Grounding Video Reasoning in Physical Signals

Alibay Osmanli Zixu Cheng Shaogang Gong
Queen Mary University of London
{a.osmanli,zixu.cheng,s.gong}@qmul.ac.uk

Abstract

Physical video understanding requires more than naming an event correctly. A model can answer a question about pouring, sliding, or collision from textual regularities while still failing to localize the event in time or space. We introduce a grounded benchmark for physical video understanding that extends the what–when–where evaluation structure of V-STaR [5] to four video sources, six physics domains, three prompt families (`physics`, `vstar_like`, and `neutral_rstr`), and four input conditions (original, shuffled, ablated, and frame-masked). The benchmark contains 1,560 base video clips from SSV2 [11], YouCook2 [32], HoloAssist [24], and Roundabout-TAU [16]. Each clip is first converted into a shared grounded event record, and the three query families are derived from that record. Temporal and spatial targets are shared across prompt families, while the non-physics families use deterministic family-appropriate semantic `a_what` targets derived from the same record. Across models and prompt families, `physics` remains the strongest regime overall, `vstar_like` is the clearest non-physics semantic comparison, and `neutral_rstr` behaves as a harder templated control. Prompt-family robustness is selective rather than universal, perturbation gains cluster in weak original cases, and spatial grounding is the weakest across settings. These results suggest that video Q&A reasoning benchmarks shall report physically grounded, prompt-aware, and perturbation-aware diagnostics alongside aggregate accuracy.

1. Introduction

In a video question answering task, a correct answer to a physical video question does not guarantee visually grounded understanding. A model can answer a question about pouring, sliding, or collision because the wording narrows the event type, even if it never identifies when the event happens or where the relevant objects are in a video. This matters in physical scenes, where motion, contact, force, and state change unfold over time and are often spatially localized. Standard answer-only benchmarks do not separate these cases. They tell us whether the final response matches the label, not whether the model uses video in a grounded way [4].

Recognising that a clip contains a "collision" from the question text is different from locating the interaction in time and grounding the participating objects in space. In robotics, video moment retrieval, and embodied decision-making, the first kind of success is not enough.

Existing benchmarks cover only part of this problem. Physical reasoning datasets such as CLEVRER [29], IntPhys [21], and PhysBench [6] probe physical understanding, but they mostly rely on categorical outputs. Grounded video benchmarks [13, 27, 31] evaluate temporal or spatial localisation, but they are not organized around physical event structure. V-STaR [5] took an important step by showing that strong performance on *what* can coexist with weak performance on *when* and *where*. That raises a useful diagnostic question for physical video understanding: when a model answers correctly, is it actually grounded in the video?

We study this question by introducing a physics-focused benchmark. Our benchmark keeps the grounded *what–when–where* structure of V-STaR, but moves it to six physics domains: Gravity, Fluids, Collisions, Deformation, Friction, and State Changes. Each sample is organized around a shared grounded event record with an event description, temporal span, and bounding box. The three prompt families—`physics`, `vstar_like`, and `neutral_rstr`—are alternative query formulations derived from that same record, and each is evaluated under four input conditions—original, shuffled, ablated, and frame-masked. The `physics` family is the main benchmark regime, `vstar_like` is a V-STaR-style semantic comparison rather than an exact reconstruction of the original benchmark prompts, and `neutral_rstr` is a neutral wording control that functions as a templated ablation.

The grounded event record stays fixed across prompt families, but the semantic field is expressed differently across them. Both `a_when` and `a_where` are scored against shared targets throughout. For `neutral_rstr` and `vstar_like`, `a_what` is scored against shorter family-appropriate semantic targets derived from the same record.

Under this protocol, `physics` remains the strongest overall regime, `vstar_like` is the main non-physics comparison, and `neutral_rstr` remains the harder control. These shifts are not uniform across models. Some remain comparatively strong across formulations, while others depend more heavily on the cues made explicit by physics framing.

Aggregate scores still hide much of this behavior. Two models with similar overall performance can react very differently when the wording changes or the visual evidence is perturbed. Figure 1 previews the setup on one shared clip. The prompt families ask about the same event, but the ground-truth and model box overlays show that plausible answers can still hide very different spatial grounding.

