68	2.63	64.55	0.73	0.72	43.17	24.43	19.17	34.12
	Emotion	44.90	78.05	32.98	48.46	33.85	20.00	32.48	32.18	17.29	18.57	10.77	4.72	10.70	68.09	0.00	1.85	43.98	26.18	32.10	37.34
InternVideo3-8B-Instruct	Lab	0.00	23.35	4.55	12.41	4.56	2.35	6.52	7.08	0.03	0.37	0.00	0.00	0.25	59.67	0.54	0.70	34.02	22.37	4.21	15.62
	Street	4.80	60.00	3.38	20.37	3.07	0.86	9.08	11.18	0.00	0.58	0.00	0.00	0.29	58.27	0.75	0.90	41.61	22.36	3.70	21.62
	Emotion	0.00	34.15	9.31	26.54	9.23	6.15	13.72	14.63	0.00	0.00	0.00	0.00	0.00	59.05	0.00	1.87	40.46	22.58	3.70	23.50
LLaVA-OV-2-8B-Instruct	Lab	33.14	54.82	2.31	3.94	2.33	1.30	2.38	2.39	3.40	2.76	1.12	0.49	1.30	5.95	0.00	0.03	3.60	2.27	0.12	4.37
	Street	45.63	66.92	5.90	12.64	6.26	2.70	6.72	6.73	5.37	3.63	1.41	0.41	1.82	4.38	0.00	0.03	3.03	1.70	0.79	3.29
	Emotion	36.73	56.10	14.17	18.46	13.85	10.77	13.16	12.80	9.06	13.08	3.07	1.91	4.98	6.94	0.00	0.14	5.11	2.81	3.70	2.95
VideoChat3-4B	Lab	29.52	51.68	14.40	36.08	14.89	6.46	18.94	20.12	2.02	3.42	2.55	0.16	1.88	4.15	0.00	0.06	2.37	1.59	0.21	1.77
	Street	40.83	75.38	13.72	40.61	14.60	5.15	19.75	21.67	2.57	0.85	0.10	0.01	0.47	3.37	0.00	0.10	2.35	1.31	0.79	1.55
	Emotion	36.73	56.10	29.29	63.08	30.77	15.77	35.04	35.55	2.90	10.80	0.41	0.00	2.96	2.43	0.00	0.00	1.62	0.99	0.00	2.00
Qwen3.8-27B	Lab	36.87	50.84	14.69	29.94	14.55	7.25	16.53	16.97	7.41	5.52	2.68	0.92	2.96	80.89	1.28	1.35	45.42	**30.31**	21.80	32.60
	Street	44.76	60.00	19.60	44.42	19.26	8.10	22.52	23.52	4.87	3.17	1.37	0.61	1.78	63.68	1.04	1.87	44.49	24.00	17.21	32.25
	Emotion	36.73	80.49	41.54	63.46	41.15	28.08	41.41	41.23	11.96	23.59	15.67	**11.41**	16.37	64.09	1.74	3.13	45.60	24.95	22.35	36.12
Open-source Small MLLMs with SFT																					
Qwen3.5-4B-SFT	Lab	59.51	86.45	27.38	52.93	28.19	13.16	29.85	30.15	12.30	5.43	2.02	0.71	2.67	62.00	32.76	62.57	59.56	24.07	**39.92**	46.77
	Street	61.57	86.92	26.94	48.47	29.45	11.41	28.19	27.97	14.70	3.04	1.58	0.69	1.66	52.15	30.27	52.38	59.01	21.16	29.78	**55.81**
	Emotion	**63.27**	85.37	43.46	76.15	44.23	24.23	45.47	45.40	**24.89**	18.25	12.88	6.51	11.17	49.89	**35.09**	**113.86**	**62.31**	20.90	37.35	52.67

main because they share relatively stable interview settings, while the remaining videos form the Street and Emotion domains. The three subsets contain 990, 144, and 80 videos, corresponding to 11,276, 1,955, and 432 queries, respectively. We evaluate all domains using the same inference settings and metrics described in Sec. 4.1.

Tab. 7 and Tab. 8 report the complete domain results. Overall, models perform relatively poorly in the Lab domain, while many obtain stronger results on Street and Emotion videos. Lab and psychiatric interviews contain restrained, low-amplitude movements against stable backgrounds. These scenes provide few contextual cues, leaving the model to distinguish subtle motion directly. By contrast, people in Street and Emotion videos tend to move more naturally and expressively. These videos also contain richer scene and interaction context, which allows MLLMs to infer actions from both motion and surrounding events. Their visual distribution may also be closer to the diverse web videos used during pretraining. In addition, rare micro-actions occur less frequently in these real-world subsets, reducing the effect of the long-tailed category distribution.

Fig. 7 further shows that the domain effect is not uniform across tasks or models. For each domain, we select one primary metric from each of the five tasks and apply row-wise min–max normalization across models. The resulting profiles vary noticeably across Lab, Street, and Emotion: a model that performs well in one scene or task does not necessarily retain the same advantage elsewhere. Emotion videos are generally more favorable for Grounding and Detection, where salient movements and contextual cues make temporal evidence easier to recover, whereas the relative rankings for Recognition and Temporal Reasoning remain more mixed. The strong and consistent profile of Qwen3.5-4B-SFT also indicates that supervision from MMA-MTU improves the transfer of micro-action understanding across scenes. These results show that domain-level evaluation exposes capability differences that are obscured when all videos are aggregated into a single score.

Table 9: Comparison between label (**L**) and label-free (**LF**) queries. Small values beside LF results denote the performance difference LF − L.

