From Motion to Meaning: A Benchmark for Physical Reasoning Through Event Decomposition

Jung-Chun Liu
Computer Science and Engineering Division
University of Michigan
Ann Arbor, MI 48109
jungchun@umich.edu

Eleanor Lin
Computer Science and Engineering Division
University of Michigan
Ann Arbor, MI 48109
elealin@umich.edu

Zichen Wang
Computer Science and Engineering Division
University of Michigan
Ann Arbor, MI 48109
zzzichen@umich.edu

1 Introduction

Understanding the physical world is essential for embodied artificial intelligence (AI) systems, enabling them to effectively interact, manipulate objects, and perform complex tasks. Vision-language models (VLMs), which integrate visual perception with linguistic reasoning, hold significant promise in achieving human-like physical reasoning capabilities. However, existing benchmarks designed to assess these capabilities often rely on simplistic question-answer (Q&A) formats, which inadequately capture the nuanced ways humans intuitively reason about physics through sequences of visual events and linguistic descriptions.

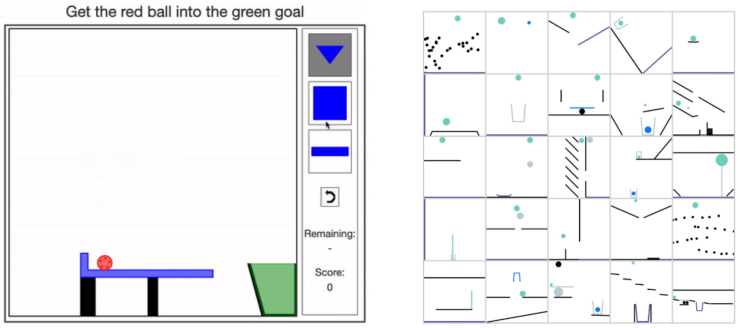
In this work, we address the limitations of traditional benchmarks by introducing a new evaluation framework specifically designed to probe VLMs' understanding of Newtonian mechanics, a fundamental component of intuitive physical reasoning.[1] Our benchmark consists of three interconnected tasks, each progressively testing and revealing different dimensions of physical reasoning capabilities:

1. **Physical Event Captioning**: Given a visual sequence depicting a physical scenario, the model must generate a structured textual summary that accurately describes the events occurring within the scene.
2. **Physical Event Prediction**: Presented with an initial partial sequence of a physical interaction, the model is tasked with predicting subsequent outcomes, explicitly stating its reasoning to arrive at a conclusion.
3. **Self-Correction of Physical Reasoning**: The model initially predicts outcomes based on partial information and then revisits its predictions after observing the complete sequence, identifying and correcting its reasoning errors.

We make three significant contributions to the evaluation of physical reasoning in VLMs. First, we employ prompts requiring models to explicitly state their reasoning in language, thereby providing clearer insights into specific points of failure and better approximating the nuanced human cognitive process of breaking down complex visual information into sequential linguistic descriptions. Second, we demonstrate that encouraging models to engage in multiple rounds of explicit reasoning enhances their performance beyond single-step inference. Third, we introduce an approach for pinpointing the

[1] See the appendices for links to our data and code.

Preprint. Under review.



(a) Virtual Tools (b) Phyre

Figure 1: Example tasks from two interactive physical reasoning benchmarks. (a) Virtual Tools tasks involve placing objects to solve physical puzzles using principles like gravity and collision [1]. (b) PHYRE tasks require placing a ball to trigger interactions that achieve a goal, testing generalization across visually simplified scenarios [3].

effect of semantic type and temporal sequencing of physical events on predicting the final outcome of a sequence of events.

Through this framework, we aim not only to benchmark VLM capabilities but also to provide a deeper understanding of the strengths and limitations of current models. Ultimately, our benchmark serves as a diagnostic tool, highlighting specific breakdowns in VLMs' physical reasoning and guiding future advancements in developing models that more closely emulate human intuitive understanding of physics.

2 Related Work

Melnik *et al*. provide a survey on benchmarks designed to evaluate physical reasoning capabilities in artificial intelligence systems [5]. Following their categorization, we focus our discussion on two primary groups: interactive problem-solving benchmarks and language-related benchmarks.

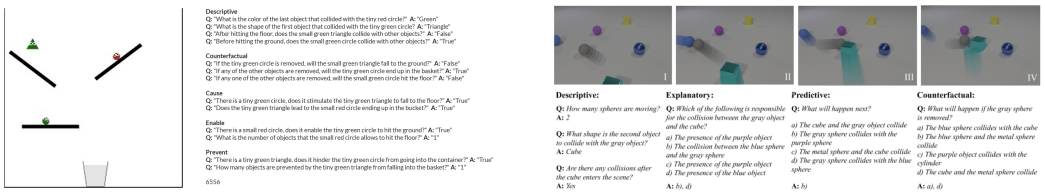
2.1 Interactive Problem-Solving Benchmarks

Interactive benchmarks allow AI systems to actively engage with the environment, offering a solid first step for testing further problem-solving through direct manipulation and exploration.