Research questions. We organize this study around four questions. RQ1 asks whether the grounded failure pattern identified by V-STaR persists in physical video understanding. RQ2 asks how stable grounded performance remains across `physics`, `vstar_like`, and `neutral_rstr` when the underlying event record is fixed. RQ3 asks what perturbation gains and losses reveal beyond original-condition performance. RQ4 asks which weaknesses remain even for strong models and favorable prompt families.

We keep the hypotheses narrow. We expect the V-STaR failure pattern to persist in physics-focused video, with spatial grounding remaining the weakest component. We also expect `physics` to be strongest overall, `vstar_like` to be the strongest non-physics comparison, and `neutral_rstr` to behave as a harder templated control. Finally, we expect perturbation gains to concentrate in weak or mid-baseline cases and to function as diagnostics of evidence sensitivity rather than as simple robustness wins.

Our main contributions are:

- A grounded benchmark for physical video understanding, built from 1,560 base clips from SSV2 [11], YouCook2 [32], HoloAssist [24], and Roundabout-TAU [16], and organized into six physics domains.
- A three-prompt evaluation design built on the same shared event record, separating the main `physics` regime from a V-STaR-style prompt family and a neutral wording control.
- A perturbation analysis framework that combines shuffled, ablated, and frame-masked inputs with component-wise metrics and diagnostic indices to interpret changes under degraded evidence.
- Empirical evidence that prompt-family robustness is selective across models, perturbation gains concentrate in weaker original cases, and spatial grounding remains the most persistent weakness across model families.

2. Related Work

2.1. Video-LLM evaluation

Recent video-LLMs combine a visual encoder with a pre-trained language model and instruction tuning over mixed-modality datasets [15, 18]. Systems such as Qwen2.5-VL [3], Qwen3-VL [2], VideoLLaMA3 [30], InternVideo2.5 [25], InternVL3.5 [23], and MiniCPM-o [28] perform well on standard benchmarks including Video-MME [8], MVBench [14], and LongVideoBench [26]. Those benchmarks are useful for broad coverage, but they score the final answer rather than the grounding process. A correct answer does not tell us whether the model located the relevant event in time and space or whether it relied on prompt cues and dataset regularities.

2.2. Physical reasoning benchmarks

Physical reasoning has been studied in synthetic and real-world settings for a long time. IntPhys [21] evaluates physical plausibility, while CLEVRER [29] probes causal and counterfactual reasoning in a controlled collision world. More recent benchmarks such as PhysBench [6] evaluate multimodal LLMs on physical concepts including gravity, collision, and material behavior. The common limitation is the answer format. Most of these benchmarks use categorical or multiple-choice outputs, so a model can score well without grounding the event itself in time or space.

2.3. Grounded video understanding

Grounded video understanding is usually studied through temporal grounding, spatio-temporal grounding, and grounded question answering. TALL and Charades-STA formalized temporal moment localization in untrimmed video [9], while ActivityNet Captions extended that setting to dense event description [12]. VidSTG [31] made spatio-temporal grounding a standard evaluation problem, and TVQA+ [13] and NExT-GQA [27] added temporal and spatial annotations to video Q&A. These datasets make grounding visible, but they are not designed around physical event structure or prompt-sensitive diagnosis.

2.4. Diagnostic evaluation and V-STaR

Several recent works argue that strong aggregate scores can conceal brittle behavior. Buch *et al*. [4] showed that some video-language benchmarks can be solved from linguistic structure alone. Bagad *et al*. [1] found that leading video-language models remain weak at chronological reasoning when frame order is manipulated. In other modalities, Winoground [22] and CheckList [20] showed the value of controlled diagnostic testing instead of relying only on headline benchmark scores.