Model	Rec.		Temp.		Grounding								Detection					
	Acc.		Acc.		F1@.5		R@1 @.5		R@1 Avg.		mIoU		F1@.5		mAP @.5		mAP Avg.	
	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF	L	LF
Proprietary MLLMs																		
GPT 5.6 Luna	35.26	40.53 +5.27	66.95	61.92 -5.03	22.89	25.67 +2.78	22.47	25.44 +2.97	24.46	27.03 +2.57	24.83	27.29 +2.46	12.79	12.75 -0.04	3.88	5.90 +2.02	5.67	6.73 +1.06
Open-source Large MLLMs																		
DeepSeek v4.1-Flash	35.26	37.23 +1.97	48.95	49.79 +0.84	6.08	7.26 +1.18	5.99	7.00 +1.01	6.32	7.47 +1.15	6.40	7.62 +1.22	2.30	2.49 +0.19	2.89	3.23 +0.34	3.57	3.42 -0.15
GLM-5.3-Flash	27.35	30.64 +3.29	61.72	65.90 +4.18	24.76	24.49 -0.27	24.68	24.34 -0.34	27.56	26.06 -1.51	28.10	26.52 -1.58	7.25	9.28 +2.03	4.50	3.97 -0.53	5.47	3.99 -1.48
Qwen3.8-Max	36.74	38.06 +1.32	68.41	65.90 -2.51	24.99	25.58 +0.59	24.25	25.01 +0.76	26.57	26.57 0.00	26.91	26.78 -0.13	11.89	11.73 -0.16	6.10	4.96 -1.14	6.78	5.98 -0.80
Open-source Small MLLMs																		
Phi-3.5-Vision-Instruct	0.16	0.66 +0.50	0.00	9.41 +9.41	3.93	4.85 +0.92	3.16	3.26 +0.10	5.35	6.25 +0.90	6.01	7.05 +1.04	0.30	0.03 -0.27	0.58	0.00 -0.58	0.67	0.27 -0.40
VideoLLaMA3-7B	23.89	22.90 -0.99	57.74	52.30 -5.44	0.00	0.00 0.00	0.00	0.00 0.00	0.01	0.01 +0.01	0.01	0.01 0.00	0.44	0.06 -0.38	0.37	0.38 +0.01	0.68	0.26 -0.42
Phi-4-Multimodal-Instruct	24.38	26.69 +2.31	43.10	49.58 +6.48	4.84	9.62 +4.78	4.89	9.68 +4.79	8.13	16.12 +7.98	8.96	17.88 +8.92	0.32	0.13 -0.19	1.68	0.13 -1.55	1.06	0.25 -0.81
InternVL3-8B	31.96	31.96 0.00	44.35	57.95 +13.60	13.99	9.66 -4.33	13.94	9.82 -4.12	19.87	12.43 -7.45	21.46	13.18 -8.28	1.32	0.60 -0.72	1.85	1.67 -0.18	2.45	1.63 -0.82
Kimi-VL-A3B-Instruct	12.19	17.63 +5.44	30.96	33.68 +2.72	5.33	4.41 -0.92	5.32	4.55 -0.77	7.12	7.42 +0.30	7.55	8.09 +0.54	0.04	0.25 +0.21	0.00	0.11 +0.11	0.14	0.30 +0.16
Mistral-Small-3.2-24B	29.82	30.81 +0.99	50.63	40.79 -9.84	14.66	14.47 -0.19	14.61	14.37 -0.24	19.54	19.25 -0.29	20.66	20.35 -0.31	1.96	0.82 -1.14	3.26	2.40 -0.86	3.47	2.42 -1.05
InternVL3.5-14B	30.81	32.45 +1.64	46.44	50.42 +3.98	8.05	8.89 +0.84	8.00	8.86 +0.86	11.09	12.07 +0.98	11.89	12.95 +1.06	1.15	0.53 -0.62	2.10	0.37 -1.73	2.47	1.10 -1.37
InternVL3.5-2B	15.32	21.75 +6.43	15.90	19.67 +3.77	5.27	0.70 -4.57	5.27	0.53 -4.74	8.97	0.96 -8.01	10.12	1.08 -9.04	0.21	0.11 -0.10	0.06	0.06 0.00	0.24	0.06 -0.18
InternVL3.5-38B	26.36	23.23 -3.13	26.15	30.75 +4.60	0.48	0.19 -0.29	0.48	0.19 -0.29	0.79	0.19 -0.61	0.84	0.21 -0.63	0.00	0.05 +0.05	0.00	0.05 +0.05	0.15	0.29 +0.14
InternVL3.5-4B	1.81	1.65 -0.16	6.69	17.36 +10.67	1.51	1.25 -0.26	1.53	1.15 -0.38	2.25	1.55 -0.70	2.43	1.63 -0.80	0.41	0.36 -0.05	1.55	0.10 -1.45	1.55	0.32 -1.23
InternVL3.5-8B	10.71	19.77 +9.06	23.22	26.57 +3.35	6.88	6.87 -0.01	6.76	6.71 -0.05	9.59	9.21 -0.39	10.34	9.87 -0.47	1.32	0.72 -0.60	1.79	0.19 -1.60	1.94	0.78 -1.16
InternVL3.5-GPT-20B-A4B	33.44	33.28 -0.16	41.00	44.14 +3.14	4.76	5.07 +0.31	4.74	4.98 +0.24	5.89	6.51 +0.62	6.30	7.06 +0.76	0.60	0.60 0.00	1.53	0.71 -0.82	1.47	1.03 -0.44
LLaVA-OV-1.5-4B-Instruct	16.47	20.10 +3.63	52.30	50.84 -1.46	0.67	0.37 -0.30	0.67	0.38 -0.29	1.01	0.76 -0.26	1.15	0.89 -0.26	0.00	0.00 0.00	0.00	0.00 0.00	0.05	0.05 0.00
Qwen3-VL-2B-Instruct	4.61	10.21 +5.60	18.62	4.39 -14.23	9.23	7.59 -1.64	9.54	7.52 -2.02	13.60	10.19 -3.41	14.50	10.62 -3.88	1.53	0.20 -1.33	1.77	0.12 -1.65	1.89	0.22 -1.67
Qwen3-VL-4B-Instruct	29.98	33.11 +3.13	44.98	43.93 -1.05	14.94	14.76 -0.18	14.95	14.57 -0.38	20.54	20.18 -0.36	22.30	21.97 -0.33	1.90	1.51 -0.39	2.55	1.34 -1.21	2.13	1.63 -0.50
GLM-4.6V-Flash	25.54	29.98 +4.44	53.56	53.35 -0.21	14.14	14.38 +0.24	13.99	14.28 +0.29	16.05	18.12 +2.07	16.53	18.83 +2.30	1.11	0.70 -0.41	1.08	0.14 -0.94	1.43	0.84 -0.59
Ministral-3-14B-Instruct	30.81	32.45 +1.64	42.26	42.26 0.00	5.12	6.25 +1.13	5.13	6.09 +0.96	5.60	6.46 +0.86	5.74	6.59 +0.85	0.81	1.07 +0.26	2.17	1.28 -0.89	1.99	1.58 -0.41
Ministral-3-3B-Instruct	17.96	21.09 +3.13	46.65	46.65 0.00	5.45	6.28 +0.83	5.51	6.23 +0.72	6.83	8.63 +1.80	7.15	9.15 +2.00	1.40	1.16 -0.24	1.27	1.37 +0.10	1.83	1.68 -0.15
Ministral-3-8B-Instruct	31.14	31.47 +0.33	45.40	47.07 +1.67	6.87	7.31 +0.44	6.80	7.09 +0.29	7.78	9.25 +1.47	8.17	9.79 +1.62	2.25	1.40 -0.85	2.69	1.07 -1.62	3.01	2.32 -0.69
MiniCPM-o-4.5	25.86	21.42 -4.44	49.16	55.02 +5.86	5.96	1.39 -4.57	5.80	1.34 -4.46	8.18	1.66 -6.53	8.96	1.80 -7.16	1.74	1.57 -0.17	1.69	2.22 +0.53	2.08	1.98 -0.10
Qwen3.5-27B	32.62	32.29 -0.33	67.57	61.92 -5.65	9.18	9.74 +0.56	8.96	9.77 +0.81	9.13	10.08 +0.95	9.15	10.11 +0.96	3.86	3.60 -0.26	2.99	2.06 -0.93	3.32	2.28 -1.04
Qwen3.5-35B-A3B	32.45	32.29 -0.16	55.23	52.30 -2.93	3.44	0.19 -3.25	3.40	0.19 -3.21	3.55	0.18 -3.37	3.57	0.18 -3.39	0.87	0.54 -0.33	1.33	0.05 -1.28	1.66	0.64 -1.02
Qwen3.5-4B	27.84	29.98 +2.14	62.34	64.44 +2.10	4.30	2.85 -1.45	4.26	2.78 -1.48	4.28	2.93 -1.35	4.33	3.00 -1.33	1.91	1.88 -0.03	3.82	1.83 -1.99	3.66	1.71 -1.95
Qwen3.5-9B	27.68	26.69 -0.99	59.83	61.09 +1.26	17.98	18.65 +0.67	17.97	18.59 +0.62	21.90	23.38 +1.48	22.77	24.53 +1.76	4.74	4.59 -0.15	4.29	2.13 -2.16	3.99	3.06 -0.93
Molmo2-4B	0.00	0.00 0.00	0.42	0.00 -0.42	6.73	3.66 -3.07	6.76	3.64 -3.12	9.10	4.71 -4.39	9.62	4.89 -4.73	0.44	0.10 -0.34	2.42	0.51 -1.91	1.94	0.58 -1.36
MiniCPM-V-4.6	5.44	6.75 +1.31	12.97	10.04 -2.93	5.17	0.95 -4.22	4.84	0.81 -4.03	7.76	1.29 -6.48	8.54	1.51 -7.03	0.05	0.13 +0.08	0.13	0.02 -0.11	0.17	0.04 -0.13
Molmo2-8B	19.28	9.39 -9.89	67.15	66.11 -1.04	2.09	2.38 +0.29	2.11	2.35 +0.24	2.42	2.93 +0.51	2.44	3.00 +0.56	0.95	0.97 +0.02	0.77	0.77 0.00	1.41	1.04 -0.37
Qwen3.6-35B-A3B	36.24	35.26 -0.98	57.32	48.95 -8.37	19.60	15.27 -4.33	18.45	14.42 -4.03	21.41	16.10 -5.31	22.16	16.52 -5.64	7.30	6.17 -1.13	4.28	2.98 -1.30	4.59	3.31 -1.28
InternVideo3-8B-Instruct	0.33	0.00 -0.33	33.89	13.60 -20.29	5.31	0.79 -4.52	5.13	0.81 -4.32	8.75	1.40 -7.35	9.76	1.60 -8.16	0.10	0.00 -0.10	0.01	0.00 -0.01	0.43	0.26 -0.17
LLaVA-OV-2-8B-Instruct	30.97	31.96 +0.99	54.39	55.02 +0.63	2.86	3.82 +0.96	2.83	3.79 +0.96	2.86	3.77 +0.91	2.85	3.75 +0.90	5.08	4.06 -1.02	1.75	1.94 +0.19	2.31	2.41 +0.10
VideoChat3-4B	31.30	34.27 +2.97	48.12	55.02 +6.90	15.52	16.32 +0.80	15.91	16.63 +0.72	20.16	20.37 +0.22	21.49	21.45 -0.04	2.60	1.83 -0.77	1.44	1.15 -0.29	2.42	1.31 -1.11
Qwen3.8-27B	36.57	37.56 +0.99	48.95	57.32 +8.37	16.78	18.35 +1.57	16.34	17.73 +1.39	17.68	19.93 +2.25	18.14	20.35 +2.21	7.91	7.16 -0.75	5.78	4.63 -1.15	5.17	6.24 +1.07
Qwen3.5-4B-SFT	**53.38**	51.40 -1.98	**93.31**	76.36 -16.95	27.79	**28.46** +0.67	27.79	**28.46** +0.67	28.93	**29.89** +0.96	29.19	**30.13** +0.94	**14.33**	13.59 -0.74	**8.45**	4.70 -3.75	**7.45**	5.69 -1.76

E.4 Full Label-Free Results

As defined in Sec. 3.3, "label-free" refers to the form of the released query: the predefined action-category name is not shown explicitly, and the action is instead specified through an observable movement description. The action label and its temporally aligned video clip are used during construction to generate this description, but the label itself is not exposed in the resulting query. Label-free evaluation is therefore not intended to remove label supervision from the data-generation process. It compares two ways in which the same action may be requested: an exact category name and a natural-language description closer to how users refer to an observed movement in practice. Each label-free query is paired with a label query that retains the same video, task conditions, query structure, and target, so their difference is limited to how the action is expressed.

Tab. 9 reports the complete comparison between label and label-free queries. Label-free queries improve Recognition for 23 of the 37 models and Temporal Reasoning for 19 models. Unlike a short category name, a movement description provides additional visual cues about how an action is performed. This information often helps models distinguish similar actions and identify their temporal relations. The gains are not universal, however, indicating that some models still rely more strongly on familiar action names than on descriptions of visible motion.

Grounding is nearly evenly divided: 18 models improve and 18 decline on F1@0.5. A label-free description can provide useful localization cues, but it also requires the model to associate a natural expression with the corresponding visual event. Detection shows a clearer decline. Among the 37 models supporting this task, 28 perform worse on F1@0.5 and 31 obtain lower average mAP. Many Detection queries use coarse action constraints to recover complete event sets. Replacing these constraints with longer and less precise movement descriptions introduces information that may be useful for recognizing one action but less effective for deciding which events belong in the predicted

set. These results show that understanding a described movement and using that description to recover a complete set of temporally bounded events remain distinct challenges.

E.5 Token Usage

We use an API to evaluate the proprietary and open-source large MLLMs on the main benchmark, consuming approximately 2.18 billion tokens in total, including 2.179 billion input tokens and 2.50 million output tokens. GPT-5.6 Luna accounted for 1.289 billion total tokens, followed by DeepSeek v4.1-Flash with 436.53 million, GLM-5.3 Flash with 303.16 million, and Qwen3.8-Max with 152.69 million. Input tokens dominate because the requests include video or sampled visual content in addition to the query and output instructions. These statistics are reported only as a reference for the scale of API usage. Tokenization and visual-input accounting differ across providers and model interfaces, so the counts should not be interpreted as a direct comparison of computational efficiency or monetary cost.

