MotionAtlas: A High-Quality Dataset and Benchmark for Dense Motion Captioning

Weisong Liu[1*], Haochen Wang[1*], Kuan Gao[2*], Yuhao Wang[4], Yikang Zhou[5], Zhongwei Ren[6], Jacky Mai[3], Anna Wang[3], Yanwei Li[2], Jason Li[3‡], and Zhaoxiang Zhang[1†]

[1]Institute of Automation, Chinese Academy of Sciences, China
[2]Shanghai Jiao Tong University, China [3]Nanyang Technological University, Singapore
[4]Peking University, China [5]Wuhan University, China [6]Beijing Jiaotong University, China
[*]Equal contribution [‡]Project lead [†]Corresponding author
Corresponding author: zhaoxiang.zhang@ia.ac.cn
Project page: https://kagura-0001.github.io/projects/MotionAtlas

Abstract. We propose **MotionAtlas**, a system for detailed captioning of *motion-centric* videos, comprising (1) a dedicated human-annotated benchmark, (2) a scalable, high-quality pipeline to construct training samples, and (3) a family of powerful Video-MLLMs. Unlike conventional global motion captioning datasets, we focus on region-aware motion captioning: given a video and a spatiotemporal mask, the model generates precise descriptions of motion *within the target region*, thereby alleviating visual clutter and motion entanglement and enabling reliable, quantifiable evaluation. Concretely, we first build MotionAtlas-Bench, a comprehensive benchmark comprising 2,073 multiple-choice questions, meticulously annotated for a curated set of high-quality, motion-centric videos, to evaluate fine-grained motion understanding of the objects in question. Second, we design a rigorous and scalable data pipeline that leverages self-bootstrap refinement to suppress fine-grained hallucinations, yielding 159k high-quality motion captioning data. Third, we design a tailored training data composition strategy, which achieves consistent and substantial performance gains across diverse baseline Video-MLLMs, including Molmo2 and Qwen3-VL. For instance, MotionAtlas-4B surpasses Qwen3-VL-4B by an average of 5.2 percentage points across general motion benchmarks. The benchmark, dataset, and code have been released.

Keywords: Video Understanding · Video Motion · Multi-modal Large Language Models

1 Introduction

Driven by high-quality video-text datasets and large-scale model development, Video Multimodal Large Language Models (Video-MLLMs) [3, 8, 30, 35] have achieved significant progress on general video understanding tasks. However, most existing models still struggle with fine-grained motion dynamics in real-world videos, particularly when precise spatiotemporal reasoning over local regions is required [37, 47]. For instance, the professional badminton scenario in

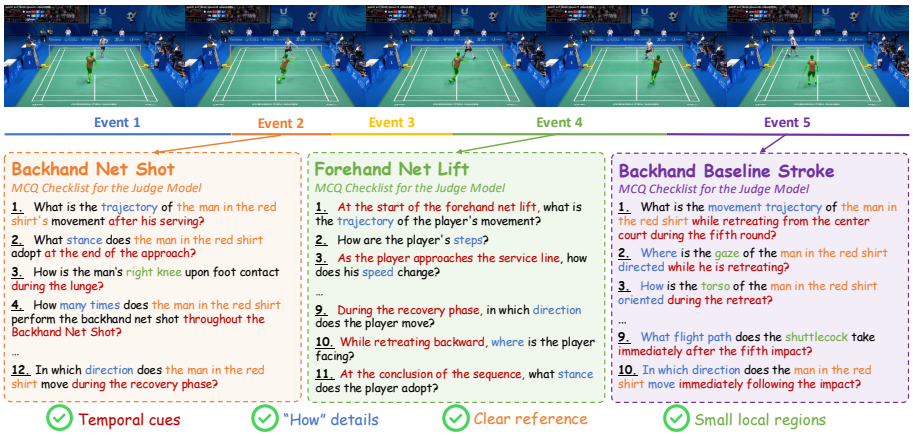

Fig. 1: Illustration of our MotionAtlas-Bench. Each video is first decomposed into events. Then, for each event, the judge model, using candidate captions, answers each multiple-choice question (MCQ) on the checklist. The questions emphasize temporal cues, how-level kinematics, clear references, and small local regions, enabling reliable and diagnostic evaluation.

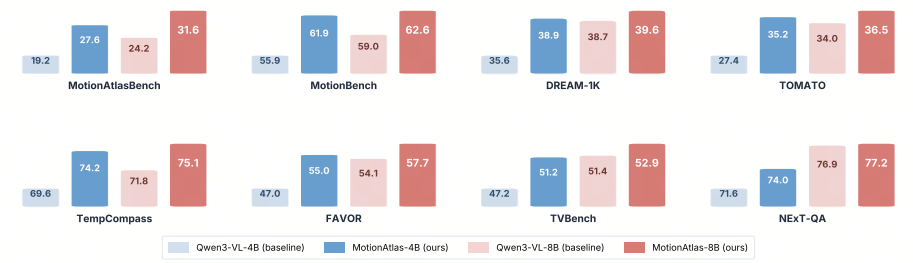

Fig. 2: Quantitative comparison between our MotionAtlas and baselines (Qwen3-VL). Our MotionAtlas brings significant improvements over baselines *consistently*.

Figure 1 requires models to answer diagnostic questions that demand fine-grained motion perception, *e.g.*, trajectory tracking, kinematic state assessment, and temporal anchoring. Such tasks remain difficult for current Video-MLLMs, which tend to prioritize high-level semantics over fine-grained, context-aware motion dynamics.

Most prior works focus on global video motion captioning, which aims to describe the motion of the entire scene or prominent objects [6, 31, 34, 43]. While this paradigm has been widely explored, its core limitation lies in its *inherent evaluation intractability*: reliably assessing the quality of generated global motion captions is extremely challenging, which makes it hard to verify their completeness (recall) and faithfulness (anti-hallucination) [4, 17, 34], leading to unreliable annotation pipelines that cannot scale.