V-STaR [5] is the closest prior benchmark to this study. It introduced a grounded *what–when–where* evaluation struc-

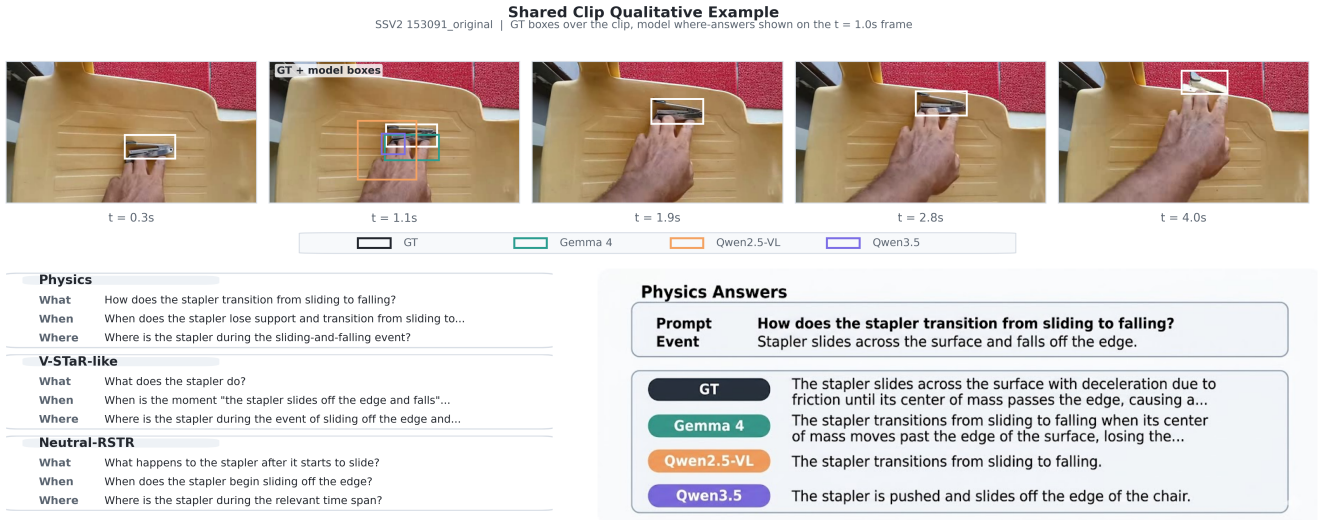

Figure 1. One shared clip across prompt families and grounded outputs. Top: matched SSV2 frames with the ground-truth box trajectory and representative model `a_where` predictions. Bottom left: the three query families derived from the same event record. Bottom right: representative physics answers from the ground truth and three models.

ture and showed that video-LLMs often perform much better on *what* than on *when* or *where*. Our benchmark uses the same diagnostic logic. Against physical reasoning benchmarks, we add grounded outputs. Against standard grounding benchmarks, we add a physics-domain layer and controlled perturbations. We also include a V-STaR-style prompt family explicitly, not leaving that comparison implicit.

3. Benchmark design

3.1. Task format and prompt families

Our benchmark inherits the grounded *what–when–where* evaluation structure introduced by V-STaR [5]. Given a video clip and a physical question, the model must produce a single structured prediction with three fields:

- `a_what`: a short text description of the physical event,
- `a_when`: a temporal interval [`start_sec`, `end_sec`],
- `a_where`: a normalised bounding box `x`, `y`, `w`, and `h`.

Requiring all three parts in one response is deliberate. Recognition, temporal grounding, and spatial grounding can each be gamed differently if they are evaluated in isolation. By forcing the model to commit to one coherent account of the event, we make it harder to score well by naming the event while ignoring where it happened or when it started.

We evaluate the same grounded event record under three prompt families:

- **`physics`**: the main benchmark regime, written to foreground physical dynamics and event descriptions.
- **`vstar_like`**: questions written in the style of the original V-STaR prompts using the existing annotations only. It is a continuity control, not an exact reconstruction of original V-STaR semantics.
- **`neutral_rstr`**: a neutral wording control that preserves the same grounded event and output schema while removing physics-specific phrasing. It is best read as a templated ablation rather than as the paper's main semantic baseline.

The prompt families are not separate annotation pipelines. All three are derived from the same grounded event record, which fixes the event identity, temporal span, and spatial reference. For cross-family evaluation, `a_when` and `a_where` are therefore scored against the same targets in every family. The text field works slightly differently. The `physics` family uses the reference event description directly, while `neutral_rstr` and `vstar_like` use shorter family-appropriate semantic `a_what` targets derived from that same record. This keeps the semantic target close to the answer style requested by the non-physics prompts while preserving the shared grounded event.

Figure 1 shows an example of this shared-record design. The wording changes across prompt families, but the grounded event, temporal span, and spatial reference do not.