Virtual Tools [1] is an interactive 2D environment designed to test physical problem-solving abilities by choosing an initial action to achieve specific goals, such as guiding a ball into a goal area (see Figure 1a). This benchmark consists of different physical concepts such as falling, launching, or bridging, with an emphasis on in-time exploratory learning and physical manipulation. The Virtual Tools Game features naturalistic physics, such that objects move and collide as if subject to real-world forces, *e.g*., gravity. The Virtual Tools Game also features objects of various colors, shapes, densities, and sizes which can be placed strategically so as to achieve creator-specified goals.

PHYRE (Physical Reasoning Environment) [3] similarly provides an interactive platform with visually simplistic 2D environments where agents must strategically place balls to meet predefined goals, like making objects touch (see Figure 1b). PHYRE is particularly notable for its structured puzzles that rigorously test physical foresight and generalization across unseen scenarios.

These interactive benchmarks, including others such as Phy-Q [7], have been instrumental in exploring the boundaries of physical interaction and reasoning. However, existing AI systems demonstrate poor performance on these benchmarks to date. We argue that this is primarily because the benchmarks and proposed methods inherently lack mechanisms for complex linguistic reasoning about sequences of events and rely purely on vision. While precisely predicting each future time frame is difficult, abstracting continuous visual information into a structured sequence of events and applying logical reasoning on top of it significantly simplifies the task and resembles human thinking more. To address



(a) CRAFT (b) CLEVERER

Figure 2: Example tasks from language-integrated video reasoning benchmarks. (a) CRAFT evaluates agents using natural language questions about causal and counterfactual events within video sequences [2]. (b) CLEVRER extends this by including descriptive, explanatory, predictive, and counterfactual questions on synthetic physics simulations, enabling multi-faceted evaluation of event understanding and reasoning [8].

this gap, we adapt the Virtual Tools environment to create our own benchmark, which tests VLMs' ability to reason in language about sequences of physical events.

2.2 Language-Related Benchmarks

Language-based benchmarks incorporate linguistic queries to assess an AI system's understanding and reasoning about physical events described or queried through natural language, offering deeper insights into cognitive aspects of physical reasoning.

CRAFT [2] integrates video sequences with linguistic question-answering tasks, probing agents' understanding of causal interactions and counterfactual scenarios (see Figure 2a). CRAFT's linguistic component allows it to evaluate descriptive, explanatory, and counterfactual reasoning, providing rich contextual evaluations.

CLEVRER (Collision Events for Video Representation and Reasoning) [8] further advances linguistic reasoning benchmarks by posing natural language questions on simulated video scenarios involving collisions (see Figure 2b). It encompasses descriptive, explanatory, predictive, and counterfactual questions, systematically examining various dimensions of causal and temporal reasoning.

Other related benchmarks, such as ComPhy [4] and CRIPP-VQA [6], also explore linguistic and causal reasoning dimensions. Nonetheless, the question-answer format of all these methods primarily targets isolated and discrete aspects of scenarios, limiting comprehensive reasoning about extended sequences of events. In contrast, our proposed benchmark aims to bridge the gaps highlighted above by incorporating detailed linguistic prompts and self-correction mechanisms, thereby facilitating holistic evaluations of how vision-language models reason through complex sequences of events using both visual and linguistic modalities.

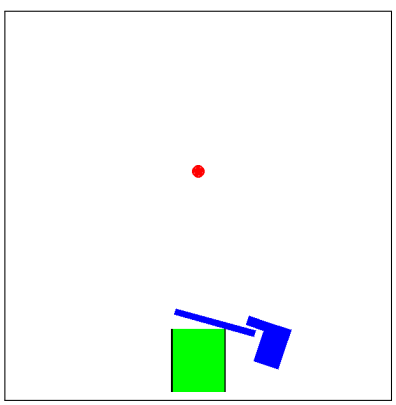
3 Approach

In this section, we present the motivation and design behind our proposed benchmark for evaluating the physical reasoning and problem-solving abilities of modern vision-language models (VLMs). Central to our approach is the idea that complex physical scenes can be broken down into a sequence of interpretable physical events (Section 3.1), which can then be reasoned about using language (Section 3.2). Our benchmark builds on this insight to bridge the gap between low-level visual prediction and high-level logical reasoning (Section 3.3).