To address this, we shift the focus to *region* captioning [5, 16, 44–46, 49] for motion-centric videos: given a video and a spatiotemporal mask, the model gen-

Table 1: Comparison of related datasets. **Part I** compares MotionAtlas-Bench against existing evaluation benchmarks, emphasizing dense QA coverage and motion attribute details. "r_t" denotes relative object size (mean / median). "Mot. Score" represents the magnitude of the Farneback optical flow. "MCQ" means multiple-choice question. *Our MotionAtlas-Bench is comprehensive and challenging.* **Part II** compares MotionAtlas-Data with existing training datasets in terms of caption complexity and verb density. "Qua. Con." indicates quality control, and our data incorporates a self-bootstrap refinement strategy introduced in §3.2. *Our MotionAtlas-Data is detailed and faithful.*

Part I. Evaluation Benchmarks

Benchmark	Ref. Obj.	Mot. Score	Anno. Type	QA / Vid.	Obj. Size (r_t)	Multi-Aspect Attr.
MotionBench [17]	×	3.6	QA	1.5	Global	×
FAVOR-Bench [31]	×	4.6	Caption / MCQ	4.6	Global	×
DREAM-1K [34]	×	4.5	Caption	–	Global	×
VideoRefer-D	✓	4.2	Caption	–	0.17 / 0.14	×
MotionAtlas-Bench	✓	**5.4**	**MCQ**	**19.4**	**0.14 / 0.10**	✓

Part II. Training Datasets

Dataset	Ref. Obj.	Anno. Method	Anno. Type	Videos	Texts	Verb / Vid.	Length	Qua. Con.
MotionSight [13]	×	Auto	QA	40K	87K	12	70	–
VideoRefer (*Detail*) [45]	✓	Auto	Caption	125K	125K	11	65	–
TarsierRecap [34]	×	Auto	Caption	585K	585K	13	74	–
MotionBench (Train) [17]	×	Human	Caption	5K	5K	17	121	Human
MotionAtlas-Data	✓	**Auto**	**Caption**	**82K**	**159K**	**23**	**212**	**Self-Bootstrap**

erates a precise description of the motion *within the target region only*. Confining the description to a spatiotemporal mask removes the visual clutter and motion entanglement inherent to global captioning, allowing annotators to focus exclusively on local motion. This yields finer-grained, more complete, and more consistent annotations, which in turn enable the reliable, quantifiable, and scalable evaluation that is otherwise infeasible for global motion description. To this end, we propose **MotionAtlas**, comprising a human-annotated evaluation benchmark, a carefully designed training corpus, and a family of powerful Video-MLLMs.

As the foundation, we first construct MotionAtlas-Bench, tailored for region-level motion captioning, to systematically validate the quality of generated motion descriptions and to ablate the effectiveness of candidate data pipelines. Because our annotations are equipped with fine-grained identity references and temporal anchors, we can reliably localize missing motion details and diagnose fine-grained hallucinations in the generated descriptions [12, 17, 34, 39]. As summarized in Table 1 (Part I), our benchmark is both:

- *Comprehensive.* It provides far denser supervision than prior benchmarks, with 19.4 questions per video, and is the only benchmark to annotate detailed multi-aspect motion attributes.
- *Challenging.* It targets smaller objects that are harder to localize and track, and exhibits the highest motion complexity among all compared benchmarks (Table 1).

Next, we design a rigorous and scalable training data pipeline that operates in three stages. First, temporal segmentation produces detailed motion caption proposals for each segment. Second, a self-bootstrap refinement mechanism

contrasts multiple rollout captions within the same temporal window to identify high-risk claims [23], and then exploits the VLM's stronger understanding than generation [23, 36, 40] to filter out implausible details and hallucinations. Third, multi-source summarization guided by global captions as temporal cues removes inconsistencies and redundancies, yielding coherent motion descriptions [7,11,17,22,29,39,43]. As shown in Table 1 (Part II), this pipeline produces richer, longer motion annotations while substantially reducing fine-grained hallucinations compared with existing datasets.

Finally, to fully leverage this data, we propose a tailored training recipe that strengthens the model's ability to capture region-specific motion dynamics. By integrating spatiotemporal region cues into the training mixture, we obtain a family of Video-MLLMs that excel at fine-grained video understanding, not limited to region motion captioning.

Empirically, as demonstrated in Figure 2, MotionAtlas brings *consistent* and significant improvements over various baselines [3, 8] across eight representative motion-related and general video understanding benchmarks.

2 MotionAtlas-Bench

MotionAtlas-Bench operationalizes region-level motion evaluation through an *event–motion–fact* annotation hierarchy: each video is split into events, each event is described in dense motion detail, and each detail is distilled into an atomic, independently verifiable fact. The resulting facts form a multiple-choice-question (MCQ) checklist for judging a caption, which turns the otherwise intractable problem of scoring free-form motion descriptions into dense, diagnostic verification.

2.1 Data Curation

We draw candidate clips from four motion-rich sources [12, 17, 34, 39] and retain only entities whose motion is genuinely difficult to describe. A VLM scores each candidate and keeps those exhibiting salient multi-atomic motion or small object/background motion, precisely the cases in which global captioning is least reliable. We further include task-oriented clips covering complex kinematics (*e.g.*, somersaults, footwork sequences) and rapid, multi-step motion, so that the benchmark concentrates on challenging dynamics rather than simple single-action clips. Where source masks are missing, annotators supplement them with SAM2 [27] to guarantee per-frame spatial grounding for every referred entity.

2.2 Evaluation Protocol

Checklist-Based Judgement. To holistically evaluate fine-grained motion captioning, we perform dense evaluation against granular factual questions (as exemplified in Figure 1). Given a model-generated caption C for video V, a judge model is presented with *each* MCQ $\hat{Q}_k$ alongside C, and selects the most

appropriate option. Based on the judge's selections across all MCQs, we report three complementary metrics:

- **Accuracy** = $N_{\text{correct}}/N_{\text{total}}$: the overall proportion of correctly described motions across all test cases.
- **Recall** = $(N_{\text{correct}} + N_{\text{error}})/N_{\text{total}}$: the tendency to describe the target motion rather than respond "not mentioned".
- **Precision** = $N_{\text{correct}}/(N_{\text{correct}} + N_{\text{error}})$: the accuracy of the descriptions when the motion is explicitly mentioned.

We design two increasingly difficult grounding settings to disentangle different aspects of model capability.