3.2. Perturbation conditions

Each base clip is evaluated under four input conditions:

- **Original**: the unmodified RGB video.
- **Shuffled**: frames are randomly permuted while the set of frames is kept fixed.
- **Ablated**: each frame is converted to greyscale and blurred to suppress color and fine texture while preserving coarse spatial structure.
- **Frame-Masked**: half of the frames are replaced by black frames while video length and frame rate remain unchanged.

Each condition targets a different component of the visual

Condition	Input change	Diagnostic role
Original	unmodified video	baseline performance
Shuffled	frame order permuted	temporal order sensitivity
Ablated	greyscale + blur	appearance dependence
Frame-Masked	50% frames replaced by black	robustness to missing evidence

Table 1. The four evaluation conditions and the diagnostic role each is intended to probe. The grounded target is held fixed across conditions.

signal. Shuffling removes temporal order while preserving the frame set. Ablation removes color and fine appearance detail while preserving coarse structure. Frame masking removes evidence intermittently while preserving the temporal structure. Together, these conditions let us ask whether a model depends on temporal order, appearance detail, or persistent visual evidence, instead of collapsing those effects into one original-input score.

Table 1 summarizes what each condition is meant to probe, and Figure 2 shows the same clip under all four variants. The ablated row is especially useful because object extent remains visible while fine appearance cues are suppressed.

3.3. Metrics and reporting conventions

We score the three output fields with grounded metrics. Let Acc be text accuracy for `a_what`, tIoU be temporal intersection-over-union for `a_when`, and sIoU be spatial intersection-over-union for `a_where`. Following V-STaR [5], we combine these three components with the Logarithmic Geometric Mean:

$$\begin{aligned}\mathrm{LGM} = -\frac{1}{3}\Big[&\log(1-\mathrm{Acc}+\epsilon)\\ &+\log(1-\mathrm{tIoU}+\epsilon)\\ &+\log(1-\mathrm{sIoU}+\epsilon)\Big]\end{aligned} \tag{1}$$

where ϵ is a small constant for numerical stability.

We use LGM for continuity with V-STaR, but interpret it together with the component scores rather than as a replacement for them. LGM is high only when all three components are strong. A model that achieves good semantic accuracy while failing at temporal or spatial localisation will still score poorly. In this setting, that behavior is intentional: strong `a_what` performance should not hide weak `a_when` or `a_where` predictions.

There is one reporting detail worth stating explicitly. V-STaR applies an additional linear readability scaling to LGM [5]. Our pipeline reports LGM directly on the normalized component scale. This changes the displayed magnitude, but not the ranking or the perturbation ratios.

3.4. Diagnostic indices

We derive three indices from the four condition scores:

$$\mathrm{SBI} = 1 - (\mathrm{LGM}_{\mathrm{orig}} - \mathrm{LGM}_{\mathrm{shuf}}), \tag{2}$$

$$\mathrm{PRI} = \frac{\mathrm{LGM}_{\mathrm{abl}}}{\mathrm{LGM}_{\mathrm{orig}}}, \tag{3}$$

$$\mathrm{SPI} = \frac{\mathrm{LGM}_{\mathrm{mask}}}{\mathrm{LGM}_{\mathrm{orig}}}. \tag{4}$$

PRI and SPI are retention ratios relative to the original condition, so a value of 1 means that the perturbed input matches the original-condition score. SBI is written on the same centered scale, with 1 again meaning no change under temporal shuffling. Larger values therefore indicate greater robustness to the corresponding perturbation, while smaller values indicate loss under the perturbed input. We use these indices as descriptive diagnostics rather than standalone evidence: they are interpreted together with component deltas, baseline strata, validity transitions, and qualitative examples.

3.5. Validation-aware scoring

Model outputs are sometimes empty, malformed, or refusal-like. The evaluation pipeline counts those cases in the denominator and scores them as zero rather than silently dropping them. This matters for interpretation. If failed outputs were excluded, perturbation deltas would look artificially favorable for models that simply stop producing valid answers under harder settings.

