

MASS: Motion-Aware Spatial-temporal Grounding for Physics Reasoning and Comprehension in Vision-Language Models

Supplementary Material

Xiyang Wu^{1,2} Zongxia Li¹ Jihui Jin² Gouthaman KV² Vishnu Raj²
Nilotpal Sinha² Jingxi Chen¹ Fan Du² Dinesh Manocha¹
¹University of Maryland ²Dolby Laboratories

A. Additional Details of MASS-Bench

This section supplements the dataset description in Section 3 of the main paper with expanded definitions and statistics.

A.1. Dataset Composition and Metadata

MASS-Bench aggregates three physics-centric video sources: 1,229 samples (14.7%) from MotionSight [2], 4,009 (47.9%) from VideoPhy2 [1], and 2,862 (34.2%) from VideoHallu [4]. In total, the dataset comprises 6,093 training and 2,268 test examples drawn from 4,350 unique videos. Videos average 545.8 frames (~ 19.6 s at 27.4 FPS) with a mean resolution of 1120×702 pixels. Due to overlap inherited from the source datasets, approximately 1.5% of unique videos appear in both splits; however, no (video, question) pairs are shared across splits, so there is no direct QA leakage. All methods are evaluated on the same fixed test set with identical evaluation scripts. Importantly, MASS relies only on depth and tracking signals extracted from the input video at inference time and does not use any ground-truth test annotations.

A.2. Question Categories

We group all questions into five categories reflecting the type of physical reasoning required, ordered from basic perception to advanced comprehension. Table 1 shows the distribution.

Category	Count	Pct.
SU (Spatial Understanding)	2,785	33.3%
TU (Temporal Understanding)	1,633	19.5%
PA (Physical Abnormality Detection)	1,432	17.1%
PC (Physics Comprehension)	1,304	15.6%
MAR (Motion & Action Recognition)	1,205	14.4%

Table 1. Distribution of question categories in MASS-Bench.

(a) *Spatial Understanding (SU)*. Identifying objects and their geometric relationships, positions, and scene layouts.

(b) *Temporal Understanding (TU)*. Interpreting how events evolve over time, including ordering, duration, and temporal dependencies.

(c) *Motion & Action Recognition (MAR)*. Detecting and characterizing object motions and agent actions across frames.

(d) *Physics Comprehension (PC)*. Applying physical principles to infer, explain, or predict real-world dynamics.

(e) *Physical Abnormality Detection (PA)*. Identifying motions or events that violate physical laws or exhibit implausible behavior.

These categories are ordered from easy to challenging. VLMs must first establish spatial and temporal awareness, then recognize motion patterns, before progressing to physics comprehension and violation detection.

A.3. Example and Question Types

We categorize each video-question-answer pair along two complementary axes.

For **example type**, we distinguish *positive* examples (3,436, 41.1%), where the video follows real-world physics, from *negative* examples (4,925, 58.9%), where at least one part of the scene violates physical plausibility. Negative examples are more challenging because the model must rely on visual evidence rather than defaulting to language priors.

For **question type** (Figure 2 of the main paper), we separate *factual* questions (5,427, 67.0%) from *critical-thinking* questions (2,673, 33.0%). Factual questions provide explicit cues that focus attention on particular entities or events. Critical-thinking questions are less direct: they omit such cues and instead require the model to infer intent, identify salient events, or reason about the underlying physical dynamics.

A.4. Motion Grounding Annotations

Beyond the question, video, and ground-truth answer, each pair includes five layers of motion-grounding annotation: (1) temporal video segmentation indexed by frame ranges; (2) visual grounding for each queried entity, specified by entity

name and bounding box; (3) an entity-level temporal profile that tracks each grounded entity across the full video, persisting once detected and left blank in segments where the entity is absent; (4) per-segment motion attributes, including the first and last observed 3D positions to capture coarse spatial layout; and (5) 3D motion vectors representing each entity’s temporal displacement. Together, these cues convert physics-intensive perceptual challenges into structured textual and mathematical representations that support more reliable physical reasoning.

B. Implementation Details

All experiments, including ablations, are conducted using full-parameter fine-tuning on 8 NVIDIA H100 (80GB) GPUs. Both Qwen2.5-VL-7B and LLaVA-OneVision-7B are trained under identical settings for fair comparison. The GRPO post-training phase for each model requires approximately 9–12 hours, while the supervised fine-tuning (SFT) stage in our ablation studies completes within 2–3 hours.