F Do MLLMs really look at micro-actions?

F.1 Semantic Priors or Micro-Actions

Table 10: Qwen3.5-4B performance with ordered video frames and text-only input.

Input	Rec. Acc.	Temp. Acc.	Grounding R@1 F1@.5	R@1 @.2	R@1 @.5	R@1 @.7	R@1 Avg.	mIoU	Detection mAP F1@.5	mAP @.2	mAP @.5	mAP @.7	mAP Avg.	Description CLIP-S	B-4	CIDEr-D	BLEURT	EM-S	AccL	tIoU	Err.
Ordered Frames	**30.65**	**64.12**	**3.65**	**6.89**	**3.74**	**1.94**	**3.93**	**4.01**	**2.14**	**2.70**	**1.43**	**0.81**	**1.48**	**52.25**	**0.26**	**0.55**	**26.63**	**19.68**	**7.72**	**16.38**	0.79
Text Only	21.32	52.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.06

To examine whether an MLLM answers from visual micro-actions or from regularities in the query, we evaluate Qwen3.5-4B under controlled inputs. The text-only setting removes the video and retains the original query. We compare the setting with the standard ordered-frame input using the same prompts and evaluation metrics.

We do not use single-frame or temporally shuffled inputs as additional controls. MMA-MTU is defined on complete untrimmed videos that may contain multiple action instances, including Recognition and Temporal Reasoning queries. A single frame cannot preserve the event coverage needed to answer these queries and therefore changes the evaluated task rather than providing a reduced version of the same visual evidence. Shuffling frames similarly removes the correspondence between visual observations and the original timeline. This makes temporal relations, timestamps, event boundaries, and multi-event outputs inconsistent with their ground truth. Performance under either setting would consequently mix the loss of visual evidence with a mismatch between the input and task definition. The text-only control instead keeps the query and target unchanged and is used specifically to measure how much can be answered from linguistic and category priors alone.

Tab. 10 shows that the text-only results remain close to chance level. Without video input, Qwen3.5-4B obtains 21.32% Accuracy on five-choice Recognition and 52.00% on binary Temporal Reasoning, close to their random baselines of 20% and 50%, respectively. This indicates that the query and option distribution provide little exploitable bias. With ordered video frames, the two scores increase to 30.65% and 64.12%, showing that the model uses visual micro-action evidence beyond the information available in the query alone. The text-only model also scores zero on Grounding, Detection, and Description, whose outputs require visible events and temporal boundaries. Together, these results support both the validity of the option construction and the contribution of visual evidence to the five tasks.

F.2 Guess or Actually See?

A model may obtain plausible localization results by predicting an action whenever it is mentioned in the query. To examine whether models can distinguish visible evidence from an absent action, MMA-MTU includes queries whose ground truth is "No_Match". The Grounding and Detection tasks contain 5 and 565 such queries, respectively. A response is correct when the model returns "No_Match" or a valid empty prediction set. We report "No_Match" Accuracy for each task. We

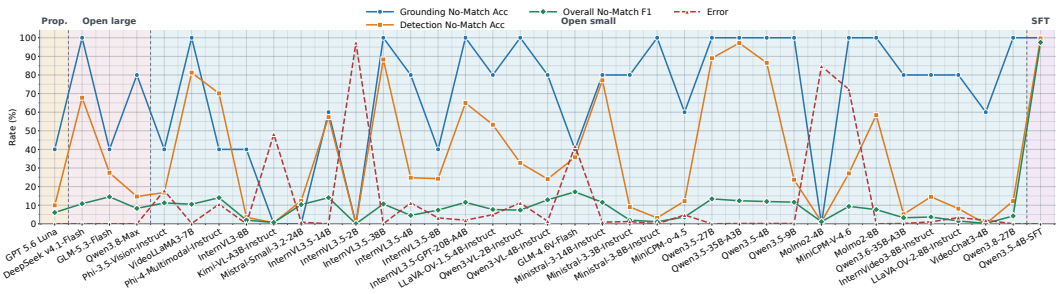


Figure 8: Performance on queries with no matching action.

also report overall "No_Match" F1 over positive and negative queries, which penalizes models that improve negative-query accuracy by rejecting actions that are actually present. Error Rate measures invalid outputs on the 570 "No_Match" queries.

Figure 8 shows large differences in how models handle absent actions. Detection "No_Match" Accuracy ranges from 0.00% to 99.65% among the zero-shot models. Several models rarely return "No_Match", suggesting that they tend to produce an event once an action is mentioned. Others reject most negative queries, but their much lower Overall "No_Match" F1 indicates that they also reject actions that are present. For example, Qwen3.5-35B-A3B reaches 97.17% Detection "No_Match" Accuracy but only 12.40% Overall "No_Match" F1. Correct rejection alone therefore does not establish that a model has grounded its answer in the video.

The fine-tuned Qwen3.5-4B shows a different pattern. It obtains 100.00% Grounding "No_Match" Accuracy, 99.65% Detection "No_Match" Accuracy, and 97.51% Overall "No_Match" F1, without invalid outputs. Compared with its zero-shot counterpart, the gain in F1 shows that fine-tuning improves both rejection of absent actions and retention of visible ones. These results expose a failure mode that positive-query localization metrics cannot capture: many MLLMs can predict where an action occurs, yet remain unreliable at deciding whether the action occurs at all. Because Grounding contains only five negative queries, its "No_Match" Accuracy should be treated as a supplementary result rather than a stable model ranking.

G Model Performance on the 12 Subtasks

G.1 More Results Across the 12 Subtasks

The Tab. 2 reports aggregate results for the five task categories. Tab. 11 divides them into 12 subtasks: VLMR and CGMR for Recognition; ORTR and OATR for Temporal Reasoning; FVMG, TCMG, and COMG for Grounding; FVMD and COMD for Detection; and FVDS, TCDS, and CCDS for Description. We use the same metrics as in the main paper.

Recognition changes only moderately under a coarse-category constraint. CGMR is lower than VLMR for 23 of the 36 models and higher for 12, with an average difference of -1.31 points. GPT 5.6 Luna leads VLMR with 42.78, while DeepSeek v4.1-Flash reaches the highest CGMR accuracy of 38.19. Temporal Reasoning shows a clearer pattern. OATR considers the overlap of up to three actions and is harder than the pairwise ordering tested by ORTR. All 36 models perform worse on OATR, with an average decrease of 12.38 points. GPT 5.6 Luna, for example, drops from 81.63 on ORTR to 55.12 on OATR, whereas Qwen3.8-Max obtains the best OATR result of 65.12.

The largest difference appears in Grounding. Every model with valid Grounding outputs performs worse on COMG than on FVMG. The best $\mathrm{F1@0.5}$ falls from 25.29 on FVMG and 26.93 on TCMG to 8.31 on COMG. Locating the overlap of two actions requires both events and their shared interval to be recovered correctly, making errors in either action directly affect the result. Restricting the temporal range has a much smaller effect: TCMG improves over FVMG for 15 models and decreases for 21, with an average change of only 0.22 points. A shorter search range therefore does not remove the main difficulty of recognizing subtle movements and aligning their boundaries. Detection follows a related pattern. COMD improves $\mathrm{F1@0.5}$ for 35 of the 36 models with valid results, but

Table 11: Performance across the 12 subtasks of MMA-MTU.