4. Dataset construction

4.1. Sources and scale

The benchmark is constructed from four video sources that differ in viewpoint, event type, and annotation style: SSV2 [11], YouCook2 [32], HoloAssist [24], and Roundabout-TAU [16]. In total, the benchmark contains 1,560 base clips. Each base clip is expanded into four perturbation conditions, producing 6,240 scored video-condition pairs.

Table 2 shows the source breakdown. SSV2 and YouCook2 provide the largest portions of the data. HoloAssist contributes egocentric manipulation scenes in which hands and objects overlap under camera motion. Roundabout-TAU contributes overhead traffic footage where appearance cues are weak and motion structure matters more than texture.

4.2. Physics domains

Every clip is assigned to one of six physics domains: Gravity, Fluids, Collisions, Deformation, Friction, and State Changes. The domain layer matters because it lets us ask not only whether a model fails, but where it fails. A model may

Figure 2. The four input conditions applied to the same YouCook2 clip at matched timestamps. Yellow boxes show the ground-truth object reference across rows. Shuffling preserves the frame set while breaking temporal order; ablation suppresses color and fine texture while preserving coarse object extent; frame masking removes evidence intermittently.

Source	Base clips
SSV2	600
YouCook2	600
HoloAssist	275
Roundabout-TAU	85
Total	1,560

Table 2. Source breakdown. Counts are for base clips before perturbation expansion.

Domain	Base clips
Gravity	248
Fluids	283
Collisions	296
Deformation	248
Friction	196
State Changes	289
Total	1,560

Table 3. Domain breakdown for the benchmark base clips.

perform well on State Changes because those events are temporally salient and semantically distinctive, yet still struggle on Friction or Collisions where localization is harder and the critical evidence can be brief or spatially small.

Table 3 shows the domain distribution across base clips. No domain is severely underrepresented, so per-domain analysis remains meaningful.

4.3. Annotation pipeline

Each base clip is converted into a common grounded record containing a reference event description, a temporal span, a bounding box, and a physics-domain label. The key design choice is that this record is prompt-family-agnostic. It defines which event should be grounded, when it happens, and where it occurs, before any particular query wording is chosen. Text annotations are produced with a local Qwen3.5-based generator [19], which rewrites source metadata and event windows into that shared event description. Temporal spans come from the source adapters and event windows rather than from a second prompt-generation stage.

Spatial annotations are generated on the original clip with GroundingDINO [17] and then reused across all four perturbation conditions. This keeps the target object reference fixed when the input is shuffled, ablated, or frame-masked. The prompt families are layered on top of the same grounded record. They do not regenerate temporal spans or spatial boxes. The semantic text target is handled more carefully: `physics` keeps the longer reference event description, while `neutral_rstr` and `vstar_like` derive shorter prompt-aligned semantic `a_what` targets from the same grounded record.

We also performed a manual audit of approximately 120 generated records sampled across sources and physics domains. For each record, we checked whether the reference event was visible, whether the temporal span covered the

relevant physical interaction, and whether the spatial target matched the relevant object or interaction region. The audit was not meant to make the benchmark fully human-verified; it was a sanity check for systematic pipeline errors and for cases where automatic supervision is fragile. The main ambiguous cases were egocentric clips, where hands and objects overlap, and traffic clips, where several vehicles may participate in the same interaction. We therefore interpret fine-grained spatial scores with this label noise in mind.

4.4. Source adaptation and Roundabout selection

The four sources are not simply merged. Each one is adapted into the same grounded annotation format and mapped into the six-domain taxonomy. SSV2 contributes short object-manipulation clips in which temporal order is often decisive. YouCook2 contributes longer procedural clips with extended state changes and fluid events. HoloAssist contributes first-person manipulation in which object visibility and grounding are harder because the camera moves with the actor.

Roundabout-TAU requires an additional filtering step. Much of the raw traffic footage shows ordinary circulation rather than a localized physical interaction. We therefore map the source event labels into our physics taxonomy and exclude normal traffic by default. Clips are retained only when the event can be grounded as a localized interaction, such as collision-like behavior or a physically meaningful maneuver conflict. This is also how we justify using Roundabout for the *Collisions* domain: we do not treat every traffic clip as a collision example, only the subset whose event annotation corresponds to an interaction that can be temporally and spatially grounded.