Single-Frame Grounding. The model receives the full video and only the target entity's mask at its first visible frame, so it must autonomously track the entity throughout the clip, jointly testing spatial tracking and motion understanding.

Full-Sequence Grounding. The model receives the full video with exact per-frame masks overlaid on every frame. This continuous localization removes tracking ambiguity and background distractors, strictly measuring the model's intrinsic motion captioning capacity.

2.3 Annotation Pipeline

Overview. Fine-grained motion is continuous, layered, and partly subjective, giving rise to three recurring failure modes of naive annotation: ambiguous event boundaries, concurrent body-part motions, and hard-to-verify attributes (*e.g.*, speed, force). Our pipeline addresses them in turn, pairing a VLM for coverage with a human for correction at every stage:

$$V, \{m_t\} \xrightarrow{\text{Segment}} \{\hat{E}_k\} \xrightarrow{\text{Caption}} \{D_k\} \xrightarrow{\text{MCQ Extraction}} \{\hat{Q}_k\}, \tag{1}$$

where V is the video, m_t the target mask at frame t, and k indexes events; $\hat{E}_k$ is the human-refined event, D_k its dense motion description, and $\hat{Q}_k$ the verified MCQ entering the checklist. Hats mark artifacts finalized by human review.

Event Annotation. Segmenting a long clip into events lets both the VLM and the annotator reason over a short window at high temporal resolution rather than over the entire video at once. We enforce a single principle, *one event corresponds to one core motion*, which directly resolves the boundary ambiguity noted above. Gemini 3 Pro proposes events with frame intervals; annotators then merge over-fragmented events, correct inaccurate boundaries, and remove hallucinated motion (*e.g.*, camera motion attributed to the subject).

Detailed Fact Proposal and Annotation. Events carry only coarse semantics, so we densify each one along a motion ontology of six aspects: Kinematics (direction, trajectory, speed change, rotation), Parts (independent body-part motions, *e.g.*, "left fist jabs forward while right hand guards the chin"), Spatial

(position relative to environmental landmarks), INTERACTION (contact, grasping, releasing), STATE (posture and occlusion changes), and CAMERA (camera motion).

For each event, we feed the VLM a *merged focal crop*, the union of the entity's bounding boxes across the whole event rather than per-frame crops, so the crop region stays fixed even under large displacement: this removes spatial jitter and raises the effective resolution on the target while suppressing background distractors. The VLM drafts a dense description, and annotators then supplement the details that VLMs systematically omit, chiefly part-level actions and precise directions.

Multiple-Choice Questions Construction. To make scoring objective, we decompose each dense description into atomic facts, each isolating a single independently verifiable attribute. Gemini 3 Pro then converts every fact into an MCQ whose distractors are alternative values *within the same aspect* (*e.g.*, a different rotation direction), requiring the judge to discriminate fine motion details rather than coarse topics; the option structure and metrics follow §2.2.

Quality Control. We carry out two-tier quality control: (1) *Blind filtering*: an LLM answers questions without video context, removing low-discriminability questions that can be guessed by commonsense. (2) *Human fact verification*: annotators verify each correct option's accuracy, correct hallucinated attributes, ensure at least one hard-case confusion option, and assign missing options to unverifiable facts. Each MCQ is reviewed by two annotators: factual-validity disagreements are removed or marked as unverifiable, while wording or option-design disagreements are resolved by adjudication.

2.4 Benchmark Statistics

After quality control, MotionAtlas-Bench contains 107 videos and 2,073 MCQs. The aspect distribution (Figure 3) is dominated by SPATIAL (29.7%), PARTS (27.7%), and KINEMATICS (19.8%), reflecting our emphasis on fine-grained body articulations and spatial dynamics over coarse action labels. Most videos (57.0%) span 30–60s and 39.3% exceed one minute, requiring models to continuously track the target and its fine-grained motion throughout the clip. Each video averages 19.4 MCQs (max 56), ensuring dense coverage of concurrent motion layers. Target entities are small, with a mean rate of 14% as shown in Table 1, requiring precise spatial grounding rather than reliance on global scene cues.

3 Data Pipeline and Training Recipe for MotionAtlas

3.1 Data Collection & Filtering

We collect videos with per-frame tracking annotations (mask or bounding box) from 7 diverse open-source datasets, totaling 126K videos and 220K region samples. Each sample is a (video, entity) pair, and a single video may contain multiple referred entities. Before annotation, we apply three-tier filtering to the 220K raw samples to ensure data quality.

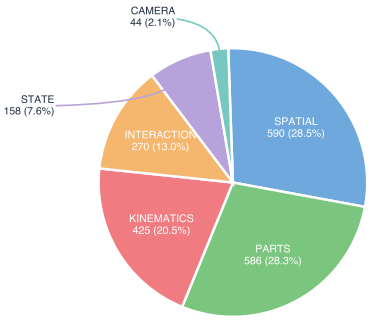

Fig. 3: Data distribution of our proposed MotionAtlas-Bench aspects.

Table 2: MotionAtlas-Data sources. Details of our filtering strategy are introduced in §3.1.

Data Source	Videos	Samples
MeVIS [12]	1.6K	8K
SAV [27]	39K	82K
TAO [11]	500	1.4K
DanceTrack [29]	37	210
ViCaS [2]	20K	65K
VastTrack [26]	46K	46K
GOT [18]	10K	18K
Total (Raw)	126K	220K
Total (Filtered)	**82K**	**159K**

Basic Quality Filtering. We first remove videos shorter than 1 second or longer than 1 minute (clips that are too short lack sufficient motion, while overly long clips exceed the VLM's effective context window), as well as samples whose spatial resolution falls below a minimum threshold.

Motion Score Filter. Unlike methods based on whole-video global motion scoring, we design an *entity-centric motion score* that measures only the referred entity's own motion intensity while suppressing camera motion and common-mode background motion. Specifically, we compute the optical flow intensity c_t within the entity mask and subtract the flow intensity β_t of the surrounding ring and background regions, yielding a residual motion score:

$$f_t = \max(0,\ c_t - \beta_t) + \alpha \cdot c_t \tag{2}$$

The α term preserves a small absolute motion signal to prevent over-suppression from eliminating genuine motion of small targets. We set $\alpha = 0.3$ in all experiments. The final entity motion score jointly considers motion intensity (90^{th} percentile) and motion persistence (fraction of frames p exceeding a fixed threshold), suppressing sporadic motion spikes.