Model	Recognition		Temporal Reasoning		Grounding									Detection				Description																									
	VLMR	CGMR	ORTR	OATR	FVMG R@1			TCMG R@1			COMG R@1			FVMD mAP		COMD mAP		FVDS								TCDS								CCDS											
	Acc.	Acc.	Acc.	Acc.	F1@.5	Avg.	mIoU	F1@.5	Avg.	mIoU	F1@.5	Avg.	mIoU	F1@.5	Avg.	F1@.5	Avg.	CLIP	BL-4	CI-D	BLRT	EMS	Acc I_c	tIoU	CLIP	BL-4	CI-D	BLRT	EMS	Acc I_c	tIoU	CLIP	BL-4	CI-D	BLRT	EMS	Acc I_c	tIoU
Proprietary MLLMs																																											
GPT 5.6 Luna	**42.78**	37.45	**81.63**	55.12	24.19	26.59	26.96	23.45	25.04	25.62	8.10	**11.42**	**12.29**	**9.25**	**5.86**	**13.15**	**5.81**	66.80	1.35	**2.91**	45.22	25.41	19.77	44.51	63.72	1.77	0.94	42.84	23.93	**18.07**	**38.81**	62.99	1.22	1.46	44.65	23.74	35.24	**22.82**
Open-source Large MLLMs																																											
DeepSeek v4.1-Flash	42.03	**38.19**	60.44	44.88	6.51	6.93	7.03	7.55	7.61	7.95	0.31	0.45	0.52	2.32	2.25	2.67	1.99	74.23	1.19	0.36	45.29	27.74	6.16	42.13	69.81	1.82	1.86	43.97	26.17	10.67	34.73	75.18	1.49	2.05	43.85	27.68	30.33	14.16
GLM-5.3-Flash	31.90	29.02	71.11	58.21	24.59	**27.62**	**28.12**	24.87	26.49	26.86	**8.31**	10.56	11.33	6.59	4.52	8.62	4.05	74.26	0.68	1.78	43.00	28.02	18.99	44.20	66.90	1.13	1.98	38.44	25.32	14.02	37.45	74.78	0.77	1.96	41.06	27.87	36.33	22.09
Qwen3.8-Max	41.52	37.30	68.60	**65.12**	**25.29**	27.05	27.33	**26.93**	**27.22**	**27.41**	5.71	9.00	9.77	7.94	5.53	11.14	4.98	76.10	1.22	0.97	46.46	28.62	**20.57**	**45.82**	72.53	**2.69**	**5.48**	45.06	27.00	14.43	37.86	**85.01**	1.44	**4.44**	**46.96**	**31.44**	**37.80**	22.01
Open-source Small MLLMs																																											
Phi-3.5-Vision-Instruct	1.01	0.50	7.06	0.12	4.24	5.66	6.32	13.11	17.19	18.13	2.87	2.94	3.27	0.00	0.01	0.29	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.07	0.00	0.00	0.66	0.40	0.17	0.11
VideoLLaMA3-7B	22.53	26.69	67.03	48.33	0.02	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.44	4.78	0.00	0.01	2.78	1.83	0.85	2.28	0.35	0.00	0.00	0.18	0.13	0.00	0.00	1.48	0.00	0.00	0.84	0.55	0.27	0.22
Phi-4-Multimodal-Instruct	26.84	26.74	50.86	38.81	6.16	10.48	11.59	14.00	18.95	20.21	1.87	4.39	5.77	0.00	0.00	0.32	0.34	66.21	0.31	0.69	35.73	24.91	1.88	26.96	64.23	0.00	0.72	34.43	23.84	1.93	14.75	53.72	0.49	0.74	28.96	19.95	7.99	6.58
InternVL3-8B	34.68	30.06	56.51	45.95	12.20	17.40	18.70	10.01	14.02	15.19	2.70	5.78	6.64	0.25	1.07	1.13	1.60	67.96	0.17	1.01	40.74	26.09	3.46	32.93	60.80	0.08	0.00	36.65	23.44	3.44	26.73	65.15	0.09	0.01	39.96	24.86	25.19	8.03
Kimi-VL-A3B-Instruct	15.70	10.52	31.87	28.69	4.58	6.96	7.42	2.44	3.32	3.61	0.41	2.46	3.61	0.36	0.17	0.63	0.28	8.20	0.00	0.04	5.13	3.09	0.30	4.58	8.00	0.00	0.08	5.06	3.02	0.21	1.97	12.58	0.00	0.19	7.59	4.68	5.11	1.58
Mistral-Small-3.2-24B	32.41	30.85	68.92	39.29	14.24	19.46	20.52	12.80	16.59	17.59	4.05	9.52	11.43	0.40	0.66	1.65	1.52	**79.00**	1.35	0.19	**47.27**	**29.61**	1.89	39.06	**75.38**	1.86	2.08	**46.21**	**28.27**	3.15	32.19	77.20	**1.79**	2.97	45.42	28.62	22.88	10.03
InternVL3.5-14B	34.18	34.72	48.67	46.55	8.25	11.96	12.78	9.81	12.84	13.83	1.87	3.18	3.61	1.07	0.96	0.95	0.98	68.28	0.26	1.15	40.45	25.96	3.74	30.31	64.83	0.69	0.26	40.22	24.73	4.27	26.99	71.94	0.73	1.41	39.50	27.02	19.21	9.67
InternVL3.5-2B	23.80	18.85	21.04	13.10	3.52	6.22	7.01	1.14	2.45	2.68	0.86	3.32	4.62	0.00	0.00	0.22	0.11	16.38	0.00	0.00	5.88	6.31	0.34	7.91	5.98	0.00	0.00	2.23	2.30	0.00	2.91	6.45	0.00	0.00	3.39	2.51	1.49	0.79
InternVL3.5-38B	22.53	30.16	35.01	27.86	0.34	0.49	0.52	0.28	0.46	0.60	0.00	0.07	0.07	0.00	0.00	0.06	0.17	13.64	0.00	0.31	8.99	5.22	0.21	8.49	21.61	0.00	0.10	13.83	8.24	0.34	9.28	22.83	0.01	0.22	14.74	8.70	5.94	2.97
InternVL3.5-4B	1.77	1.84	15.70	6.07	1.48	2.01	2.16	0.46	0.64	0.69	0.00	0.00	0.00	0.27	0.22	0.48	0.53	64.89	0.24	1.39	35.34	24.67	3.17	28.21	40.90	0.06	0.04	22.06	15.53	2.62	13.85	30.00	0.01	0.12	15.96	11.28	5.27	3.74
InternVL3.5-8B	20.51	11.46	28.89	22.74	6.86	9.74	10.47	4.06	7.56	8.51	1.97	3.88	4.39	0.44	0.31	0.90	0.74	58.91	0.15	1.30	34.84	22.66	4.14	28.84	52.44	0.09	0.10	31.49	20.15	3.40	16.02	55.99	0.10	0.21	34.03	21.45	20.11	7.24
InternVL3.5-GPT-20B-A4B	34.94	36.76	48.04	40.95	4.68	6.40	6.92	6.67	7.20	7.54	0.31	0.62	0.72	0.24	0.42	0.81	0.79	69.94	0.25	1.00	39.29	26.60	5.08	33.02	69.59	0.00	0.22	38.51	26.43	3.26	28.04	59.69	0.17	0.13	33.12	22.47	17.72	7.34
LLaVA-OV-1.5-4B-Instruct	20.00	17.66	62.32	45.71	0.41	0.85	1.01	2.84	3.52	3.72	0.00	0.00	0.00	0.00	0.00	0.13	0.07	17.60	0.00	0.01	10.21	6.69	0.35	11.21	22.48	0.00	0.00	13.74	8.55	0.00	11.05	22.81	0.00	0.00	15.53	8.77	5.08	2.19
Qwen3-VL-2B-Instruct	11.39	5.90	21.98	7.74	8.47	12.56	13.26	7.71	8.70	9.03	1.21	2.01	2.22	0.22	0.48	0.75	0.78	10.38	0.00	0.07	6.11	3.92	0.07	3.70	4.61	0.00	0.08	2.58	1.74	0.03	0.92	10.10	0.00	0.13	5.44	3.82	2.25	0.97
Qwen3-VL-4B-Instruct	35.70	31.05	58.24	40.12	14.15	20.06	21.80	15.66	22.17	23.64	3.84	7.93	9.64	0.93	1.53	1.93	1.24	74.12	0.36	0.72	41.39	27.71	5.52	22.91	55.99	0.28	0.52	32.31	21.21	2.84	15.18	63.45	0.32	0.75	34.00	23.71	12.93	6.98
GLM-4.6V-Flash	31.39	30.06	64.84	47.38	13.27	16.62	17.19	13.18	13.25	13.54	4.05	4.78	5.23	0.70	0.57	1.08	1.18	46.72	0.11	1.70	30.57	17.68	6.67	33.87	39.83	0.07	0.07	26.06	15.25	3.72	20.94	31.35	0.03	0.08	21.40	11.89	14.79	9.15
Ministral-3-14B-Instruct	32.91	35.47	46.62	43.45	5.95	6.70	6.88	2.62	2.70	2.78	2.18	2.11	2.29	0.37	0.27	0.86	0.88	31.55	0.42	0.24	22.59	11.94	0.23	20.96	17.68	0.75	0.00	13.55	6.78	0.41	10.41	14.68	0.26	0.32	10.42	5.40	3.71	2.88
Ministral-3-3B-Instruct	22.03	18.65	50.55	44.64	6.04	8.15	8.54	6.82	8.31	8.98	1.25	2.21	2.35	0.94	0.34	1.32	0.79	32.41	0.22	0.00	23.36	12.36	0.87	19.14	6.98	0.00	0.20	4.79	2.60	0.46	2.94	4.94	0.00	0.29	3.27	1.88	0.89	0.88
Ministral-3-8B-Instruct	31.65	36.41	50.39	45.00	7.20	8.95	9.44	5.63	6.40	6.87	0.93	2.08	2.29	1.12	0.24	2.00	1.56	73.39	0.80	0.65	46.96	27.30	1.15	28.05	66.94	1.12	0.36	43.91	25.00	1.39	25.63	54.70	1.50	1.26	35.01	19.95	8.18	5.34
MiniCPM-o-4.5	23.80	21.68	63.27	43.69	3.84	5.74	6.35	5.88	7.00	7.35	1.03	1.35	1.61	1.02	0.38	2.32	1.20	63.38	0.32	1.06	37.78	23.85	5.25	33.37	39.20	0.15	0.44	24.68	14.98	2.58	19.01	61.89	0.49	0.62	37.13	22.97	23.76	9.78
Qwen3.5-27B	35.95	34.33	72.21	60.36	9.74	10.36	10.44	5.41	6.41	6.39	0.00	0.41	0.48	3.45	3.01	3.53	1.90	54.80	**1.52**	1.12	34.93	20.70	10.87	35.11	27.35	0.42	1.27	18.24	10.27	3.65	15.04	18.11	0.06	0.50	12.43	6.88	9.74	6.79
Qwen3.5-35B-A3B	33.92	33.58	62.32	51.43	2.17	2.32	2.37	0.32	0.30	0.34	0.00	0.00	0.00	0.28	0.18	0.67	0.48	76.04	0.57	0.78	42.05	28.74	10.79	39.74	69.59	0.76	0.96	39.81	26.35	8.93	33.27	63.06	0.62	0.54	36.45	24.01	23.57	12.24
Qwen3.5-4B	33.16	30.16	77.86	53.69	3.86	4.02	4.08	3.69	5.04	5.19	0.00	0.35	0.52	1.96	1.37	2.18	1.88	66.59	0.33	0.78	33.03	25.03	7.55	23.71	58.73	0.00	0.60	30.59	22.10	4.79	20.81	34.04	0.07	0.60	17.80	12.88	9.84	6.32
Qwen3.5-9B	28.86	28.87	67.66	57.50	18.38	23.18	24.17	15.09	17.58	18.22	1.77	3.04	3.63	3.73	3.79	5.13	3.02	75.05	0.63	1.44	42.27	28.39	7.71	31.90	69.23	1.11	0.38	40.07	26.29	6.13	25.37	72.43	1.03	0.59	41.95	27.36	24.57	15.48
Molmo2-4B	0.00	0.00	0.63	0.24	5.03	7.04	7.43	5.02	6.45	6.84	0.62	1.90	2.43	0.07	0.06	0.28	0.71	37.79	0.10	0.53	23.15	14.49	2.20	12.99	10.70	0.00	0.08	6.53	4.10	0.21	2.82	4.28	0.00	0.09	2.76	1.61	1.68	0.73
MiniCPM-V-4.6	8.35	5.95	17.90	9.29	3.68	5.75	6.40	2.43	3.48	3.97	0.31	0.93	1.38	0.00	0.04	0.26	0.14	27.60	0.00	0.40	17.54	10.42	0.42	8.61	37.48	0.00	0.40	25.28	14.04	0.41	12.22	49.24	0.13	0.31	32.17	18.44	12.07	6.44
Molmo2-8B	11.14	19.44	73.16	57.14	1.91	2.40	2.46	2.49	3.11	3.14	0.00	0.03	0.07	0.48	0.09	0.88	0.85	5.60	0.00	0.02	3.41	2.18	0.85	3.94	1.05	0.00	0.01	0.71	0.41	0.08	0.14	0.45	0.00	0.00	0.36	0.18	0.41	0.30
Qwen3.6-35B-A3B	35.44	37.35	65.78	49.05	18.71	20.64	21.24	18.95	19.15	19.57	2.70	4.95	5.98	5.16	3.55	7.22	2.84	78.22	0.70	1.28	43.81	29.46	17.36	42.48	73.04	0.54	0.30	40.86	27.87	11.89	36.29	80.57	1.13	1.16	44.48	30.14	32.76	17.04
InternVideo3-8B-Instruct	0.00	1.09	38.62	17.98	4.19	6.95	7.76	10.50	12.49	13.31	0.00	0.00	0.03	0.00	0.03	0.03	0.33	71.16	0.56	0.50	41.95	26.74	2.78	28.75	45.62	0.30	0.62	29.13	17.41	0.69	16.46	57.34	0.61	1.38	32.68	21.45	7.74	4.99
LLaVA-OV-2-8B-Instruct	34.43	35.81	69.07	45.95	3.44	3.50	3.49	3.28	4.00	4.05	0.00	0.07	0.08	2.09	1.07	4.30	1.57	15.30	0.00	0.09	9.47	5.85	0.85	11.30	0.55	0.00	0.00	0.33	0.21	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VideoChat3-4B	40.25	30.16	65.93	44.76	15.53	20.39	21.62	16.62	21.94	23.49	1.27	3.67	4.52	1.06	1.03	2.39	1.13	4.70	0.00	0.18	2.97	1.79	0.42	3.00	7.61	0.00	0.01	4.13	2.92	0.41	2.35	0.92	0.00	0.00	0.55	0.36	0.05	0.14
Qwen3.8-27B	40.25	38.00	53.53	51.67	17.14	19.23	19.71	16.52	16.81	17.14	3.74	5.57	6.48	5.69	2.77	7.41	3.76	78.30	1.08	1.11	45.73	29.46	13.81	44.12	73.45	1.37	1.25	43.46	27.79	12.57	34.33	81.15	1.34	2.41	46.11	30.23	34.14	20.41

