InPhyRe Discovers: Large Multimodal Models Struggle in Inductive Physical Reasoning

Gautam Sreekumar *sreekum1@msu.edu*
Department of Computer Science and Engineering, Michigan State University

Vishnu Naresh Boddeti *vishnu@msu.edu*
Department of Computer Science and Engineering, Michigan State University

Reviewed on OpenReview: https://openreview.net/forum?id=TGmridJZQo

Abstract

Large multimodal models (LMMs) encode physical laws observed during training, such as momentum conservation, as *parametric knowledge*. It allows LMMs to answer physical reasoning queries, such as the outcome of a potential collision event from visual input. However, since parametric knowledge includes only the physical laws seen during training, it is insufficient for reasoning in inference scenarios that follow physical laws unseen during training. In such novel physical environments, humans could adapt their physical reasoning based on provided demonstrations. This *inductive physical reasoning* ability is indispensable for LMMs if they are to replace human agents in safety-critical applications. Despite its importance, existing visual benchmarks do not evaluate inductive physical reasoning and only consider the parametric knowledge in LMMs. To this end, we propose InPhyRe, the first visual question answering benchmark to measure inductive physical reasoning in LMMs. InPhyRe evaluates LMMs' ability to predict the outcome of collision events in algorithmically generated synthetic videos. By inspecting over 13 open-source and proprietary LMMs, InPhyRe informs us that **(1)** LMMs struggle to apply their limited parametric knowledge about universal physical laws to reasoning, **(2)** inductive physical reasoning in LMMs is weak when the physical laws underlying inference scenarios were unseen during training, and **(3)** inductive physical reasoning in LMMs suffers from language bias and may ignore the visual inputs, questioning the trustworthiness of LMMs regarding visual inputs.

Project page: https://gautamsreekumar.github.io/project_pages/inphyre/index.html

1 Introduction

Case Study

Premise: A large multimodal model (LMM) is used to determine whether a car crash will occur on a snowy road from a video. To ensure the LMM understands that the physical coefficients of the snowy road differ from those of a dry road, a few demonstration videos of snowy roads with and without collisions are provided as context. The LMM predicts that a crash is unlikely.

Question: Did the LMM account for the unseen physical coefficients using the demonstration videos, or did it use only its parametric physical knowledge to make its prediction?

Large multimodal models (LMMs) are known to encode universal physical laws (*e.g.*, momentum conservation) observed during training as *parametric knowledge* to answer physical reasoning queries (*e.g.*, whether a collision occurs or not) from visual input (Chen et al., 2024d; Cherian et al., 2024; Mudur et al., 2025). However, since parametric knowledge includes only the physical laws seen during training, it is insufficient in scenarios that potentially follow unobserved physical laws and conditions, such as the case study of the snowy

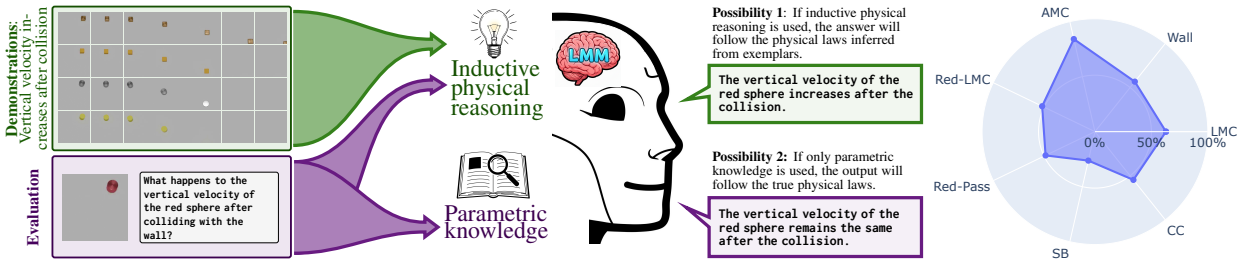


Figure 1: (*Left*) A large multimodal model (LMM) is asked to predict the change in vertical velocity of an object colliding with a vertical wall. The model will output "possibility 2" if it uses its **parametric knowledge** that encodes the universal physical laws (in this case, the momentum conservation principle). However, parametric knowledge would be insufficient if the collision event violated the physical laws encoded in the model. For the model to infer the underlying physical laws, we provide the model with exemplar videos of collisions that violate the momentum conservation principle. The model may now rely on its **inductive physical reasoning** capabilities to generate "possibility 1". (*Right*) InPhyRe shows that LMMs struggle with inductive physical reasoning.

road. Editing the parametric knowledge (Wang et al., 2024d) to suit the scenario is expensive. In cases where the underlying physical laws are hard to enunciate, knowledge editing may not even be feasible. In contrast, humans would easily adapt their physical knowledge about collisions to snowy road conditions with the help of demonstration videos to predict any collision, if presented with the same case study. This crucial skill, which we refer to as ***inductive physical reasoning***[1], is a hallmark of intelligence that humans develop at a very young age (Hayes, 2007; Ricco, 2015). Inductive physical reasoning, where the model infers the underlying physical laws from as few demonstration samples as possible without additional training and applies them for physical reasoning, is an indispensable capability that LMMs must possess in addition to parametric knowledge if they are to be deployed in safety-critical applications such as autonomous driving (Zhou et al., 2024; Zhang, 2025).

Despite its vital nature, there are no visual benchmarks that quantitatively evaluate inductive physical reasoning in LMMs. Existing benchmarks typically do not consider physical scenario changes between training and testing or do not require the evaluated model to infer the physical scenario during test-time (see Tab. 1). For instance, Baradel et al. (2020); Chen et al. (2022); Tung et al. (2023); Chow et al. (2025) evaluate only the parametric knowledge of LMMs, particularly about universal physical concepts observable in natural videos, such as friction and gravity, and not the inductive physical reasoning ability. However, creating inductive physical reasoning benchmarks is challenging.

To evaluate inductive physical reasoning separated from parametric knowledge in LMMs, the benchmark must exclude scenarios seen during LMMs' training. Since we cannot always know which scenarios were seen by LMMs, we opt to build our benchmark using physically impossible scenarios that violate universal physical laws, as they are less likely to have been seen by LMMs during training. However, it is inherently impossible to find such scenarios in natural videos, and prohibitively expensive to manually edit and repurpose existing benchmarks that measure parametric knowledge. Although intuitive physics reasoning benchmarks contain videos that violate true physics, their underlying violations are not articulated, as their goal is not to evaluate how well models can infer the physical laws (Riochet et al., 2021; Bordes et al., 2025). Thus, algorithmically generated synthetic videos become the only viable option. Algorithmic generation of physical event videos violating universal physical laws still requires careful design and manual interventions on environments, physical laws, and object trajectories.

We propose **InPhyRe** (Inductive Physical Reasoning), the ***first*** visual question answering benchmark that evaluates how well LMMs can infer the underlying physics from demonstration samples and use it to make physical reasoning predictions. InPhyRe comprises algorithmically generated synthetic videos of collision events. During evaluation, LMMs are given the first frame of a video and asked questions about the outcome of a textually described collision involving the objects in this frame. However, some of these scenarios violate universal physical laws. In these scenarios, where parametric knowledge is ineffective, LMMs must infer

[1]Inductive physical reasoning is closer to inductive reasoning than general visual reasoning. See § D.

the underlying physics from demonstration videos taken from the same scenarios for physical reasoning. InPhyRe quantifies inductive physical reasoning in LMMs as the performance disparity between scenarios that follow the true physical laws and those that do not. InPhyRe's goal is not to evaluate the utility of LMMs in unrealistic scenarios that violate universal physical laws, but rather to evaluate their adaptability in scenarios with unseen physical laws. This goal also differs from that of "intuitive physics understanding," where violation-of-expectation is used to gauge the parametric knowledge (see § C for more details).

Scope of InPhyRe: Since it is difficult to create a benchmark visualizing the violations of every universal physical law exhaustively, we limit our studies to the laws of mechanics, such as momentum and energy conservation principles. Nonetheless, the outcomes of the collision events in InPhyRe violate the most fundamental laws of mechanics. Therefore, we believe that the conclusions from InPhyRe are likely valid for inductive physical reasoning in LMMs about other branches of physics.

What does InPhyRe find? Our results show that InPhyRe is a formidable benchmark for current LMMs, despite its visual simplicity. Our results indicate that both **open-source and proprietary LMMs struggle to infer and utilize unseen physical laws from demonstration samples** (Fig. 1). From empirical evidence, we conjecture that LMMs do not understand physical laws as transferable mathematical models, but rather as a fixed set of rules that all objects obey. Moreover, we show that LMMs primarily derive their inductive physical reasoning capabilities from their language components and ignore the visual inputs in the demonstration samples. Chain-of-thought prompting and fine-tuning are also futile for conditional physical reasoning tasks in InPhyRe (§§ E.2 and E.3). This means that it is possible that the LMM in our case study did not account for the snowy conditions.

Contributions and Findings

- We introduce InPhyRe, the ***first*** visual question answering benchmark to evaluate the inductive physical reasoning capabilities of large multimodal models (LMMs).
- Finding 1: LMMs have only limited parametric knowledge about universal physical laws and struggle to apply these laws even in scenarios that follow these physical laws (§ 4.2).
- Finding 2: Demonstration samples improve LMMs' predictions only when the samples agree with the models' parametric knowledge, resulting in poor inductive physical reasoning on scenarios that violate the true physical laws (§§ 4.3, 4.4 and 5).
- Finding 3: Inductive physical reasoning in LMMs suffers from strong language bias and prefers textual reasoning. Many LMMs struggle to infer visual information and relate it to the query when it is necessary (§ 4.5).

2 Related works

Our research question of inductive physical reasoning differs from the existing physical reasoning tasks, which we list here and then collectively distinguish inductive physical reasoning from them. Commonsense physical reasoning evaluates LLMs' ability to reason and provide instructions for everyday tasks such as picking up objects or cutting fruits (Bisk et al., 2020; Aroca-Ouellette et al., 2021; Wang et al., 2023). Intuitive physics understanding evaluates parametric knowledge of models through their ability to detect, but not articulate or understand, violations of universal physical laws (Riochet et al., 2021; Garrido et al., 2025). Counterfactual physical reasoning evaluates the ability to reason about alternative outcomes for a given scenario and change under the same physical laws (Yi et al., 2020; Baradel et al., 2020; Chen et al., 2022). Other physical reasoning tasks focus on the theoretical physics knowledge in LLMs (Pang et al., 2025; Mudur et al., 2025; Yu et al., 2025) and their ability to reason about the latent physical properties (Chen et al., 2024d; Chow et al., 2025), sometimes using interactive simulators (Cherian et al., 2024). Synthetic collision events are also commonly used in physical reasoning benchmarks (Yi et al., 2020; Baradel et al., 2020; Chen et al., 2022). The above-mentioned tasks either do not include test scenarios that differ from training scenarios in the underlying physics or do not require the model to infer the underlying physics if such test scenarios with unseen physics exist. In essence, these tasks only evaluate the parametric knowledge of LMMs.

Benchmark	What are the physical reasoning tasks	Physical conditions change between training and testing?	Require test-time inference of physical conditions?
CLEVRER [1]	Factual and counterfactual physical reasoning	No ✗	No ✗
ComPhy [2]	Physical reasoning requiring latent property prediction	No ✗	Yes ✓
CoPhy [3]	Counterfactual physical reasoning	No ✗	No ✗
PhysBench [4]	Physical reasoning about object properties and dynamics	No ✗	No ✗
IntPhys [5]	Physical plausibility prediction	Yes ✓	No ✗
Physion [6]	Object contact prediction	No ✗	No ✗
Physion++ [7]	Object contact prediction involving property prediction	No ✗	No ✗
ContPhy [8]	Physical property and dynamics prediction	No ✗	No ✗
InPhyRe	Infer physical laws from demo samples and apply them.	**Yes ✓**	**Yes ✓**

Table 1: **How does InPhyRe differ from prior physical reasoning benchmarks?** Unlike prior benchmarks, the test scenarios in InPhyRe follow different physical laws compared to the training set and require the evaluated model to infer these unseen physical laws from demonstration samples. [1] (Yi et al., 2020), [2] (Chen et al., 2022), [3] (Baradel et al., 2020), [4] (Chow et al., 2025), [5] (Riochet et al., 2021; Bordes et al., 2025), [6] (Bear et al., 2021), [7] (Tung et al., 2023), [8] (Zheng et al., 2024)

Existing benchmarks cannot be repurposed to evaluate inductive physical reasoning. Most physical reasoning benchmarks use only true physical laws that are possibly encoded in the parametric knowledge of LMMs, and thus may confound the evaluation of inductive physical reasoning. **Example:** Counterfactual physical reasoning benchmarks do not verify if the model prediction used its memorized physical laws/conditions, or was based on the provided factual scenario. In contrast, by relying on impossible physics, we ensure that LMMs can only answer the queries using the physics dynamics inferred from demonstration samples. Although intuitive physics understanding benchmarks comprise impossible physical scenarios, they do not articulate the corresponding underlying physical laws, as their goal is to evaluate only whether the model can detect a violation of physics, and not whether they can infer the new physical laws. The key differences between InPhyRe and some prior physical reasoning benchmarks are shown in Tab. 1. See § C for more discussion on related works.

3 InPhyRe: Inductive Physical Reasoning Benchmark

In this section, we will describe our proposed benchmark, InPhyRe – Inductive Physical Reasoning and how we will utilize it to evaluate the inductive physical reasoning capabilities of large multimodal models (LMMs). Our benchmark comprises collision event videos that violate real-world physical laws, such as momentum conservation and object continuity. During evaluation, the inputs to the LMM are the first frame from a collision video and a multiple-choice question about the outcome of a described collision event. The model is then evaluated in zero-shot and few-shot settings to quantify its relative strengths of parametric knowledge and inductive physical reasoning. **The dataset can be found here.**

Research Question: We will first explain the key prerequisite concepts and then state the research question addressed by InPhyRe. As mentioned before, each evaluation sample consists of the first frame of a collision video and a multiple-choice question about the collision's outcome. The model may answer this question using its implicit knowledge of the physical laws that it obtained during training. We refer to this as physical reasoning using parametric knowledge. However, if the physical reasoning required to correctly answer the question does not follow the physical laws embedded in the model, parametric knowledge is insufficient. To let the model infer the physical laws required to answer the question, we provide demonstration samples[2] containing videos of similar events governed by the same physical laws required to answer the question. We refer to this ability to infer the physical laws from exemplars and answer the question as inductive physical reasoning. InPhyRe is designed to answer the following research question about physical reasoning in LMMs:

[2]Henceforth referred to as "exemplars."

(RQ) Can LMMs flexibly switch between parametric knowledge and inductive physical reasoning by comparing the physical laws in the exemplars to those encoded in the models' parameters?

To answer **(RQ)**, InPhyRe evaluates LMMs in scenarios that follow the true physical laws and those that do not. An LMM with strong inductive physical reasoning will perform identically in both situations, as parametric knowledge and inductive physical reasoning are not competing qualities.

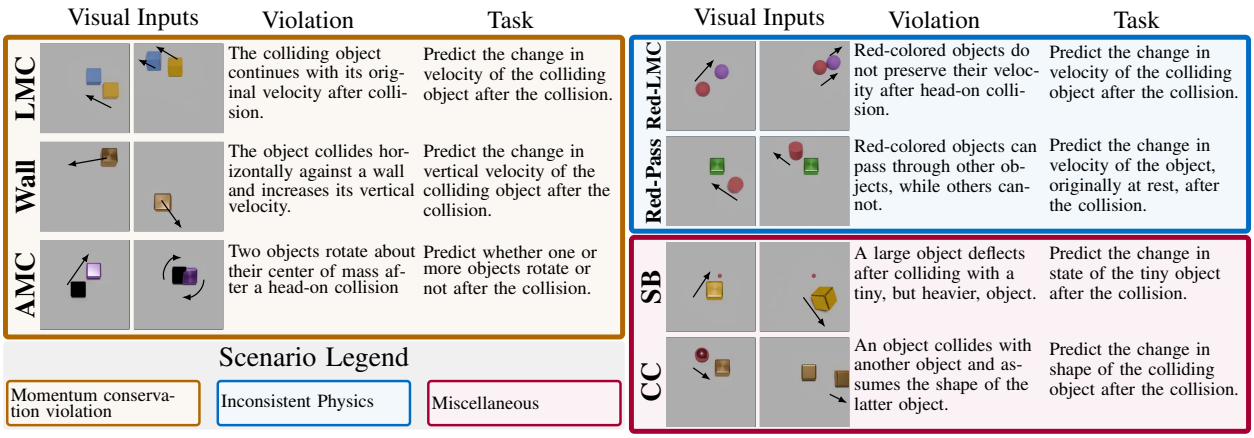


Figure 2: **InPhyRe** comprises videos ("visual inputs") of collision events that violate a real-world physical law ("violation"). LMMs must predict state changes in objects due to the collisions, while accounting for the violated physical law ("task"). The videos are grouped into "scenarios", which are further grouped into three categories based on the nature of physical law they violate. Arrows indicate object motion and are not part of the actual images in the dataset.

Task Description: InPhyRe is a visual question answering benchmark for physical reasoning. Each evaluation sample includes an image of a scene with one or more objects and a question about the outcome of a described collision event involving the objects in that image. Similar to (Johnson et al., 2017; Yi et al., 2020; Chen et al., 2022), the objects are primitive shapes (cubes, cylinders, and spheres) of various colors and textures lying on a plain surface (Fig. 2). The question is about the change in the state of an object after the collision. *E.g.*, `What happens to the velocity of the red cube after colliding with the blue sphere?` To answer the question, we provide four options as possible answers to the model. E.g., "`A. red cube's velocity increases, B. red cube's velocity decreases, C. cannot be determined, D. no change in velocity.`" The questions may also contain additional information about the objects or the collision event. E.g., "`red cube and yellow cylinder have equal mass`" or "`blue sphere collides elastically against the wall.`" The option chosen by the model is parsed from its generation output. See §§ B.2 and B.3 for more details about prompting and parsing, respectively.

Irregular Scenarios: The collision event videos in InPhyRe are grouped into "scenarios." Each scenario is characterized by the true physical law that it violates. We call them "irregular scenarios" and denote them with their shorthand notations shown in Fig. 2. The irregular scenarios are further grouped into three categories based on the nature of the violated physical laws (see Fig. 2):

(1) Scenarios in the momentum conservation violation category evaluate the inductive physical reasoning of LMMs when the principle of momentum conservation is violated. It comprises three scenarios: linear momentum conservation (LMC), angular momentum conservation (AMC), and directional linear momentum conservation (Wall). In LMC, a moving object collides elastically with an object of equal mass at rest, and, instead of losing its momentum, continues with the same velocity. A similar collision event occurs in AMC, except the objects rotate about their center of mass, despite the collision being head-on, violating the principle of angular momentum conservation. In Wall, the vertical velocity of an object increases after colliding with a vertical wall, violating the linear momentum conservation principle, but only along the vertical direction.

(2) In the inconsistent physics category, objects with certain visual properties follow physical laws different from other objects. In the real world, these visual properties would not have affected the modified physical laws. The objective is to examine whether LMMs can logically combine parametric knowledge and inductive physical reasoning based on the object's visual properties. For this category, few-shot evaluations will include exemplars with both sets of physical laws. This category includes two scenarios: (i) Red-LMC, where red-colored objects violate linear momentum conservation, and (ii) Red-Pass, where only red-colored objects can physically pass through other objects.

(3) InPhyRe also includes some miscellaneous scenarios to evaluate whether LMMs have a visually biased perception of physical laws. For instance, in size-bias (SB), a dimensionally large object deflects after colliding with a dimensionally small, but much heavier, object. We include information about mass in the question. SB evaluates whether LMMs conflate the concepts of volume and mass. In color-constancy (CC), a moving object collides with an object at rest and assumes the visual appearance of the latter object. This collision obeys linear momentum conservation. CC evaluates the object permanence of LMMs as it may seem to the model that the colliding object has disappeared.

Regular Scenarios: For each irregular scenario, there is a corresponding "regular" scenario depicting similar collisions while following the true physical laws. Since multiple types of violations are possible for every universal physical law, regular scenarios are fewer than irregular scenarios. InPhyRe includes regular versions of LMC, SB, and AMC that act as real-world counterparts of various irregular scenarios. These regular scenarios are used to evaluate the parametric knowledge of LMMs. InPhyRe uses both regular and irregular scenarios jointly to answer **(RQ)**.

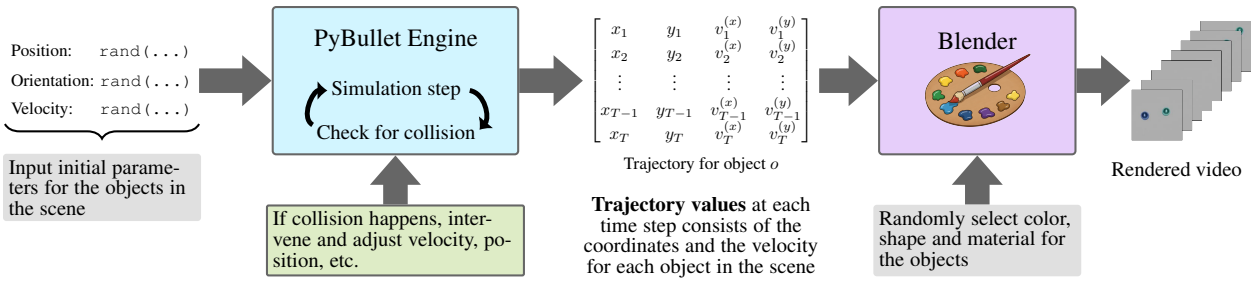


Figure 3: We initialize the object states in PyBullet. When a collision occurs during the simulation, we intervene and manually adjust the objects' states such that the resulting trajectory violates some real-world physical law. The object trajectories are then used by Blender to render the final video.