After three-tier filtering, 159K samples are retained for subsequent annotation. Detailed sources are illustrated in Table 2.

3.2 Dataset Annotation Pipeline

Overview. The benchmark pipeline in §2.3 achieves high annotation quality through AI-Human collaboration, but its human refinement (event refinement) and human fact verification (MCQ verification) steps limit scalability. To expand the annotation scale from the benchmark's 107 videos to 159K samples, we need to replace these manual steps with automated mechanisms while maintaining comparable annotation quality.

The core challenge is that, once human review is removed, single-pass VLM captioning introduces systematic hallucinations, particularly at the temporal boundaries between events. To address this, we propose a four-stage pipeline,

segment–caption–bootstrap–summarize, which inherits the event segmentation and spatial crop design from §2.3 and introduces *dual-rollout differential verification* to replace human review and *multi-source narrative synthesis* to eliminate temporal incoherence between events. The entire pipeline uses Qwen3-VL-235B as the unified VLM backbone:

$$V, \{m_t\} \xrightarrow{\text{Segment}} \{E_k\} \xrightarrow{\text{Caption}} \{c_k^{(1)}, c_k^{(2)}\} \xrightarrow{\text{Bootstrap}} \{\hat{c}_k\} \xrightarrow{\text{Summary}} N \tag{3}$$

where V is the input video, $\{m_t\}$ denotes the per-frame entity masks, E_k is the k-th segmented motion event, $\hat{c}_k$ is the refined caption for event E_k, and N is the final structured motion narrative.

Event Segmentation. Following §2.3, we segment videos by motion semantics to achieve temporal zoom-in, obtaining event sequences $\{E_k = (s_k, e_k, d_k^{\text{seg}})\}$, where s_k and e_k are the start and end timestamps of event E_k, and d_k^{seg} is a coarse event descriptor generated from sparsely sampled global frames. Here, *no human refinement* is performed.

Local & Global Captioning. We query the VLM backbone for both local and global captioning. (1) At the local level, for each event E_k we construct a merged focal crop V_k^{crop} from the union of bounding boxes across all masks within the event's frame interval, and the VLM generates a local description c_k rich in motion details. (2) At the global level, the VLM generates a full-video caption c_{full} that provides scene context and a global timeline. Local captions offer higher spatiotemporal resolution on motion details, while the global caption provides a reliable overall temporal framework.

Self-Bootstrap Refinement. This is the core mechanism that replaces human review in §2.3. Its design is based on the following observation: *divergences between two independent rollouts are more likely to originate from hallucinations, while agreement points are more likely to reflect genuine visual evidence.* This procedure includes four phases:

1. Dual Rollout: independently generate two descriptions $\{c_k^{(1)}, c_k^{(2)}\}$ for the same visual input V_k^{crop}.
2. Differential Extraction: an LLM extracts the set of conflicting claims, which carry the highest uncertainty and are thus the most informative to verify.
3. Visual Grounded Judgment: for each conflicting claim, the two assertions plus a distractor option are randomly shuffled and presented to the VLM for 3-way blind adjudication.
4. Caption Correction: verified judgments are applied to the original description to produce the refined result $\hat{c}_k$.

Multi-Source Narrative Synthesis. Finally, this stage integrates local captions $\{\hat{c}_k\}$ with the full-video caption c_{full}, removing speculative descriptions at event boundaries and redundant content between adjacent events, while using c_{full} as a global timeline reference to establish temporal coherence.

Table 3: Training data mixture for the unified motion SFT stage, spanning general and region-level sources.

Category	Source	Scale
General Motion Caption	Tarsier2-Recap [43]	417K
General Image/Text QA	LLaVA-OneVision-1.5 [1]	320K
General Motion QA	MotionSight [13]	130K
Region Motion Caption	**MotionAtlas**	**159K**

Table 4: Transfer ablation of the SFT data composition on general motion benchmarks (Qwen3-VL-4B).

MotionAtlas SFT variant	MotionBench	TOMATO	FAVOR
Base	55.9	27.4	47.0
+ General caption	60.5 ↑ 4.6	28.4 ↑ 1.0	52.2 ↑ 5.2
+ Region detail, text ref.	61.7 ↑ 5.8	31.9 ↑ 4.5	55.7 ↑ 8.7
+ Region detail, visual cue	**61.9** ↑ 6.0	**35.2** ↑ 7.8	55.0 ↑ 8.0

3.3 Data Statistics

Comparison with Existing Training Datasets. As shown in Table 1 (Part II), MotionAtlas-Data substantially surpasses existing video training datasets in motion density, with an average of 23 action verbs and 212 words per sample. Crucially, whereas most large-scale datasets provide global descriptions, our annotations are strictly region-level, ensuring that these rich motion details are explicitly grounded to specific target entities.

3.4 Training Recipe

To validate the effectiveness of MotionAtlas-Data for both general motion and referring (region-level) motion understanding, we mix the following four categories of data during a unified SFT stage, as demonstrated in Table 3.

The mixing strategy pursues three goals: (1) injecting entity-centric fine-grained motion captioning through MotionAtlas-Data; (2) preserving general motion understanding through Tarsier2-Recap [43] and MotionSight [13]; and (3) preventing catastrophic forgetting through general QA data from LLaVA-OneVision-1.5 [1]. In the unified SFT stage, we concatenate the four sources and sample examples proportionally to dataset size for one epoch. We analyze the transfer contribution of each data component (Table 4) in §4.3.

4 Experiments

4.1 Experimental Setup

Implementation Details. We adopt Qwen3-VL (4B, 8B) [3] and Molmo2 (4B, 8B) [8] as our base models. For all models, we perform full parameter fine-tuning for 1 epoch using the data mixture detailed in §3.4. During training, we uniformly sample 16 frames per video and limit the maximum sequence length to 16,384 tokens. The models are optimized using AdamW with a peak learning rate of 1×10^{-5} and a cosine learning rate schedule with a 3% linear warmup.