the average gain is only 0.82 points. Its effect on average mAP is smaller, indicating that category or temporal constraints help recover relevant events more than they improve boundary precision.

Description produces different rankings under language-based and event-based metrics. On FVDS and TCDS, Mistral-Small-3.2-24B leads several reference-based metrics, including CLIPScore, BLEURT, and EMScore. Qwen3.8-Max instead obtains the best FVDS label accuracy and temporal IoU, while GPT 5.6 Luna leads these two measures on TCDS. Qwen3.8-Max also achieves the best results on most CCDS metrics. A description can therefore resemble the reference text while assigning an incorrect action label or temporal boundary. Reporting language similarity together with label accuracy and temporal IoU separates these sources of error.

The subtask results also show that temporal reasoning does not imply temporal localization. Qwen3.5-4B reaches 77.86 on ORTR but only 3.86 F1@0.5 on FVMG, 0.00 on COMG, and 1.96 on FVMD. Molmo2-8B similarly obtains 73.16 on ORTR and 57.14 on OATR, yet reaches only 1.91 on FVMG and 0.00 on COMG. Models can infer a temporal relation without recovering the intervals that support it. The changing rankings across the 12 subtasks confirm that Recognition, Temporal Reasoning, Grounding, Detection, and Description capture related but distinct parts of micro-action understanding.

G.2 More Model Performance Analysis

Analysis protocol. We use Accuracy for Recognition and Temporal Reasoning, F1@0.5 for Grounding and Detection, and label accuracy for Description as the primary subtask metrics. Fig. 9 applies min-max normalization to model scores within each subtask. Fig. 10 computes pairwise Spearman correlations from the model rankings produced by these primary metrics. We use these analyses to examine whether performance on one task can represent a model's broader micro-action understanding ability.

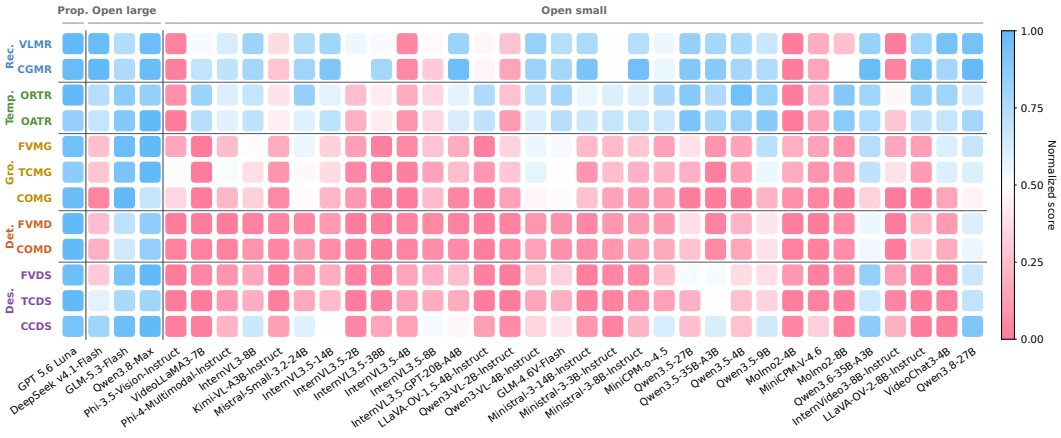


Figure 9: Normalized performance of evaluated MLLMs across the 12 MMA-MTU subtasks. The results show variation in model strengths across Recognition, Temporal Reasoning, Grounding, Detection, and Description.

Capability profiles across subtasks. Fig. 9 shows substantial variation in model strengths across the 12 subtasks. GPT 5.6 Luna, Qwen3.8-Max, and GLM-5.3-Flash perform strongly across several task families, whereas many models exhibit more specialized profiles. DeepSeek v4.1-Flash is competitive on Recognition but weaker on Grounding and Detection. Molmo2-8B and Qwen3.5-4B perform well on Temporal Reasoning without comparable performance in temporal localization. A model that recognizes actions or reasons about their relations well may therefore still fail to recover the corresponding events and temporal boundaries. Evaluating only one of these abilities would provide an incomplete and potentially misleading account of model performance.

Relationships between subtasks. Fig. 10 shows substantially stronger correlations within task families than across them. The average correlation is 0.87 between subtasks from the same family, compared with 0.54 across different families. Recognition, Temporal Reasoning, Detection, and

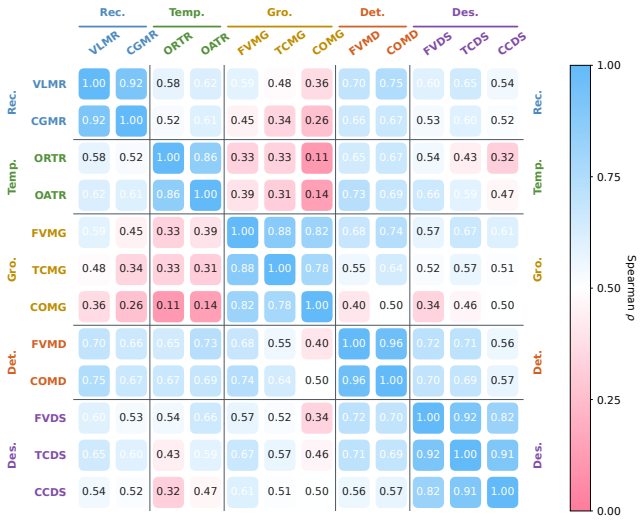


Figure 10: Spearman rank correlation of model performance across the 12 subtasks of MMA-MTU. Subtasks within the same task show strong correlations, while correlations across task are substantially weaker.

Description have within-family correlations of 0.92, 0.85, 0.96, and 0.82–0.92, respectively, while correlations among the three Grounding subtasks range from 0.78 to 0.88. These results support the internal consistency of the five task families while showing that they measure different aspects of micro-action understanding.

The clearest separation occurs between Temporal Reasoning and Grounding, whose average cross-family correlation is only 0.27. COMG correlates weakly with ORTR (0.11) and OATR (0.14), showing that judging temporal relations does not imply an ability to locate the evidence establishing them. Detection is more broadly correlated with the other families because it combines action recognition with temporal localization, but its strongest correlation remains between FVMD and COMD (0.96).

Together with the heterogeneous model profiles, these results provide direct empirical motivation for MMA-MTU: **Recognition, Temporal Reasoning, Grounding, Detection, and Description are connected but non-equivalent capabilities, and a unified benchmark is needed to evaluate them on the same video evidence.**

H More Experiments of Supervised Fine-Tuning

H.1 Supervised Fine-Tuning Setup

We fine-tune Qwen3.5-4B and Qwen3-VL-2B Bai et al. (2025a) on the MMA-MTU training split using QLoRA. Both models are loaded with 4-bit NF4 quantization and bfloat16 computation, with double quantization additionally enabled for Qwen3-VL-2B. We freeze their visual encoders and insert rank-16 LoRA adapters with $\alpha = 32$ and a dropout rate of 0.05. For Qwen3.5-4B, the adapters are applied to the attention, linear-attention, and feed-forward projection layers, resulting in 32.5M trainable parameters. For Qwen3-VL-2B, they are applied to the self-attention and feed-forward projection layers, resulting in 17.4M trainable parameters.