Video generation details: We modify the video generation pipeline from (Yi et al., 2020) to generate synthetic videos using PyBullet (Coumans & Bai, 2016–2021) and Blender (Community, 2025). First, we define objects and their properties such as mass and lateral friction. Then we randomly initialize their state variables, such as position and velocity, in a PyBullet environment. The trajectories are obtained by running the simulation. However, unlike in (Yi et al., 2020), our trajectories are governed by custom physical laws that differ from real-world physical laws. To simulate these custom physical laws, we intervene when collisions occur during simulation and manually adjust states of objects, such as linear/angular velocity and direction. These trajectories are then used by Blender to render the final video. Visual object properties such as color and texture are randomly chosen. See Fig. 3.

Question generation details: Each scenario comprises around 2000 samples. Exemplars for each evaluation sample are randomly chosen from the rest, and the choices are affected by the random seed. Question-answer pairs are generated for each sample from pre-defined templates. Since each scenario concerns a particular query (e.g., change in the vertical velocity in Wall), we use multiple templates for the questions and the answer options to avoid lexical repetition. For instance, in Wall, the templates for the question are {"What happens to the vertical velocity of `<obj>` when it collides with wall?", "What is the outcome of `<obj>` colliding with wall?", "What occurs to the vertical velocity of `<obj>` when `<obj>` and wall collide?"}. A similar multi-template approach is used to generate answers. The answer options are also shuffled for each sample so that the model may not simply repeat the answer options from the exemplars.

4 What does InPhyRe Discover about Physical Reasoning in LMMs?

Before we examine the physical reasoning abilities of LMMs, we will describe the evaluation setup and codify the procedure to answer specific queries about physical reasoning using InPhyRe.

4.1 Evaluation Setup

Evaluated LMMs: We use InPhyRe to evaluate the quality of physical reasoning (both parametric and inductive) in a diverse cohort of LMMs. To represent the variety of choices in model design and training datasets, we include **13 open-source LMMs**: LLaVA-NeXT-Video (Zhang et al., 2024), LLaVA-OneVision (Li et al., 2024a), LLaVA-NeXT-Interleave (Li et al., 2024c), Gemma 3 (Kamath et al., 2025) herd, Aria (Li et al., 2024b), VideoLLaMA3 (Zhang et al., 2025) herd, InternVL3 (Zhu et al., 2025) herd, Qwen2-VL (Wang et al., 2024a), and Qwen2.5-Omni (Xu et al., 2025). The chosen models use different vision encoders and language models, and cover an extensive parameter count range (from 1B to nearly 25B). The image encoders in the chosen LMMs are pre-trained and then fine-tuned. The majority of the chosen LMMs fine-tune separately trained LLMs, while Gemma and Aria train their language models from scratch. Aria uses a mixture-of-experts (MoE) architecture. All evaluations in this section were run with three random seeds. The statistical reliability of our findings is measured using the standard deviation of accuracy over different runs with these random seeds (see Fig. 8). More details of the LMMs are listed in Tab. 4. **Results on closed models** such as GPT-4 and Gemini are in § 5.

Evaluation tools: InPhyRe contains regular scenarios that follow real-world physical laws and irregular scenarios that violate one or more real-world physical laws. For each LMM, we conduct zero-shot evaluation in regular scenarios and few-shot evaluations in both regular and irregular scenarios. The few-shot setting is further categorized into two sub-settings: (i) "visual-text", where the exemplars contain collision videos along with a question-answer pair, and (ii) "visual-only", where the exemplars include only videos. In all settings, the evaluation metric is the model's accuracy in choosing the correct option for the multiple-choice question. Below, we enunciate our specific questions about physical reasoning in LMMs and how to quantitatively answer them using InPhyRe:

- § 4.2 *How much parametric knowledge do LMMs have about universal physical laws?* We answer this using the zero-shot predictive accuracy of the model in regular scenarios that follow universal physical laws.
- § 4.3 *Can LMMs augment their parametric knowledge with exemplars?* To answer this question, we compare the few-shot performance of the model in regular scenarios with the zero-shot accuracy. Here, exemplars are taken from the evaluated regular scenario.
- § 4.4 *How strong is inductive physical reasoning in LMMs?* Inductive physical reasoning in LMMs is evaluated by comparing their few-shot performances in regular and irregular scenarios.
- § 4.5 *How much of this inductive physical reasoning is aided by language?* We answer this question using three experiments that measure the LMMs' preference for textual reasoning and their ability for visual reasoning when necessary.

4.2 How much parametric knowledge do LMMs have about universal physics?

InPhyRe includes regular versions of LMC, SB, and Wall that obey universal physical laws of mechanics, such as principles of momentum and energy conservation. We measure the parametric knowledge about these universal laws in LMMs as their zero-shot accuracy in these regular scenarios.

Tab. 2 shows the zero-shot accuracy of LMMs in regular scenarios. In each scenario, the task was to predict the change in velocity of an object after colliding with another object at rest from the initial image frame of the collision video. Surprisingly, **many LMMs struggle to answer even these simple questions using the momentum conservation principle**, and the models achieve above 80% accuracy in only 6 out of 39 scenarios, mainly in SB (Reg.). The performances of LMMs also vary greatly between scenarios, despite these scenarios following the same physical laws, *e.g.*, most models performed poorly in Wall (Reg.) compared to other scenarios. Qualitative inspection of their outputs in § E.5 reveals that LMMs can state

LMM	LMC (Regular)		SB (Regular)		Wall (Regular)		Average	
	Zero-shot	3-shot	Zero-shot	3-shot	Zero-shot	3-shot	Zero-shot	3-shot
IVL3-1B (Zhu et al., 2025)	71.33	34.77 (-36.57)	86.98	44.89 (-42.08)	5.67	8.10 (+ 2.43)	54.66	29.25 (-25.41)
VL3-2B (Zhang et al., 2025)	56.47	64.25 (+ 7.78)	52.99	62.76 (+ 9.78)	3.67	40.98 (+37.32)	37.71	56.00 (+18.29)
IVL3-2B (Zhu et al., 2025)	77.73	88.35 (+10.62)	66.35	73.14 (+ 6.80)	29.33	88.18 (+58.85)	57.80	83.23 (+25.42)
Gem3-4B (Kamath et al., 2025)	20.00	70.70 (+50.70)	57.85	71.43 (+13.58)	78.12	72.88 (- 5.23)	51.99	71.67 (+19.68)
LLaVA-NV (Zhang et al., 2024)	50.22	58.40 (+ 8.18)	15.95	33.54 (+17.58)	2.33	8.77 (+ 6.43)	22.83	33.57 (+10.73)
IVL3-8B (Zhu et al., 2025)	61.68	99.87 (+38.18)	94.21	99.88 (+ 5.67)	56.72	96.17 (+39.45)	70.87	98.64 (+27.77)
LLaVA-OV (Li et al., 2024a)	75.85	99.15 (+23.30)	83.81	99.11 (+15.30)	4.52	99.35 (+94.83)	54.73	99.20 (+44.48)
VL3-7B (Zhang et al., 2025)	77.52	98.30 (+20.78)	69.29	94.78 (+25.49)	7.22	48.70 (+41.48)	51.34	80.59 (+29.25)
LLaVA-NIL (Li et al., 2024c)	83.93	97.80 (+13.87)	64.92	70.60 (+ 5.69)	0.68	89.98 (+89.30)	49.84	86.13 (+36.28)
Qwen2-VL (Wang et al., 2024a)	67.18	97.57 (+30.38)	72.89	99.01 (+26.11)	6.78	98.67 (+91.88)	48.95	98.41 (+49.46)
Qwen2.5-O (Xu et al., 2025)	56.98	35.77 (-21.22)	94.62	83.11 (-11.51)	5.08	52.05 (+46.97)	52.23	56.97 (+ 4.75)
Gem3-12B (Kamath et al., 2025)	35.90	87.65 (+51.75)	85.46	83.41 (- 2.05)	68.80	72.65 (+ 3.85)	63.39	81.24 (+17.85)
Aria (Li et al., 2024b)	39.97	56.70 (+16.73)	78.18	91.40 (+13.23)	35.37	81.98 (+46.62)	51.17	76.69 (+25.53)

Table 2: Zero-shot and 3-shot evaluation results on regular scenarios.

universal physical laws (*e.g.*, "kinetic energy is conserved in an elastic collision") but struggle to apply them for physical reasoning. They also hallucinate irrelevant assumptions (*e.g.*, about material) and incorrect physical laws that further hurt their reasoning. We conclude that LMMs memorize the laws of mechanics and can recollect them as factual information, but fail to apply this knowledge for physical reasoning. A similar conclusion was made in (Yu et al., 2025), but for abstract physical reasoning.

Conclusion

LMMs have limited parametric knowledge about the laws of mechanics. Although LMMs can state these universal laws, they often struggle to apply them for physical reasoning.

4.3 Can LMMs augment their parametric knowledge with exemplars?

Before evaluating the inductive physical reasoning of LMMs in irregular scenarios, we must verify that exemplars can improve physical reasoning in LMMs. Therefore, we evaluate LMMs in regular scenarios in the few-shot setting with exemplars that do not contradict any universal physical laws encoded in the models' parameters. Specifically, we consider the "visual-text" setting, where question-answer pairs accompany videos in exemplars. Then we compare the few-shot performance of LMMs in regular scenarios with their corresponding zero-shot performance.

Tab. 2 compares the 3-shot performance of LMMs in regular scenarios with their corresponding zero-shot performance. **All LMMs significantly improved their performance when provided with exemplars in at least one scenario.** On average, we observe that all models except InternVL3-1B improved their performance with exemplars. Among the LMMs evaluated, Qwen2-VL achieved the highest average increase in performance with exemplars. LLaVA-Onevision, Qwen2-VL, and InternVL3-8B also achieved nearly 100% average accuracy over all scenarios. These results clearly demonstrate that LMMs can use exemplars to improve their prediction accuracy.

Conclusion

Exemplars that obeyed universal physical laws support parametric knowledge in LMMs successfully. With only three exemplars, several LMMs achieve nearly 100% prediction accuracy.

4.4 How strong is inductive physical reasoning in LMMs?

We established that exemplars that obey universal physical laws improve the performance of LMMs in regular scenarios. We will now evaluate whether LMMs can leverage exemplars that do not follow the true physical laws to reason in irregular scenarios. Following the previous experiments, the task is to predict the outcome of a potential collision event from an image with the help of exemplar videos from the same scenario. Exemplars also include question-answer pairs. However, these collision events (and the provided exemplars) violate the true physical laws. Therefore, to correctly answer the questions, LMMs must infer the underlying physical laws from exemplars through inductive physical reasoning.

	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC	Average over scenarios
InternVL3-1B	45.55	18.73	79.90	51.98	55.12	9.05	41.12	43.06
VideoLLaMA3-2B	34.58	29.32	63.42	33.83	43.15	7.72	45.40	36.77
InternVL3-2B	17.97	1.07	97.90	17.37	24.90	0.18	6.75	23.73
Gemma 3-4B	81.33	50.17	62.80	64.88	66.13	26.70	36.00	55.43
LLaVA-NeXT-Vid	16.82	37.80	48.93	11.62	12.88	19.08	32.13	25.61
InternVL3-8B	93.47	99.60	100.00	69.47	60.02	45.83	68.42	76.69
LLaVA-OneVision	96.65	97.85	99.50	67.87	66.13	28.72	64.95	74.52
VideoLLaMA3-7B	83.37	82.65	96.08	43.17	44.88	61.17	67.25	68.37
LLaVA-NeXT-IL	91.80	95.90	94.42	71.62	50.55	42.33	97.65	77.75
Qwen2-VL	94.60	99.78	96.97	97.53	92.58	15.07	87.22	83.39
Qwen2.5-Omni	70.35	66.67	91.03	58.53	52.95	58.25	42.22	62.86
Gemma 3-12B	62.38	89.30	94.32	23.05	22.43	16.47	86.00	56.28
Aria	30.15	1.55	74.10	31.35	21.40	9.77	9.73	25.44
Average over LMMs	63.00	59.26	84.57	49.41	47.16	26.18	52.68	

(a) Predictive accuracy in irregular scenarios

	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC	Average over scenarios
InternVL3-1B	-25.78	10.63	8.57	-19.35	-16.22	-77.93	-30.22	-21.47
VideoLLaMA3-2B	-29.67	-11.67	-0.83	-30.42	-21.10	-55.05	-18.85	-23.94
InternVL3-2B	-70.38	-87.12	9.55	-70.98	-63.45	-72.96	-81.60	-62.42
Gemma 3-4B	10.63	-27.95	-7.90	-5.82	-4.57	-44.73	-34.70	-16.43
LLaVA-NeXT-Vid	-41.58	29.03	-9.47	-46.78	-45.52	-14.45	-26.27	-22.15
InternVL3-8B	-6.40	3.43	0.13	-30.40	-39.85	-54.05	-31.45	-22.65
LLaVA-OneVision	-2.50	-1.50	0.35	-31.28	-33.02	-70.39	-34.20	-24.65
VideoLLaMA3-7B	-14.93	33.95	-2.22	-55.13	-53.42	-33.62	-31.05	-22.35
LLaVA-NeXT-IL	-6.00	5.92	-3.38	-26.18	-47.25	-28.27	-0.15	-15.05
Qwen2-VL	-2.97	1.12	-0.60	-0.03	-4.98	-83.94	-10.35	-14.54
Qwen2.5-Omni	13.37	14.62	34.05	1.55	-4.03	-36.37	-14.77	1.20
Gemma 3-12B	-25.27	16.65	6.67	-64.60	-65.22	-69.00	-1.65	-28.92
Aria	-26.55	-80.43	17.40	-25.35	-35.30	-81.63	-46.97	-39.83
Average over LMMs	-17.54	-7.18	4.02	-31.14	-33.38	-55.57	-27.86	

(b) Difference in accuracy between irregular and regular scenarios

Figure 4: Performance of LMMs in irregular scenarios when exemplars contain both videos and QA pairs.

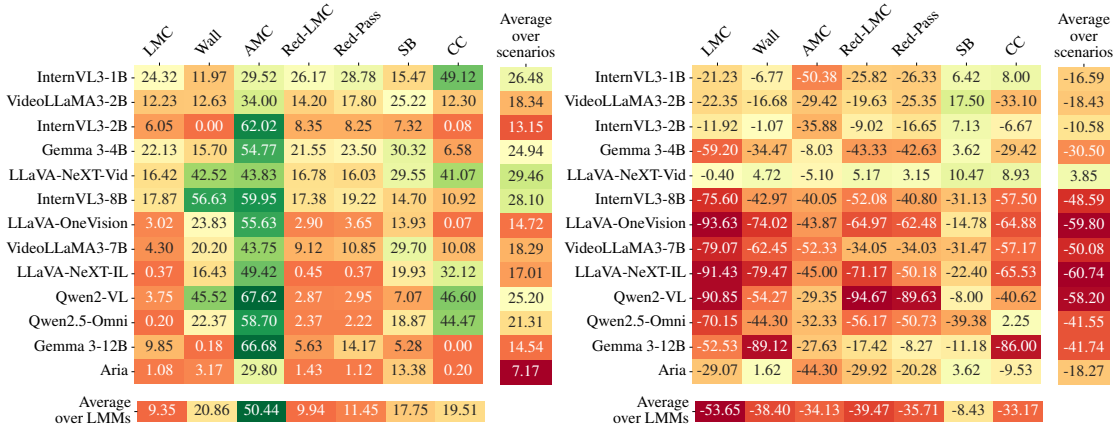
Fig. 4b compares the 3-shot accuracy of each combination of model and irregular scenario against that model's best performance among zero-shot and few-shot evaluations in the corresponding regular scenario that depicts similar events while following universal physical laws. For SB and Wall, the corresponding regular scenarios are SB (Reg.) and Wall (Reg.), respectively. For all other irregular scenarios, LMC (Reg.) is the corresponding regular scenario. A negative value means poor inductive physical reasoning in that scenario. In Fig. 4b, we observe that **most models show a drop in accuracy compared to regular scenarios, indicating weak inductive physical reasoning**. However, the drop in accuracy varies with the models. InternVL3-2B has the highest average drop in accuracy. Performance deterioration also varies with the scenario. Almost all models suffer a considerable drop in accuracy in SB, indicating that LMMs struggle to differentiate between volume and mass. Surprisingly, several LMMs performed better in AMC than LMC (Reg.). However, as we show in § E.6, this apparent inductive physical reasoning is due to LMMs possessing the wrong parametric knowledge about angular momentum conservation, thus inadvertently performing well in irregular scenarios. This incorrect parametric knowledge also allows the LMMs to perform well in AMC when only videos are included in the demonstration samples. The absolute accuracy values are shown in Fig. 4a and also provided in § B.4. Fig. 4a shows the larger LMMs achieve high accuracy in LMC and Wall categories as well. In the next section, we will show that this inductive physical reasoning is guided by the QA pairs in the demonstration samples, rather than by visual understanding of the underlying physics.

We note that Qwen2-VL performed the best among the evaluated LMMs, achieving $> 90\%$ accuracy in nearly all irregular scenarios, except SB. As a reminder, in SB, a dimensionally large object with low mass collides with a dimensionally small object with a much higher mass and fails to displace the dimensionally smaller object, in accordance with the true momentum conservation principle when the relative difference between the objects' mass is large. We include the information about the relative mass of the colliding objects in SB evaluation queries. E.g., "<obj 1> has more mass than <obj 2>." Although our prompt phrasing does not explicitly specify the magnitude of the mass difference, this ambiguity is intentionally resolvable through the demonstration samples. Based on the results, Qwen2-VL's failure suggests a breakdown in inductive inference. The general drop in performance across LMMs in the SB scenario is also evidence that the observed drop in accuracy in InPhyRe cannot be attributed to the poor physics prior that LMMs have over the irregular scenarios, as the SB scenario does not violate any physical laws.

Conclusion

LMMs demonstrate only weak inductive physical reasoning when exemplars violate parametric knowledge. Almost all LMMs showed significant deterioration in performance.

4.5 How much of this inductive physical reasoning is aided by language?



(a) Predictive accuracy in video-only setting

(b) Difference in accuracy between video-only and video-text settings

Figure 5: Performance of LMMs in irregular scenarios when exemplars contain only videos.

In our previous experiments, exemplars included both videos and question-answer pairs. However, multimodal models are well known to exploit their language bias in visual question answering (VQA) tasks (Goyal et al., 2017). This raises the possibility that the observed inductive physical reasoning, limited as it may be, originated primarily from the language component of LMMs. The existence of language bias can undermine the trustworthiness of LMMs on visual inputs. To detect language bias, we devise three experiments to measure different aspects of language bias.

Experiment 1: We repeat our few-shot experiments from § 4.4 but with exemplars that contain *only* the collision videos. Through this experiment, we verify whether LMMs can visually infer the physical laws from demonstration samples without explicit textual guidance in the form of question-answer pairs. Following (Min et al., 2022), we include randomly chosen options in the exemplars to guide the model to choose an option instead of open-ended reasoning (see § B.2).

Fig. 5b shows the difference between performances under video-only and video-text settings. We find that the inductive physical reasoning in LMMs, unfortunately, arises largely from the language components of LMMs. When the exemplars contained only videos, almost all LMMs showed a significant drop in accuracy compared to their performance when the exemplars contained both videos and question-answer pairs. In certain scenarios, LMMs could achieve only near-zero accuracy, *e.g.*, Qwen2.5-Omni in LMC and LLaVA-Next-Interleave in Red-LMC and Red-Pass. The accuracy drop was higher for larger models. Our findings echo the recent evidence of language bias in LMMs reported by (Baldassini et al., 2024; Chen et al., 2025), although their findings were not about physical reasoning in LMMs.

Experiment 2: In CC, one object assumes the color and the shape of the other object after collision. In this experiment, we modify the CC scenario to omit the target color and shape required to answer the query from the question. For example, an original question "what happens to the black cube after collision with the blue sphere?" becomes "what happens to the black cube after collision with the other object?" The options will include both "changes to blue" and, say, "changes to red." To choose the right option, LMMs must infer this information and its relation to the query from the visual inputs. Through this experiment, we verify whether LMMs can obtain visual information and relate it to the query if necessary. In Fig. 6a, we compare LMMs' accuracy in modified CC against those in § 4.4. LMMs that can extract the required visual information and its relation to the query will not show any accuracy drop. We find that 7 out of 13 LMMs show a significant drop (> 5 pp) in their accuracy. Only four LMMs – Gemma3-4B, LLaVA-Next-Video, InternVL3-8B, and Qwen2.5-Omni – maintain their accuracy. The remaining two achieved less than random chance accuracy in both scenarios.