Evaluation Benchmarks. To systematically assess fine-grained motion understanding and general video reasoning, we evaluate on 8 representative benchmarks: MotionAtlas-Bench, MotionBench [17], DREAM-1K [34], TOMATO [28], NExT-QA [38], TempCompass [20], FAVOR-Bench [31], and TVBench [10].

Table 5: Main results on our **MotionAtlas-Bench**. We report results in terms of **Accuracy (acc.)**. "Single-Frame" denotes Single-Frame Grounding and "Full-Seq." denotes Full-Sequence Grounding. Based on Qwen3-VL-8B [3], MotionAtlas-Data even surpasses Qwen3-VL-235B [3], demonstrating the effectiveness of our data.

Model	Single-Frame Grounding (acc.)							Full-Sequence Grounding (acc.)						
	Overall	Spatial	Parts	Kin.	Inter.	State	Cam.	Overall	Spatial	Parts	Kin.	Inter.	State	Cam.
Gemini 3 Pro	36.4	**36.6**	**34.7**	32.0	**43.2**	46.0	24.0	36.5	**36.8**	33.5	**38.1**	38.1	44.0	22.0
GPT-5.2 [25]	**36.9**	36.5	34.0	**34.2**	42.1	**51.3**	24.0	**37.6**	36.0	**38.8**	36.6	**38.5**	46.7	22.0
Gemini 2.5 Pro [9]	29.8	29.8	28.9	27.5	29.3	43.3	20.0	31.9	33.4	30.5	25.5	34.1	47.3	26.0
Gemini 2.5 Flash [9]	28.9	29.0	27.1	25.3	33.0	40.0	24.0	35.4	34.7	36.3	32.3	35.5	47.3	22.0
Qwen3-VL-235B [3]	30.5	31.8	27.8	28.9	31.5	38.7	30.0	33.7	35.0	33.2	31.1	34.4	38.7	28.0
Qwen3-VL-32B [3]	29.7	30.3	27.5	26.0	31.5	41.6	**32.0**	35.7	32.8	33.9	34.4	38.4	**53.4**	**36.0**
Molmo2-4B [8]	13.7	13.6	12.9	13.7	14.5	16.1	13.0	15.6	15.6	15.0	15.3	16.1	18.1	14.6
+ MotionAtlas-Data	21.6 ↑ 7.9	21.6	20.9	21.2	22.7	24.4	20.3	25.5 ↑ 9.9	25.4	24.8	25.2	26.4	28.3	24.5
Molmo2-8B [8]	19.2	19.1	18.3	19.3	20.0	21.6	18.4	21.5	21.5	21.0	21.4	21.9	23.9	20.4
+ MotionAtlas-Data	24.4 ↑ 5.2	24.2	23.4	24.3	25.4	27.4	23.0	27.5 ↑ 6.0	27.4	26.7	27.3	28.4	30.4	26.7
Qwen3-VL-4B [3]	19.3	19.5	20.0	14.1	20.2	28.7	16.3	21.7	21.9	22.4	16.5	22.6	31.1	18.7
+ MotionAtlas-Data	27.7 ↑ 8.4	25.9	27.9	26.9	28.3	35.3	28.2	30.1 ↑ 8.4	28.3	30.3	29.3	30.7	37.7	30.6
Qwen3-VL-8B [3]	24.3	24.5	23.9	20.3	24.8	37.6	18.0	26.7	27.2	24.6	26.7	26.4	34.7	22.0
+ MotionAtlas-Data	31.6 ↑ 7.3	31.1	31.2	30.6	29.0	42.8	34.5	34.1 ↑ 7.4	33.5	33.6	33.0	31.4	45.2	36.9

Evaluation Protocol. We follow the standard evaluation settings of lmms-eval [48] and VLMEvalKit [14], uniformly sampling 16 frames for all short-video benchmarks. All models, including proprietary ones (*e.g.*, GPT-5.2 and Gemini), are evaluated under the same frame budget to ensure a fair comparison.

4.2 Main Results

Results on MotionAtlas-Bench. Table 5 reports the performance on our MotionAtlas-Bench under both Single-Frame and Full-Sequence Grounding settings. Overall, region-level motion captioning proves highly challenging: even the most capable proprietary models struggle to achieve high recall on fine-grained details. In contrast, fine-tuning on MotionAtlas-Data yields consistent and substantial improvements across all base models. Moreover, the Full-Sequence setting consistently outperforms Single-Frame Grounding, indicating that continuous spatial prompting effectively strengthens the model's capacity to caption fine-grained motion.

Results on General Video Benchmarks. Table 6 evaluates all models on 7 public motion-related video understanding benchmarks. On most benchmarks, proprietary models lead by a moderate margin in temporal and fine-grained motion perception (*e.g.*, TempCompass [20], FAVOR-Bench [31]). However, open-source models of comparable size (Molmo2-8B [8], Qwen3-VL-8B [3]) already achieve competitive or even superior scores on several tasks such as NExT-QA [38], indicating that the gap largely lies in fine-grained motion perception rather than general video QA.

A key finding is that training on MotionAtlas-Data, which consists entirely of *region-level* motion captions, *consistently* improves performance on *general* (non-region) motion benchmarks. For example, Qwen3-VL-4B [3] gains +6.0 on MotionBench [17], +7.8 on TOMATO [28], +8.1 on FAVOR-Bench [31], and +4.6

Table 6: Main results on motion-related video understanding benchmarks. Based on Qwen3-VL-8B [3], our MotionAtlas achieves competitive performance even compared with closed-source models, especially on MotionBench [17] and TempCompass [20].