Each of the 86,501 training queries is treated as an independent multimodal instruction-following sample comprising a video, a task-specific query and output schema, and a structured target. Supervision is applied only to the assistant response. The targets follow the same interfaces used during evaluation: option IDs for Recognition and Temporal Reasoning, temporal intervals for Grounding, action–interval event sets for Detection, and structured events with natural-language descriptions for Description. Videos are sampled at 4 FPS and resized to 336×336 using aspect-ratio-preserving padding. Each member of a paired label and label-free query receives a weight of 0.5, while all

Table 12: Effect of supervised fine-tuning on MMA-MTU. The Δ row reports SFT minus Base.

Model	Rec.	Temp.	Grounding						Detection					Description							Err.↓
	Acc.	Acc.	F1@.5	R@1 @.2	R@1 @.5	R@1 @.7	R@1 Avg.	mIoU	F1@.5	mAP @.2	mAP @.5	mAP @.7	mAP Avg.	CLIP	BL-4	CI-D	BLRT	EMS	Acc_L	tIoU	
Qwen3.5-4B	30.65	64.12	3.65	6.89	3.74	1.94	3.93	4.01	2.14	2.70	1.43	**0.81**	1.48	52.25	0.26	0.55	26.63	19.68	7.72	16.38	0.79
Qwen3.5-4B-SFT	**59.98**	**86.46**	**27.97**	**53.30**	**29.00**	**13.39**	**30.27**	**30.49**	**13.03**	**4.83**	**1.95**	**0.81**	**2.41**	**60.36**	**32.55**	**62.20**	**59.56**	**23.60**	**38.51**	**48.12**	**0.13**
Δ	+29.33	+22.34	+24.32	+46.41	+25.26	+11.45	+26.34	+26.48	+10.89	+2.13	+0.52	0.00	+0.93	+8.11	+32.29	+61.65	+32.93	+3.92	+30.79	+31.74	-0.66
Qwen3-VL-2B-Instruct	6.80	13.88	8.03	22.74	8.33	4.26	11.65	12.30	0.65	1.11	0.28	0.10	0.51	8.82	0.00	0.08	4.97	3.33	0.89	1.96	27.83
Qwen3-VL-2B-Instruct-SFT	**66.90**	**89.51**	**37.79**	**64.66**	**38.90**	**20.54**	**38.83**	**38.69**	**17.25**	**7.97**	**3.94**	**0.81**	**4.30**	**60.04**	**34.13**	**72.84**	**60.17**	**23.48**	**44.32**	**51.60**	**0.04**
Δ	+60.10	+75.63	+29.76	+41.92	+30.57	+16.28	+27.18	+26.39	+16.60	+6.86	+3.66	+0.71	+3.79	+51.22	+34.13	+72.76	+55.20	+20.15	+43.43	+49.64	-27.79

other samples retain a weight of 1.0, preventing the same semantic fact from being counted twice. We optimize both models for two epochs using fused AdamW with a learning rate of 2×10^{-4}.

H.2 More Results and Analysis

Tab. 12 shows that supervised fine-tuning substantially improves both models across all five tasks. Qwen3-VL-2B-Instruct-SFT performs better in Recognition, Temporal Reasoning, Grounding, and Detection, including gains of 9.82 points on Grounding F1@0.5 and 4.22 points on Detection F1@0.5 over the fine-tuned 4B model. Its lower Error Rate also indicates more reliable compliance with the required output structures. Since the difference in Error Rate cannot explain the overall gap, the improvement mainly reflects better action understanding and temporal localization rather than output validity alone.

Description improves markedly after fine-tuning, but its metrics capture different aspects of the output. Reference-based metrics such as BLEU-4, CIDEr-D, BLEURT, and EMScore measure whether the generated text follows the content and expression patterns of the ground-truth descriptions. Their large increases show that the models learn the expected description structure and produce language that more closely matches the annotation style. CLIPScore provides complementary evidence by measuring visual-semantic alignment rather than surface-form overlap. Its improvement for both models indicates that the generated descriptions also become more consistent with the observed micro-actions. Meanwhile, Acc_L and tIoU evaluate whether the descriptions recover the correct action categories and temporal intervals. Improvements in these structured metrics show that the gains are not limited to linguistic imitation: the models more accurately identify which events occur and when they occur.

The two fine-tuned models nevertheless exhibit different Description profiles. Qwen3-VL-2B-Instruct-SFT performs better on BLEU-4, CIDEr-D, BLEURT, Acc_L, and tIoU, whereas Qwen3.5-4B-SFT retains small advantages on CLIPScore and EMScore. Thus, no single metric is sufficient to characterize description quality. Reference-based language metrics reflect conformity to the released descriptions, CLIPScore measures broader visual-semantic agreement, and the structured metrics assess action identity and temporal alignment. Their joint improvement provides stronger evidence that supervised fine-tuning enhances both description generation and the underlying micro-action understanding. Under the current QLoRA configuration, the smaller Qwen3-VL-2B adapts more effectively than Qwen3.5-4B, although this comparison should not be interpreted as a general advantage of smaller models because their architectures differ.

I More Qualitative Examples

Fig. 11 illustrates the gap between fluent language generation and complete micro-action understanding. The ground truth contains five events: Pointing, Scratching arms, Touching nose, Playing or tidying hair, and Touching legs. Most models recognize the broad progression of hand–arm and hand–face motion, but they frequently confuse the precise interaction target. Touching nose is variously predicted as Covering mouth, Rubbing lip, Touching cheek, or Touching forehead, while Playing or tidying hair is often merged with other hand-head actions. Several models also omit Pointing and Touching legs. These errors show that detecting motion near a body region is easier than determining the exact contact target and assigning the corresponding fine-grained category.

The examples also reveal different temporal failure modes. Qwen3.8-Max recovers Scratching arms and Touching nose with relatively accurate intervals but misses several events. GPT-5.6-Luna produces a detailed account yet adds extra actions, whereas InternVideo3-8B collapses most of the video into one long event. Phi-4-Multimodal instead emphasizes scene appearance and assigns a single coarse label to the full video. Supervised fine-tuning substantially improves output structure

MMA-MTU Description Ground-Truth

Motion-Evidence-Guided MLLM Annotation



Frame Difference
Flow
Skeleton
Relative Trajectory

Query: Identify every fine-grained action in the 17.06-second video. For each event, give its action label, describe the visible movement, and report its start and end times.
Answer: A total of **5** events match. From 1.20 to 2.55 seconds, the person extends their right arm and points with their hand while the left arm moves upward and the right elbow bends. The action label is "Pointing". From 2.75 to 6.68 seconds, the left hand moves first to scratch the right arm, followed by the right hand, with both elbows bending and wrists approaching each other. The action label is "Scratching arms". From 6.80 to 9.35 seconds, the person touches their nose while their head moves downward. The action label is "Touching nose". From 9.51 to 12.63 seconds, the person plays with their hair, moving their right hand first to approach the head while the left hand follows. The action label is "Playing or tidying hair". From 14.71 to 16.80 seconds, the person brings both hands down towards their legs, with the wrists moving in a mixed direction while the elbows bend to facilitate the contact. The action label is "Touching legs".

Predictions from MLLMs

Chat GPT 5.6 Luna

Description: The woman raises her right hand and gestures near her head, then lowers it and rubs her left upper arm with her right hand. She continues rubbing the left arm before bringing her right hand to cover her mouth and nose. She then touches and rubs around her mouth and cheek, moves her hand through the front and side of her hair, touches near her ear, and briefly rubs around her eye and forehead. She lowers her hand and remains seated, turning her head and making small head movements while speaking.
Labels & Timestamps: [Other hand movement, 1.50~2.25]; [Scratching arms, 3.25~7.25]; [Covering mouth, 7.50~8.50]; [Rubbing lip, 8.50~9.00]; [Playing or tidying hair, 10.25~11.50]; [Touching ears, 11.00~12.00]; [Rubbing eyes, 11.50~12.25]; [Other hand movement, 12.00~12.50]; [Turning head, 13.00~16.75]

Deep Seek V4.1 Flash

Description: The video shows a woman seated at a metal mesh table engaging in a continuous sequence of hand-to-face and upper-body movements. From about 0.75 seconds to 3.0 seconds she raises her right hand and touches her neck and shoulder area. Between roughly 3.5 seconds and 6.5 seconds she folds her arms and rests her hands against her body. From about 7.0 seconds to 9.0 seconds she brings her right hand to her mouth and covers it. Around 10.5 seconds to 12.5 seconds she moves her hand up to scratch and touch her forehead and hair. Finally, from about 12.5 seconds to 13.5 seconds she touches her cheek and ear.
Labels & Timestamps: [Scratching or touching neck, 0.75~3.00]; [Folding arms, 3.50~6.50]; [Covering mouth, 7.00~9.00]; [Scratching or touching forehead, 10.50~12.50]; [Scratching or touching cheek, 12.50~13.50]

GLM 5.3 Flash

Description: A woman with pink curly hair sits at a metal outdoor table talking. Around 3 seconds she lifts her left hand and rubs and holds her right upper arm and shoulder, moving her hand slightly along the sleeve until about 7 seconds. She then raises her left hand to her face, briefly touching her cheek and covering her mouth and nose with her hand from about 7 to 9 seconds before lowering it. Between about 10.5 and 12.5 seconds she again lifts her left hand to her face, touching her forehead and cheek while speaking, then drops her hand. Throughout the clip she also turns her head from side to side and makes small illustrative hand gestures while talking.
Labels & Timestamps: [Scratching or touching shoulder, 3.00~7.00]; [Covering mouth, 7.00~9.00]; [Scratching or touching forehead, 10.50~12.50]; [Illustrative gestures, 13.00~17.00]

Qwen 3.8 MAX

Description: The video shows a woman sitting at an outdoor table performing several distinct fine-grained actions. Initially, she rests her arm on the table before lifting her hand to rub and scratch her upper right arm between 3.6 and 7.2 seconds. Shortly after, she raises her left hand to touch and cover her nose and mouth area from 7.8 to 9.2 seconds. Finally, she lifts her left hand again to scratch or touch her forehead and temple region between 10.6 and 12.8 seconds.
Labels & Timestamps: [Scratching arms, 3.60~7.20]; [Touching nose, 7.80~9.20]; [Scratching or touching forehead, 10.60~12.80]