LMM	CC	CC (Mod.)	Diff.
InternVL3-1B (Zhu et al., 2025)	41.12	28.20	-12.92
VideoLLaMA3-2B (Zhang et al., 2025)	45.40	15.23	-30.17
InternVL3-2B (Zhu et al., 2025)	6.75	26.50	19.75
Gemma 3-4B (Kamath et al., 2025)	36.00	39.18	3.18
LLaVA-NeXT-Vid (Zhang et al., 2024)	32.13	28.77	-3.37
InternVL3-8B (Zhu et al., 2025)	68.42	63.30	-5.12
LLaVA-OneVision (Li et al., 2024a)	64.95	18.92	-46.03
VideoLLaMA3-7B (Zhang et al., 2025)	67.25	71.92	4.67
LLaVA-NeXT-IL (Li et al., 2024c)	97.65	66.85	-30.80
Qwen2-VL (Wang et al., 2024a)	87.22	79.15	-8.07
Qwen2.5-Omni (Xu et al., 2025)	42.22	42.23	0.02
Gemma 3-12B (Kamath et al., 2025)	86.00	70.17	-15.83
Aria (Li et al., 2024b)	9.73	10.12	0.38

(a) Results for experiment 2

	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC	Average over scenarios
InternVL3-1B	-0.07	0.37	-0.05	-0.50	0.23	0.00	-0.37	-0.05
VideoLLaMA3-2B	-0.37	1.07	-0.83	-0.38	-0.72	0.03	-0.03	-0.18
InternVL3-2B	0.25	0.32	0.28	0.20	-0.62	0.08	0.80	0.19
Gemma 3-4B	-0.07	-0.35	1.33	-0.70	-0.32	0.80	0.42	0.16
LLaVA-NeXT-Vid	-0.25	-0.25	0.17	0.07	0.40	0.78	0.70	0.23
InternVL3-8B	0.12	0.25	0.00	-1.20	0.53	-0.08	1.82	0.20
LLaVA-OneVision	0.30	-0.12	0.13	-1.13	-0.52	-0.43	-0.10	-0.27
VideoLLaMA3-7B	-0.07	-0.73	-0.13	-0.13	0.00	0.73	-0.05	-0.05
LLaVA-NeXT-IL	0.15	0.07	0.62	-0.27	0.25	-1.25	-0.43	-0.12
Qwen2-VL	0.88	0.02	0.10	-0.15	-0.73	0.43	-0.35	0.03
Qwen2.5-Omni	-1.85	0.78	-0.58	-1.20	-2.00	-0.72	-0.42	-0.85
Gemma 3-12B	-1.25	0.45	0.33	-0.42	-1.12	-0.23	-1.28	-0.50
Aria	-0.77	-0.28	-0.37	0.60	-0.05	-0.08	0.87	-0.01
Average over LMMs	-0.23	0.12	0.08	-0.40	-0.36	0.01	0.12	

(b) Results for experiment 3

Figure 6: (left) Comparing the performance of LMMs in CC when it is necessary for them to extract visual information and relations to answer the queries. Accuracy drops indicate a lower ability to extract and align visual information with queries. (right) Change in accuracy when random mismatched videos accompany question-answer pairs in demonstration samples

Experiment 3: We use arbitrary frames from randomly chosen videos to accompany question-answer pairs in demonstration samples. Here, we check whether LMMs rely on shortcuts when the answer can be obtained through pattern-matching over the textual parts of the prompt alone.

In Fig. 6b, we show the change in LMMs' accuracy w.r.t. their corresponding accuracy in irregular scenarios in § 4.4. A trivial change in accuracy indicates the model used textual shortcuts. We note that the accuracy change is less than 1 pp in 86.8% of combinations, indicating shortcut learning for inductive physical reasoning.

Conclusion

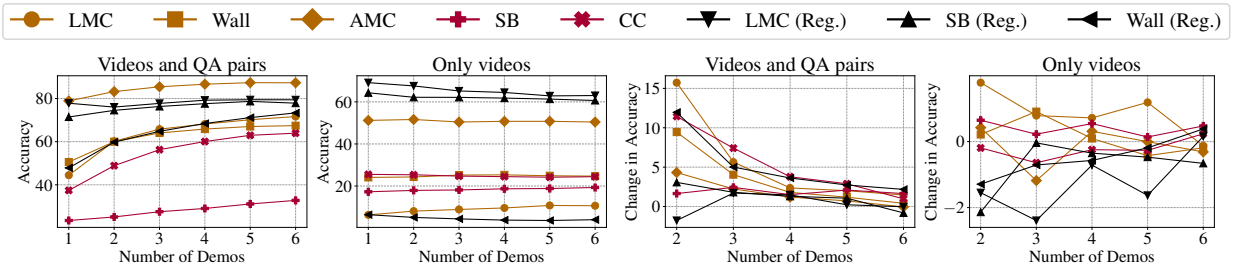
Inductive physical reasoning in the evaluated LMMs has a strong language bias, relying preferably on the textual content of the exemplars to answer the question. Only about half of the evaluated LMMs could infer visual information when it was necessary.

5 Further Discussion

LMM	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC
GPT 4.1 Mini	5.08 (-69.05)	31.61 (-61.26)	68.89 (- 5.23)	6.97 (-67.15)	6.57 (-67.55)	90.25 (+31.40)	52.36 (-21.76)
GPT 4.1 Nano	37.54 (+ 1.16)	32.91 (-39.55)	53.77 (+17.39)	43.71 (+ 7.33)	30.28 (- 6.11)	29.05 (-21.94)	18.34 (-18.04)
Gemini 2.5 Flash	16.93 (-20.55)	34.22 (-12.91)	59.75 (+22.26)	16.99 (-20.49)	16.19 (-21.30)	32.61 (-57.04)	68.24 (+30.75)
Gemini 2.5 Flash Lite	48.49 (- 3.57)	54.82 (-28.39)	82.66 (+30.60)	30.13 (-21.93)	24.76 (-27.30)	50.75 (-34.69)	22.31 (-29.75)
Gemini 2 Flash	70.75 (-21.51)	98.19 (- 0.50)	94.72 (+ 2.46)	15.59 (-76.67)	10.68 (-81.58)	87.14 (- 9.36)	98.34 (+ 6.08)

Table 3: Performance of some mainstream models in irregular scenarios. The numbers in parentheses show the difference between their performances in irregular scenarios and those in the corresponding regular scenarios.

Evaluation of proprietary LMMs: While proprietary LMMs generally do not allow flexible output parsing through their APIs, they are more sophisticated and generally more accurate. In this section, we evaluate the inductive physical reasoning in these mainstream models. Specifically, we consider models from GPT (Achiam et al., 2023) and Gemini (Comanici et al., 2025) families. Due to the length limits of the prompts (since the API encodes images as strings), we passed only 2 demonstration samples to the API. The results are shown in Tab. 3. The numbers show the accuracy in irregular scenarios, and the numbers in the parentheses show their difference with the accuracy in the corresponding regular scenario. We observe that mainstream models perform similarly to open-source models in inductive physical reasoning. The evaluated mainstream models fail in all scenarios except AMC.



(a) Accuracy trend (b) Trend in accuracy change

Figure 7: Effect of increasing number of demonstration samples on the accuracy for video-text and video-only settings. When the number of demonstration samples increases, accuracy increases in the video-text setting and shows a saturating trend. However, accuracy remains mostly stagnant in the video-only setting when the number of demonstration samples increases.

Effect of number of exemplars: The number of exemplars in an evaluation sample is limited by the context length in which the models were trained/fine-tuned. Our few-shot experiments in § 4 used three exemplars per evaluation sample. In this section, we will examine if the number of exemplars significantly affects performance on InPhyRe. We vary the number of exemplars from 1 to 3 and average the prediction accuracy over the evaluated LMMs for each scenario. We evaluate on both "video-text" and "video-only" settings to investigate potential language bias. Fig. 7a (left) shows the change in prediction accuracy with the number of exemplars when the exemplars contain both videos and question-answer pairs. Across all scenarios, we observe a clear improvement in predictive performance. In particular, LMC and CC show significant improvements in accuracy. However, we note that the accuracy drops between regular and irregular scenarios continue to persist in LMMs despite this accuracy improvement. We also observe a trend of diminishing returns in Fig. 7b, where accuracy saturates and accuracy gains become lower, as the number of demonstration samples increases. These results indicate that merely increasing the number of demonstration samples is not a strong solution to improve inductive physical reasoning, as it attains diminishing accuracy improvements with higher inference cost. However, similar improvement is not present in Fig. 7a (right), where exemplars contained only videos. Performance even deteriorated in LMC (Reg.) as the number of exemplars increased, although the remaining scenarios were unaffected. These results reaffirm our observation about the language bias in LMMs.

	LMC (Reg.)	SB (Reg.)	Wall (Reg.)	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC	Average over scenarios
InternVL3-1B	0.30	0.56	0.11	0.64	0.37	0.92	3.58	3.19	0.33	0.12	1.01
VideoLLaMA3-2B	0.58	1.05	0.97	0.22	0.87	1.16	0.79	0.82	0.52	0.39	0.74
InternVL3-2B	0.78	0.94	0.41	0.21	0.08	0.24	0.37	0.82	0.05	0.32	0.42
Gemma 3-4B	1.30	0.43	1.01	0.06	1.17	0.92	2.75	3.33	0.54	0.15	1.16
LLaVA-NeXT-Vid	0.66	0.49	0.09	0.29	0.50	0.59	0.57	0.56	0.72	0.34	0.48
InternVL3-8B	0.06	0.10	0.19	0.38	0.04	0.00	1.42	1.70	0.59	0.82	0.53
LLaVA-OneVision	0.22	0.13	0.12	0.39	0.11	0.08	0.53	1.00	0.44	0.23	0.32
VideoLLaMA3-7B	0.19	0.52	1.01	0.24	1.39	0.21	0.16	0.99	0.63	0.08	0.54
LLaVA-NeXT-IL	0.36	0.75	0.70	0.36	0.15	0.15	0.95	0.82	0.76	0.35	0.53
Qwen2-VL	0.12	0.06	0.18	0.11	0.10	0.48	0.49	0.26	0.39	0.37	0.26
Qwen2.5-Omni	0.50	0.21	0.86	0.15	0.57	0.33	0.77	0.42	0.32	0.55	0.47
Gemma 3-12B	0.42	1.26	0.76	0.92	0.43	0.41	0.55	0.36	0.17	0.51	0.58
Aria	0.64	0.39	0.37	0.04	0.14	1.19	0.29	1.37	0.25	0.38	0.51
Average over LMMs	0.47	0.53	0.52	0.31	0.46	0.52	1.02	1.20	0.44	0.36	

(a) Video-text setting

	LMC (Reg.)	SB (Reg.)	Wall (Reg.)	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC	Average over scenarios
InternVL3-1B	0.29	0.13	0.09	0.18	0.43	0.37	0.25	0.39	0.41	0.37	0.29
VideoLLaMA3-2B	0.70	0.78	0.30	0.24	0.43	0.29	0.66	0.62	0.88	0.22	0.51
InternVL3-2B	0.40	0.20	0.29	0.23	0.00	0.38	0.29	0.19	0.16	0.05	0.22
Gemma 3-4B	0.36	0.17	0.40	0.46	0.32	0.32	0.34	1.03	0.85	0.28	0.45
LLaVA-NeXT-Vid	0.02	0.58	0.05	0.25	0.12	0.13	0.20	0.14	0.40	0.05	0.19
InternVL3-8B	0.29	0.27	0.69	0.08	0.33	0.36	0.36	0.27	0.14	0.06	0.28
LLaVA-OneVision	0.25	0.25	0.26	0.28	0.14	0.54	0.11	0.32	0.24	0.05	0.24
VideoLLaMA3-7B	0.50	0.26	0.54	0.43	0.51	1.10	0.06	0.00	0.52	0.36	0.43
LLaVA-NeXT-IL	0.15	0.21	0.11	0.06	0.12	0.12	0.08	0.06	0.25	0.28	0.15
Qwen2-VL	0.25	0.19	0.08	0.19	0.44	0.41	0.22	0.20	0.37	0.25	0.26
Qwen2.5-Omni	0.56	0.18	0.14	0.02	0.19	0.38	0.15	0.08	0.28	0.29	0.23
Gemma 3-12B	0.64	0.38	0.30	0.07	0.08	0.66	0.15	0.19	0.59	0.00	0.31
Aria	0.41	0.25	0.25	0.16	0.12	0.15	0.22	0.17	0.08	0.04	0.19
Average over LMMs	0.37	0.30	0.27	0.20	0.25	0.40	0.24	0.28	0.40	0.18	

(b) Video-only setting

Figure 8: Standard deviation in accuracy over different runs in video-text and video-only settings.

We measure the statistical reliability of InPhyRe and our findings through the standard deviation of predictive accuracy for each LMM-scenario combination over runs with different random seeds. The heatmaps

showing the standard deviation under video-text and video-only settings are in Fig. 8. We note that the standard deviation is less than 1 percentage point (pp) for most LMM-scenario combinations, indicating the robustness of InPhyRe and our findings.

We evaluate human performance on InPhyRe using ten human subjects. Each subject was provided with one demonstration sample from each scenario, without the accompanying question-answer pair. For Red-LMC and Red-Pass, they were provided four demonstration samples since they were required to infer conditional reasoning rules. Then, they were asked to answer one evaluation query from the same scenario. We asked them to answer only one query since all queries in a scenario shared the underlying physical logic. Despite being provided only demonstration samples without textual information, the subjects scored above 90% accuracy in many scenarios. They struggled relatively more in Red-LMC, Red-Pass, and CC. Detailed subject-wise results are provided in § E.8.

Effect of exemplar retrieval method: The choice of retrieved samples could affect the performance of LLMs (Liu et al., 2022; Peng et al., 2024). We now verify if this proposition holds for visual inductive physical reasoning. By adapting the textual exemplar retriever from (Liu et al., 2022) for vision, we design a "nearest-neighbor exemplar retriever" (NNER) that finds the top-k video samples closest to the initial frame of the evaluation sample according to their cosine distance in the feature space of CLIP-L (Radford et al., 2021). To evaluate the effect of retrievers on visual inductive physical reasoning, we include only videos in exemplars. We do not evaluate on Red-LMC and Red-Pass since NNER does not guarantee that the retrieved samples include videos with and without red-colored objects. In Fig. 9, we observe either an insignificant or no change in performance between a random retriever and NNER, agreeing with our observation that LMMs rely primarily on language for inductive physical reasoning.

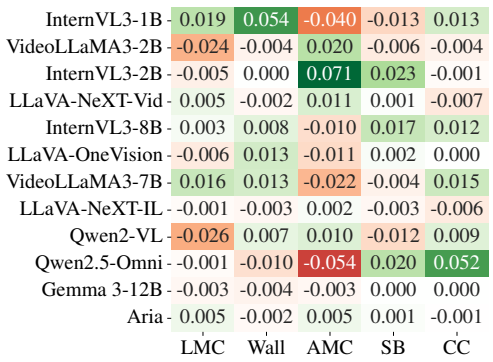


Figure 9: Change in accuracy when NNER is used to find exemplars.

Effect of CoT prompting and fine-tuning: Chain-of-thought (CoT) prompting (Wei et al., 2022) and fine-tuning (FT) have been shown to improve reasoning in LMMs (Buschoff et al., 2025). However, the underlying physical law and the samples themselves must be available in advance for CoT prompting and fine-tuning, respectively. Therefore, these paradigms are not suitable for the premise of inductive physical reasoning, which posits that we have access to only visual samples from unseen scenarios immediately before inference. Additionally, FT involves an expensive training stage. Nonetheless, we still evaluate the effects of CoT and FT on InPhyRe to obtain a "soft upper bound" on the LMMs' performance. The results with CoT prompting and FT are shown in Tabs. 8 and 9, respectively. We find that both CoT and FT can generally improve performance in most scenarios, except Red-Pass and Red-LMC, which, unlike other scenarios, require conditional physical reasoning, signaling the utility limits of CoT prompting and FT. As a reminder, in these scenarios, only red-colored objects violate the true physical laws.

Causes of poor performance in irregular scenarios: We provide a preliminary analysis of the causes of poor performance in irregular scenarios in § F. To this end, we use linear probes on the hidden states from pre-trained and fine-tuned InternVL3-1B, and visualize attention values over the tokens in the last layer of Gemma3-12B. We find that the hidden states of both pre-trained and fine-tuned models carry sufficient information to classify the underlying scenario (Fig. 27). Moreover, after fine-tuning, the hidden states adaptively include attribute information from demonstration and evaluation samples depending on the underlying scenario (Figs. 28 and 29). Visualization of attention values from the last layer of Gemma3-12B shows that the model spends an order of magnitude less attention over image tokens compared to text tokens (Fig. 26). We believe that these findings will assist in developing methods to explicitly improve inductive physical reasoning. Results on further experiments on quantized models, a visually more complex version of InPhyRe, evaluation with all evaluation sample frames included, effects of prompt perturbations, and qualitative results on InPhyRe are in § E.

6 Conclusion

When inference scenarios violate the physical laws encoded in the model parameters, LMMs must ideally derive their physical reasoning from demonstration samples. Therefore, to ensure their trustworthiness, it is critical to evaluate how well LMMs can infer physical laws from these exemplars. To this end, we introduced InPhyRe, the ***first*** visual question answering benchmark to quantify parametric knowledge and inductive physical reasoning in LMMs. InPhyRe evaluates LMMs in collision scenarios that violate universal physical laws such as momentum conservation. Through zero-shot and few-shot experiments in these scenarios, we found that LMMs have limited parametric knowledge of universal physical laws and struggle to apply these laws during physical reasoning. LMMs demonstrated only weak inductive physical reasoning when the inference scenario violates universal physical laws. LMMs also struggled in visual inductive physical reasoning when QA pairs were absent in demonstration samples. Many LMMs failed to infer visual information and relate it to the query, even when it was necessary. Although increasing the number of demonstration samples improved accuracy, the accuracy improvement showed a diminishing return trend, while the accuracy gap between regular and irregular scenarios persisted. This suggests the need for targeted solutions to improve inductive physical reasoning.

Limitations: Although InPhyRe proved to be a formidable benchmark for LMMs in terms of physical laws, its simple synthetic scene could not have posed visual challenges to the models. Inductive physical reasoning in LMMs could be worse in a more crowded and visually realistic scene that also challenges the perception abilities of LMMs. Since real-world videos cannot violate true physical laws, the closest alternative to evaluate inductive physical reasoning in LMMs is a hyperrealistic video benchmark generated using advanced video generation models (OpenAI, 2024; Kondratyuk et al., 2024; Chen et al., 2024b). However, such endeavors can be fruitful only with absolute and precise control over physical realism, which has not yet been fully achieved in current video generation models (Cho et al., 2024; Motamed et al., 2025).

7 Acknowledgments

This work was supported in part by the National Science Foundation (award #2500983) and the Office of Naval Research (award #N00014-23-1-2417). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or ONR.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. PROST: Physical Reasoning about Objects through Space and Time. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

Tayfun Ates, M Ateşoğlu, Çağatay Yiğit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Advances in Neural Information Processing Systems*, 2023b.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A New Benchmark for Physical Reasoning. In *Advances in Neural Information Processing Systems*, 2019.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What Makes Multimodal In-Context Learning Work? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2024.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual Learning of Physical Dynamics. In *International Conference on Learning Representations*, 2020.

Daniel Bear, Elias Wang, Damian Mrowca, Felix Jedidja Binder, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin A Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems*, 2021.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI Conference on Artificial Intelligence*, 2020.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments. *arXiv preprint arXiv:2506.09849*, 2025.

Chen Bowen, Rune Sætre, and Yusuke Miyao. A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Language Models. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2024.

Luca M Schulze Buschoff, Konstantinos Voudouris, Elif Akata, Matthias Bethge, Joshua B Tenenbaum, and Eric Schulz. Testing the Limits of Fine-Tuning for Improving Visual Cognition in Vision Language Models. *arXiv preprint arXiv:2502.15678*, 2025.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric Depth From Vision Language Models. *arXiv preprint arXiv:2509.25413*, 2025.

Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. PhysGame: Uncovering Physical Commonsense Violations in Gameplay Videos. *arXiv preprint arXiv:2412.01800*, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *European Conference on Computer Vision*, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In *Advances in Neural Information Processing Systems*, 2024b.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can Multimodal Large Language Models Truly Perform Multimodal In-Context Learning? In *Winter Conference on Applications of Computer Vision*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024c.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. ComPhy: Compositional Physical Reasoning of Objects and Events from Videos. In *International Conference on Learning Representations*, 2022.

Zhenfang Chen, Shilong Dong, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Compositional Physical Reasoning of Objects and Events from Videos. *arXiv preprint arXiv:2408.02687*, 2024d.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, et al. Inductive or deductive? Rethinking the fundamental reasoning abilities of LLMs. In *ACL Workshop on Natural Language Reasoning and Structure Explanations*, 2024.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. LLMPhy: Complex Physical Reasoning Using Large Language Models and World Models. *arXiv preprint arXiv:2411.08027*, 2024.

Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an AGI World Model? A Complete Survey on Text-to-Video Generation. *arXiv preprint arXiv:2403.05131*, 2024.