4.5. Prompt and perturbation reuse

The query families are rendered automatically from the shared record rather than rewritten manually for each sample. This keeps supervision aligned across prompt families, but some prompts are terser or less natural than others. We therefore interpret cross-family results primarily through aggregate behavior over shared targets. Changes across prompt families or perturbation conditions reflect question formulation, visual evidence, and answer-style alignment rather than drift in the underlying grounded event.

5. Experimental Setup

5.1. Models

Table 4 lists the ten-model suite, spanning general-purpose VLMs, video-native models, a compact multimodal model, and recent open multimodal LLMs.

We evaluate Qwen2.5-VL [3], Qwen3-VL and Qwen3-VL-Thinking [2], VideoLLaMA3 [30], InternVideo2.5 [25], InternVL3.5 [23], MiniCPM-o [28], Qwen3.5 [19], Gemma 4 [10], and Molmo2 [7]. The suite gives us useful contrasts: two Qwen generations, two video-native models, one image-centric VLM, a compact multimodal model, and several recent open multimodal LLMs.

5.2. Inference setup

All models return one JSON object containing `a_what`, `a_when`, and `a_where`. We evaluate all four perturbation conditions, and cross-family comparisons use the full completed model set. Frame budgets follow native presets rather than a single shared budget because the benchmark is meant to measure model behavior in its normal operating mode. The most informative comparisons are therefore within-model changes across prompt families and perturbations, not perfectly matched frame counts across architectures.

5.3. Evaluation and analysis protocol

For each sample, the prediction is compared to the reference annotation using text accuracy, temporal IoU, spatial IoU, and LGM. Condition-level means are computed first, and SBI, PRI, and SPI are derived from those means. Missing, malformed, or refusal-like outputs are scored as zero and kept in the denominator. Otherwise the benchmark would overstate models that simply stop returning usable answers under harder conditions.

The main original-condition physics comparison uses all ten integrated models. Prompt-family comparisons use the same suite, with shared `a_when` and `a_where` targets and prompt-aligned non-physics `a_what` targets derived from the same event record.

We interpret perturbation behavior with analyses beyond the aggregate indices. For selected models and prompt families, we use per-sample ΔLGM, bootstrap confidence intervals, sign tests, baseline-stratified summaries, leave-one-dataset-out checks, and qualitative examples to interpret gains and losses.

6. Results

6.1. Main physics benchmark

Table 5 reports original-condition grounded metrics in the `physics` prompt family together with the three perturbation indices.

Table 5 shows the same failure pattern that motivated V-STaR: strong semantic accuracy does not guarantee equally strong temporal or spatial grounding. Qwen3-VL leads Acc, Qwen2.5-VL leads tIoU and sIoU, and VideoLLaMA3 leads LGM. These models arrive there in different ways: VideoLLaMA3 is strongest temporally, Qwen3-VL is strongest semantically, and Qwen2.5-VL is the most balanced on temporal and spatial localization. Molmo2 joins this top group with a distinct perturbation profile. Spatial grounding remains the clearest weakness, with no model exceeding 0.073 mean sIoU.

Model	Params	Family
VideoLLaMA3-7B	7B	video-centric model
Qwen3-VL-8B-Instruct	8B	general-purpose VLM
Molmo2-8B	8B	multimodal LLM
Qwen2.5-VL-7B-Instruct	7B	general-purpose VLM
Gemma4-26B-A4B-IT	26B	multimodal LLM
MiniCPM-o 2.6	2.6B	compact multimodal model
Qwen3-VL-8B-Thinking	8B	thinking-variant VLM
InternVideo2.5-Chat-8B	8B	video-native model
InternVL3.5-8B	8B	image-centric VLM
Qwen3.5-9B	9B	native multimodal LLM

Table 4. Ten-model suite spanning general-purpose VLMs, video-native models, compact multimodal models, and open multimodal LLMs.