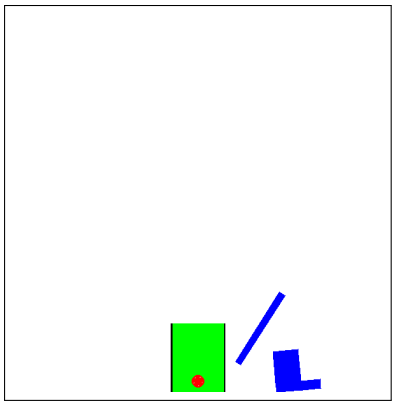
3.1 Problem Formulation

Physical reasoning tasks often require an agent to understand not only what is visually happening but why it happens and what might happen next. Instead of requiring pixel-level future prediction, which is often brittle and low in semantic clarity, we propose formulating the task as reasoning over a structured sequence of discrete events.

This aligns well with the nature of the classical Newtonian mechanics underpinning Virtual Tools. In this environment, rigid bodies remain stationary unless acted upon, accelerate under gravity, and



(a) **Caption:** A collision occurs between the l-shaped object and the lid.



(b) **Caption:** The ball reaches the goal.

Figure 3: Examples of event captions in our benchmark. (a) A collision between two blocks, which removes the block obstructing the ball from reaching the green goal region. (b) The red ball enters the goal region. Each benchmark instance is a video showing the composition of one or more physical events.

interact through collisions. This physical structure provides a principled basis for decomposing continuous motion into discrete, interpretable segments. We say an ***event*** occurs when there is a transition between two such segments. More formally, an event is defined as a moment when one or more objects experience a change in their force interaction state—typically due to collisions or the initiation or termination of contact. In addition to these physical transitions, we also define the red ball reaching the goal region as a distinct, task-specific event type.

We end up with the following five event types in our benchmark:

- **Collision:** Two objects bump into each other and cease contact within a specified duration.
- **Start Touch:** Two objects come into contact for more than the specified duration.
- **Stop Touch:** Two objects that were previously touching (*i.e.*, the state after a **Start Touch** event) cease contact.
- **Goal:** The red ball successfully enters the green goal region. A benchmark instance ending with this event is a "success."
- **End:** The simulation concludes without the red ball reaching the goal. A benchmark instance ending with this event is a "failure."

Start Touch and **End Touch** represent the start and end of a touching event, respectively, so these 2 event types always appear as a pair. In our work, the specified duration used to distinguish collisions from start/stop touch events is 0.2 seconds.

3.2 Event Captioning

Having segmented the continuous motion into discrete events, we next assign natural language descriptions to each event. This design choice reflects how humans typically reason about physical processes—not at the pixel or frame level, but through abstract interpretations of cause and effect expressed in language.

To generate meaningful captions, we consider two levels of semantic granularity. At a basic level, each object can be described generically using role-based labels such as "the red ball" or "the blue object." At a more expressive level, objects are annotated with semantically rich names like "platform," "lid," or "bar,"' which provide additional context to help guide the model's reasoning. To balance clarity and expressiveness, we adopt a hybrid strategy: object references are grounded in visually distinctive features (*e.g.*, color and shape) while incorporating domain-relevant semantic labels when available. Captions are generated using fixed templates tailored to each event type:

- **Collision:** A collision occurs between the `[object A]` and the `[object B]`.
- **Start Touch:** The `[object A]` and the `[object B]` start touching.
- **Stop Touch:** The `[object A]` and the `[object B]` stop touching.
- **Goal:** The `[object A]` reaches the goal.
- **End:** The `[object A]` never reaches the goal until the end of the video.

This templated approach ensures consistency across the dataset while allowing sufficient flexibility to capture meaningful distinctions between object roles and interactions.

3.3 Benchmark Overview

With the above preparation, we are finally ready to offer an overview of our new benchmark. We repurpose the Virtual Tools Game simulator [1] to generate our benchmark dataset. Each instance in the benchmark consists of a video clip depicting a simulated sequence of physical events. We also provide a decomposition of the full video into individual video clips, each capturing one physical event. Additionally, we provide one templated caption for each event. Each video contains a red ball, a green goal region, and potentially various static black blocks or movable blue blocks that interact according to Newtonian mechanics. An instance is labeled as a "success" if the red ball reaches the goal, and a "failure" otherwise.

Our benchmark includes ***simple*** and ***general*** subsets. The simple subset features videos with only one physical interaction preceding the terminal goal or end event. The general subset includes more complex compositions of multiple events, testing the model's ability to reason over extended causal chains. Across both subsets, we include in total 78 distinct event instances. Finally, we adapt and modify canonical configurations from the Virtual Tools Game—including Catapult, Gap, Unbox, Launch, and Table levels—to construct diverse and interpretable evaluation scenarios.