Model	Motion Bench [17]	DREAM-1K F1-Score [34]	TOMATO [28]	NExT-QA MCQ [38]	Temp Compass [20]	FAVOR [31]	TVBench [10]
Closed-source Models							
GPT-5.2 [25]	65.4	42.2	53.0	79.9	73.0	56.8	53.8
Gemini 2.5 Pro [9]	62.0	42.7	48.6	79.8	73.7	58.8	59.9
Gemini 3 Pro	62.6	42.5	48.3	79.6	74.1	58.5	56.8
Open-source Models							
Molmo2-4B [8]	59.9	25.7	30.3	82.4	70.2	55.8	50.6
+ MotionAtlas-Data	60.6 ↑ 0.7	32.7 ↑ 7.0	35.9 ↑ 5.6	83.6 ↑ 1.2	75.1 ↑ 4.9	57.4 ↑ 1.6	52.1 ↑ 1.5
Molmo2-8B [8]	60.5	29.6	33.8	83.6	72.1	57.9	54.1
+ MotionAtlas-Data	61.8 ↑ 1.3	34.0 ↑ 4.4	36.8 ↑ 3.0	84.3 ↑ 0.7	75.7 ↑ 3.6	58.7 ↑ 0.8	54.8 ↑ 0.7
Qwen3-VL-4B [3]	55.9	35.6	27.4	71.6	69.6	47.0	47.2
+ MotionAtlas-Data	61.9 ↑ 6.0	38.9 ↑ 3.3	35.2 ↑ 7.8	74.0 ↑ 2.4	74.2 ↑ 4.6	55.0 ↑ 8.1	51.2 ↑ 4.0
Qwen3-VL-8B [3]	59.0	38.7	34.0	76.9	71.8	54.1	51.4
+ MotionAtlas-Data	62.6 ↑ 3.6	39.6 ↑ 0.9	36.5 ↑ 2.5	77.2 ↑ 0.2	75.1 ↑ 3.3	57.7 ↑ 3.6	52.9 ↑ 1.5

on TempCompass [20], indicating that learning to describe *fine-grained motion generalizes to broader video tasks*. The largest gains arise on benchmarks emphasizing temporal dynamics and action-attribute recognition (TOMATO [28], FAVOR-Bench [31]), consistent with the fine-grained motion coverage of our data. Qwen3-VL-8B [3] and the Molmo-2 series [8] follow the same trend, with consistent gains across all benchmarks and no degradation, confirming that our data does *not* induce catastrophic forgetting.

4.3 Transfer Component Ablation

To understand *where* this cross-benchmark transfer comes from, we incrementally add each data component and evaluate on general motion benchmarks using Qwen3-VL-4B [3] (Table 4). Adding general captioning already lifts all three benchmarks, but region-detail captioning yields a markedly larger gain (*e.g.*, TOMATO [28] 27.4 → 31.9 and FAVOR-Bench [31] 47.0 → 55.7), indicating that the improvements are *not* merely a byproduct of more verbs or generic captioning. Training with explicit visual region cues further improves transfer to high-dynamic reasoning (TOMATO [28] reaches 35.2, ↑ 7.8 over the base model). This suggests that region-grounded supervision teaches models to exploit the temporal density that base models underuse.

4.4 Data Scale Ablation

Using Qwen3-VL-4B [3] as the base model, Table 7 and Figure 4 compare training with and without MotionAtlas-Data at three data scales (32K, 95K, 159K).

Scaling with MotionAtlas-Data. When MotionAtlas-Data is included, performance scales monotonically across all benchmarks: MotionAtlas-Bench rises from 22.9 (32K) to 28.3 (159K), a +5.4 gain, and external benchmarks follow the same trend (TOMATO [28] 28.4 → 35.2, FAVOR-Bench [31] 48.1 → 55.0).

Table 7: Performance comparisons with varying training data scales. Model performance continues to improve as the scale of training data increases. However, these performance improvements are significantly diminished when training proceeds without our proposed MotionAtlas-Data.

Method (Data Scale)	Motion Atlas	Motion Bench [17]	DREAM-1K F1-Score [34]	TOMATO [28]	NExT-QA MCQ [38]	Temp Compass [20]	FAVOR [31]	TVBench [10]
Qwen3-VL-4B [3]	19.2	55.9	35.9	27.4	71.6	69.6	47.0	47.2
w/ MotionAtlas-Data								
20% (32K)	22.9	58.9	36.9	28.4	72.2	71.2	48.1	47.0
60% (95K)	24.6	59.5	37.0	30.1	73.0	72.3	50.9	49.0
100% (159K)	**28.3**	**61.9**	**38.9**	**35.2**	**74.0**	**74.2**	**55.0**	**51.2**
w/o MotionAtlas-Data								
20% (32K)	12.9	57.4	37.3	29.0	70.9	70.8	47.4	46.7
60% (95K)	12.9	58.8	36.9	30.7	71.3	72.3	50.6	47.4
100% (159K)	12.2	60.5	38.3	28.4	71.9	73.3	52.2	48.5

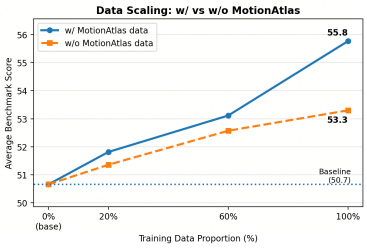

Fig. 4: Data scaling curves. Adding MotionAtlas-Data brings more significant improvements.

Table 8: Ablation of the MotionAtlas data pipeline (MA Pipeline) components, evaluated on MotionAtlas-Bench via caption quality judging.

Method	Acc (↑)	Recall (↑)	Precision (↑)
MA Pipeline (full)	**39.9**	**68.2**	**58.5**
w/o Self-Bootstrap	36.4	64.1	56.8
w/o Full-Video Caption	33.2	58.9	56.4
w/o Spatial Crop	32.7	60.9	53.6

Necessity of MotionAtlas-Data. In contrast, replacing MotionAtlas-Data with an equal amount of TarsierRecap data [43] yields essentially no improvement on MotionAtlas-Bench (∼12–13% at all scales), showing that its *distinctive fine-grained details* have no viable substitute. External benchmarks may still improve without our data, but the gains are consistently *smaller* (*e.g.*, 37.3 → 38.3 *vs.* 36.9 → 38.9 on DREAM-1K [34]), further validating its complementary value.

4.5 Pipeline Component Ablation

To isolate the contribution of each component in our MotionAtlas (MA) data pipeline (§3.2), we ablate three key designs, self-bootstrap refinement, full-video captioning, and spatial cropping, and evaluate the resulting annotation quality on MotionAtlas-Bench (Table 8).

Self-Bootstrap Refinement. Removing the dual-rollout self-bootstrap stage leads to a −3.5 accuracy drop (39.9 → 36.4). Without this module, single-pass VLM descriptions retain more uncorrected hallucinations, raising error rates and reducing the model's tendency to describe fine-grained details.