Model	Acc	tIoU	sIoU	LGM	SBI	PRI	SPI
VideoLLaMA3	0.319	0.547	0.042	**2.634**	0.959	0.926	1.032
Qwen3-VL	**0.687**	0.183	0.056	2.551	1.076	1.231	**1.337**
Molmo2	0.548	0.322	0.023	2.515	1.326	1.087	0.974
Qwen2.5-VL	0.378	**0.560**	**0.073**	2.399	1.076	1.073	1.091
Gemma4	0.481	0.455	0.034	2.150	1.173	1.047	0.862
MiniCPM-o 2.6	0.232	0.373	0.033	1.396	0.947	0.946	0.706
Qwen3-VL-Thinking	0.635	0.126	0.051	1.332	1.106	1.304	1.132
InternVideo2.5	0.189	0.285	0.027	1.223	1.290	1.064	0.705
InternVL3.5	0.393	0.263	0.036	0.828	1.302	1.058	1.001
Qwen3.5	0.310	0.323	0.037	0.766	1.309	1.093	0.965

Table 5. Physics-prompt results averaged over the four sources. Acc, tIoU, sIoU, and LGM are original-condition scores; SBI, PRI, and SPI summarize perturbation response. Values above 1 mean the perturbed condition outscored the original. Models are ordered by LGM values.

6.2. Prompt-family changes and V-STaR-style prompting

Table 6 reports original-condition LGM across prompt families on the full ten-model set.

Table 6 and Figure 1 show the same ordering: `physics` is strongest overall, `vstar_like` sits between it and `neutral_rstr`, and mean LGM falls from 1.779 to 1.139 to 0.547, making `vstar_like` the main non-physics comparison.

Cross-family shifts are model-specific rather than uniform. VideoLLaMA3 remains strong under `vstar_like`, and MiniCPM-o 2.6 is the clearest positive case: its `vstar_like` score exceeds its `physics` score. Molmo2 also recovers meaningfully under `vstar_like`, while Gemma4 is the cleanest `neutral_rstr` case. The strongest negative cases are Qwen3-VL and Qwen3-VL-Thinking, which both drop sharply outside `physics`. Qwen3.5 follows the same pattern at a lower level. High performance under physics-framed queries therefore does not necessarily transfer to alternative semantic formulations.

6.3. Interpreting perturbation indices

By construction, PRI and SPI compare ablated and masked performance to the original condition, and SBI compares shuffled performance to the original on the same centered scale. Values above 1 mean that the perturbed input outscored the original, but they should be read as diagnostics rather than blanket robustness. Across the analyses, low- and mid-baseline rows are more likely to improve than high-baseline rows.

Figure 2 and Table 7 show why the sign alone is not enough under `physics`. The same sign can arise in different regimes, so positive ΔLGM should not be read as one uniform notion of robustness. Qwen3-VL and VideoLLaMA3 make the contrast clear: their original-condition LGM values are close, but Qwen3-VL often improves under perturbation whereas VideoLLaMA3 is flatter or negative on shuffled and ablated inputs.

The case-level analysis points to a more specific failure pattern. Perturbations often rescue rows that were already weak on the original input, for example by making a temporal guess easier or by reducing distracting appearance detail. They usually do not improve rows where the model was already well grounded. In those stronger rows, shuffling tends to remove useful order information and masking removes evidence the model was using. This is why we avoid describing above-1 indices as better physical reasoning. They are better read as evidence sensitivity: the model's score changes when order, appearance, or frame availability is changed, but the mechanism depends on the original baseline and on which component, `a_what`, `a_when`, or `a_where`, moved.

Model	physics	vstar_like	neutral_rstr
VideoLLaMA3	2.634	2.171	1.181
Qwen3-VL	2.551	0.116	0.114
Molmo2	2.515	1.384	0.866
Qwen2.5-VL	2.399	1.058	0.750
Gemma4	2.150	0.945	1.315
MiniCPM-o 2.6	1.396	1.824	0.515
Qwen3-VL-Thinking	1.332	0.242	0.102
InternVideo2.5	1.223	0.839	0.304
InternVL3.5	0.828	0.448	0.297
Qwen3.5	0.766	0.524	0.214
Mean over 10-model set	1.779	1.139	0.547

Table 6. Original-condition LGM across prompt families on the full ten-model set. Non-physics rows use prompt-aligned semantic a_what; a_when and a_where remain shared across families. Models are ordered by physics LGM values.