4 Evaluation

We evaluate the zero- and few-shot physical reasoning capabilities of vision-language models (VLMs) on our benchmark across three tasks designed to test different aspects of event understanding and prediction. The models we benchmark include: **GPT-4o** and **o3-mini**,[2] **LLaVA-Video-72B-Qwen2**,[3], **Claude 3.7 Sonnet**,[4] and **Gemini 2.0 Flash**.[5] These models were selected based on their state-of-the-art performance in vision-language tasks involving temporal and physical reasoning. LLaVA-Video-72B-Qwen2 and Gemini 2.0 Flash take video inputs, while the remaining models take image inputs. Except for LLaVA-Video-72B-Qwen2 (which we host ourselves) and Gemini 2.0 Flash (which we access through Google's API), we access all models through their publicly available browser-based chat interfaces. For LLaVA-Video-72B-Qwen2, running inference takes a total of 1.6 hours using a single Nvidia RTX A6000; all other models evaluated do not require any special hardware. We include our prompts for each task in the appendix.

4.1 Task 1 (Pilot): Physical Event Captioning

In this pilot task, we test a model's ability to decompose a visual scene into interpretable physical segments and describe them in consistent, structured language. The model observes a full video composed of multiple events and is asked to generate a caption for each event in the style of a reference example.

We evaluate model performance by comparing model output to the gold-standard caption sequence, using GPT-4 as a judge. The judge scores the similarity between the reference and predicted descriptions on a scale from 1 to 5, considering consistency in object interaction descriptions and accurate identification of final outcomes.

[2] `https://chatgpt.com/`

[3] `https://huggingface.co/lmms-lab/LLaVA-Video-7B-Qwen2`

[4] `https://claude.ai`

[5] `https://aistudio.google.com/prompts/new_chat?model=gemini-2.0-flash-exp`

Table 1: **Physical Event Captioning from Video Evaluation.** Average GPT-4 judge scores (1 to 5) for Gemini 2.0 Flash on Task 1, across the simple subset, general subset, and all instances.

Model	Simple	General	Overall
Gemini 2.0 Flash	1.86	2.13	2.05

We present the results for Task 1 in Table 1. Overall, Gemini 2.0 Flash performs poorly on Task 1, scoring < 3 on average on the entire benchmark, as well as on the simple and general subsets. Manual analysis of the model generations reveals basic errors in video understanding which likely account for the low scores, *e.g.*, stating that "The ball touches the goal" when in fact the ball never reaches the goal. We also conjecture that more detailed instruction and examples (beyond the 1-shot example we prompt with) may be necessary for models to generate descriptions in the desired structure.

As an additional, exploratory step for this pilot task, we also experiment with converting the video input to a single still image (depicting the starting configuration of the scene), and prompting those models which take only image inputs (Claude 3.7 Sonnet, o3-mini, and GPT-4o) with this image to perform the same task. For the image-only models, Task 1 is thus an even more challenging prediction (as opposed to mere description) task, since the models need to generate captions for unseen future physical events. As seen in Table 2, the image-only models perform equally poorly to the Gemini 2.0 Flash video model on Task 1, all scoring < 3 on all splits of the dataset.

Table 2: **Physical Event Captioning from Image Evaluation.** Average GPT-4 judge scores (1 to 5) for models which take only image inputs, across the simple subset, general subset, and all instances.

Model	Simple	General	Overall
Claude 3.7 Sonnet	1.86	1.73	1.77
o3-mini	**2.57**	2.33	**2.41**
GPT-4o	2.14	**2.53**	**2.41**

Based on our findings from the Task 1 pilot, we shift our focus from structured event description to event description and prediction through unstructured language in the subsequent tasks. We also prioritize evaluating models that can process video input, over models that can only process still images, in subsequent tasks.

4.2 Task 2.1: Predicting Future Physical Events from Initial Event

In this task, we test the model's ability to perform language-based reasoning for physical event prediction. The model receives a video clip depicting the first of a series of physical events. In other words, given a benchmark video instance which shows a sequence of multiple physical events, we take a video clip starting from the beginning of the video and ending at the moment when the first event occurs. The model is asked to predict whether the ball will reach the goal ("success") or not ("failure"). We ask the model to generate its reasoning before providing its final prediction.

Table 3: **Physical Event Prediction Accuracy.** Accuracy of each model on simple and general cases, as well as overall accuracy. Baseline is computed by always predicting the majority class ("Failure").