François Chollet. On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In *International Conference on Learning Representations*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

The Blender Community. Blender 4.4. `https://docs.blender.org/manual/en/4.4/index.html`, 2025.

Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2021.

Stephanie Denison and Fei Xu. Integrating physical constraints in statistical inference by 11-month-old infants. *Cognitive science*, 34(5):885–908, 2010.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models. *arXiv preprint arXiv:2411.14432*, 2024.

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! Predicting Unintentional Action in Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Laura Faßbender, Johannes Falck, Francisco M López, Yee Lee Shing, Jochen Triesch, and Gudrun Schwarzer. A comparison of force adaptation in toddlers and adults during a drawer opening task. *Scientific Reports*, 15(1):3699, 2025.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In *Advances in Neural Information Processing Systems*, 2022.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. Large Language Models Are Not Strong Abstract Reasoners. In *International Joint Conference on Artificial Intelligence*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Brett K Hayes. The Development of Inductive Reasoning. *Inductive Reasoning Experimental, Developmental, and Computational Approaches*, pp. 25–54, 2007.

Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, and Zhiyu Zoey Chen. IDEA: Enhancing the Rule Learning Ability of Large Language Model Agent through Induction, Deduction, and Abduction. *arXiv preprint arXiv:2408.10455*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. GRASP: A Novel Benchmark for Evaluating Language GRounding and Situated Physics Understanding in Multimodal Language Models. In *International Joint Conference on Artificial Intelligence*, 2024.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. VideoPoet: A Large Language Model for Zero-Shot Video Generation. In *International Conference on Machine Learning*, 2024.

Laura Kotovsky and Renée Baillargeon. The development of calibration-based reasoning about collision events in young infants. *Cognition*, 67(3):311–351, 1998.

Adam Lerer, Sam Gross, and Rob Fergus. Learning Physical Intuition of Block Towers by Example. In *International Conference on Machine Learning*, 2016.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. ARIA: An Open Multimodal Native Mixture-of-Experts Model. *arXiv preprint arXiv:2410.05993*, 2024b.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*, 2024c.

Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. MIRAGE: Evaluating and Explaining Inductive Reasoning Process in Language Models. In *International Conference on Learning Representations*, 2025.

Jiaoda Li, Yifan Hou, Mrinmaya Sachan, and Ryan Cotterell. What Do Language Models Learn in Context? The Structured Task Hypothesis. In *Annual Meeting of the Association for Computational Linguistics*, 2024d.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What Makes Good In-Context Examples for GPT-3? In *Deep Learning Inside Out (DeeLIO): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 2022.

Francesco Margoni, Luca Surian, and Renée Baillargeon. The Violation-of-Expectation Paradigm: A Conceptual Overview. *Psychological Review*, 131(3):716, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Conference on Empirical Methods in Natural Language Processing*, 2022.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large Language Models as General Pattern Machines. In *Conference on Robot Learning*, 2023.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian Image Understanding: Unfolding the Dynamics of Objects in Static Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Nayantara Mudur, Hao Cui, Subhashini Venugopalan, Paul Raccuglia, Michael P Brenner, and Peter Norgaard. FEABench: Evaluating Language Models on Multiphysics Reasoning Ability. *arXiv preprint arXiv:2504.06260*, 2025.

Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. Learning vs Retrieval: The Role of In-Context Examples in Regression with LLMs. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2025.

OpenAI. Sora – Creating video from text. `https://openai.com/index/sora/`, 2024.

Xinyu Pang, Ruixin Hong, Zhanke Zhou, Fangrui Lv, Xinwei Yang, Zhilong Liang, Bo Han, and Changshui Zhang. Physics Reasoner: Knowledge-Augmented Reasoning for Solving Physics Problems with Large Language Models. In *International Conference on Computational Linguistics*, 2025.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting Demonstration Selection Strategies in In-Context Learning. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021.

Robert B Ricco. The Development of Reasoning. *Handbook of Child Psychology and Developmental Science*, pp. 1–52, 2015.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys 2019: A Benchmark for Visual Intuitive Physics Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2021.

Mohammad Reza Taesiri, Finlay Macklon, and Cor-Paul Bezemer. CLIP meets GamePhysics: Towards bug identification in gameplay videos using zero-shot transfer learning. In *International Conference on Mining Software Repositories*, 2022a.

Mohammad Reza Taesiri, Finlay Macklon, Yihe Wang, Hengshuo Shen, and Cor-Paul Bezemer. Large Language Models are Pretty Good Zero-Shot Video Game Bug Detectors. *arXiv preprint arXiv:2210.02506*, 2022b.

Mohammad Reza Taesiri, Tianjun Feng, Cor-Paul Bezemer, and Anh Nguyen. Glitchbench: Can large multimodal models detect video game glitches? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models. In *Advances in Neural Information Processing Systems*, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating Physical Scene Understanding that Requires Online Inference of Different Physical Properties. In *Advances in Neural Information Processing Systems*, 2023.

Robert Vacareanu, Vlad-Andrei Negru, Vasile Suciu, and Mihai Surdeanu. From Words to Numbers: Your Large Language Model Is Secretly A Capable Regressor When Given In-Context Examples. In *Conference on Language Modeling*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. Can In-context Learning Really Generalize to Out-of-distribution Tasks? *arXiv preprint arXiv:2410.09695*, 2024b.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah Goodman. Hypothesis Search: Inductive Reasoning with Language Models. In *International Conference on Learning Representations*, 2024c.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge Editing for Large Language Models: A Survey. *ACM Computing Surveys*, 57(3):1–37, 2024d.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha S Srinivasa. NEWTON: Are Large Language Models Capable of Physical Reasoning? In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking Progress to Infant-Level Physical Reasoning in AI. *Transactions in Machine Learning Research*, 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025.

Kai Yan, Zhan Ling, Kang Liu, Yifan Yang, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. MIR-Bench: Benchmarking LLM's Long-Context Intelligence via Many-Shot In-Context Inductive Reasoning. *arXiv preprint arXiv:2502.09933*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024b.

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*, 2020.

Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. The Stochastic Parrot on LLM's Shoulder: A Summative Assessment of Physical Concept Understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2025.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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. URL https://arxiv.org/abs/2501.13106.

Xingjian Zhang. How Vision Language Models Will Shape The Future Of Self-Driving Cars, March 2025. URL https://www.forbes.com/councils/forbestechcouncil/2025/03/18/how-vision-language-models-will-shape-the-future-of-self-driving-cars/.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. ContPhy: Continuum Physical Concept Learning and Reasoning from Videos. In *International Conference on Machine Learning*, 2024.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

Table of Contents

A Data-generation Details **21**
A.1 Rendering the Videos . . . 21

B Details of Experimental Setup and Evaluation **23**
B.1 Evaluated LMMs . . . 23
B.2 Prompting Methods . . . 23
B.3 Parsing the Chosen Option from Generated Output . . . 25
B.4 Absolute Few-shot Accuracy of LMMs in Irregular Scenarios . . . 25
B.5 Other Details . . . 25

C Other Related Works **26**

D Various Types of Reasoning **28**

E Additional Experimental Evaluation **29**
E.1 Effect of weight quantization on inductive physical reasoning . . . 29
E.2 Can Chain-of-Thought Prompting Improve Inductive Physical Reasoning? . . . 29
E.3 Effect of Fine-tuning on Inductive Physical Reasoning . . . 30
E.4 Do the Results from CoT Prompting and Fine-Tuning Experiments Invalidate Inductive Physical Reasoning? . . . 30
E.5 Qualitative Inspection of Generated Outputs . . . 31
E.6 Explaining the apparent inductive physical reasoning in AMC . . . 41
E.7 Do the findings hold for more complex scenes? . . . 45
E.8 Human Evaluation for InPhyRe . . . 46
E.9 Does using all evaluation frames improve performance? . . . 46
E.10 Effect of Structural Perturbations to Prompts . . . 48

F Why Do LMMs Perform Poorly in Irregular Scenarios **49**
F.1 Analysis of Attention Maps in Gemma3-12B . . . 49
F.2 Analysis of Hidden States in InternVL3-1B . . . 50

G Use of LLMs and Generative AI **53**

A Data-generation Details

A.1 Rendering the Videos

We use Blender's Python wrapper[3] (v4.4.0) to render the video from the trajectories. We modified the image rendering code[4] from (Johnson et al., 2017). We first designed a base scene with lamp and camera positions suitable for capturing entire object trajectories in InPhyRe. The object textures were taken from (Johnson et al., 2017), but more hues were added. Each video consists of 240 frames from which 8 frames are uniformly sampled to form the video tokens by the corresponding video processors of the LMMs.

[3] https://pypi.org/project/bpy/
[4] https://github.com/facebookresearch/clevr-dataset-gen

A sample video from each scenario is shown below: regular scenarios in Figs. 10 to 12, momentum conservation violation scenarios in Figs. 13 to 15, inconsistency physics scenarios in Figs. 18 and 19, and miscellaneous irregular scenarios in Figs. 16 and 17.

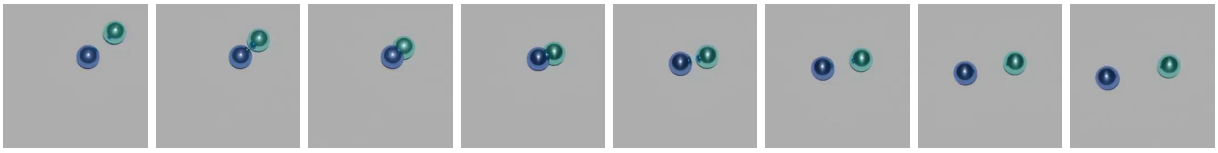


Figure 10: Regular scenario where linear momentum conservation is followed – LMC (Regular)

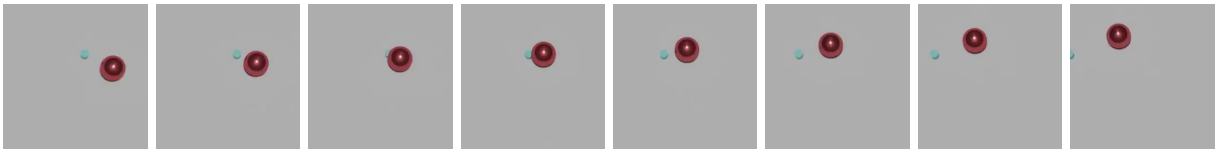


Figure 11: Regular scenario where the larger object has more mass – SB (Regular)

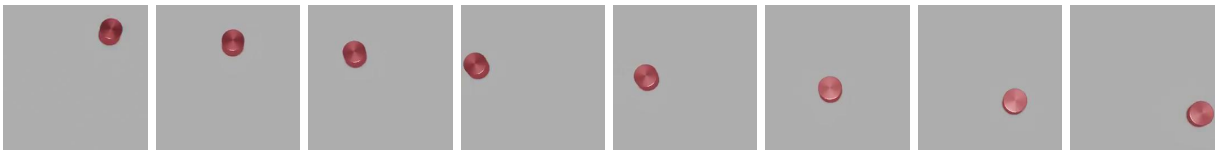


Figure 12: Regular scenario where linear momentum is conserved along the vertical direction – Wall (Regular)

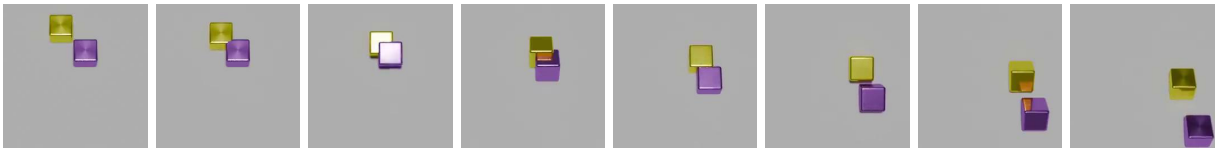


Figure 13: Irregular scenario where linear momentum conservation is violated – LMC

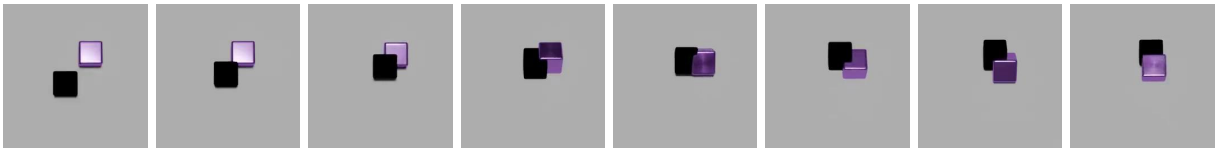


Figure 14: Irregular scenario where angular momentum conservation is violated – AMC

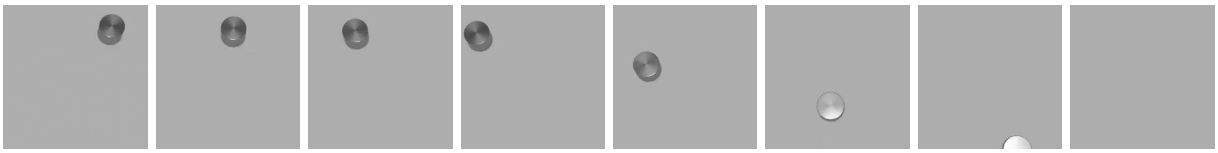


Figure 15: Irregular scenario where linear momentum conservation is violated along the vertical direction – Wall

Figure 16: Irregular scenario where only red-colored objects violate linear momentum conservation – Red-LMC

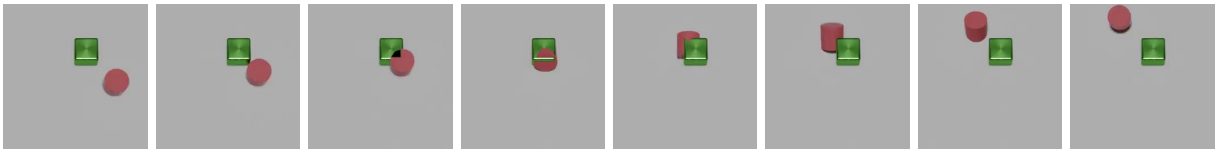


Figure 17: Irregular scenario where red-colored objects can pass through other objects – Red-Pass

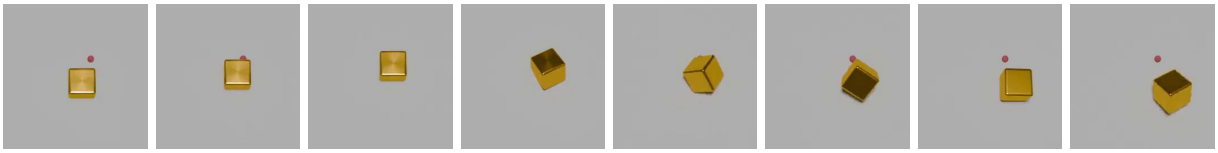


Figure 18: Irregular scenario where a large object deflects after colliding with a tiny, much heavier, object – SB

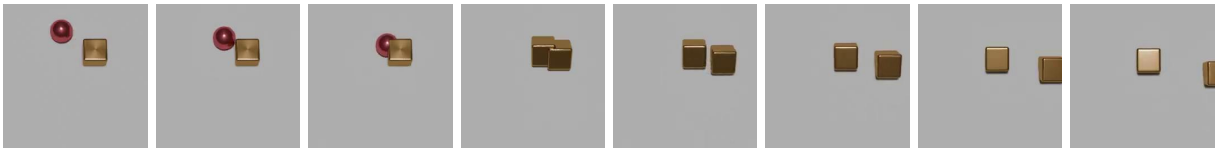


Figure 19: Irregular scenario where the colliding object assumes the hue and the shape of the other object after collision – CC

B Details of Experimental Setup and Evaluation

B.1 Evaluated LMMs

We evaluate a diverse collection of LMMs trained on both public and proprietary datasets with different architectural design choices. The model weights were taken from Huggingface[5]. The models are listed in Tab. 4.

B.2 Prompting Methods

For prompting, we used the `apply_chat_template` method to convert conversations from a list of Python dictionaries to LMM-specific prompts. This helps in reusing the code. An example from LMC (regular) with one exemplar containing both video and question-answer pair in the conversation style of LLaVA-NeXT-Video is given below:

```
[
    {
    "role": "system",
    "content": [{"type": "text",
                 "text": "Understand the underlying physics from the following videos and choose only an option among
                      A, B, C and D to answer the question. Do not provide reasoning."}]
    },
    {
```

[5] https://huggingface.co/models

Model	HF ID	#Params (B)	Vision encoder	Language model	MoE
InternVL3 (1B) (Zhu et al., 2025)	`OpenGVLab/InternVL3-1B-hf`	0.9	InternViT (Chen et al., 2024c)	Qwen2.5 (Yang et al., 2024b)	✗
InternVL3 (2B) (Zhu et al., 2025)	`OpenGVLab/InternVL3-2B-hf`	1.9	InternViT (Chen et al., 2024c)	Qwen2.5 (Yang et al., 2024b)	✗
VideoLLaMA3 (2B) (Zhang et al., 2025)	`DAMO-NLP-SG/VideoLLaMA3-2B`	2	SigLIP (Zhai et al., 2023)	Qwen2.5 (Yang et al., 2024b)	✗
Gemma 3 (4B) (Kamath et al., 2025)	`google/gemma-3-4b-it`	4	SigLIP (Zhai et al., 2023)	*	✗
LLaVA-NeXT-Video (Zhang et al., 2024)	`llava-hf/LLaVA-NeXT-Video-7B-hf`	7	CLIP-L (Radford et al., 2021)	Vicuna 1.5 (Zheng et al., 2023)	✗
LLaVA-OneVision (Li et al., 2024a)	`llava-hf/llava-onevision-qwen2-7b-si-hf`	7	SigLIP (Zhai et al., 2023)	Qwen2 (Yang et al., 2024a)	✗
LLaVA-NeXT-Interleave (Li et al., 2024c)	`llava-hf/llava-interleave-qwen-7b-hf`	7	SigLIP (Zhai et al., 2023)	Qwen1.5 (Bai et al., 2023a)	✗
VideoLLaMA3 (7B) (Zhang et al., 2025)	`DAMO-NLP-SG/VideoLLaMA3-7B`	7	SigLIP Zhai et al. (2023)	Qwen2.5 (Yang et al., 2024b)	✗
Qwen2-VL (Wang et al., 2024a)	`Qwen/Qwen2-VL-7B-Instruct`	7	Qwen-VL (Bai et al., 2023a)	Qwen2 (Yang et al., 2024a)	✗
Qwen2.5-Omni (Xu et al., 2025)	`Qwen/Qwen2.5-Omni-7B`	7	Qwen2.5-VL (Bai et al., 2025)	Qwen2.5 (Yang et al., 2024b)	✗
InternVL3 (8B) (Zhu et al., 2025)	`OpenGVLab/InternVL3-8B-hf`	8.1	InternViT (Chen et al., 2024c)	Qwen2.5 (Yang et al., 2024b)	✗
Gemma 3 (12B) (Kamath et al., 2025)	`google/gemma-3-12b-it`	12	SigLIP (Zhai et al., 2023)	*	✗
Aria (Li et al., 2024b)	`rhymes-ai/Aria`	24.9	SigLIP (Zhai et al., 2023)	*	✓

Table 4: We evaluate a diverse assortment of LMMs with varying architectural and data choices. * indicates that the language component was trained from scratch. "HF ID" is the identifier for the model weights on Huggingface.

```
    "role": "user",
    "content": [{"type": "video", "path": "path/to/exemplar_video.mp4"},
                {"type": "text", "text": "cyan cylinder and red cylinder have equal mass. How will the speed of cyan
                     cylinder change after colliding with red cylinder? A: Speed does not change, B: cyan cylinder's
                     speed decreases, C: cyan cylinder's speed will increase, D: Not enough data"}]
    },
    {
    "role": "assistant",
    "content": [{"type": "text", "text": "B"}]
    }
    {
    "role": "user",
    "content": [{"type": "video", "path": "path/to/evaluation_image.png"},
                {"type": "text", "text": "yellow cube and purple sphere have equal mass. How will the speed of yellow
                     cube change after colliding with purple sphere? A: Not enough data, B: Speed does not change, C:
                     yellow cube's speed will decrease, D: yellow cube's speed will increase"}]
    }
]
```

The exact style of the dictionary differs with the LMM. All prompts include a system prompt. When exemplars are provided, the system prompt is "`Understand the underlying physics from the following videos and choose an option among A, B, C and D to answer the question. Do not provide reasoning.`" When exemplars are absent, the system prompt is "`Choose an option among A, B, C and D to answer the question based on your understanding of physical laws. Do not provide reasoning.`"

However, these instructions are insufficient to restrict the model's output to options. Since LLMs are first pre-trained for next-token prediction, they tend to complete the prompt instead of answering the question in it. Thus, the generated output tends to be descriptive, especially when there are no exemplars. For instance, suppose the model wants to choose option "D: Both objects move". Instead of simply outputting "D", the model may output "... Therefore, both objects may move." Even instruction-tuned models sometimes fail to format their outputs, despite including the instruction "`Do not provide reasoning`" in the system prompt. Parsing the chosen option from such descriptive outputs is difficult.