Full-Video Caption. Relying on local captions only causes the largest accuracy decline (−6.7, to 33.2). The global caption resolves temporal inconsistencies between adjacent events, removes speculative descriptions at event boundaries, and provides a coherent timeline that anchors the local narratives.

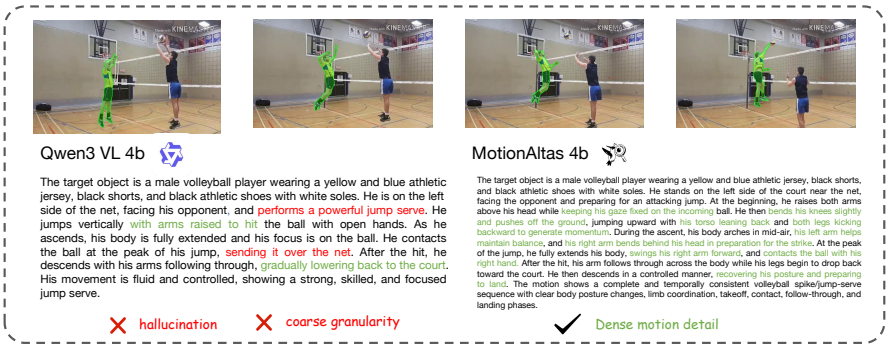

Fig. 5: Qualitative comparison between different datasets.

Spatial Crop. Feeding only the original video segments, without spatial zoom-in, drops accuracy by −7.2 (to 32.7). This confirms that cropping to the target entity's bounding box is essential for directing the VLM's attention to the relevant motion and away from background clutter and other moving objects.

4.6 Qualitative Results

Figure 5 compares Qwen3-VL-4B with MotionAtlas-4B. The base model tends to produce coarse motion descriptions and occasionally hallucinated events, whereas MotionAtlas-4B generates denser and more precise descriptions that capture fine-grained body dynamics throughout the action.

5 Related Work

MLLMs for video understanding. Driven by large-scale datasets and architectural advancements, MLLMs have achieved remarkable progress in general video understanding [3, 8, 35]. Recent foundation models (e.g., Qwen3-VL [3], InternVL3.5 [35], Molmo2 [8], VideoLLaMA 3 [47], InternVideo2.5 [37]) and instruction-tuning efforts (e.g., LLaVA-Video [51], LLaVA-OneVision-1.5 [1]) demonstrate strong capabilities in temporal scene comprehension and user alignment. However, while excelling globally, these models still struggle with fine-grained motion dynamics and precise spatiotemporal reasoning over localized regions [24, 37, 47].

Video motion understanding. High-quality datasets are a primary driver of multimodal foundation models [6, 21, 51], yet large-scale resources for motion understanding remain scarce. Traditional benchmarks predominantly target global action recognition or coarse temporal reasoning (*e.g.*, ActivityNet-QA [41], NExT-QA [38], MVBench [19], Video-MME [15]). Unlike atomic or fine-grained action recognition, which primarily asks what action occurs at the

clip level, MotionAtlas evaluates how a specified region moves, covering trajectory, part articulation, interaction, state change, and camera-related motion under explicit spatial grounding. To push towards detailed motion perception, recent works like MotionBench [17], Tarsier [34, 43], MotionSight [13], FAVOR-Bench [31], and TVBench [10] introduce fine-grained motion annotations. Despite this, global motion captioning still suffers from evaluation intractability and severe hallucinations [4, 6], complicating data quality maintenance and faithfulness verification.

Region-level detailed understanding. Concurrently, region-level captioning has proven effective for describing detailed objects and their attributes [5, 16, 26, 32, 44, 50], with recent works extending this to spatial-temporal video domains [2, 33, 42, 45, 46]. However, these methods predominantly address visual appearance and static attributes, leaving region-level fine-grained motion dynamics largely unexplored. MotionAtlas bridges this gap by confining motion descriptions to region-level spatiotemporal masks, decoupling visual clutter from motion entanglement, and thereby establishing a scalable, hallucination-resistant paradigm for both data generation and reliable evaluation.

6 Conclusion

In this work, we presented **MotionAtlas**, a framework that systematically addresses region-level motion understanding in Video-MLLMs. Recognizing the evaluation intractability of global video captioning, we shift the focus to region-level reasoning and introduce **MotionAtlas-Bench**, a diagnostic benchmark that evaluates event-centric, region-grounded motion through a reliable checklist-style MCQ protocol with explicit identity references, temporal anchors, and multi-aspect attributes. We further proposed a rigorous and scalable data pipeline that combines temporal segmentation, VLM-based self-difference judgment, and multi-source summarization, yielding **MotionAtlas-Data**, a large-scale corpus with exceptionally dense action verbs and minimal hallucinations. Powered by this data and a tailored training recipe, our Video-MLLMs achieve consistent and significant improvements over strong baselines on both motion-specific and general video understanding tasks. We believe MotionAtlas establishes a solid foundation for fine-grained video motion understanding and points toward more robust and precise video-language modeling.

Limitations and future work. MotionAtlas currently targets the single-target setting; extending our region-grounded formulation to multi-object scenarios with interactions and identity correspondence is a natural next step. Moreover, the large VLM backbone is used only for offline data construction, while deployment relies on the distilled MotionAtlas-4B/8B models.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Athar, A., Deng, X., Chen, L.C.: Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. In: CVPR (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
4. Chai, W., Song, E., Du, Y., Meng, C., Madhavan, V., Bar-Tal, O., Hwang, J.N., Xie, S., Manning, C.D.: Auroracap: Efficient, performant video detailed captioning and a new benchmark. In: ICLR (2025)
5. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic (2023), https://arxiv.org/abs/2306.15195
6. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Lin, B., Tang, Z., Yuan, L., Qiao, Y., Lin, D., Zhao, F., Wang, J.: ShareGPT4video: Improving video understanding and generation with better captions. In: NeurIPS Datasets and Benchmarks Track (2024)
7. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., Martin, M., Wang, H., Rasheed, H.A., Sun, P., Huang, P.Y., Bolya, D., Ravi, N., Jain, S., Stark, T., Moon, S., Damavandi, B., Lee, V., Westbury, A., Khan, S., Kraehenbuehl, P., Dollar, P., Torresani, L., Grauman, K., Feichtenhofer, C.: PerceptionLM: Open-access data and models for detailed visual understanding. In: NeurIPS (2025)
8. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
9. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
10. Cores, D., Dorkenwald, M., Mucientes, M., Snoek, C.G.M., Asano, Y.M.: TVBench: Redesigning video-language evaluation (2025), https://openreview.net/forum?id=DrNN5qx66Z
11. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: ECCV (2020)
12. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: ICCV (2023)
13. Du, Y., Fan, T., Nan, K., Xie, R., Zhou, P., Li, X., Yang, J., Yang, Z., Tai, Y.: Motionsight: Boosting fine-grained motion understanding in multimodal llms. arXiv preprint arXiv:2506.01674 (2025)