C. Visual Feature Representation Template

Figure 1 illustrates the prompt template used during both post-training and inference for VLMs in MASS-Bench. This template is designed to elicit structured reasoning for free-form video question answering. The motion-grounding information incorporated into the prompt is derived from the spatial–motion feature extraction module and the visual representation pipeline described in Section 4 of the main paper.

For each detected entity, denoted as `<Entity Name>`, we generate motion-grounding descriptors for every video segment in which the entity appears. These descriptors include the entity’s first and last 3D positions within the segment (`<first_position>` and `<last_position>`), the corresponding 3D motion vector (`<motion>`), the bounding box in the segment’s first frame (`<bbox>`), and the segment’s temporal extent indicated by the starting and ending frame indices (`<first_frame>` and `<last_frame>`).

D. Evaluation Template

We provide the LLM-as-a-judge evaluation template used in our experiments in Figure 2. This template standardizes how we assess the correctness of model predictions by comparing each VLM’s answer against the ground-truth annotation and the corresponding question.

E. Experiments on Real-world Video QA

As a supplement to the experiments in the main paper, we provide additional results on two real-world video question-answering benchmarks, MMVU [5] and MVBench [3], to further assess the generalization ability of VLMs augmented

Model	MMVU (%)	MVBench (%)
<i>Baselines</i>		
InternVL3.5-8B-Flash	55.04	45.78
Qwen2.5-VL-7B	64.00	51.30
LLaVA-OneVision-7B	50.41	55.24
<i>Larger / Close-Source</i>		
GPT-4o	75.96	58.50
Gemini-2.5-Flash	75.27	54.32
<i>VLMs with MASS</i>		
LLaVA-OneVision-7B + MASS	48.95	48.33
Qwen2.5-VL-7B + MASS	65.76	61.59

Table 2. **Overall accuracies (%) on MMVU and MVBench.** We report overall performance on two real-world video question-answering benchmarks, MMVU [5] and MVBench [3]. GPT-4o achieves the highest accuracy on both datasets. Qwen2.5-VL-7B+MASS improves over its baseline and narrows the gap to close-source models, while LLaVA-OneVision-7B+MASS is more sensitive to detection quality on these out-of-distribution tasks.

with MASS. MMVU evaluates models on expert-level temporal, procedural, and interaction-centric reasoning, while MVBench is a widely used benchmark for temporal video understanding and action-centric tasks. Both datasets contain diverse video-based QA pairs captured from real-world scenarios.

Table 2 reports the performance of VLMs integrated with MASS alongside state-of-the-art baselines. On MMVU, large-scale close-source models achieve substantially higher accuracy due to the benchmark’s emphasis on expert-level, interdisciplinary reasoning—highlighting the limitations of smaller 7B models. Qwen2.5-VL-7B+MASS improves over its baseline on both benchmarks and narrows the gap to close-source models, while LLaVA-OneVision-7B+MASS shows a performance drop on these out-of-distribution tasks. As also shown in Table 4, LLaVA is more sensitive to detection and tracking quality: when salient entities are missed or motion cues are noisy—common in MMVU and MVBench—performance can degrade, motivating more selective motion-profile extraction in future work.

F. Noise-Induced Robustness

We inject two noise types into the motion profile to test spatial–motion grounding robustness: (1) **BBox/position noise**, perturbing bounding boxes and initial/final 3D positions (SAM-2 & Grounding-DINO); (2) **Trajectory Noise**, perturbing 3D motion vectors (Co-Tracker). We apply a relatively large 10% Gaussian perturbation to the motion vectors, positions, and bounding boxes, both separately and jointly. Table 3 shows slight accuracy drops, yet MASS-augmented backbones still substantially outperform their baselines (Table 2 in the main paper) across all noise settings, even after RFT. Because our physical comprehension is mostly quali-

Prompt Template for Video-Language Models

Task Description:

The model receives a system instruction enforcing explicit reasoning and final answer formatting. Given a video, a question, and motion-grounding metadata, the model must produce detailed reasoning inside `<think>` tags and a concise final answer inside `<answer>` tags.