Therefore, to avoid descriptive outputs, we provide demonstration samples in the expected output format since LLMs can understand output formatting from exemplars (Min et al., 2022). Even when exemplars do not contain question-answer pairs, we include randomly chosen options among {"A", "B", "C", "D"} to condition the model to output only the option index. An example of such a conversation for zero-shot evaluation is provided below:

```
[
    {
    "role": "system",
    "content": [{"type": "text",
                "text": "Choose an option among A, B, C and D to answer the question based on your understanding of
                     physical laws. Do not provide reasoning."}]
```

```
    },
    {
    "role": "assistant",
    "content": [{"type": "text", "text": "A"}]
    }
    {
    "role": "assistant",
    "content": [{"type": "text", "text": "D"}]
    }
    {
    "role": "assistant",
    "content": [{"type": "text", "text": "C"}]
    }
    {
    "role": "user",
    "content": [{"type": "video", "path": "path/to/evaluation_image.png"},
               {"type": "text", "text": "yellow cube and purple sphere have equal mass. How will the speed of yellow
                    cube change after colliding with purple sphere? A: Not enough data, B: Speed does not change, C:
                    yellow cube's speed will decrease, D: yellow cube's speed will increase"}]
    }
]
```

Here, options in the assistant dictionaries do not have any relation to the video in the exemplar. Therefore, we still refer to this approach as "zero-shot evaluation" in the sense that no useful samples are provided as exemplars to the model.

B.3 Parsing the Chosen Option from Generated Output

We typically ask the model to generate a maximum of 10 new tokens. In certain cases, when the model output was more verbose, we allowed 100 tokens to be generated. For our chain-of-thought experiments in § E.2, we allowed 300 tokens. The tokens generated by the model are decoded using its corresponding tokenizer. To obtain the chosen option from this decoded output, we "clean" the decoded string first until the first character is the option index. We remove the following substrings in our "cleaning" procedure:

1. Placeholders for image and video tokens. E.g., `<|im_start|>`, `<fim_suffix>`.
2. Partial placeholders for image and video tokens. E.g., only `<|im_st` from `<|im_start|>`.
3. Strings such as "The correct answer is" prepending the chosen option. Such strings are first collected manually and then removed during the cleaning procedure.
4. Strings that follow the chosen option. E.g., "Human: Which movie...". Similar to the previous case, these strings can be collected for each model and then removed during the experimental evaluation.

If multiple options are chosen, then the model's output is marked to be incorrect. In some cases where not enough exemplars are available for the model, we allow the model to generate more tokens and attempt to find the chosen option based on textual overlap between the generated output and the given options.

B.4 Absolute Few-shot Accuracy of LMMs in Irregular Scenarios

Tabs. 5 and 6 show the absolute accuracy values for the experiments in §§ 4.4 and 4.5.

B.5 Other Details

Compute: Almost all experiments were run on individual A6000 GPUs on a server with 128 AMD EPYC 7502 (32-core) processors. Very few experiments were run on H200 and A100 GPUs.

Modifications to Huggingface: Some of the evaluated LMMs did not account for multiple videos in the prompt. Even the latest version of the Transformers library[6] had this bug. So we made minor changes to the

[6] https://github.com/huggingface/transformers

LMM	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC
InternVL3-1B (Zhu et al., 2025)	45.55 (-25.78)	18.73 (+10.63)	79.90 (+ 8.57)	51.98 (-19.35)	55.12 (-16.22)	9.05 (-77.93)	41.12 (-30.22)
VideoLLaMA3-2B (Zhang et al., 2025)	34.58 (-29.67)	29.32 (-11.67)	63.42 (- 0.83)	33.83 (-30.42)	43.15 (-21.10)	7.72 (-55.05)	45.40 (-18.85)
InternVL3-2B (Zhu et al., 2025)	17.97 (-70.38)	1.07 (-87.12)	97.90 (+ 9.55)	17.37 (-70.98)	24.90 (-63.45)	0.18 (-72.96)	6.75 (-81.60)
Gemma 3-4B (Kamath et al., 2025)	81.33 (+10.63)	50.17 (-27.95)	62.80 (- 7.90)	64.88 (- 5.82)	66.13 (- 4.57)	26.70 (-44.73)	36.00 (-34.70)
LLaVA-NeXT-Vid (Zhang et al., 2024)	16.82 (-41.58)	37.80 (+29.03)	48.93 (- 9.47)	11.62 (-46.78)	12.88 (-45.52)	19.08 (-14.45)	32.13 (-26.27)
InternVL3-8B (Zhu et al., 2025)	93.47 (- 6.40)	99.60 (+ 3.43)	100.00 (+ 0.13)	69.47 (-30.40)	60.02 (-39.85)	45.83 (-54.05)	68.42 (-31.45)
LLaVA-OneVision (Li et al., 2024a)	96.65 (- 2.50)	97.85 (- 1.50)	99.50 (+ 0.35)	67.87 (-31.28)	66.13 (-33.02)	28.72 (-70.39)	64.95 (-34.20)
VideoLLaMA3-7B (Zhang et al., 2025)	83.37 (-14.93)	82.65 (+33.95)	96.08 (- 2.22)	43.17 (-55.13)	44.88 (-53.42)	61.17 (-33.62)	67.25 (-31.05)
LLaVA-NeXT-IL (Li et al., 2024c)	91.80 (- 6.00)	95.90 (+ 5.92)	94.42 (- 3.38)	71.62 (-26.18)	50.55 (-47.25)	42.33 (-28.27)	97.65 (- 0.15)
Qwen2-VL (Wang et al., 2024a)	94.60 (- 2.97)	99.78 (+ 1.12)	96.97 (- 0.60)	97.53 (- 0.03)	92.58 (- 4.98)	15.07 (-83.94)	87.22 (-10.35)
Qwen2.5-Omni (Xu et al., 2025)	70.35 (+13.37)	66.67 (+14.62)	91.03 (+34.05)	58.53 (+ 1.55)	52.95 (- 4.03)	58.25 (-36.37)	42.22 (-14.77)
Gemma 3-12B (Kamath et al., 2025)	62.38 (-25.27)	89.30 (+16.65)	94.32 (+ 6.67)	23.05 (-64.60)	22.43 (-65.22)	16.47 (-69.00)	86.00 (- 1.65)
Aria (Li et al., 2024b)	30.15 (-26.55)	1.55 (-80.43)	74.10 (+17.40)	31.35 (-25.35)	21.40 (-35.30)	9.77 (-81.63)	9.73 (-46.97)

Table 5: 3-shot evaluation results of LMMs on irregular scenarios. The exemplars contain **both** videos and QA pairs.

LMM	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC
InternVL3-1B (Zhu et al., 2025)	24.32 (-21.23)	11.97 (- 6.77)	29.52 (-50.38)	26.17 (-25.82)	28.78 (-26.33)	15.47 (+ 6.42)	49.12 (+ 8.00)
VideoLLaMA3-2B (Zhang et al., 2025)	12.23 (-22.35)	12.63 (-16.68)	34.00 (-29.42)	14.20 (-19.63)	17.80 (-25.35)	25.22 (+17.50)	12.30 (-33.10)
InternVL3-2B (Zhu et al., 2025)	6.05 (-11.92)	0.00 (- 1.07)	62.02 (-35.88)	8.35 (- 9.02)	8.25 (-16.65)	7.32 (+ 7.13)	0.08 (- 6.67)
Gemma 3-4B (Kamath et al., 2025)	22.13 (-59.20)	15.70 (-34.47)	54.77 (- 8.03)	21.55 (-43.33)	23.50 (-42.63)	30.32 (+ 3.62)	6.58 (-29.42)
LLaVA-NeXT-Vid (Zhang et al., 2024)	16.42 (- 0.40)	42.52 (+ 4.72)	43.83 (- 5.10)	16.78 (+ 5.17)	16.03 (+ 3.15)	29.55 (+10.47)	41.07 (+ 8.93)
InternVL3-8B (Zhu et al., 2025)	17.87 (-75.60)	56.63 (-42.97)	59.95 (-40.05)	17.38 (-52.08)	19.22 (-40.80)	14.70 (-31.13)	10.92 (-57.50)
LLaVA-OneVision (Li et al., 2024a)	3.02 (-93.63)	23.83 (-74.02)	55.63 (-43.87)	2.90 (-64.97)	3.65 (-62.48)	13.93 (-14.78)	0.07 (-64.88)
VideoLLaMA3-7B (Zhang et al., 2025)	4.30 (-79.07)	20.20 (-62.45)	43.75 (-52.33)	9.12 (-34.05)	10.85 (-34.03)	29.70 (-31.47)	10.08 (-57.17)
LLaVA-NeXT-IL (Li et al., 2024c)	0.37 (-91.43)	16.43 (-79.47)	49.42 (-45.00)	0.45 (-71.17)	0.37 (-50.18)	19.93 (-22.40)	32.12 (-65.53)
Qwen2-VL (Wang et al., 2024a)	3.75 (-90.85)	45.52 (-54.27)	67.62 (-29.35)	2.87 (-94.67)	2.95 (-89.63)	7.07 (- 8.00)	46.60 (-40.62)
Qwen2.5-Omni (Xu et al., 2025)	0.20 (-70.15)	22.37 (-44.30)	58.70 (-32.33)	2.37 (-56.17)	2.22 (-50.73)	18.87 (-39.38)	44.47 (+ 2.25)
Gemma 3-12B (Kamath et al., 2025)	9.85 (-52.53)	0.18 (-89.12)	66.68 (-27.63)	5.63 (-17.42)	14.17 (- 8.27)	5.28 (-11.18)	0.00 (-86.00)
Aria (Li et al., 2024b)	1.08 (-29.07)	3.17 (+ 1.62)	29.80 (-44.30)	1.43 (-29.92)	1.12 (-20.28)	13.38 (+ 3.62)	0.20 (- 9.53)

Table 6: 3-shot evaluation results of LMMs in irregular scenarios. The exemplars contain **only** videos.

codebase of LLaVA-NeXT-Video, LLaVA-OneVision, and InternVL3 models. The modified "transformers" library is included in the codebase.

C Other Related Works

In this section, we include some recent works that evaluated various reasoning aspects of LLMs and LMMs. We will also clarify that our objective has never been explored in any of these works.

Intuitive Physics Understanding: Physically impossible scenarios have been employed to evaluate physical reasoning in learned models, following the "violation of expectation" principle from cognitive theory (Margoni et al., 2024). Here, the key hypothesis is that a model with excellent physical reasoning can also understand when the underlying physical laws in the given scenario violate the known physical laws. However, violation of expectation in infants and children is often the initial step towards adaptation to a new physical environment (Kotovsky & Baillargeon, 1998; Denison & Xu, 2010; Faßbender et al., 2025). For instance, (Faßbender et al., 2025) showed that infants and children adapted their force while opening and closing drawers, whose friction was temporally altered. Violation of expectation in learned models in the context of physical laws can be used as a proxy for physical reasoning. An early example of such work is the IntPhys (Riochet et al., 2021) benchmark that quantified the physical reasoning abilities of models trained on visual datasets that obeyed universal physical laws using their next-frame prediction errors on physically impossible scenarios. More works on intuitive physics understanding have emerged since (Epstein et al., 2020; Weihs et al., 2022; Jassim et al., 2024; Garrido et al., 2025).

Intuitive physics understanding is not inductive physical reasoning: Intuitive physics understanding differs from InPhyRe in the final objective, as the underlying assumption in intuitive physics understanding does not impact our setting. InPhyRe evaluates how well a large multimodal model can infer the underlying physical laws from the demonstration samples and apply them for physical reasoning when given a scenario the model has not seen during its training. This property, which we refer to as inductive physical reasoning

in the main paper, is the key question we pose. The absolute physical reasoning ability (that we refer to as parametric knowledge) of this model on regular physical tasks that the model might have seen during training is not of interest to us. However, since we do not know which scenarios were observed during training and which were not, we rely on impossible scenarios to evaluate the inductive physical reasoning.

Evaluation of physical reasoning in learned models: Research interests in learning visual physical reasoning predate the era of large models. Early works generated synthetic vision datasets that depicted physical events such as collisions and falls, and trained models to predict future events (Lerer et al., 2016; Baradel et al., 2020; Bear et al., 2021), answer questions about cue events Mottaghi et al. (2016), or interact with the physical simulator to achieve an end goal (Bakhtin et al., 2019). Other tasks involved visually inferring latent physical properties, such as mass and friction, from the physical interactions (Chen et al., 2022; Tung et al., 2023) and causal physical reasoning (Yi et al., 2020; Ates et al., 2022).

Reasoning from demonstration samples: Prior efforts have attempted to reason *how* LLMs utilize demonstration samples. These works consider both parametric knowledge (Min et al., 2022; Li et al., 2024d; Nafar et al., 2025) and inductive reasoning hypotheses (Garg et al., 2022; Bai et al., 2023b; Wang et al., 2024b; Vacareanu et al., 2024; Nafar et al., 2025). However, their findings are usually limited to synthetic regression tasks on LLMs (Garg et al., 2022; Bai et al., 2023b; Wang et al., 2024b), and they do not consider physical reasoning tasks on LMMs.

Glitch detection using LMMs: Another task similar to intuitive physics understanding is "glitch detection." Here, the intuition is that a model that understands the underlying physics can also detect glitches in a given scenario. Some examples of glitch detection using LLMs are (Taesiri et al., 2022b;a; 2024; Cao et al., 2024).

Use of synthetic data for physical reasoning: Synthetically generated images and videos are commonly used for physical reasoning, as collecting visual data on real physical events is both taxing and time-consuming. Several of the works that we listed above and in § 2 also use synthetic data. In Tab. 7, we list some additional works that use synthetically generated collision events for physical reasoning. We also include CLEVR (Johnson et al., 2017) dataset in the table due to its visual resemblance with InPhyRe. Note: The tasks in ComPhy vary in terms of the underlying law required for reasoning (e.g., objects with the same charge repelling after a collision). All the events are still collision events.

Benchmark	# of Tasks	Physics Engine	Renderer	Events other than collision events
CLEVR (Johnson et al., 2017)	-	No physics, only images	Blender	No events, only images
CLEVRER (Yi et al., 2020)	4 question types, 5 description types	Bullet	Blender	None
ComPhy (Chen et al., 2022)	4 (mass, charge, color, collision)	Bullet	Blender	None
CoPhy (Baradel et al., 2020)	3 (BlocktowerCF, BallsCF, CollisionCF)	PyBullet	PyBullet	Stability of stacked objects
InPhyRe (ours)	10 (spanning linear and angular momentum conservation, object permanence)	PyBullet	Blender	None

Table 7: A tabular comparison of InPhyRe with other works that use synthetically generated collision events for physical reasoning.

Different w.r.t. ContPhy and PhysBench: ContPhy (Zheng et al., 2024) and PhysBench (Chow et al., 2025) are among the most comprehensive physical reasoning benchmarks that appeared recently. They include both real and synthetic videos that show physical events governed by a wide span of physical laws, such as Newton's laws of motion, friction, fluid mechanics, etc. Therefore, they successfully evaluate the knowledge of LMMs on diverse topics required for physical reasoning. The key difference between InPhyRe and these works is the objective. InPhyRe evaluates the ability of the models to adapt to an unseen scenario, while these works evaluate the parametric knowledge about physics in these models. As a consequence, ContPhy and PhysBench arrive at conclusions different from ours. Zheng et al. (2024) find that these models "struggle to perform well on our benchmark, highlighting their limited physical commonsense for the continuum, especially soft bodies, and fluids." Similarly, Chow et al. (2025) "identified significant gaps in physical world understanding, particularly in open-source models, due to inadequate training data" and postulated that these models "struggle with understanding the physical world – likely due to the absence of physical knowledge

in their training data and the lack of embedded physical priors." In contrast, we find that LMMs struggle to adapt to an unseen scenario. It is also not clear if more training data can improve inductive physical reasoning from demonstration samples.

D Various Types of Reasoning

In this section, we will distinguish between general visual reasoning, inductive reasoning, and inductive physical reasoning (our work).

General visual reasoning refers to the broad set of tasks that involve answering questions from visual signals (one or more images and/or videos). To address these tasks, the model must extract information from the visual signal and apply auxiliary information that the model already has about the content of the visual signal. This information is generally factual. For example, the input image may contain a knife and a fruit, and the auxiliary information corresponding to this content is that knives are sharp and can be used to cut fruits (Aroca-Ouellette et al., 2021; Bisk et al., 2020; Wang et al., 2023; Dong et al., 2024). The skills required for general visual reasoning include localization, understanding, and information retrieval.

Visual reasoning is different from physical reasoning, although they share the input modalities and skill set partially. In physical reasoning, the task is to apply the physical knowledge possessed by the model. Unlike factual information, physical knowledge is a framework that is actionable only when applied to a specific context. For example, in the previous example of a knife next to a fruit, a relevant physical knowledge is that an object remains at rest unless acted upon by an external force (Newton's first law of motion). To use this physical knowledge, the model must not only localize and understand the objects in the scene, but also realize that Newton's first law of motion applies to the objects. In contrast, if one of the objects in the scene were a fluid, the model must realize that the laws of fluid dynamics also apply to that object. In summary, physical reasoning involves an additional application of mathematical frameworks over visual reasoning and is, therefore, more challenging.

Inductive reasoning is the ability of an agent to infer the underlying rules from a few samples and then apply these rules in a new evaluation scenario. Since the samples may not fully inform the agent about every underlying rule, inductive reasoning involves a degree of uncertainty. Existing works that evaluate inductive reasoning in LLMs follow this premise. As an example of inductive reasoning, consider the following sequence: A000, B001, C010, D011, E100. Which is the next element in this sequence? A possible (not necessarily unique) underlying rule of the sequence that we can infer from the premise is that the alphabets follow their canonical alphabetical order, while the remaining numbers encode the position of the alphabets in binary format. According to this rule, the next two elements in the sequence are F101 and G110.

Is evaluating inductive physical reasoning similar to evaluating inductive reasoning? Unlike our work on inductive physical reasoning, evaluating inductive reasoning does not contradict any existing knowledge in the models. For instance, it is unlikely that an evaluated model has any knowledge regarding the above sequence example of inductive reasoning. It is also possible that the model has never seen any sequence like that during training. In contrast, we are evaluating the ability of the model to adapt any existing physical knowledge that it might have to the evaluation scenario. Thus, inductive physical reasoning is not only inferring the underlying physics from demonstration samples but also doing so when the inferred physics potentially contradicts the parametric knowledge of the model.

Evaluating inductive reasoning in LLMs: There exists a long line of works that evaluate the abstract reasoning abilities of LLMs. Most of the other prior efforts to evaluate inductive reasoning are restricted to reasoning from textual inputs about abstract tasks (Mirchandani et al., 2023; Gendron et al., 2024; Wang et al., 2024c; Cheng et al., 2024; He et al., 2024; Bowen et al., 2024; Yan et al., 2025; Li et al., 2025). One notable work is Abstract and Reasoning Corpus (ARC) (Chollet, 2019), containing abstract tests similar to traditional IQ tests to evaluate "general artificial intelligence" in large models. It provides a training set that allows the candidate (human or machine learning agent) to understand the reasoning required to solve the tasks. Unlike our work, these tasks are largely symbolic and used to evaluate LLMs.

E Additional Experimental Evaluation

E.1 Effect of weight quantization on inductive physical reasoning

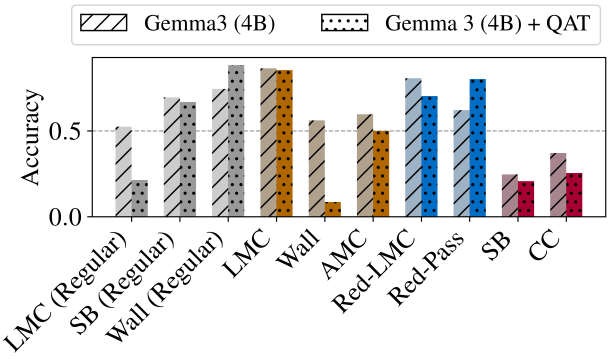


Figure 20: Effect of weight quantization on inductive physical reasoning

LLMs and LMMs are billion-parameter models with expensive inference. Therefore, weight quantization is used to reduce their resource needs. An important concern with weight quantization is the drop in performance due to lower precision. In this subsection, we examine whether weight quantization adversely affects inductive physical reasoning in LMMs. We compare the prediction accuracy of Gemma 3 (4B) with its quantized version, obtained by fine-tuning using Quantization Aware Training (QAT) (Jacob et al., 2018). The models are evaluated in both regular and irregular scenarios with exemplars including both videos and question-answer pairs.

Fig. 20 compares the accuracy of the quantized model against its non-quantized counterpart. Out of the ten evaluated scenarios, the quantized model performs worse than the full-precision model on all but two scenarios. In LMC, they perform comparably, while in Wall and LMC (regular), the quantized model shows a stark drop in accuracy. Notably, the quantized model outperforms the full precision model in Wall (regular) and Red-Pass. The results indicate that quantization indeed affects inductive physical reasoning, but its impact varies with the physical reasoning task.

E.2 Can Chain-of-Thought Prompting Improve Inductive Physical Reasoning?