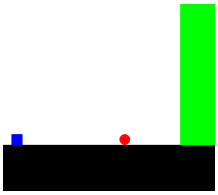
Model	Simple Cases	General Cases	All Cases
Gemini 2.0 Flash	0.286	0.375	0.348
LLaVA-Video-7b-Qwen	0.429	0.375	0.391
Baseline (always predict "Failure")	**0.571**	**0.563**	**0.565**

Our results for Task 2.1 are presented in Table 3. (Additional results for GPT-4o and o3-mini, which require reformulating this task to use image instead of video inputs, are available in the Appendix A.9.) Both Gemini 2.0 Flash and LLaVA-Video-7b-Qwen fail to outperform a naive baseline. To understand why, we annotate the generated model reasoning traces for correctness and present our results in Table 4. We see that for both models, most incorrect predictions are due to incorrect

reasoning traces. In fact, when the reasoning is correct, Gemini 2.0 Flash always generates a correct prediction. Similarly, LLaVA-Video-7b-Qwen generates incorrect predictions from correct reasoning only in rare cases (4.3% of the time). Interestingly, in a significant proportion of cases (21.7% for Gemini 2.0 Flash and 17.4% for LLaVA-Video-7b-Qwen), the model generates the correct prediction, but *for the wrong reason* (see Figure 4 for an example). These observations motivate our Task 3, where we test VLMs' ability to correct their own reasoning and predictions about physical events (see Section 4.4).

Table 4: **Physical Event Reasoning Accuracy** Proportions of benchmark instances with correct (✓)/incorrect (✗) reasoning/prediction.

Model	✓ **Reasoning** ✓ **Prediction** (↑)	✗ **Reasoning** ✗ **Prediction** (↓)	✗ **Reasoning** ✓ **Prediction** (↓)	✓ **Reasoning** ✗ **Prediction** (↓)
Gemini 2.0 Flash	0.130	0.652	0.217	0
LLaVA-Video-7b-Qwen	0.304	0.478	0.174	0.043



Reasoning: The red ball is located relatively far away from the goal on the right. The position of the blue object in front of the red ball further impedes it from reaching the goal.

Answer: Failure ✓

(a) The video.

(b) Model reasoning and outcome prediction.

Figure 4: A frame from a benchmark video instance depicting a "failure" case (the red ball remains stationary and never enters the green goal, due to lack of any external forces acting upon it). While Gemini 2.0 Flash correctly predicts that the video ends in failure, the model exhibits errors in its reasoning, stating that the blue block "impedes" the ball from reaching the goal. In fact, the blue block is on the wrong side of the ball for it to block the ball from reaching the goal.

4.3 Task 2.2: Predicting Future Physical Events from Different Event Types

In this task, we evaluate the models' ability to predict future physical events, given prior physical events of different temporal and semantic types. For each benchmark instance depicting a series of physical events, we segment the instance into a series of video clips, each depicting a single event. We then prompt the model separately on each clip to predict whether the ball will reach the goal ("success") or not "failure"). For example, the instance in Figure 3 can be decomposed into a **collision** event and a **goal** event, for a total of 2 video clips. For Task 2.2, we would separately prompt the model to predict the final outcome from each of the 2 clips. As in Task 2.1, we ask the model to generate its reasoning before providing its final prediction.

We present the resulting prediction accuracy by event type in Table 5. As expected, given a clip of the **goal** event (which shows the ball reaching the goal at end of the video), both Gemini 2.0 Flash and LLaVA-Video-7b-Qwen can correctly predict the video outcome 100% of the time. However, contrary to expectation, neither model is good at predicting the video outcome from clips of the **end** event (where the video ends without the ball reaching the goal). Upon closer inspection, both models sometimes hallucinate a "success" prediction despite receiving video input that clearly shows the ball coming to a halt without reaching the goal, *e.g.*, "After falling, the red ball enters the green goal region."

Now we consider cases beyond the trivial case of predicting the end state from a **goal** video clip. For Gemini 2.0 Flash, the **start touch** event is the most useful for predicting the final outcome, with the model achieving 50.0% accuracy. In contrast, for LLaVA-Video-7b-Qwen, the **stop touch** event is most useful for predicting the final outcome, with the model achieving 50.0% accuracy.

Overall, the Task 2.2 results suggest that not only the temporal sequencing of events, but also event semantics, affect models' physical event prediction accuracy. If only temporal sequencing

Table 5: **Prediction Accuracy by Input Event Type.** Accuracy of each model in Task 2.2, given input video clips showing different event types.