Core Requirements:

- Use natural internal dialogue in the `<think>` section (e.g., “let me think”, “hmm”, “wait”).
- Perform step-by-step reasoning validating spatial-temporal cues.
- Place the final answer **only** inside `<answer>` tags.
- Use **free-form answer format**: Provide a short text answer within the answer tags.

Motion-Grounding Information:

We use the following template to represent the motion-grounding information generated for VLMs. This template is used to fill the `{motion_grouding_info}` in the QA prompt template below:

Entity #1: `<Entity Name>`

* Segment #1 First Position `<first_position>`, Motion Vector `<motion>`, Last Position `<last_position>`, Bounding Box `<bbox>`, Frame `<first_frame>...<last_frame>`

* Segment #2 First Position `<first_position>`, Motion Vector `<motion>`, Last Position `<last_position>`, Bounding Box `<bbox>`, Frame `<first_frame>...<last_frame>`

...

Entity #2: `<Entity Name>` ...

Actual Prompt Used for Video QA:

Conversation Setup:

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process inside `<think>` `</think>` tags, then provides the final answer inside `<answer>` `</answer>` tags.

Question:

`<Question>` {question} `</Question>`

Motion-Grounding Information:

`{motion_grouding_info}`

Reasoning Instruction:

Please think about this question as if you were a human pondering deeply. Use internal dialogue such as “let me think”, “wait”, “hmm”, “I see”, and include verification or self-reflection in the reasoning process. Provide detailed reasoning in `<think>` `</think>`, then provide the final answer in `<answer>` `</answer>`.

Free-form Answer Instruction:

Please provide your text answer within the `<answer>` `</answer>` tags.

Figure 1. **Prompt template used for motion-aware video question answering.** The template first serializes entity-level motion grounding (positions, motion vectors, bounding boxes, and frame ranges) into text, then injects this context into a chain-of-thought style prompt that guides the VLM to reason in `<think>` tags and output its final prediction in standardized `<answer>` tags.

Prompt Template for Evaluation

Task Description:

You are an intelligent teacher whose task is to evaluate the correctness of a model’s answer to a question, given a reference ground-truth answer.

Inputs:

- **Question:** wrapped in `<Question> ... </Question>`
- **Ground-truth answer:** wrapped in `<GT> ... </GT>`
- **Model prediction:** wrapped in `<Answer> ... </Answer>`

Evaluation Criteria:

- If the prediction **does not conflict** with the ground truth, output `<Eval> Correct </Eval>`.
- If the prediction **conflicts** with the ground truth, output `<Eval> Incorrect </Eval>`.
- If the correctness of the prediction is **unclear**, output `<Eval> Unclear </Eval>`.
- Reason carefully about the relationship between the prediction and the ground truth, but keep the final evaluation **very brief**.

Output Format:

Produce *only* one of the following tokens as the final output:

```
<Eval> Correct </Eval>
<Eval> Incorrect </Eval>
<Eval> Unclear </Eval>
```

Figure 2. **Prompt template used for automatic evaluation of model answers against ground-truth references.** The template presents the question, ground truth, and model output provided for LLM-as-a-judge evaluation and guides the evaluator to produce one of three outcomes, *Correct*, *Incorrect*, or *Unclear*, ensuring reliable and consistent scoring across all predictions.

Model (+ MASS)	Baselines (MASS)		Noise Injection		
	w/	w/o	BBox	Motion	All
Qwen2.5-VL-7B	61.93	52.30	58.64	58.51	58.42
LLaVA-OneVision-7B	59.24	45.79	54.50	54.19	54.19

Table 3. Noise-induced robustness evaluation on MASS-augmented models. BBox, Motion, and All columns report accuracy under 10% Gaussian perturbations applied to bounding-box/position, 3D trajectory, and both components jointly. Baseline columns (*w/* and *w/o* MASS) are provided as reference.

tative and the motion profiles serve as supplementary cues rather than exact measurements, the model tolerates moderate value-level perturbations. Failures mainly occur when noise flips the sign of motion components, *i.e.*, reversing positive/negative values, leading the model to make wrong assumptions about motion directions and further cascading into a misinterpretation of the entire physical process.