Intern Video3 8B Instruct

Description: The video shows a person sitting at a table outdoors, with their right hand raised to their head, fingers interlaced or touching the hair. The hand remains in this position throughout the clip, with slight movements indicating hair adjustment or scratching. The person's head is slightly tilted, and their gaze is directed downward or to the side. There is no significant change in posture or other body movements observed.
Labels & Timestamps: [Scratching or touching head, 0.00~17.06]

Mini stral3 8B Instruct

Description: In the provided 17.06-second video, the individual performs a series of observable fine-grained actions. Initially, from approximately 0.00s to 1.00s, the person is seen holding a laptop bag on her lap while her right hand interacts with the bag, which can be labeled as 'Other hand movement'. Between 1.00s and 2.00s, she begins to cover her mouth with her right hand, which aligns with the action 'Covering mouth'. From around 2.00s to 7.00s, she repeatedly covers her mouth and then touches her face near her cheek, which can be categorized as 'Scratching or touching cheek'. Between 7.00s and 10.00s, she continues to cover her mouth and occasionally touches her face, maintaining the 'Covering mouth' and 'Scratching or touching cheek' actions. From 10.00s to 17.06s, she predominantly keeps her hands near her face, with her right hand covering her mouth and occasionally touching her cheek, continuing the 'Covering mouth' and 'Scratching or touching cheek' actions. No other fine-grained actions from the MMA82 ontology are distinctly observable within the video.
Labels & Timestamps: [Rubbing eyes, 2.75~3.75]; [Rubbing hands, 5.50~6.50]; [Rubbing eyes, 8.50~10.50]; [Rubbing hands, 12.50~13.50]; [Rubbing eyes, 15.50~16.50]

Phi4 Multimodal

Description: The woman is sitting at a table with a grill in front of her. She is talking and gesturing with her hands. She appears to be explaining something to the camera. The background shows a park with trees and a gazebo. The woman is wearing a black top and jeans. She has pink hair and is wearing a necklace. The table has a black bag on it. The woman is sitting on a metal chair.
Labels & Timestamps: [Holding hands, 0.00~17.06]

Qwen 3.5 4B

Description: A woman with pink hair sits at a metal table outdoors. She is talking and gesturing with her hands. At one point, she touches her face with her hand, specifically near her mouth and cheek area.
Labels & Timestamps: [Illustrative gestures, 2.00~7.00]; [Covering mouth, 7.50~9.00]; [Scratching or touching cheek, 10.00~12.00]

Qwen 3.5 4B SFT

Description: A total of 4 events match. From 0.00 to 1.80 seconds, the head turns horizontally while the right wrist moves upward and approaches the face. The action label is "Turning head". From 1.80 to 3.00 seconds, the person scratches or touches their forehead while the left elbow bends and the right elbow extends. The action label is "Scratching or touching forehead". From 3.00 to 5.00 seconds, the person scratches or touches their chest or abdomen, with the left hand moving first and the right hand following as the elbows bend. The action label is "Scratching or touching chest or abdomen". From 5.00 to 8.30 seconds, the person rubs their eyes with both hands moving approximately simultaneously, with the left hand moving first while the right elbow bends. The action label is "Rubbing eyes"
Labels & Timestamps: [Turning head, 0.00~1.80]; [Scratching or touching forehead, 1.80~3.00]; [Scratching or touching chest or abdomen, 3.00~5.00]; [Rubbing eyes, 5.00~8.30]

Qwen 3 2B

No Match

Qwen 3 2B SFT

Description: A total of 4 events match. From 0.00 to 2.80 seconds, the head moves horizontally to the left while the shoulders shift vertically in opposite directions. The action label is "Nodding". From 2.80 to 6.10 seconds, the person scratches their arms, with the left hand moving first and the right hand following as the elbows bend and the wrists approach each other. The action label is "Scratching arms". From 6.10 to 7.60 seconds, the person touches their nose with their right hand while their left hand moves away. The action label is "Touching nose". From 7.60 to 10.00 seconds, the person touches their nose with their right hand while their left hand moves away. The action label is "Touching nose".",
Labels & Timestamps: [Nodding, 0.00~2.80]; [Scratching arms, 2.80~6.10]; [Touching nose, 6.10~7.60]; [Touching nose, 7.60~10.00]

Figure 11: Qualitative comparison on the MMA-MTU Description task. The top presents the motion-evidence-guided ground truth, followed by event descriptions, labels, and temporal boundaries predicted by representative MLLMs.

and event-level descriptions. In particular, Qwen3-VL-2B changes from "No_Match" to a structured multi-event prediction that correctly identifies Scratching arms and Touching nose and describes their visible mechanics. Nevertheless, both fine-tuned models still miss or misclassify some events. The qualitative results are therefore consistent with the quantitative metrics: fine-tuning improves description format, action semantics, and temporal alignment, but complete recovery of subtle, consecutive micro-actions remains difficult.

J More Future Exploration Results

Table 13: Statistics of the 80 emotion-annotated videos.

Emotion	Videos	Queries	Avg. Queries	Proportion	Total Duration	Avg. Duration	Action Instances	Avg. Instances
Anger	11	74	6.73	13.75%	39.49s	3.59s	23	2.09
Disgust	10	34	3.40	12.50%	34.58s	3.46s	15	1.50
Fear	12	53	4.42	15.00%	38.37s	3.20s	20	1.67
Happy	7	37	5.29	8.75%	19.26s	2.75s	16	2.29
Melancholy	13	62	4.77	16.25%	81.41s	6.26s	22	1.69
Neutral	8	43	5.38	10.00%	25.39s	3.17s	19	2.38
Sad	11	76	6.91	13.75%	64.34s	5.85s	23	2.09
Surprise	8	53	6.63	10.00%	26.89s	3.36s	15	1.88
Overall	**80**	**432**	**5.40**	**100.00%**	**329.72**	**4.12**	**153**	**1.91**

Tab. 13 summarizes the emotion subset used in this exploration. It contains 80 videos from eight emotion categories, with 432 queries and 153 annotated micro-action instances. The videos have a total duration of 329.72 seconds and contain 5.40 queries and 1.91 action instances per video on average. The number of videos in each category ranges from 7 to 13, while the number of queries ranges from 34 to 76. Video duration and action density also vary across categories: Melancholy and Sad contain longer videos, whereas Neutral and Happy have more action instances per video.

The prediction (Pre) and ground-truth (GT) conditions aggregate outputs from Recognition, Temporal Reasoning, Grounding, Detection, and Description. For each video, Pre collects the parsed responses produced by MLLM for every associated query and concatenates the query, its choices when applicable, and the predicted result in their original order. Predicted option IDs are mapped back to their choice text to make the evidence readable. A query is omitted if its prediction cannot be parsed, although all 432 predictions for the 80 emotion videos are available in this experiment. GT follows the same construction but replaces each predicted result with the corresponding annotated answer. Pre and GT therefore contain identical queries, choices, and query counts; they differ in whether their results come from model predictions or ground-truth annotations.

Tab. 14 reports the 33 complete emotion-recognition results for the models evaluated in the main benchmark, together with two models fine-tuned using SFT. We separately examine complete and face-masked RGB inputs, and report how predicted or ground-truth micro-action descriptions change the corresponding visual baseline.

The effect of micro-action evidence depends strongly on the available visual information. With complete RGB frames, predicted descriptions improve 17 of the 35 models with available results and reduce performance for 14, producing only a small average gain of 0.43 points. Ground-truth descriptions are more reliable, but their average improvement over RGB remains limited to 0.79 points. Facial appearance and other visual context already provide strong emotion cues in this setting, so additional micro-action information is useful only when it is sufficiently accurate and complementary.

The benefit becomes clearer after facial information is removed. Adding predicted micro-actions to RGB (Mask) improves 23 of 35 models by 2.39 points on average, while ground-truth descriptions improve 25 of 35 models by 2.89 points. Micro-actions therefore provide behavioral evidence that is not contained in facial appearance and can partly compensate for its removal. The two SFT models show different transfer patterns: fine-tuning improves some masked-input combinations but does not consistently improve emotion recognition. Learning the MMA-MTU tasks strengthens micro-action understanding, but the downstream benefit still depends on the quality and compatibility of the predicted evidence.

Table 14: Emotion recognition accuracy (%) using complete and face-masked RGB frames. **Pre** denotes model-predicted micro-action outputs, and **GT** denotes ground-truth micro-action descriptions. Small values beside combined-input results show their difference from the corresponding RGB baseline.