Model	Collision	Start Touch	Stop Touch	Goal	End
Gemini 2.0 Flash	0.487	0.500	0.375	**1.000**	0.385
LLaVA-Video-7b-Qwen	0.333	0.375	0.500	1.000	0.000
Baseline (always predict "Failure")	**0.667**	**0.625**	**0.625**	0.000	**1.000**

mattered, we would expect models to predict outcomes most accurately using the **goal** and **end** events temporally closest to the outcome. However, we have seen that this is not necessarily the case.

4.4 Task 3: VLM Physical Reasoning Self-Correction

In this task, we evaluate the model's ability to identify and correct errors in its reasoning process and predictions of whether a given benchmark instance ends in success or failure. The model is provided with a video clip showing the full sequence of physical events, its reasoning and prediction generated for this benchmark instance from Task 2.1 (see Section 4.2), and is asked to regenerate its reasoning and prediction, correcting any errors. Importantly, we do not tell the model whether the original reasoning and prediction are correct; we thus test not only model reasoning correction capabilities, but also models' ability to abstain from over-correcting responses that are already correct.

Comparing Tables 3 and 6, we see that both Gemini 2.0 Flash and LLaVA-Video-7b-Qwen show improvements in prediction accuracy, demonstrating both models' ability to identify and correct incorrect predictions. For Gemini 2.0 Flash, the improvement is statistically significant. A Wilcoxon signed-rank test indicated that Gemini 2.0 Flash's prediction accuracy over all cases was significantly higher after applying self-correction in Task 3 (0.913) compared to Task 2.1 (0.348), $V = 27.5$, $p = 0.0005$. Additionally, Gemini 2.0 Flash significantly outperforms the naive baseline, as indicated by a Wilcoxon signed-rank test, achieving an accuracy of 0.913 over all cases, compared to the naive baseline's accuracy of 0.565, $V = 68.0$, $p = 0.028$.

Table 6: **Prediction Self-Correction Accuracy.** Accuracy of each model in Task 3 after seeing the full video, evaluating their ability to revise earlier predictions.

Model	Simple Cases	General Cases	All Cases
Gemini 2.0 Flash	**1.000**	**0.875**	**0.913**
LLaVA-Video-7b-Qwen	0.429	0.500	0.478
Baseline (always predict "Failure")	0.571	0.563	0.565

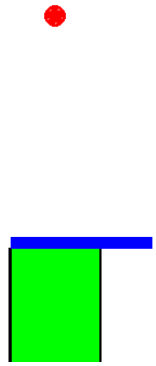
Comparing Tables 4 and 7, we see that Gemini 2.0 Flash is more successful at correcting its reasoning than LLaVA-Video-7b-Qwen. Gemini 2.0 Flash boosts the proportion of instances with correct reasoning leading to a correct prediction, from 13.0% in Task 2.1 to 60.9% in Task 3. The increase for LLaVA-Video-7b is much more modest (30.4% to 34.8%).

Table 7: **Physical Event Reasoning Correction Accuracy** Proportions of benchmark instances with correct (✓)/incorrect (✗) reasoning/prediction in Task 3.

Model	✓ **Reasoning** ✓ **Prediction** (↑)	✗ **Reasoning** ✗ **Prediction** (↓)	✗ **Reasoning** ✓ **Prediction** (↓)	✓ **Reasoning** ✗ **Prediction** (↓)
Gemini 2.0 Flash	0.609	0.087	0.304	0
LLaVA-Video-7b-Qwen	0.348	0.478	0.130	0.043

In Figure 5, we show an example of how Gemini 2.0 Flash is able to correct its reasoning, leading to a corrected final prediction. For the depicted benchmark instance, Gemini 2.0 Flash originally predicted in Task 2.1 (section 4.2) that the red ball would enter the green goal, disregarding the blue

block blocking the ball from entering the goal. However, in Task 3, Gemini 2.0 Flash corrects its reasoning by making note of this blue block as an obstacle, leading the model to correctly answer that the instance represents a failure case.



(a) The video

Pre-correction (Task 2.1)	**Reasoning:** The red ball is positioned directly above the green goal region. Given that no external forces appear to act on the ball, it will drop vertically into the goal. **Answer:** Success ✗
Post-correction (Task 3)	**Reasoning:** The red ball drops directly down and lands on the blue platform that is just above the green goal. Therefore the ball will not enter the green goal region. **Answer:** Failure ✓

(b) Model reasoning and outcome prediction pre- and post-correction.

Figure 5: A frame from a benchmark video instance depicting a "failure" case (the red ball is blocked from entering the green goal by the blue block). Gemini 2.0 Flash successfully corrects errors in its reasoning, leading to correction of its prediction of the final outcome from "Success" to "Failure."