G. Inference Time and Memory Usage

We report wall-clock inference time on 100 randomly sampled test cases on an NVIDIA L40S (Table 4). MASS increases per-query latency due to motion-profile extraction, while GPU memory increases only modestly. Our goal is to evaluate the benefit of explicit spatial–motion cues for physical understanding, not an optimized pipeline. Importantly,

Model	Time (s/case)		GPU Mem (GB)	
	w/o	w/	w/o	w/
LLaVA-OV-7B	1.86	6.49	14.97	15.69
Qwen2.5-VL-7B	5.83	10.46	15.45	16.18

Table 4. Per-query inference time (s) and GPU memory usage (GB) on an NVIDIA L40S over 100 randomly sampled test cases, with and without MASS.

tantly, the motion profile is modular and cacheable per video, so it can be reused across multiple questions without re-extraction. Overhead can be further reduced via faster upstream modules and standard efficiency knobs (*e.g.*, lower frame stride/resolution, smaller top- K grounded entities), which we leave to future work.

H. Multi-object Tracking and Camera Motion

We agree that dense tracking in crowded scenes increases compute and may introduce failures. MASS-Bench is entity-centric: we track only grounded target entities (top- K , tunable; default 5), and our QA design and video filtering avoid highly crowded, many-entity cases. Camera ego-motion remains an open challenge; the current benchmark largely contains limited camera motion to focus on entity-centric spatial/motion grounding. MASS does not assume perfect

tracks and degrades gracefully under noisy motion cues (as shown in the noise robustness analysis in Table 3). We plan to extend the benchmark and method to crowded multi-entity tracking and stronger camera motion handling in future work.

I. Case Study

In this section, we present additional qualitative examples of video question answering produced by several SoRA VLMs across the categories defined in MASS-Bench, as shown in Figures 3–7. Hallucinated or incorrect predictions are highlighted in red. Each example is accompanied by expert human annotations describing the ground-truth physics-driven dynamics, providing a clear reference for evaluating model behavior and identifying failure modes.

References

- [1] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. 1
- [2] Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. Motion-sight: Boosting fine-grained motion understanding in multi-modal llms. *arXiv preprint arXiv:2506.01674*, 2025. 1
- [3] 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*, pages 22195–22206, 2024. 2
- [4] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025. 1
- [5] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 2

Video 

Human Expert
Event Annotation: *The airplane is flying low over a wildland area, but there is no visible evidence of an airport or runway nearby.*
Categorization: Spatial Understanding

Question
*Is **the airplane** in the video close to **the airport**?*


GPT-4o
 Yes, the airplane is close to the airport.

Gemini-2.5-Flash
 The plane is flying very low with its landing gear down, but the ground below is just open wildland with no runways or airport structures. So despite its low altitude, the visuals show it is not close to an airport.

Qwen-2.5-VL
 Based on the video alone, **it appears that the airplane is very close to the ground, which might suggest it is near the airport.**

Qwen-2.5-VL + MASS
 No, the airline is not necessarily close to the airport. We need more context to make a definitive answer.

Figure 3. **Video question-answering example from the Spatial Understanding (SU) category.** We present physics reasoning and comprehension cases from state-of-the-art VLMs evaluated on spatial understanding tasks. Each example includes the video-generation prompt and human expert annotations, with visual grounding annotated (**Red**), the corresponding questions (**Purple**), and model responses from GPT-4o (**Orange**), Gemini-2.5-Flash (**Blue**), Qwen2.5-VL (**Green**), and Qwen2.5-VL + MASS (**Gray**). Hallucinated content and critical contextual errors are highlighted in **Red**.

Video 

Human Expert
Event Annotation: *The video only shows a square hoop with both ice and fire, and the tiger does not jump through it.*
Categorization: Temporal Understanding.

Question
*Does the **tiger** jump through a **circular hoop** with fire before a **square hoop** with ice?*


GPT-4o
 No, the tiger jumps through **a square hoop with fire.**

Gemini-2.5-Flash
 No, the tiger does not jump through a circular hoop with fire before a square hoop with ice.

Qwen-2.5-VL
 The tiger **jumps through a circular hoop with fire** before a square hoop with ice.

Qwen-2.5-VL + MASS
 No, the tiger does not jump through a circular hoop with fire before a square hoop with ice.