The experiments in § 4 show that LMMs struggle to infer the underlying physical laws from demonstration samples and apply them for physical reasoning. These experiments also revealed the underlying language bias of these models. In this section, we evaluate if chain-of-thought (CoT) prompting (Wei et al., 2022) can help with inductive physical reasoning. Intuitively, CoT can alleviate the burden of inferring the underlying physical laws from video frames and multiple-choice question-answer pairs by enunciating the physical laws required for reasoning in the prompt.

LMM	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC	LMC (Regular)	SB (Regular)	Wall (Regular)
InternVL3-1B (Zhu et al., 2025)	93.20 (+47.65)	72.57 (+53.83)	93.13 (+13.23)	40.98 (-11.00)	68.30 (+13.18)	10.45 (+ 1.40)	36.88 (- 4.23)	63.77 (+29.00)	96.97 (+52.08)	87.72 (+79.62)
VideoLLaMA3-2B (Zhang et al., 2025)	99.62 (+65.03)	69.07 (+39.75)	97.58 (+34.17)	4.95 (-28.88)	37.17 (- 5.98)	31.72 (+24.00)	88.48 (+43.08)	98.25 (+34.00)	98.74 (+35.98)	97.13 (+56.15)
InternVL3-2B (Zhu et al., 2025)	99.30 (+81.33)	42.28 (+41.22)	96.32 (- 1.58)	0.65 (-16.72)	7.90 (-17.00)	35.70 (+35.52)	82.20 (+75.45)	99.52 (+11.17)	99.78 (+26.64)	99.82 (+11.63)
Gemma 3-4B (Kamath et al., 2025)	99.28 (+17.95)	99.03 (+48.87)	94.58 (+31.78)	15.02 (-49.87)	56.05 (-10.08)	33.97 (+ 7.27)	71.50 (+35.50)	95.93 (+25.23)	97.78 (+26.35)	99.32 (+26.43)
LLaVA-NeXT-Vid (Zhang et al., 2024)	76.08 (+59.27)	89.45 (+51.65)	67.88 (+18.95)	8.78 (- 2.83)	23.92 (+11.03)	28.43 (+ 9.35)	63.80 (+31.67)	74.52 (+16.12)	59.84 (+26.30)	78.15 (+69.38)
InternVL3-8B (Zhu et al., 2025)	100.00 (+ 6.53)	99.98 (+ 0.38)	100.00 (0)	3.28 (-66.18)	24.47 (-35.55)	51.10 (+ 5.27)	100.00 (+31.58)	100.00 (+ 0.13)	100.00 (+ 0.12)	99.68 (+ 3.52)
LLaVA-OneVision (Li et al., 2024a)	99.17 (+ 2.53)	99.67 (+ 1.82)	99.45 (- 0.05)	11.90 (-55.97)	33.98 (-32.16)	10.88 (-17.84)	99.70 (+34.75)	99.83 (+ 0.68)	99.90 (+ 0.79)	99.52 (+ 0.17)
VideoLLaMA3-7B (Zhang et al., 2025)	99.75 (+16.38)	99.28 (+16.63)	97.95 (+ 1.87)	3.30 (-39.87)	21.20 (-23.68)	47.38 (-13.78)	99.87 (+32.62)	99.90 (+ 1.60)	99.70 (+ 4.91)	94.02 (+45.32)
LLaVA-NeXT-IL (Li et al., 2024c)	99.33 (+ 7.53)	92.80 (- 3.10)	96.58 (+ 2.17)	7.00 (-64.62)	45.98 (- 4.57)	53.53 (+11.20)	99.80 (+ 2.15)	100.00 (+ 2.20)	91.76 (+21.15)	94.85 (+ 4.87)
Qwen2-VL (Wang et al., 2024a)	99.92 (+ 5.32)	99.93 (+ 0.15)	97.32 (+ 0.35)	3.12 (-94.42)	17.50 (-75.08)	30.98 (+15.92)	99.98 (+12.77)	100.00 (+ 2.43)	100.00 (+ 0.99)	99.85 (+ 1.18)
Qwen2.5-Omni (Xu et al., 2025)	99.40 (+29.05)	99.68 (+33.02)	97.78 (+ 6.75)	5.47 (-53.07)	30.27 (-22.68)	61.60 (+ 3.35)	94.18 (+51.97)	96.83 (+61.07)	99.48 (+16.37)	92.43 (+40.38)
Gemma 3-12B (Kamath et al., 2025)	84.82 (+22.43)	96.98 (+ 7.68)	96.57 (+ 2.25)	26.10 (+ 3.05)	69.60 (+47.17)	11.38 (- 5.08)	99.23 (+13.23)	83.38 (- 4.27)	94.08 (+10.67)	77.98 (+ 5.33)
Aria (Li et al., 2024b)	99.92 (+69.77)	98.65 (+97.10)	97.77 (+23.67)	3.65 (-27.70)	8.17 (-13.23)	30.73 (+20.97)	98.15 (+88.42)	99.33 (+42.63)	99.26 (+ 7.86)	99.85 (+17.87)

Table 8: Performance of various LMMs in both regular and irregular scenarios when evaluated using chain-of-thought (CoT) prompting, where the underlying physical law is explicitly included along with the answers in demonstration samples.

Tab. 8 shows the reasoning accuracies for various LMMs in both regular and irregular scenarios. The red and green numbers in the parentheses show the decrease and increase in the accuracy against the corresponding accuracy without CoT. We note a significant improvement in the performance across almost all model-scenario combinations, except Red-Pass and Red-LMC. CoT prompting fails to help multiple models in Red-Pass and Red-LMC scenarios, even worsening the performance of some models. In Red-Pass and Red-LMC, all except red colored objects follow the true physical laws. That is, unlike the remaining scenarios, Red-Pass and Red-LMC require the LMM to apply conditional reasoning depending on the color of the object. Although the demonstration samples include collisions with and without red colored objects, it appears that LMMs are unable to infer the conditional nature of the underlying reasoning.

E.3 Effect of Fine-tuning on Inductive Physical Reasoning

In this section, we explore the performance improvement on InPhyRe that we can obtain through fine-tuning. Fine-tuning through direct supervision and reinforcement learning has been shown to improve visual reasoning across diverse tasks (Zhai et al., 2024; Tan et al., 2025; Cai et al., 2025). Note that we cannot evaluate inductive physical reasoning by fine-tuning LMMs on InPhyRe samples, as it becomes unclear if the LMM's output is due to the fine-tuning or due to the inductive physical reasoning. We conduct this experiment as a proxy way to obtain an "upper bound" on what an LMM can achieve on InPhyRe. Since fine-tuning is an expensive process, we limit our experiment to the smallest LMM in the evaluated cohort, InternVL3-1B. We use supervised fine-tuning without any low-rank adaptation techniques such as LoRA (Hu et al., 2022). Since no single hyperparameter combination worked consistently well for all scenarios, presumably due to the diversity between the tasks, we report the results for all hyperparameter combinations from a grid search.

Epochs	LR	LMC (Reg.)	SB (Reg.)	Wall (Reg.)	LMC	Wall	AMC	Red-LMC	Red-Pass	SB	CC
20	2×10^{-5}	12.01 (-21.26)	42.97 (- 4.57)	18.29 (+ 8.99)	56.38 (+ 8.14)	23.72 (+ 5.23)	62.56 (-20.65)	68.57 (+10.13)	48.27 (- 7.72)	25.38 (+14.92)	58.44 (+10.50)
50	2×10^{-5}	32.36 (- 0.90)	35.16 (-12.38)	31.16 (+21.86)	54.27 (+ 6.03)	46.68 (+28.19)	66.48 (-16.73)	29.32 (-29.12)	35.79 (-20.20)	28.94 (+18.49)	70.40 (+22.46)
100	2×10^{-5}	38.09 (+ 4.82)	54.19 (+ 6.65)	42.21 (+32.91)	75.03 (+26.78)	93.47 (+74.97)	89.25 (+ 6.03)	42.76 (-15.69)	33.28 (-22.71)	56.48 (+46.03)	83.97 (+36.03)
120	2×10^{-5}	83.12 (+49.85)	92.85 (+45.31)	58.69 (+49.40)	56.83 (+ 8.59)	68.34 (+49.85)	75.63 (- 7.59)	54.94 (- 3.51)	50.08 (- 5.91)	38.94 (+28.49)	90.10 (+42.16)
20	2×10^{-4}	37.64 (+ 4.37)	25.72 (-21.82)	24.37 (+15.08)	0.70 (-47.54)	0.00 (-18.49)	7.09 (-76.13)	23.91 (-34.54)	28.02 (-27.97)	0.00 (-10.45)	28.44 (-19.50)
50	2×10^{-4}	73.57 (+40.30)	95.08 (+47.54)	27.99 (+18.69)	43.22 (- 5.03)	30.40 (+11.91)	81.86 (- 1.36)	27.42 (-31.03)	28.92 (-27.07)	23.92 (+13.47)	70.60 (+22.66)
100	2×10^{-4}	94.07 (+60.80)	30.70 (-16.84)	39.30 (+30.00)	27.84 (-20.40)	31.11 (+12.61)	38.84 (-44.37)	31.68 (-26.77)	34.49 (-21.50)	35.58 (+25.13)	42.11 (- 5.83)
120	2×10^{-4}	42.46 (+ 9.20)	71.08 (+23.54)	87.04 (+77.74)	30.80 (-17.44)	69.55 (+51.06)	59.90 (-23.32)	42.26 (-16.19)	24.91 (-31.08)	32.31 (+21.86)	35.43 (-12.51)
20	1×10^{-3}	0.00 (-33.27)	0.00 (-47.54)	14.42 (+ 5.13)	0.00 (-48.24)	0.00 (-18.49)	0.00 (-83.22)	0.00 (-58.45)	0.00 (-55.99)	0.00 (-10.45)	0.00 (-47.94)
50	1×10^{-3}	0.00 (-33.27)	1.78 (-45.76)	0.00 (- 9.30)	0.00 (-48.24)	0.00 (-18.49)	0.00 (-83.22)	0.00 (-58.45)	25.56 (-30.43)	0.00 (-10.45)	0.00 (-47.94)
100	1×10^{-3}	18.39 (-14.87)	24.15 (-23.39)	19.70 (+10.40)	29.10 (-19.15)	11.51 (- 6.98)	26.38 (-56.83)	6.37 (-52.08)	27.42 (-28.57)	15.78 (+ 5.33)	23.12 (-24.82)
120	1×10^{-3}	10.00 (-23.27)	8.73 (-38.81)	28.19 (+18.89)	29.95 (-18.29)	20.95 (+ 2.46)	23.27 (-59.95)	26.42 (-32.03)	0.05 (-55.94)	24.82 (+14.37)	21.36 (-26.58)
Best		94.07 (+60.80)	95.08 (+47.54)	87.04 (+77.74)	75.03 (+26.78)	93.47 (+74.97)	89.25 (+ 6.03)	68.57 (+10.13)	50.08 (- 5.91)	56.48 (+46.03)	90.10 (+42.16)

Table 9: Results of fine-tuned InternVL3-1B on regular and irregular scenarios for different hyperparameter combinations. The best results for each scenario are given in the last row.

Tab. 9 shows the accuracy of fine-tuned InternVL3-1B on both regular and irregular scenarios for different hyperparameter combinations. We observe that fine-tuning improves the performance in all scenarios with at least one hyperparameter combination, except for the Red-Pass scenario. These observations are similar to those from our experiments on CoT prompting. As we mentioned in the previous section, Red-Pass and Red-LMC require conditional reasoning, and it seems that fine-tuning cannot improve conditional physical reasoning in LMMs. The absolute accuracies are comparatively small for Red-LMC and SB.

E.4 Do the Results from CoT Prompting and Fine-Tuning Experiments Invalidate Inductive Physical Reasoning?

In the previous sections, we evaluated the effects of CoT prompting and fine-tuning on the prediction accuracy for InPhyRe benchmark. Although both CoT prompting and fine-tuning improved the performance of multiple LMMs on most scenarios (except Red-Pass and Red-LMC), we emphasize that neither of these solutions is viable when the inductive physical reasoning ability is put to the test in practice. Specifically, both these techniques require prior access to the underlying physical laws or the inference samples themselves.

In CoT prompting, we included the underlying physical law in the impossible scenario as part of the prompt. This would not be possible in unseen scenarios where we do not possess any information about the physical laws involved. In contrast, throughout this paper, we had instead made a milder assumption that we only had access to prior visual samples and the right response in those samples. The more arduous and crucial

task of inferring the underlying physical law was left to the model itself. Similarly, the models are trained on these samples in the fine-tuning experiments. Once these samples become part of the training data, it becomes nearly impossible to know if the model predictions are based on their inductive physical reasoning or because they remember their fine-tuning samples. Nonetheless, we included these experiments for the sake of completion, as they provide a surrogate for an upper bound on the performance achievable in InPhyRe.

E.5 Qualitative Inspection of Generated Outputs

In § 4.2, we evaluated the parametric knowledge of LMMs in regular scenarios. We used the accuracy with which the model chose the correct option as our evaluation metric. Although this metric is useful for quantifying physical reasoning, it does not inform us how the model arrived at the option. This lack of clarity is outside the scope of a quantitative benchmark like InPhyRe. Nonetheless, we look at some of the open-ended generation output of these LMMs in regular scenarios to obtain insights into the underlying physical reasoning in these models. For this experiment, we provide the LMMs with only the input image and the question. Answer options are **not** included in the prompt so that the model will resort to open-ended generation to answer the question. We do not conduct this experiment on irregular scenarios as (1) video-only sub-setting did not show any competent performance, and (2) video-text sub-setting cannot be conducted without providing the options in exemplars, which would lead to the model predicting just the option alone.

The following subsubsections show the open-ended generation outputs for LMC (regular), SB (regular), and Wall (regular). Horizontal lines are used to demarcate the outputs of each LMM. In each output, incorrect physical reasoning is highlighted in red, correct physical reasoning relevant to the task is highlighted in green, and partially true physical reasoning in yellow. If any information from the prompt is ignored or if new assumptions are hallucinated, we highlight them in blue. After each generation output, we describe our observations. We do not list observations for the models that do not provide sufficient information about their reasoning. Note that the outputs of the same LMM could vary between individual samples. Therefore, the outputs of an LMM for this specific evaluation sample may not be similar to the model's outputs on other evaluation samples from the same scenario.

E.5.1 Open-ended generation output for LMC (regular)

Frames: 

System prompt: Answer the question based on your understanding of physical laws.

Question: yellow cube and purple sphere have equal mass. How will the speed of yellow cube change after colliding with purple sphere?

InternVL3-1B (Zhu et al., 2025): If the yellow cube and the purple sphere have equal mass, the collision will result in both objects moving in opposite directions. Since the collision is elastic, the speed of the yellow cube will be halved. This is because the kinetic energy is conserved in an elastic collision.

Observations: InternVL3-1B hallucinated that the collision was elastic, instead of stating it as an assumption. Moreover, the speed will not be halved even if the collision is elastic.

VideoLLaMA3-2B (Zhang et al., 2025): After the collision, the yellow cube will move at a slower speed than before.

InternVL3-2B (Zhu et al., 2025): When the yellow cube collides with the purple sphere, the speed of the yellow cube will decrease. This is because the collision will transfer momentum from the yellow cube to the purple sphere, causing the yellow cube to slow down. The exact change in speed depends on the masses and the nature of the collision (elastic or inelastic). In an elastic collision, both momentum and kinetic energy are conserved, while in an inelastic collision, kinetic energy is not conserved, and some energy is converted into other forms, such as heat or deformation.

Observations: The reasoning is correct and derived from accurate physical laws.

Gemma3-4B (Kamath et al., 2025): Okay, let's analyze this collision based on physics principles.

Here's the breakdown:

1. **Conservation of Momentum:** In a closed system (like this one, assuming no external forces), the total momentum *before* a collision equals the total momentum *after* the collision. Momentum is calculated as mass (m) times velocity (v): p = mv.

2. **Newton's Third Law (Action-Reaction):** For every action, there is an equal and opposite reaction. This means that when the sphere hits the cube, the cube exerts an equal and opposite force back on the sphere.

3. **Collision Dynamics:** Because the cube and sphere have equal mass, and the collision is likely to be fairly brief, we can assume it's an *elastic* collision (or close to it – perfectly elastic collisions are rare in reality). In an elastic collision, kinetic energy is also conserved.

Now, let's apply this:

* **Before the collision:** Both the cube and sphere are at rest. Total momentum = 0. * **During the collision:** The sphere transfers its momentum to the cube. Since the masses are equal, the *change* in momentum for each object will be the same, but the *velocity* will change differently. * **After the collision:** The cube will move in one direction, and the sphere will move in the opposite direction. Because the masses are equal, the *speeds* of the cube and sphere will be equal, but in opposite directions.

Therefore, the speed of the yellow cube will change, and it will move with a speed equal in magnitude, but opposite in direction, to the speed the purple sphere had before the collision.

Important Note: This assumes a *perfect* elastic collision. In a real-world scenario, some kinetic energy would be lost as heat or sound due to friction between the objects, so the speeds after the collision wouldn't be *exactly* equal. However, the principle of conservation of momentum and Newton's Third Law still apply.

Would you like me to explain any part of this in more detail or consider a slightly different collision scenario?

Observations: Gemma3-4B incorrectly assumed that objects would be at rest before the collision. It also contradicts itself later and states that the purple sphere was in motion before the collision.

LLaVA-NeXT-Video (Zhang et al., 2024): The speed of the yellow cube will not change after colliding with the purple sphere because they have equal mass. The conservation of momentum is a fundamental principle in physics that states that the total momentum of a closed system remains constant. When two objects collide, the total momentum of the system is the sum of the momenta of the individual objects. If the objects have equal mass, then the momentum of the system is the same before and after the collision.

Observations: The stated law of momentum conservation is accurate but is applied incorrectly. Additionally, momentum conservation is a universal law and does not depend on the masses of the individual objects.

InternVL3-8B (Zhu et al., 2025): To determine how the speed of the yellow cube will change after colliding with the purple sphere, we need to consider the principles of conservation of momentum and kinetic energy, assuming an elastic collision.

1. **Conservation of Momentum**: In an elastic collision, the total momentum before the collision is equal to the total momentum after the collision. If the masses of the yellow cube and the purple sphere are equal, and they are initially at rest, the momentum before the collision is zero. After the collision, the momentum of the cube and the sphere will be equal in magnitude but opposite in direction.

2. **Conservation of Kinetic Energy**: In an elastic collision, the total kinetic energy before the collision is equal to the total kinetic energy after the collision. Since the masses are equal and initially at rest, the kinetic energy before the collision is zero. After the collision, the kinetic energy will be distributed between the cube and the sphere.

Given these principles, let's analyze the situation:

- **Initial State**: Both the yellow cube and the purple sphere are at rest. Therefore, the initial momentum and kinetic energy are zero.

- **Collision**: When the yellow cube collides with the purple sphere, they will exchange velocities due to the conservation of momentum and kinetic energy. Since the masses are equal, the velocities will be equal in magnitude but opposite in direction.

- **Final State**: After the collision, the yellow cube will move in the opposite direction of the purple sphere with the same speed it had before the collision, and the purple sphere will move in the opposite direction of the yellow cube with the same speed it had before the collision.

Therefore, if the yellow cube and the purple sphere have equal mass and the collision is perfectly elastic, the speed of the yellow cube will remain the same after the collision, but its direction will be reversed.

Observations: InternVL3-8B states the physical laws correctly. However, it incorrectly believes that the objects are at rest before collision. This hallucinated assumption led it to provide the wrong reasoning.

LLaVA-Onevision (Li et al., 2024a): The speed of the yellow cube will decrease after colliding with the purple sphere. This is because the cube will transfer some of its kinetic energy to the sphere, causing it to slow down. The exact change in speed will depend on the masses and velocities of the cube and sphere before the collision, as well as the angle of the collision.

Observations: The stated law of kinetic energy transfer is accurate. The model seems to have ignored the information about the mass of the objects stated in the prompt.

VideoLLaMA3-7B (Zhang et al., 2025): decrease

LLaVA-NeXT-Interleave (Li et al., 2024c): The speed of the yellow cube will change after colliding with the purple sphere. The exact change depends on the nature of the collision, which is not specified in the image. If the collision is elastic and the yellow cube and purple sphere have equal mass, the speed of the yellow cube will change to a value that depends on the angle of impact and the initial speeds of the two objects.

Observations: The reasoning is accurate, although the reasoning principle is not explicitly stated. It additionally states that the exact change in velocity will depend on factors that are unavailable from the prompt.

Qwen2-VL (Wang et al., 2024a): The speed of the yellow cube will decrease after colliding with the purple sphere.

Qwen2.5-Omni (Xu et al., 2025): The speed of the yellow cube will decrease after colliding with the purple sphere. This is because, according to the law of conservation of momentum, the total momentum of a closed system remains constant. When the yellow cube collides with the purple sphere, they exert equal and opposite forces on each other. Since they have equal mass, the yellow cube will experience a force that will slow it down, while the purple sphere will experience a force that will accelerate it.

Observations: Qwen2-5-Omni reasoned that the yellow cube will slow down due to the objects having equal mass. In reality, the mass of the objects do not decide whether the colliding object slows down or not, but only how much it slows down.