5 Discussion

Our benchmark offers a modular and interpretable framework for evaluating physical reasoning in vision-language models. By breaking down dynamic scenes into discrete, captioned events, it opens several promising directions for future work:

- **Automatic Scene Generation:** In this work, we propose a formal definition of key physical events and an approach to automatically segment simulated videos using these events. However, the initial configurations of objects in the dataset are hand-crafted, and the corresponding results provide qualitative insights. To further evaluate VLMs' robustness quantitatively, the dataset could be generated automatically by formalizing the configuration and semantics of the scenes.
- **Whole-Sequence Reasoning:** The decomposition of motion into events provides a structured basis for reasoning about the sequence of events as a whole. Future iterations of Task 1 could formalize and evaluate how well models capture causality across full event chains.
- **Planning and Chain-of-Thought:** The structured event sequence enables VLMs to simulate mental rollouts or engage in multi-step physical reasoning. This format aligns well with prompting techniques like chain-of-thought reasoning and could be extended for physical planning tasks.
- **Interactive Problem Solving:** Although this work excludes active tool placement (in contrast to *e.g.*, [1]), our framework could be adapted for decision-making benchmarks where models must select or place tools to achieve goals. Event-based reasoning could support upstream problem-solving by predicting the effects of hypothetical actions.

Together, these extensions highlight the potential of our benchmark not only for evaluating physical understanding but also for enabling richer, multi-step reasoning grounded in physical intuition.

6 Conclusion

We introduce a new benchmark to evaluate the physical reasoning capabilities of vision-language models (VLMs) through the lens of structured event decomposition. Unlike traditional question-answering approaches, our benchmark emphasizes holistic understanding by breaking complex physical interactions into discrete, captioned events grounded in Newtonian mechanics.

Our evaluation across three tasks—event captioning, future event prediction, and self-correction—reveals critical limitations in current VLMs' physical reasoning abilities. While models like Gemini 2.0 Flash show promising self-correction performance given sufficient additional information, their ability to perform physical reasoning given limited observations remains far from human-level.

By framing visual motion as a sequence of interpretable events, our benchmark provides a new foundation for studying causal understanding, planning, and chain-of-thought reasoning in physical domains. We hope this work catalyzes further research into event-centric benchmarks, prompting strategies, and multimodal models that reason more like humans.

7 Division of Work

Jung-Chun Liu created the videos and captions for the benchmark, contributed to project ideation, benchmarked the OpenAI GPT models for the Task 1 video captioning pilot, and benchmarked the LLaVA-Video-7B-Qwen2 model for Tasks 2.1, 2.2, and 3 (event prediction and error correction tasks), and contributed to the project proposal writing and literature survey presentation. She also provided revisions for the final project presentation and write-up.

Eleanor Lin contributed to dataset validation and captioning design and project ideation, benchmarked the Gemini 2.0 Flash model for all tasks (1, 2.1, 2.2, and 3) and all models except LLaVA-Video-7B-Qwen2 for Task 2.1, and contributed to the project proposal writing and literature survey presentation. She prepared the final project presentation and contributed to revising all sections of the final write-up.

Zichen Wang contributed to project ideation, prompted the Claude 3.7 Sonnet model for the Task 1 video captioning pilot, and contributed to the project proposal writing and literature survey presentation. He also wrote the introduction, related work, benchmark, evaluation, discussion, and conclusion sections of the final write-up.

References

[1] K. R. Allen, K. A. Smith, and J. B. Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020.

[2] T. Ates, M. Ateşoğlu, Ç. Yiğit, I. Kesen, M. Kobas, E. Erdem, A. Erdem, T. Goksun, and D. Yuret. CRAFT: A benchmark for causal reasoning about forces and inTeractions. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2602–2627, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Z. Chen, K. Yi, Y. Li, M. Ding, A. Torralba, J. B. Tenenbaum, and C. Gan. Comphy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations*, 2022.

[5] A. Melnik, R. Schiewer, M. Lange, A. I. Muresanu, M. Saeidi, A. Garg, and H. Ritter. Benchmarks for physical reasoning AI. *Transactions on Machine Learning Research*, 2023. Survey Certification.

[6] M. Patel, T. Gokhale, C. Baral, and Y. Yang. CRIPP-VQA: Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[7] C. Xue, V. Pinto, C. Gamage, E. Nikonova, P. Zhang, and J. Renz. Phy-q as a measure for physical reasoning intelligence. *Nature Machine Intelligence*, 5(1):83–93, 2023.

[8] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020.

A Appendix / supplemental material

A.1 Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced: Yes. We describe our methods for creating the dataset and benchmarking models.
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results: Yes.
- Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper: Yes.

Does this paper make theoretical contributions?: No.