Figure 4. **Video question-answering example from the Temporal Understanding (TU) category.** We present physics reasoning and comprehension cases from state-of-the-art VLMs evaluated on temporal understanding tasks. Each example includes the video-generation prompt and human expert annotations, with visual grounding annotated (**Red**), the corresponding questions (**Purple**), and model responses from GPT-4o (**Orange**), Gemini-2.5-Flash (**Blue**), Qwen2.5-VL (**Green**), and Qwen2.5-VL + MASS (**Gray**). Hallucinated content and critical contextual errors are highlighted in **Red**.

Video 

Event Annotation: Upon entering the scene, the second person walks towards the first person, who remains absorbed in examining the object.
Categorization: Motion & Action Recognition.

Question What does the **second person** do upon entering the scene?

GPT-4o The second person enters the scene, walks past the first person, and **continues moving forward**, holding an item in their hand.

Gemini-2.5-Flash Upon entering the scene, the second person (an older man in a maroon robe) walks from the left... He is also **holding a white object in his right hand**. He takes a few steps and then stops, facing the camera.

Qwen-2.5-VL The second person walks into the room from the left side, holding something in their hand, **and appears to be showing or using** it while the first person continues their activity near the door.

Qwen-2.5-VL + MASS The second person enters the scene and walks towards the first person.

Figure 5. **Video question-answering example from the Motion and Action Recognition (MAR) category.** We present physics reasoning and comprehension cases from state-of-the-art VLMs evaluated on motion and action recognition tasks. Each example includes the video-generation prompt and human expert annotations, with visual grounding annotated (**Red**), the corresponding questions (**Purple**), and model responses from GPT-4o (**Orange**), Gemini-2.5-Flash (**Blue**), Qwen2.5-VL (**Green**), and Qwen2.5-VL + MASS (**Gray**). Hallucinated content and critical contextual errors are highlighted in **Red**.

Video 

Human Expert **Event Annotation:** The fluttering of the racer's clothing and the subtle body adjustments of both the horse and the rider indicate the presence of wind resistance.
Categorization: Physics Comprehension

Question Is **wind resistance** effectively depicted in the video during the **horse race**?

GPT-4o No, wind resistance **is not effectively depicted** in the images from the video.

Gemini-2.5-Flash Wind resistance is present but only subtly shown: the horses' manes and tails flow backward from their speed.

Qwen-2.5-VL The image **does not provide sufficient visual evidence** to effectively depict wind resistance during the horse race.

Qwen-2.5-VL + MASS Yes, wind resistance is effectively depicted in the video during the horse race, as the behavior of the horses and jockeys indicates the presence of wind resistance.

Figure 6. **Video question-answering example from the Physics Comprehension (PC) category.** We present physics reasoning and comprehension cases from state-of-the-art VLMs evaluated on physics comprehension tasks. Each example includes the video-generation prompt and human expert annotations, with visual grounding annotated (**Red**), the corresponding questions (**Purple**), and model responses from GPT-4o (**Orange**), Gemini-2.5-Flash (**Blue**), Qwen2.5-VL (**Green**), and Qwen2.5-VL + MASS (**Gray**). Hallucinated content and critical contextual errors are highlighted in **Red**.

Event Annotation: *The basketball player is shooting, but the ball moves upward across the hoop instead of descending through it, exhibiting a counterfactual, physics-violating motion.*

Categorization: Physical Abnormality Detection.

Does the *basketball* pass through the hoop from bottom to top?

No, the basketball **does not pass through the hoop from bottom to top.**

No, the basketball does not pass through the hoop from bottom to top. The player performs a dunk, which means the ball enters the hoop from the top and passes downwards through the net.

No, the basketball does not pass through the hoop from bottom to top. **It passes through from top to bottom.**

Yes, the basketball passes through the hoop from bottom to top.

Figure 7. **Video question-answering example from the Physics Abnormality Detection (PA) category.** We present physics reasoning and comprehension cases from state-of-the-art VLMs evaluated on physics abnormality detection tasks. Each example includes the video-generation prompt and human expert annotations, with visual grounding annotated (**Red**), the corresponding questions (**Purple**), and model responses from GPT-4o (**Orange**), Gemini-2.5-Flash (**Blue**), Qwen2.5-VL (**Green**), and Qwen2.5-VL + MASS (**Gray**). Hallucinated content and critical contextual errors are highlighted in **Red**.