Gemma3-12B (Kamath et al., 2025): Here's an analysis of what would happen, based on the laws of physics, assuming an elastic collision:

Understanding the Principles

* **Conservation of Momentum:** In a closed system (meaning no external forces like friction), the total momentum before a collision is equal to the total momentum after the collision. Momentum is mass x velocity.
* **Elastic Collision:** This is a collision where kinetic energy is conserved. In reality, collisions are rarely perfectly elastic (some energy is typically lost as heat or sound), but we'll assume it for simplicity.

Analysis of the Scenario

1. **Initial State:** The cube is moving, and the sphere is stationary. Let's say the cube's velocity is *v*. The sphere's velocity is 0. 2. **Collision:** During the collision, the cube transfers some of its momentum to the sphere. 3. **Final State (Ideal Elastic Collision):** Since the masses are equal, the collision will result in the cube stopping and the sphere moving off with a velocity equal to the initial velocity of the cube (*v*).

Therefore, the speed of the yellow cube will change and it will come to a complete stop.

Important Considerations:

* **Perfect Elasticity:** The assumption of a perfectly elastic collision is rarely true. In a real-world scenario, the cube would likely not stop completely. Some of the cube's kinetic energy would be converted to other forms (sound, heat due to deformation). * **Direction:** This analysis only considers speed (magnitude of velocity). The direction of motion after the collision would also change.

Observations: Gemma3-12B states the physical principles accurately and provides the correct reasoning using these principles.

Aria (Li et al., 2024b): In a collision between two objects of equal mass, assuming an elastic collision, the speed of each object will change such that the yellow cube will come to a stop, and the purple sphere will move forward at the initial speed of the yellow cube. This is due to the principle of conservation of momentum and kinetic energy. The yellow cube transfers all its momentum and kinetic energy to the purple sphere.

Observations: The laws stated by Aria are correct, and the laws are applied accurately during physical reasoning.

E.5.2 Open-ended generation output for SB (regular)

Frames: 

System prompt: Answer the question based on your understanding of physical laws.

Question: green cylinder has more mass than blue cube. What occurs when green cylinder and blue cube collide?

InternVL3-1B (Zhu et al., 2025): When the green cylinder and the blue cube collide, the following physical law applies:

The collision between the two objects will result in a change in their velocities. Specifically, the blue cube will move forward with a higher velocity than the green cylinder, and the green cylinder will move backward with a lower velocity. This is because the blue cube has more mass and thus a greater inertia, causing it to move faster than the green cylinder.

The direction of the velocities can be determined using the principle of conservation of momentum. The change in momentum of the blue cube will be equal to the change in momentum of the green cylinder. Mathematically, this can be expressed as:

$$\Delta p_{\text{blue}} = \Delta p_{\text{green}}$$

Where: - Δp_{blue} is the change in momentum of the blue cube. - Δp_{green} is the change in momentum of the green cylinder.

Since the blue cube has more mass, it will have a greater change in momentum, leading to a higher velocity for the blue cube. Conversely, the green cylinder will have a lower change in momentum and thus a lower velocity.

Thus, the blue cube will move forward with a higher velocity than the green cylinder.

Observations: InternVL3-1B confused the unequal mass relation between the objects from the prompt. The prompt said that the green cylinder has more mass than the blue cube, but the model inferred that the blue cube had more mass than the green cylinder. It additionally reasoned that an object with more mass will undergo a greater change in momentum. However, these mistakes canceled each other and accidentally led the model to the right answer.

VideoLLaMA3-2B (Zhang et al., 2025): green cylinder and blue cube collide

---

InternVL3-2B (Zhu et al., 2025): When the green cylinder and the blue cube collide, the green cylinder will likely deform or compress due to the force exerted by the blue cube. The blue cube will exert a force on the green cylinder, causing it to move or change shape. The exact outcome will depend on the masses, velocities, and other physical properties of the objects involved.

Observations: InternVL3-2B assumed about the material properties of the objects and reasoned that the green cylinder could potentially deform. It also incorrectly reasoned that the outcome would depend on the mass and the velocities. Since information was provided about the mass, the correct answer is that both objects would move in the same direction.

Gemma3-4B (Kamath et al., 2025): Here's what will likely happen when the green cylinder and blue cube collide, based on our understanding of physics:

1. Impulse and Momentum Conservation:

* **Impulse:** When the objects collide, they exert forces on each other for a brief period. This exchange of forces is called impulse. * **Momentum:** The total momentum of the system (cylinder + cube) before the collision must equal the total momentum after the collision. Momentum is mass times velocity (p = mv).

2. The Outcome – The Cylinder Will Move Less

* Since the green cylinder has more mass than the blue cube, it will have more momentum before the collision. * During the collision, the cylinder will decelerate (slow down) more quickly than the cube. * Because of its greater initial momentum, it will transfer more momentum to the cube. * As a result, the blue cube will experience a larger change in velocity, resulting in a greater deflection from its initial direction, while the green cylinder will experience a smaller change in velocity and will move less significantly.

In simpler terms: The heavier object (cylinder) will have a harder time changing its motion during the collision compared to the lighter object (cube).

Important Note: This assumes an *elastic* collision (no energy loss due to heat, sound, or deformation). If the collision was inelastic (e.g., the objects deform significantly and some energy is lost), the outcome would be slightly different.

If you would like, you can provide another scenario and I can analyze it using the same principles.

Observations: Gemma3-4B stated several physical laws accurately but applied them incorrectly. It also incorrectly stated that an object with a larger initial momentum will transfer more momentum.

LLaVA-NeXT-Video (Zhang et al., 2024): When a green cylinder and a blue cube collide, the collision will result in a change in their motion and possibly a change in their position relative to each other. The exact outcome depends on the relative speed and direction of the cylinder and the cube at the time of collision. If the cylinder is moving faster or in the same direction as the cube, the cylinder will push the cube to one side, and if the cylinder is moving in the opposite direction, the cube will push the cylinder to the other side. If the cylinder is moving at a right angle to the cube, the cylinder will change direction and the cube will change direction. The collision will also result in a transfer of momentum from the cylinder to the cube or vice versa, depending on the direction of the collision.

If the cylinder is moving faster than the cube, the cylinder will push the cube to one side, and the cube will move in the opposite direction. If the cylinder is moving slower, the cube will push the cylinder to the other side. If the cylinder is moving at a right angle to the cube, the cylinder will change direction and the cube will change direction.

Observations: Although the stated reasoning is true, the model does not use the information about the mass of the objects provided in the prompt. The reasoning is incomplete due to the model ignoring useful information from the prompt.

InternVL3-8B (Zhu et al., 2025): When the green cylinder and the blue cube collide, several factors determine the outcome, primarily based on the principles of physics, specifically Newton's laws of motion and the concept of momentum.

1. **Mass and Momentum**: The green cylinder has more mass than the blue cube. Momentum is the product of mass and velocity (p = mv). If both objects are moving with the same velocity before the collision, the green cylinder will have more momentum due to its greater mass.

2. **Elastic vs. Inelastic Collision**: The nature of the collision (whether it is elastic or inelastic) also plays a role. In an elastic collision, both momentum and kinetic energy are conserved. In an inelastic collision, momentum is conserved, but kinetic energy is not. The objects may stick together or deform upon impact.

3. **Resultant Motion**: - **Elastic Collision**: If the collision is perfectly elastic, the green cylinder will exert a greater force on the blue cube, potentially causing the blue cube to rebound with some velocity. The green cylinder might slow down slightly due to the transfer of momentum. - **Inelastic Collision**: If the collision is inelastic, the objects may stick together after the collision. The combined mass of the green cylinder and the blue cube will move with a velocity determined by the conservation of momentum.

4. **Energy Considerations**: In an elastic collision, the kinetic energy before and after the collision remains the same. In an inelastic collision, some kinetic energy is converted into other forms of energy, such as heat or deformation energy.

In summary, when the green cylinder and the blue cube collide, the green cylinder, having more mass, will exert a greater force on the blue cube. Depending on whether the collision is elastic or inelastic, the blue cube may rebound or stick to the green cylinder, and the green cylinder may slow down. The exact outcome depends on the initial velocities and the nature of the collision.

Observations: InternVL3-8B states the principles in physics correctly and applies them accurately.

LLaVA-Onevision (Li et al., 2024a): The green cylinder will likely push the blue cube away from it due to its greater mass and momentum. The blue cube will also experience a force in the opposite direction of the green cylinder's motion. The exact outcome of the collision will depend on the speed and angle of the green cylinder's impact with the blue cube.

Observations: LLaVA-Onevision states and applies the physical laws correctly.

VideoLLaMA3-7B (Zhang et al., 2025): green cylinder will not move

---

LLaVA-NeXT-Interleave (Li et al., 2024c): When the green cylinder and blue cube collide, the blue cube will likely move and change its position as a result of the interaction. The green cylinder may also change its position or velocity depending on the nature of the collision.

Observations: LLaVA-NeXT-Interleave reasons that the green cylinder *may* change its position or velocity, while, in reality, it would always change its position due to its larger mass, irrespective of the nature of the collision.

Qwen2-VL (Wang et al., 2024a): The green cylinder will push the blue cube.

Qwen2.5-Omni (Xu et al., 2025): When the green cylinder collides with the blue cube, the cylinder will exert a force on the cube, causing the cube to accelerate and move in the direction of the force. The cylinder will also experience a reaction force, which will cause it to slow down slightly. The exact outcome of the collision depends on the masses and velocities of the objects, as well as the nature of the collision (e.g., elastic or inelastic).

Observation: Qwen2.5-Omni correctly states and applies the relevant physical laws.

Gemma3-12B (Kamath et al., 2025): Here's what would likely happen based on physical laws, considering the image and the premise that the green cylinder has more mass than the blue cube:

The Collision

When the green cylinder and the blue cube collide, the following will generally occur:

1. **Momentum Transfer:** Momentum is a measure of mass in motion (mass x velocity). Since the green cylinder has more mass, it will have more momentum. During the collision, some of the green cylinder's momentum will be transferred to the blue cube.

2. **Green Cylinder's Slowdown:** The green cylinder, having lost some momentum, will slow down after the collision.

3. **Blue Cube's Speedup:** The blue cube, gaining momentum, will speed up after the collision. It will move in roughly the direction the green cylinder was traveling.

4. **Elasticity Matters:** The specifics will depend on how *elastic* the collision is: * **Perfectly Elastic Collision:** (Rare in the real world). Kinetic energy is conserved. The cylinder would slow down significantly, the cube would speed up significantly, and the total kinetic energy would remain the same. * **Inelastic Collision:** (More common). Some kinetic energy is lost as heat or sound during the impact. The cylinder will slow down more, and the cube will speed up less than in a perfectly elastic collision.

In summary: The green cylinder will slow down, and the blue cube will speed up. The greater the mass difference, the more pronounced these effects will be.

Important Note: This explanation assumes a simplified scenario. Factors like friction, the exact shapes of the objects, and how they bounce off each other would influence the outcome in a real-world situation.

Observations: The reasoning is accurate, although the laws were not explicitly stated.

Aria (Li et al., 2024b): When the green cylinder and blue cube collide, the green cylinder, having more mass, will likely exert a greater force on the blue cube due to its greater momentum. This could cause the blue cube to move or change direction, depending on the angle and speed of the collision. The outcome will also depend on factors such as the material properties and the speed of the objects. If the collision is elastic, both objects might bounce off each other; if inelastic, some energy may be lost as heat or sound.

Observations: Aria restates the information about the mass of the objects from the prompt, but reasons that the objects may bounce off each other, which cannot happen when one object has a higher mass than the other.

E.5.3 Open-ended generation output for Wall (regular)

Frames: 

System prompt: Answer the question based on your understanding of physical laws.

Question: black cylinder collides elastically with the wall. What is the outcome of black cylinder colliding with wall?

InternVL3-1B (Zhu et al., 2025): When a black cylinder collides elastically with the wall, the outcome depends on the specific nature of the collision. Here are a few possible scenarios:

1. **Inelastic Collision:** - If the collision is inelastic, meaning the two objects stick together after the collision, the black cylinder will move with some velocity after the collision. The wall will exert a force on the moving cylinder, causing it to deform or change direction.

2. **Elastic Collision:** - If the collision is elastic, meaning the two objects bounce off each other, the black cylinder will move with some velocity after the collision. The wall will exert a force on the moving cylinder, causing it to change direction and possibly deform.

3. **Perfectly Inelastic Collision:** - If the collision is perfectly inelastic, meaning the two objects stick together, the black cylinder will move with some velocity after the collision. The wall will exert a force on the moving cylinder, causing it to change direction and possibly deform.

Without additional information about the specific nature of the collision, it is not possible to determine the exact outcome. However, in general, the outcome will depend on whether the collision is inelastic or elastic.

Observations: InternVL3-1B ignores the information in the prompt that the collision was elastic. It also generates a contradictory statement that the objects will move with some velocity after sticking to each other if the collision is inelastic.

VideoLLaMA3-2B (Zhang et al., 2025): black cylinder bounces

InternVL3-2B (Zhu et al., 2025): When a black cylinder collides elastically with a wall, the outcome is that the cylinder will bounce back with the same speed and direction it initially had. This is because the collision is perfectly elastic, meaning there is no loss of kinetic energy. The wall exerts an equal and opposite force on the cylinder, causing it to reverse its motion.

Observation: InternVL3-2B states and applies physical laws accurately.

Gemma3-4B (Kamath et al., 2025): Here's an analysis of what happens when a black cylinder collides elastically with a wall, based on the principles of physics:

Understanding Elastic Collisions

* **Elastic Collision:** In an elastic collision, both momentum and kinetic energy are conserved. This means the total momentum of the system before the collision equals the total momentum after the collision, and the total kinetic energy before equals the total kinetic energy after. * **Conservation of Momentum:** The law of conservation of momentum states that the total momentum of a closed system remains constant. * **Conservation of Kinetic Energy:** The law of conservation of kinetic energy states that the total kinetic energy of a system remains constant.

What Happens in the Collision

1. **Momentum Transfer:** As the cylinder hits the wall, momentum is transferred from the cylinder to the wall. The cylinder initially has a certain velocity (let's call it *v*), and the wall is initially at rest. After the collision, the cylinder will likely have a velocity in the opposite direction (let's call it *-v*), and the wall will have a velocity (*v*). 2. **Angle of Reflection:** The cylinder won't bounce straight back. It will bounce at an angle. The angle of reflection depends on the relative velocities of the cylinder and the wall *before* the impact. A more detailed calculation would require knowing the initial velocity and angle, and the coefficient of restitution (which describes the "bounciness" of the collision). 3. **Kinetic Energy is Preserved:** Since it is an elastic collision, the kinetic energy of the system will remain the same, meaning no energy will be lost as heat or sound.

Outcome

The black cylinder will bounce off the wall at an angle, and the wall will move in the opposite direction with the same speed as the cylinder.

Important Note: Without knowing the initial velocity and angle of the cylinder, it's impossible to predict the exact angle of reflection.

Let me know if you'd like a more detailed explanation or want to explore scenarios with specific initial conditions!

Observations: Although Gemma3-4B states and uses the appropriate physical laws, it treats the wall as an object that can move, leading to a wrong prediction.

LLaVA-NeXT-Video (Zhang et al., 2024): The image you've provided is a simple illustration of a black cylinder colliding with a wall, but it doesn't provide enough information to determine the outcome of the

collision. The outcome would depend on the properties of the cylinder and the wall, such as their mass, size, material, and the speed at which they collide.

If the cylinder is a solid object and the collision is elastic, it would bounce off the wall without any permanent deformation or damage to either object. The cylinder would lose some of its kinetic energy during the collision, and the wall would also experience a deformation, but it would return to its original state after the collision.

If the cylinder is a liquid or a gas, the collision would be more complex, and the outcome would depend on the properties of the liquid or gas, such as its viscosity and the pressure. The cylinder would likely spread out upon impact, and the wall would experience a deformation, but the cylinder would not bounce back.

If the cylinder is a solid and the collision is inelastic, the cylinder would break or deform upon impact, and the wall would also experience a deformation.

Without more information, it's impossible to provide a specific outcome.

Observations: As one of the possibilities, LLaVA-NeXT-Video assumes the collision is elastic and correctly predicts that the colliding object would bounce off the wall. However, it then adds wrong physical reasoning that the object would lose its kinetic energy and the wall would be deformed, neither of which is possible during an elastic collision. It also hallucinates the extremely unlikely situation that the object could be a liquid or a gas as another possibility.

InternVL3-8B (Zhu et al., 2025): When a black cylinder collides elastically with a wall, the outcome can be described using the principles of elastic collisions and conservation of momentum and kinetic energy.

1. **Elastic Collision**: In an elastic collision, both momentum and kinetic energy are conserved. This means that the total momentum before the collision is equal to the total momentum after the collision, and the total kinetic energy before the collision is equal to the total kinetic energy after the collision.

2. **Collision with a Wall**: When the black cylinder collides with the wall, the wall is considered to be stationary and much more massive than the cylinder. This simplifies the analysis because the wall does not move significantly.

3. **Reversal of Velocity**: Since the wall is stationary and much more massive, the black cylinder will reverse its direction of motion upon collision. The velocity of the cylinder will change direction but not magnitude, assuming no other forces act on it (like friction or air resistance).

4. **Conservation of Momentum**: The momentum of the cylinder before the collision (p_initial = m * v_initial) will be equal to the momentum after the collision (p_final = m * v_final). Since the wall is stationary, the momentum of the wall remains zero. Therefore, the momentum of the cylinder before the collision is equal to the momentum of the cylinder after the collision but in the opposite direction.

5. **Conservation of Kinetic Energy**: The kinetic energy of the cylinder before the collision (KE_initial = 0.5 * m * v_initial$\hat{2}$) will be equal to the kinetic energy after the collision (KE_final = 0.5 * m * v_final$\hat{2}$). Since the magnitude of the velocity remains the same, the kinetic energy is conserved.

In summary, the outcome of the black cylinder colliding elastically with the wall is that the cylinder will rebound with the same speed but in the opposite direction. The wall remains stationary, and the collision conserves both momentum and kinetic energy.

Observation: InternVL3-8B states and applies the true physical laws to obtain the accurate output.

LLaVA-Onevision (Li et al., 2024a): The black cylinder will bounce back and move away from the wall.

VideoLLaMA3-7B (Zhang et al., 2025): The black cylinder will bounce back with the same speed it had before colliding with the wall.

LLaVA-NeXT-Interleave (Li et al., 2024c): The black cylinder collides elastically with the wall, which means that the kinetic energy is converted into potential energy, and the cylinder rebounds in the opposite direction. The final velocity of the cylinder after the collision is zero, as it is moving at the same speed as the wall before the collision.

Observations: LLaVA-NeXT-Interleave incorrectly states that kinetic energy is converted to potential energy during an elastic collision. Then it generates contradictory statements that the cylinder would rebound while having zero final velocity. The model also hallucinates that the wall was at the same speed as the cylinder before the collision.

Qwen2-VL (Wang et al., 2024a): The black cylinder will bounce off the wall.

Qwen2.5-Omni (Xu et al., 2025): When a black cylinder collides elastically with a wall, the outcome depends on the initial velocity and direction of the cylinder. In an elastic collision, both momentum and kinetic energy are conserved. If the cylinder is moving towards the wall, it will bounce back with the same speed but in the opposite direction. If the cylinder is moving away from the wall, it will continue moving in the same direction with the same speed. The exact outcome can be calculated using the conservation of momentum and kinetic energy equations.

Observations: Qwen2.5-Omni claims that the outcome would depend on the initial velocity of the cylinder and its direction. While the exact outcome would depend on the initial velocity, the general outcome that the cylinder would bounce back is independent of the velocity when the collision is elastic. After making the correct prediction with the appropriate reasoning, it also considers the scenario where the object was initially moving away from the wall, without realizing that the collision would not have occurred in that scenario.

Gemma3-12B (Kamath et al., 2025): Here's an analysis of what happens when the black cylinder collides elastically with the wall, based on physical laws:

Understanding Elastic Collisions

* **Elastic Collision:** In an elastic collision, both momentum and kinetic energy are conserved. This means: * The cylinder will bounce off the wall. * No energy is lost during the collision (no heat, sound, or deformation). * The wall is assumed to be rigid, meaning it doesn't move or deform.

Outcome of the Collision

* **Reversal of Velocity:** The key outcome is that the cylinder will reverse its direction of motion. If it was traveling towards the wall, it will now travel away from the wall. * **Speed is the same:** Because kinetic energy is conserved, the cylinder's speed immediately *after* the collision will be the same as its speed *before* the collision. * **Angle:** The angle at which it bounces off will depend on the angle of incidence (the angle at which it hit the wall). The angle of incidence equals the angle of reflection.

Therefore, the black cylinder will bounce back away from the wall with the same speed it had before the collision, and with a trajectory determined by the laws of reflection.

Important Note: True, perfectly elastic collisions are rare in the real world. However, for a solid, rigid object like a cylinder colliding with a solid wall, we often *approximate* the collision as elastic for simplicity in analysis.

Observations: Gemma3-12B generates correct physical laws and uses them for accurate reasoning.

Aria (Li et al., 2024b): In an elastic collision, the black cylinder would bounce off the wall with the same speed but in the opposite direction. This means if the cylinder was moving towards the wall, it would move away from the wall at the same speed after the collision. The direction of motion would be reversed, but the magnitude of the velocity would remain unchanged.