Model / family	Perturb.	Mean ΔLGM	Reading
Qwen3-VL / physics	ablated	+0.589	strong-model gain
Molmo2 / physics	shuffled	+0.326	mid-baseline temporal gain
Gemma4 / neutral_rstr	shuffled	+0.347	control-family gain
InternVL3.5 / physics	shuffled	+0.302	low-baseline gain

Table 7. Representative positive-response cases. Positive ΔLGM is descriptive, not a general robustness claim.

6.4. Domain-level findings

Domain leaders vary by domain: Qwen3-VL leads Gravity and Friction, VideoLLaMA3 leads Fluids, and Qwen2.5-VL leads Collisions, Deformation, and State Changes. There is no winner-take-all leader. Domain-level LGM changes across models, the underlying limitation does not: spatial grounding remains weak even when a model leads a domain.

7. Limitations

Automatic supervision. Event descriptions, temporal spans, and spatial boxes are produced automatically rather than fully verified by human annotators. That gives the benchmark scale, but it also introduces noise. Our manual audit of approximately 120 records helped catch obvious pipeline errors, but it is not a substitute for exhaustive human verification. The hardest cases are egocentric and traffic videos, where the grounded target can be ambiguous even for a human reader. A larger human validation pass would improve confidence in the labels, especially for spatial grounding.

Annotation-style bias. The text annotations are generated with an LLM-based stage, and one evaluated model family is closely related to that generator. We therefore use shorter non-physics semantic a_what targets derived from the shared event record instead of reusing the longer physics-style answers verbatim. That keeps cross-family semantics closer to the answer style requested by the prompts, but the non-physics targets remain deterministic derivations rather than a manually curated multi-family gold set. Large cross-family drops should therefore not be over-attributed to any single cause: instruction-following preferences, answer-style priors, and deeper failures in grounded video reasoning can all contribute.

Source imbalance. The benchmark is reasonably balanced by domain, but not by source. Roundabout-TAU contributes 85 clips, compared with 600 each from SSV2 and YouCook2, so source-level findings should be read cautiously.

Scope of the task. The benchmark evaluates grounded event understanding: what happened, when it happened, and where it happened. It does not directly test causal forecasting, counterfactual physics, or long-horizon planning. A model that localises a collision correctly can still fail to predict what follows or to reason about alternative physical outcomes.

Diagnostic, not causal, interpretation. Perturbations expose behavior that original-condition scores hide, but they do not identify a single underlying cause. A gain under ablation can reflect reduced distractors, weak original baselines, or a cleaner temporal guess, so we treat perturbations as diagnostics, not causal explanations.

8. Conclusion

We introduced a grounded benchmark for physical video understanding that extends the *what–when–where* diagnostic idea of V-STaR [5] to physics-focused video, prompt-family shifts, and controlled perturbations. The experiments show selective prompt robustness and persistent spatial failures: answer accuracy alone still hides too much behavior.

References

[1] Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2503–2516, 2023.

[2] 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*, 2025.

[3] 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. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2907–2917, 2022.

[5] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

[6] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

[7] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

[8] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025.

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017.

[10] Google DeepMind. gemma-4-26b-a4b-it. https://huggingface.co/google/gemma-4-26B-A4B-it, 2026.

[11] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017.

[12] 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 (ICCV)*, pages 706–715, 2017.

[13] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8211–8225, 2020.

[14] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024.

[15] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[16] Yuqiang Lin, Kehua Chen, Sam Lockyer, Arjun Yadav, Mingxuan Sui, Shucheng Zhang, Yan Shi, Bingzhang Wang, Yuang Zhang, Markus Zarbock, Florian Stanek, Adrian Evans, Wenbin Li, Yinhai Wang, and Nic Zhang. TAU-R1: Visual language model for traffic anomaly understanding. *arXiv preprint arXiv:2603.19098*, 2026.

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz 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)*, pages 12585–12602, 2024.

[19] Qwen Team. Qwen3.5-9b-base. https://huggingface.co/Qwen/Qwen3.5-9B-Base, 2026.

[20] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912, 2020.

[21] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[22] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross.

Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, 2022.

[23] 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.

[24] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20270–20281, 2023.

[25] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

[26] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

[27] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214, 2024.

[28] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-o: A gpt-4o level mllm for vision, speech and multimodal live streaming. *arXiv preprint arXiv:2408.01800*, 2024.

[29] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

[30] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[31] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10668–10677, 2020.

[32] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.