Does this paper rely on one or more datasets?: Yes.

If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets: Yes.
- All novel datasets introduced in this paper are included in a data appendix: Yes. Our dataset can be accessed here: `https://drive.google.com/drive/folders/1KMSoE3V0lUY25tackfPz9zyMUiJ4VrTG?usp=sharing`.
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes: Yes.
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations: NA. We only benchmark on our own dataset.
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available: NA. We only benchmark on our own dataset.
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying: NA. We only benchmark on our own dataset.

Does this paper include computational experiments?: Yes

If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix: Yes. Our code is available here: `https://github.com/jungchun1013/CSE-692`.
- All source code required for conducting and analyzing the experiments is included in a code appendix: Yes. Our code is available here: `https://github.com/jungchun1013/CSE-692`.
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes: Yes.
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from: Yes.
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results: NA.
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks: Yes.
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics: Yes.
- This paper states the number of algorithm runs used to compute each reported result: Yes.

- Analysis of experiments goes beyond single-dimensional summaries of performance (*e.g.*, average; median) to include measures of variation, confidence, or other distributional information: Yes.
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (*e.g.*, Wilcoxon signed-rank): Yes.
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments: Yes.
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting: Yes.

A.2 Data Appendix

Our dataset can be accessed here: `https://drive.google.com/drive/folders/1KMSoE3V0lUY25tackfPz9zyMUiJ4VrTG?usp=sharing`.

A.3 Code Appendix

Our code is available here: `https://github.com/jungchun1013/CSE-692`.

A.4 Task 1 Prompt

{video}

Here is the reference description of the previous video clip:

Start
A collision occurs between the support and the platform.
The platform and the ball start touching.
The platform and the ball stop touching.
The ball reaches the goal.
End

Now describe what happens in the next clip. Adhere strictly to the reference format.

{video}

A.5 Task 1 GPT-4 Judge Prompt

Here is a reference description of a video clip:
Description 1:
{description here}

Here is another description of the same video clip:
Description 2:
{description here}

Please rate how well description 2 follows description 1 on a scale from 1 to 5, where 1 means "not at all," 2 means "slightly," 3 means "moderately," 4 means "very," and 5 means "extremely."

Key aspects to consider are how similarly interactions between objects are described and whether or not reaching the goal state is described. The exact naming of objects may differ between the descriptions.

Respond with only your numerical rating. Respond in the format "Answer: "

Answer:

A.6 Task 2.1 Prompt

Based on the initial event shown in the following video, predict whether the red ball will subsequently enter the green goal region. Provide your reasoning before answering. Answer "Success" if the ball will successfully enter the goal. Answer "Failure" if the ball will fail to enter the goal.

Answer in the following format:

"Reasoning: {reasoning}
Answer: {answer}"

{video}

A.7 Task 2.2 Prompt

Based on the initial event shown in the following video, predict whether the red ball will subsequently enter the green goal region. Pay attention to the physical events in the scene. Provide your reasoning before answering. Answer "Success" if the ball will successfully enter the goal. Answer "Failure" if the ball will fail to enter the goal.

Answer in the following format:

"Reasoning: {reasoning}
Answer: {answer}"

{video}

A.8 Task 3 Prompt

Here is your previous reasoning and answer for predicting whether the red ball will subsequently enter the green goal region, based on a partial video clip. "Success" means the ball will successfully enter the goal. "Failure" means the ball will fail to enter the goal.

Reasoning: {previously generated reasoning}
Answer: {previously generated answer}

Please reconsider the attached video which shows the full sequence of events and generate your reasoning and answer again, correcting any errors.

"Reasoning: {reasoning}
Answer: {answer}"

{video}

A.9 Additional Task 2.1 Results

Here, we present the result of prompting VLMs to predict whether a benchmark instance will end in success/failure, based only on an initial image (as opposed to an initial video clip). This setting reflects the capabilities of VLMs such as GPT-4o and o3-mini, which cannot process video input. As expected, predicting from image inputs is more challenging than predicting from video inputs, as demonstrated by the higher overall F1 score for Gemini 2.0 Flash compared to GPT-4o and o3-mini. In the initial video clip, the model has access to information such objects' trajectories and accelerations, whereas this information is absent from a still image.

Table 8: **Physical Event Prediction F1 Scores** Comparison of physical event prediction F1 scores for GPT-4o and o3-mini models on image inputs with Gemini 2.0 Flash model on video inputs.

Model	Simple Cases	General Cases	All Cases
GPT-4o	**0.667**	0.353	0.435
o3-mini	0.500	0.250	0.333
Gemini 2.0 Flash	0.286	**0.545**	**0.483**