Model	RGB					RGB (Mask)				
	RGB	Pre	RGB+Pre	GT	RGB+GT	RGB	Pre	RGB +Pre	GT	RGB +GT
Proprietary MLLMs										
GPT 5.6 Luna	31.25	12.50	36.25 +5.00	15.00	36.25 +5.00	**15.00**	12.50	15.00 0.00	15.00	17.50 +2.50
Open-source Large MLLMs										
DeepSeek v4.1-Flash	30.00	11.25	27.50 -2.50	11.25	32.50 +2.50	10.00	11.25	13.75 +3.75	11.25	12.50 +2.50
GLM-5.3-Flash	32.50	18.75	**37.50** +5.00	17.50	30.00 -2.50	11.25	18.75	18.75 +7.50	17.50	20.00 +8.75
Qwen3.8-Max	**38.75**	12.50	36.25 -2.50	12.50	**38.75** 0.00	8.75	12.50	18.75 +10.00	12.50	20.00 +11.25
Open-source Small MLLMs										
Phi-3.5-Vision-Instruct	20.00	10.00	15.00 -5.00	10.00	16.25 -3.75	12.50	10.00	12.50 0.00	10.00	15.00 +2.50
VideoLLaMA3-7B	23.75	10.00	22.50 -1.25	12.50	25.00 +1.25	12.50	10.00	12.50 0.00	12.50	15.00 +2.50
Phi-4-Multimodal-Instruct	16.25	11.25	13.75 -2.50	10.00	10.00 -6.25	11.25	11.25	10.00 -1.25	10.00	6.25 -5.00
InternVL3-8B	25.00	13.75	30.00 +5.00	13.75	28.75 +3.75	11.25	13.75	13.75 +2.50	13.75	15.00 +3.75
Kimi-VL-A3B-Instruct	22.50	12.50	17.50 -5.00	10.00	18.75 -3.75	11.25	12.50	8.75 -2.50	10.00	10.00 -1.25
Mistral-Small-3.2-24B	27.50	**22.50**	28.75 +1.25	25.00	30.00 +2.50	12.50	**22.50**	17.50 +5.00	25.00	17.50 +5.00
InternVL3.5-14B	22.50	11.25	25.00 +2.50	11.25	23.75 +1.25	7.50	11.25	12.50 +5.00	11.25	11.25 +3.75
InternVL3.5-38B	11.25	10.00	12.50 +1.25	11.25	13.75 +2.50	11.25	10.00	8.75 -2.50	11.25	11.25 0.00
InternVL3.5-4B	23.75	15.00	23.75 0.00	17.50	23.75 0.00	**15.00**	15.00	16.25 +1.25	17.50	20.00 +5.00
InternVL3.5-8B	23.75	15.00	26.25 +2.50	13.75	27.50 +3.75	10.00	15.00	11.25 +1.25	13.75	10.00 0.00
InternVL3.5-GPT-20B-A4B	18.75	11.25	16.25 -2.50	10.00	13.75 -5.00	10.00	11.25	10.00 0.00	10.00	7.50 -2.50
LLaVA-OV-1.5-4B-Instruct	20.00	8.75	22.50 +2.50	7.50	22.50 +2.50	**15.00**	8.75	13.75 -1.25	7.50	15.00 0.00
Qwen3-VL-2B-Instruct	32.50	11.25	31.25 -1.25	10.00	31.25 -1.25	13.75	11.25	20.00 +6.25	10.00	15.00 +1.25
Qwen3-VL-4B-Instruct	22.50	12.50	27.50 +5.00	10.00	27.50 +5.00	**15.00**	12.50	20.00 +5.00	10.00	15.00 0.00
GLM-4.6V-Flash	6.25	20.00	0.00 -6.25	18.75	1.25 -5.00	1.25	20.00	0.00 -1.25	18.75	0.00 -1.25
Ministral-3-14B-Instruct	28.75	16.25	26.25 -2.50	21.25	26.25 -2.50	12.50	16.25	17.50 +5.00	21.25	**22.50** +10.00
Ministral-3-3B-Instruct	22.50	17.50	22.50 0.00	**28.75**	27.50 +5.00	11.25	17.50	13.75 +2.50	**28.75**	15.00 +3.75
Ministral-3-8B-Instruct	28.75	21.25	28.75 0.00	20.00	30.00 +1.25	**15.00**	21.25	**21.25** +6.25	20.00	21.25 +6.25
MiniCPM-o-4.5	30.00	17.50	32.50 +2.50	15.00	32.50 +2.50	12.50	17.50	16.25 +3.75	15.00	16.25 +3.75
Qwen3.5-27B	28.75	16.25	26.25 -2.50	18.75	30.00 +1.25	12.50	16.25	15.00 +2.50	18.75	17.50 +5.00
Qwen3.5-4B	26.25	11.25	27.50 +1.25	11.25	28.75 +2.50	12.50	11.25	11.25 -1.25	11.25	13.75 +1.25
Qwen3.5-9B	30.00	16.25	31.25 +1.25	11.25	30.00 0.00	13.75	16.25	15.00 +1.25	11.25	17.50 +3.75
Molmo2-4B	22.50	15.00	22.50 0.00	18.75	25.00 +2.50	11.25	15.00	11.25 0.00	18.75	10.00 -1.25
MiniCPM-V-4.6	21.25	11.25	26.25 +5.00	11.25	23.75 +2.50	8.75	11.25	12.50 +3.75	11.25	12.50 +3.75
Molmo2-8B	27.50	15.00	26.25 -1.25	8.75	31.25 +3.75	10.00	15.00	13.75 +3.75	8.75	15.00 +5.00
InternVideo3-8B-Instruct	21.25	11.25	26.25 +5.00	10.00	22.50 +1.25	12.50	11.25	15.00 +2.50	10.00	15.00 +2.50
LLaVA-OV-2-8B-Instruct	28.75	11.25	27.50 -1.25	8.75	27.50 -1.25	12.50	11.25	13.75 +1.25	8.75	12.50 0.00
VideoChat3-4B	22.50	13.75	27.50 +5.00	11.25	27.50 +5.00	8.75	13.75	15.00 +6.25	11.25	16.25 +7.50
Qwen3.8-27B	28.75	17.50	31.25 +2.50	17.50	30.00 +1.25	12.50	17.50	18.75 +6.25	17.50	18.75 +6.25
Open-source Small MLLMs + SFT										
Qwen3.5-4B-SFT	12.50	15.00	13.75 +1.25	16.25	13.75 +1.25	12.50	15.00	15.00 +2.50	16.25	16.25 +3.75
Qwen3-VL-2B-Instruct-SFT	21.25	5.00	18.75 -2.50	10.00	20.00 -1.25	13.75	5.00	12.50 -1.25	10.00	15.00 +1.25

K LIMITATIONS

MMA-MTU derives its 5 tasks and 12 subtasks from the action categories and temporal annotations of MMA-82. This shared foundation enables consistent evaluation across tasks, domains, and scenes, but also creates dependencies between their errors. An ambiguous action category or temporal boundary may affect multiple tasks simultaneously. The benchmark also evaluates micro-action understanding within a predefined set of 82 categories. Real-world movements may be ambiguous, combine several micro-actions, or fall outside any single category. Extending MMA-MTU toward open-vocabulary and compositional micro-action understanding would better test how models generalize beyond a fixed category set.

Many queries in MMA-MTU are generated from structured templates. For label-free queries, we select label queries containing an action name and use an MLLM to produce a motion description aligned with the corresponding video segment. The generated description replaces only the action label, while the temporal scope and task conditions remain unchanged. This paired design does not fully reproduce how users describe unfamiliar or visually similar micro-actions. Generated queries may repeat vocabulary and sentence patterns, and their descriptions are often more standardized

than spontaneous language. Future versions could incorporate human-authored queries with broader variation, incomplete descriptions, and natural ambiguity.

Negative coverage is uneven across the temporal tasks. Detection contains 565 "No_Match" queries, whereas Grounding contains only five. Consequently, the current Grounding evaluation primarily measures localization when the queried action is present and provides limited evidence about whether a model can reject an absent action. This imbalance may also encourage the prior that an action mentioned in a Grounding query must occur in the video. Expanding Grounding with visually similar and temporally plausible hard negatives would provide a more reliable test of action presence and localization.

The Description references are generated with trusted labels, temporal intervals, and motion evidence, followed by automatic checks and two rounds of human review. This process anchors each description to an annotated event, but the resulting text may still reflect the generator's linguistic style or preference for particular motion details. The current quality-control statistics measure revision and acceptance rates rather than independent attribute-level correctness for laterality, direction, contact, and movement order. In addition, BLEU-4, CIDEr-D, BLEURT, CLIPScore, and EMScore may reward reference similarity without reliably separating a valid paraphrase from a fluent description containing an incorrect fine-grained detail. Independent human evaluation and correlation analysis between automatic metrics and human judgments would strengthen the validation of this task.

Finally, the five tasks share videos and annotations but are evaluated independently. The protocol does not explicitly penalize a model whose Recognition, Grounding, Detection, and Description outputs contradict one another on the same event. Cross-task consistency scores, chained evaluation, or multi-turn settings could more directly test whether the model forms a coherent account of the video. All data also inherit the domains and source distribution of MMA-82. Although the test split is cross-subject, it is not a cross-dataset evaluation, so generalization to new camera conditions, populations, cultures, and action taxonomies remains unresolved. Because some source videos are publicly available, overlap with model pretraining data is also a potential concern, although no contamination is established by the current evidence.

L Future Work

A natural next step is to extend MMA-MTU beyond its predefined 82-category action set. Real-world micro-actions may not correspond to a single known category and often combine several movements. Future benchmarks could therefore include unseen-action grounding, hierarchical open-set recognition, and compositional queries such as "turning the head while touching the face." For actions outside the predefined category set, models could describe the observed movement directly instead of mapping it to the closest known label. Such settings would test whether models learn transferable representations of subtle motion rather than only the category structure of MMA-82.

Future versions should also move closer to how people refer to actions in practice. Label-free queries currently provide a controlled alternative to action names, but their expressions remain relatively complete and standardized. Queries written by users who have not seen the category set would introduce colloquial, incomplete, and uncertain descriptions, such as "the hand moves somewhere near the face." These queries could be paired with more systematic hard negatives. Useful cases include visually similar actions, incorrect directions or body parts, actions absent from a specified temporal window, and compositional conditions for which only one component is present. This would extend Grounding from localizing an assumed event to jointly deciding whether the described event occurs and, if so, when.

Description evaluation can also be made more closely aligned with motion understanding. Beyond overall text similarity, future protocols could decompose a description into verifiable propositions about the active body part, movement direction, interaction target, contact, and temporal order. Human judgments and proposition-level factuality measures could then be used to evaluate these elements directly and to study how well existing automatic metrics correlate with description correctness. Cross-dataset experiments are equally important: training on MMA-MTU and evaluating

on independently collected micro-action data would test generalization across category definitions, viewpoints, populations, occlusion patterns, and recording conditions.

Finally, the emotion-recognition experiment provides an initial example of using micro-action predictions beyond benchmark evaluation. Future work may explore their value for affective computing, human–computer interaction, conversational engagement, social-interaction understanding, and privacy-preserving behavioral analysis. These applications should focus on observable behavior and be evaluated with appropriate consent and privacy safeguards. Micro-action predictions should not be treated as evidence for clinical diagnosis, personality assessment, or other unsupported psychological inferences.