14. Duan, H., Fang, X., Yang, J., Zhao, X., Qiao, Y., Li, M., Agarwal, A., Chen, Z., Chen, L., Liu, Y., Ma, Y., Sun, H., Zhang, Y., Lu, S., Wong, T.H., Wang, W., Zhou, P., Li, X., Fu, C., Cui, J., Chen, J., Song, E., Mao, S., Ding, S., Liang, T., Zhang, Z., Dong, X., Zang, Y., Zhang, P., Wang, J., Lin, D., Chen, K.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models (2025), https://arxiv.org/abs/2407.11691
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR (2025)
16. Guo, Q., De Mello, S., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model. In: CVPR (2024)
17. Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In: CVPR (2025)
18. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1562–1577 (2019)
19. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024)
20. Li, L., Liu, Y., Yao, L., Zhang, P., An, C., Wang, L., Sun, X., Kong, L., Liu, Q.: Temporal reasoning transfer from text to video. In: ICLR (2025)
21. Li, X., Zhang, T., Li, Y., Yuan, H., Chen, S., Zhou, Y., Meng, J., Sun, Y., Xu, S., Qi, L., Cheng, T., Lin, Y., Huang, Z., Huang, W., Feng, J., Shi, G.: Denseworld-1m: Towards detailed dense grounded caption in the real world. arXiv preprint arXiv:2506.24102 (2025)
22. Lin, W., Wei, X., An, R., Ren, T., Chen, T., Zhang, R., Guo, Z., Zhang, W., Zhang, L., Li, H.: Perceive anything: Recognize, explain, caption, and segment anything in images and videos. In: NeurIPS (2025)
23. Manakul, P., Liusie, A., Gales, M.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In: EMNLP (2023)
24. Meng, J., Li, X., Wang, H., Tan, Y., Zhang, T., Kong, L., Tong, Y., Wang, A., Teng, Z., Wang, Y., Wang, Z.: Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. arXiv preprint arXiv:2510.20579 (2025)
25. OpenAI: Introducing gpt-5.2 (2025), https://openai.com/index/introducing-gpt-5-2/, accessed 11 November 2025
26. Peng, L., Gao, J., Liu, X., Li, W., Dong, S., Zhang, Z., Fan, H., Zhang, L.: Vasttrack: Vast category visual object tracking. NeurIPS (2024)
27. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
28. Shangguan, Z., Li, C., Ding, Y., Zheng, Y., Zhao, Y., Fitzgerald, T., Cohan, A.: TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In: ICLR (2025)
29. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: CVPR (2022)
30. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I.,

Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786

31. Tu, C., Zhang, L., Chen, P., Ye, P., Zeng, X., Cheng, W., YU, G., Chen, T.: FAVOR-bench: A comprehensive benchmark for fine-grained video motion understanding. In: NeurIPS Datasets and Benchmarks Track (2025)
32. Wang, H., Wang, Y., Zhang, T., Zhou, Y., Li, Y., Wang, J., Tian, Y., Meng, J., Huang, Z., Mai, G., Wang, A., Tong, Y., Wang, Z., Li, X., Zhang, Z.: Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. ICLR (2026)
33. Wang, J., Yuan, Y., Zheng, R., Lin, Y., Gao, J., Chen, L.Z., Bao, Y., Zhang, Y., Zeng, C., Zhou, Y., et al.: Spatialvid: A large-scale video dataset with spatial annotations. arXiv preprint arXiv:2509.09676 (2025)
34. Wang, J., Yuan, L., Zhang, Y., Sun, H.: Tarsier: Recipes for training and evaluating large video description models (2024), https://arxiv.org/abs/2407.00634
35. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
36. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)

37. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., Dou, M., Chen, K., Wang, W., Qiao, Y., Wang, Y., Wang, L.: Internvideo2.5: Empowering video mllms with long and rich context modeling (2025), https://arxiv.org/abs/2501.12386
38. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR (2021)
39. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: CVPR (2022)
40. Yang, S., Liu, Y., Zhai, B., Sun, X., Liu, Z., Barsoum, E., Li, M., Xu, C.: Captionqa: Is your caption as useful as the image itself? arXiv preprint arXiv:2511.21025 (2025)
41. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI (2019)
42. Yuan, H., Li, X., Zhang, T., Sun, Y., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint (2025)
43. Yuan, L., Wang, J., Sun, H., Zhang, Y., Lin, Y.: Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. arXiv preprint arXiv:2501.07888 (2025)
44. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: CVPR (2024)
45. Yuan, Y., Zhang, H., Li, W., Cheng, Z., Zhang, B., Li, L., Li, X., Zhao, D., Zhang, W., Zhuang, Y., et al.: Videorefer suite: Advancing spatial-temporal object understanding with video llm. In: CVPR (2025)
46. Yuan, Y., Zhang, W., Li, X., Wang, S., Li, K., Li, W., Xiao, J., Zhang, L., Ooi, B.C.: Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity (2025), https://arxiv.org/abs/2510.23603
47. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: Videollama 3: Frontier multimodal foundation models for image and video understanding (2025), https://arxiv.org/abs/2501.13106
48. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., Liu, Z.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of NAACL (2025)
49. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Liu, Y., Chen, K., Luo, P.: GPT4roi: Instruction tuning large language model on region-of-interest (2024), https://openreview.net/forum?id=DzxaRFVsgC
50. Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Chen, C.L., Yan, S.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In: NeurIPS (2024)
51. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)