Observation: Aria provides the correct reasoning, although it does not state the underlying physical laws.

Conclusion: Most models can state the underlying physical laws, but often hallucinate assumptions such as the wall moving or the objects deforming (despite the collision described as "elastic" in the prompt). In some cases, such as Gemma3-4B and InternVL3-1B, the generated output has contradicting statements about the outcome of the collision.

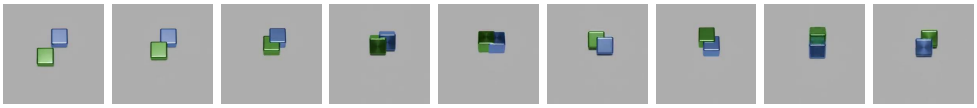
Overall Conclusion about Parametric Knowledge: Most evaluated LMMs can state the principles of momentum and energy conservation accurately. However, they struggle to apply these laws to reasoning. In

some cases, their reasoning contradicts the earlier parts of their output. Sometimes, they also hallucinate new assumptions (*e.g.*, about the object material) that stray their output from the correct reasoning. We conjecture that the LMMs know these physical laws as they know factual information, and not as mathematical models that must be applied flexibly.

E.6 Explaining the apparent inductive physical reasoning in AMC

In § 4.4, we observed that LMMs performed better in AMC compared to LMC (Reg.) when exemplars containing videos and question-answer pairs were provided. This is surprising since LMC (Reg.) followed universal physical laws and AMC did not. An LMM that can correctly infer custom physical laws from exemplars must also be able to do so when the scenarios follow the true physical laws. To explain this peculiar behavior of LMMs, we will manually inspect the open-ended generation outputs of LMMs on AMC. We follow the procedure from § E.5: the evaluation sample consists of the video and the question without any answer options. We also add "Will any object rotate?" at the prompt's end to elicit an output about the angular momentum conservation from the LMMs. We follow the highlighting style in the previous section. The outputs of various LMMs are given below.

E.6.1 Open-ended generation output for AMC (regular)

Frames: 

System prompt: Answer the question based on your understanding of physical laws.

Question: blue cube and green cube have equal mass. What is the outcome of green cube colliding with blue cube? Will any object rotate?

InternVL3-1B (Zhu et al., 2025): When two cubes collide, the outcome depends on the nature of the collision. In this case, the blue cube and the green cube have equal mass.

1. **Collision Outcome**: - If the collision is elastic, the two cubes will stick together after the collision. This means they will form a single larger cube. - If the collision is inelastic, the two cubes will separate and move apart.

2. **Rotation of Objects**: - In an elastic collision, the two cubes will stick together, so they will not rotate. - In an inelastic collision, the two cubes will separate, and they will not rotate.

Given that the collision is elastic, the outcome is that the two cubes will stick together. Therefore, the outcome of the collision is that the two cubes will stick together.

Observations: InternVL3-1B states illogical physical laws that objects stick together after an elastic collision and that objects stay separate after an inelastic collision. In reality, elastic collision ensures that the objects stay apart, while inelastic collision may result in objects sticking together. Interestingly, the output of the model is independent of this assumption as the model believes that neither elastic nor inelastic collision will result in the objects rotating.

VideoLLaMA3-2B (Zhang et al., 2025): When the green cube collides with the blue cube, both objects will rotate around their respective axes of symmetry. The collision is perfectly elastic and frictionless, meaning no energy or momentum is lost during the interaction.

Observation: VideoLLaMA3-2B incorrectly reasons that objects will rotate about the axes of symmetry due to the collision, but it does not provide any reasoning.

InternVL3-2B (Zhu et al., 2025): When the green cube collides with the blue cube, the outcome depends on the nature of the collision. If the collision is perfectly elastic, both cubes will continue to move after the collision, and no rotation will occur. If the collision is perfectly inelastic, the two cubes will stick together, and one of them will rotate while the other moves.

Since the problem states that the green cube and the blue cube have equal mass, we can infer that the collision is likely to be perfectly elastic. In this case, both cubes will continue to move after the collision, and no rotation will occur.

Observations: InternVL3-2B reasons correctly that no rotation will happen if the collision is elastic. However, it reasons that one of the objects may rotate while the other moves in the case of an inelastic collision. This is not true. Angular momentum is not preserved in the model's reasoning.

Gemma3-4B (Kamath et al., 2025): Okay, let's analyze this collision based on physical principles.

Here's a breakdown of what will likely happen:

1. **Conservation of Momentum:** The most important principle here is the conservation of momentum. In a closed system (assuming no external forces like friction or air resistance), the total momentum *before* the collision equals the total momentum *after* the collision. Momentum is calculated as mass x velocity.

2. **Equal Masses:** We are told that both cubes have the same mass. This simplifies the calculation considerably.

3. **Collision Dynamics:** Because both cubes have equal mass, when the green cube hits the blue cube, they will exert equal and opposite forces on each other. This means the green cube will recoil backward and the blue cube will move forward.

4. **Rotation:** Here's the key: Because the cubes are colliding with equal and opposite forces and they have the same mass, the system will conserve angular momentum. This means the overall rotation of the two cubes will be *zero*. While *each individual cube might rotate slightly* during the impact (due to the unevenness of the collision surface), the *combined* rotation of the pair will cancel out.

Therefore, the outcome is:

* **The green cube will recoil backward, and the blue cube will move forward.** * **Neither cube will rotate noticeably.** The rotation will be minimal and quickly damped out.

Important Note: This analysis assumes an *elastic* collision, where kinetic energy is conserved. In reality, some energy will be lost as heat and sound due to the inelastic deformation of the cubes during the impact. However, the fundamental principle of angular momentum conservation still holds.

Would you like me to delve deeper into a specific aspect of this, such as calculating the resulting velocities (if we knew the initial velocities)?

Observations: Gemma3-4B states the principle of angular momentum conservation and uses it correctly to predict that no rotation will happen.

LLaVA-NeXT-Video (Zhang et al., 2024): The image you've provided shows two cubes, one blue and one green, with equal mass. If they were to collide, the outcome would depend on the direction of the collision and the relative orientation of the cubes. If they were to collide head-on, the collision would be perfectly elastic, and no rotation would occur. If they were to collide at an angle, the cubes could rotate due to the conservation of momentum and the transfer of momentum.

In an elastic collision, the total momentum of the system is conserved, so if the blue cube is moving faster than the green cube, it would transfer its momentum to the green cube, causing it to move in the opposite direction. If the green cube is moving faster, it would transfer its momentum to the blue cube, causing it to move in the same direction. If they are moving at the same speed, they would pass each other without any change in momentum.

If the cubes were to collide at an angle, the collision would be inelastic, and the cubes would deform and possibly change shape. The cube that is struck would absorb the energy of the collision, and the other would rebound. The cube that absorbed the energy would rotate due to the conservation of angular momentum. The cube that rebound would have a change in its orientation, and the cube that struck it would have a change in its position.

Without more information about the specifics of the collision, it's impossible to predict the exact outcome.

Observations: LLaVA-NeXT-Video correctly states that no rotation will happen during an elastic collision. However, it additionally outputs that rotation is possible due to the conservation of momentum. Later, the model states that an object that absorbs energy could rotate due to the conservation of angular momentum. Here, the model states the right principle that objects could rotate to preserve angular momentum. But its output is wrong as the principle does not affect the rotation in this particular example.

InternVL3-8B (Zhu et al., 2025): When two objects with equal mass collide, the outcome depends on the nature of the collision (elastic or inelastic) and the initial velocities of the objects.

In an elastic collision, both momentum and kinetic energy are conserved. If the green cube and the blue cube have equal mass and they collide head-on with equal and opposite velocities, they will exchange velocities after the collision. This means the green cube will stop and the blue cube will continue moving with the velocity of the green cube before the collision.

In an inelastic collision, some kinetic energy is converted into other forms of energy (like heat or deformation), and the objects may stick together after the collision. If the green cube and the blue cube stick together after the collision, they will move as a single combined mass with a velocity determined by the conservation of momentum.

In either case, if the collision is perfectly head-on and the objects are rigid, they will likely rotate if they are not perfectly smooth or if there is friction between them. The rotation will depend on the distribution of mass and the point of contact during the collision.

So, yes, both objects could rotate depending on the specifics of the collision and the properties of the cubes.

Observation: The model outputs the true outcome of the collision first. To answer our specific question about the rotation, it outputs that the objects could rotate if there is friction between them. However, this is not possible if the collision is head-on as the model states.

LLaVA-Onevision (Li et al., 2024a): The outcome of the green cube colliding with the blue cube will depend on the angle and speed of the collision. If the collision is head-on and at a high speed, the cubes may not rotate at all. However, if the collision is at an angle, the cubes may rotate around their point of contact. The exact outcome would depend on the specific details of the collision, such as the mass and velocity of the cubes, as well as the surface they are colliding on.

Observation: This is correct reasoning since the model is only stating, under proper assumptions, the possibility of objects rotating around their point of contact.

VideoLLaMA3-7B (Zhang et al., 2025): yes

LLaVA-NeXT-Interleave (Li et al., 2024c): The outcome of the green cube colliding with the blue cube would depend on the initial velocity and angle of impact. If the green cube is moving towards the blue cube, it could potentially cause the blue cube to rotate. However, without knowing the initial conditions, it's not possible to predict the exact outcome.

Observation: LLaVA-NeXT-Interleave is wrong in stating that rotation can happen if one object moves towards (and collides) with another object.

Qwen2-VL (Wang et al., 2024a): Yes, the green cube will collide with the blue cube. When the green cube collides with the blue cube, the cubes will rotate around their center of mass. The exact amount of rotation will depend on the initial velocities and masses of the cubes.

Observations: Qwen2-VL provides the wrong output that the objects would rotate around their center of mass. It also supports this wrong reasoning with factually incorrect physical laws.

Qwen2.5-Omni (Xu et al., 2025): When the green cube collides with the blue cube, both cubes will experience a force that causes them to move in opposite directions. Since they have equal mass, the force exerted on each cube will be equal in magnitude but opposite in direction. This means that both cubes will

rotate after the collision. The rotation will be due to the force of impact, which causes the cubes to change their orientation in space.

Observations: Qwen2.5-Omni generates the correct outcome first (linear motion) and then incorrectly outputs that both objects will rotate after the collision. It also provides invalid reasoning for this output.

Gemma3-12B (Kamath et al., 2025): Here's an analysis of what would likely happen when the green cube collides with the blue cube, based on physical laws:

Outcome of the Collision:

Given that the cubes have equal mass and assuming an elastic collision (meaning no energy is lost due to friction or deformation), here's what would happen:

1. **Momentum Exchange:** Momentum is conserved in the collision. The green cube is initially moving, and the blue cube is stationary. During the collision, the green cube will transfer some of its momentum to the blue cube. 2. **Equal and Opposite Velocities:** Because the masses are equal, ideally the green cube will come to a stop and the blue cube will move away with approximately the same velocity that the green cube originally had. 3. **Equal Velocity:** As a result of the above, both cubes would move in a similar direction with equal velocity.

Will the Objects Rotate?

It's highly unlikely that either cube will rotate significantly during the collision. Here's why:

* **Impulse Applied:** The force during the collision is applied along a line. For rotation to occur, there would need to be an off-center force (a torque). Unless the collision isn't head-on, it's unlikely to generate such a torque. * **Shape:** Cubes are not ideal shapes for rotating due to collisions. Spheres would rotate the most.

Important Considerations/Simplifications:

* **Elasticity:** We've assumed an elastic collision. In reality, some energy would likely be lost as heat or sound, making the outcome slightly different. * **Perfect Alignment:** A perfectly head-on collision is assumed. If the cubes collide at an angle, the situation becomes more complex, and there could be some rotation, but it would still likely be minimal. * **Friction:** Any friction with the surface would also influence the outcome, slowing the cubes down.

Let me know if you'd like a deeper dive into any of these aspects!

Observations: Gemma3-12B provides both correct and incorrect reasoning about the outcome regarding the motion of the objects after the collision. It also provides accurate reasoning about the possibility of rotation, although it adds a partially true statement that cubes are less likely to rotate.

Aria (Li et al., 2024b): When the green cube collides with the blue cube, assuming they have equal mass and the collision is head-on, the outcome will depend on the nature of the collision. If it is an elastic collision, both cubes will bounce off each other with the same speed they approached with. If it is an inelastic collision, they might stick together and move as one object.

In either case, since the cubes are symmetrical and have no initial angular momentum, they will not rotate as a result of the collision. The collision will primarily affect their linear momentum.

Observations: Aria states and applies the principles of linear and angular momentum conservation.

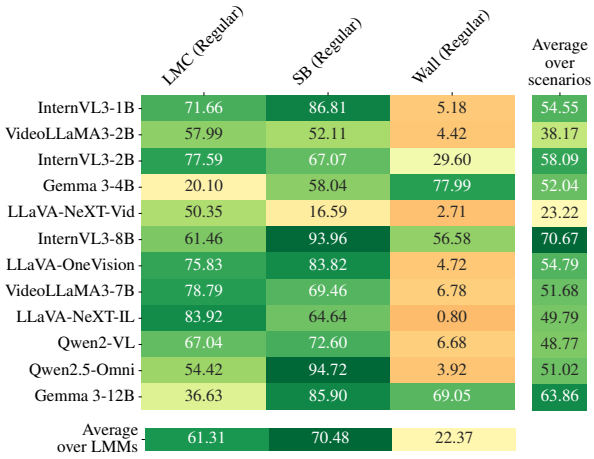
Conclusion: Some LMMs accurately predict the change in linear velocity, and then incorrectly add that the objects may rotate after the collision. Moreover, the laws stated to support their rotation argument are irrelevant to rotation. Comparing these outputs with those in § E.5.1, we conjecture that these models generated incorrect reasoning since they were explicitly prompted to make predictions about the rotation of the objects. However, we also note that some models accurately predicted that the colliding objects could rotate if certain conditions, such as colliding at an angle or frictional surfaces, were met. We believe that the

apparent inductive physical reasoning in § 4.4 was due to these inaccurate beliefs that objects could rotate after a collision.

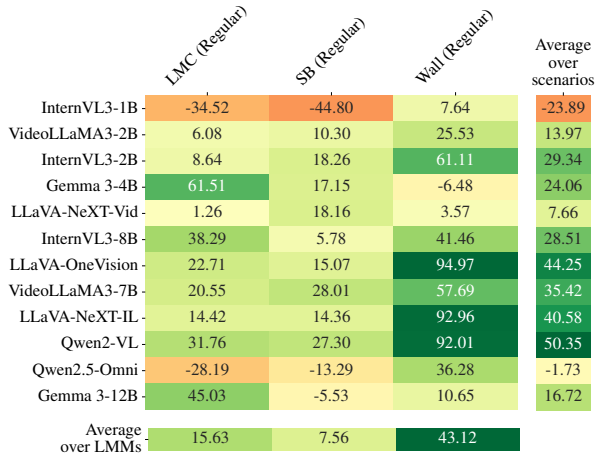
E.7 Do the findings hold for more complex scenes?

In this section, we increase the scene complexity of InPhyRe by including background objects and randomly positioning illumination sources and cameras during the rendering stage. Our intuition is that, since LMMs failed to demonstrate inductive physical reasoning in a simple environment such as InPhyRe, they should fail worse as scene complexities increase. However, since the nature of object attributes in the complex version of InPhyRe (which we refer to as InPhyRe$_{com}$) is similar to InPhyRe, we may also expect the LMMs to at least partially overcome the scene complexities.

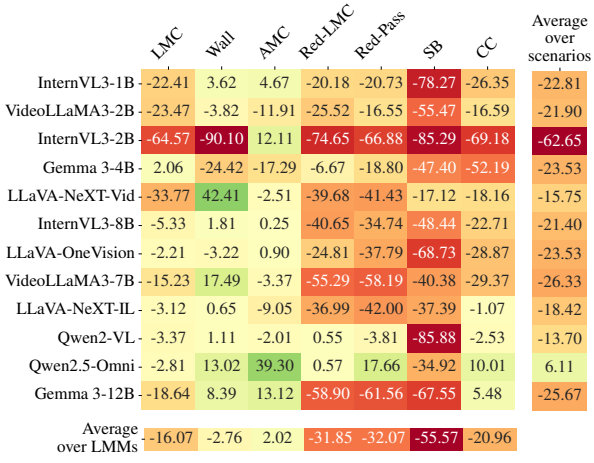
Similar to our main experiments in § 4, we will evaluate the LMMs on (1) regular scenarios without any demonstration samples, (2) regular scenarios with demonstration samples, and (3) irregular scenarios with demonstration samples. As before, inductive physical reasoning will be measured as the difference in performance between irregular and regular scenarios when demonstration samples were available. Our findings are discussed below.



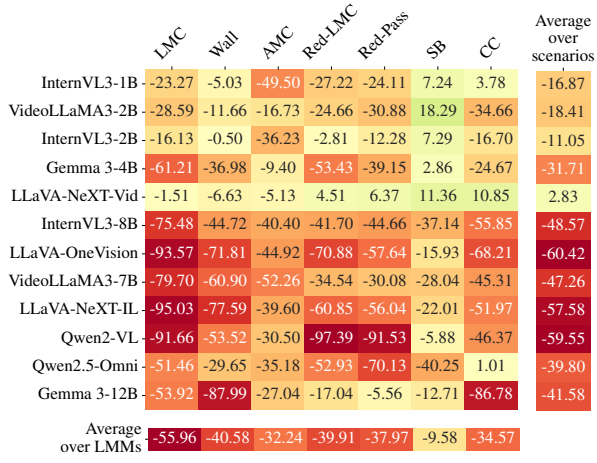
(a) Zero-shot performance in regular scenarios



(b) Change in performance in regular scenarios under 3-shot setting



(c) Change in performance in irregular scenarios under 3-shot setting



(d) Change in performance in irregular scenarios under 3-shot and video-only setting

Figure 21: Results on InPhyRe$_{com}$ – a version of InPhyRe with more scene complexities in the form of changes in camera pose, lighting conditions, and background objects.

Findings: Fig. 21a shows the zero-shot performance of LMMs in the regular scenarios of InPhyRe$_{com}$. Similar to our results on InPhyRe in Tab. 2, the LMMs perform fairly well in LMC (Reg.) and SB (Reg.), and poorly in Wall (Reg.). When demonstration samples comprising video frames and question-answer pairs are provided, the results improve significantly, and multiple LMMs achieve >90% accuracy in many scenarios (Fig. 21b). We measure inductive physical reasoning in Fig. 21c by comparing the 3-shot performance of LMMs in irregular scenarios against the LMMs' best corresponding performance in regular scenarios under zero-shot and few-shot settings. The results appear similar to those in Fig. 4b with many LMMs struggling in Red-LMC, Red-Pass, SB, and CC. Like in Fig. 4b, the largest average drop in performance was observed for InternVL3-2B. One notable difference between the results on InPhyRe and InPhyRe$_{com}$ is that Gemma3-4B showed a larger drop in performance in Fig. 21c compared to Fig. 4b. Also, although Qwen2.5-Omni has a smaller drop in Fig. 21c compared to Fig. 4b, we also point out that Qwen2.5-Omni's performance in the regular scenarios was proportionately worse on InPhyRe$_{com}$. Fig. 21d shows the further drop in accuracy when the demonstration samples include only the video frames, and not the question-answer pairs. Similar to our previous observations in Fig. 5b, we find that most LMMs show considerable language bias. LLaVa-NeXT-Video is the only model that did not show language bias in both InPhyRe and InPhyRe$_{com}$.

Although InPhyRe$_{com}$ had more scene complexities than InPhyRe, the evaluation results were similar. This indicates that added complexities did not challenge most LMMs. In future works, we aspire to generate more complex scenes with realistic degradations that pose significant perception challenges to the LMMs. Nonetheless, the absence of inductive physical reasoning in a visually simple dataset, such as InPhyRe that does not confound our core evaluation, indicates the absence of inductive physical reasoning in more complex scenarios.

E.8 Human Evaluation for InPhyRe

In this section, we evaluate how well humans fare on InPhyRe. We believe that future works on inductive physical reasoning can aspire to match or outperform this human baseline, and draw cognitive insights from our subject's explanations. To obtain the human baseline, we recruited 10 subjects and informed them about our research problem. We evaluated these subjects under a more difficult setting compared to our evaluation of LMMs – each subject was provided only one demonstration sample without the associated question-answer pair. Since the underlying logic to be inferred is the same for all samples in a given scenario, we evaluated the subjects on only one sample from each scenario. The demonstration samples and the evaluation samples from all scenarios were put together in a slideshow and shared with the subjects, along with the instructions on how to complete their task. They were provided with the same information about the tasks as the LMMs. The only additional assistance provided during the test was regarding the meaning of certain questions.

The results are shown in Fig. 22. We observe that most subjects had little trouble in both regular and irregular scenarios, despite being provided only one visual demonstration without any textual description. Here, the human subjects inferred laws from demonstration samples (*e.g.*, "the speed remained constant after collision") and applied them to the evaluation sample, irrespective of any contradictions between the inferred law and the premise given with the evaluation sample (*e.g.*, "the objects have equal mass and undergo elastic collision"). However, the subjects did struggle with Red-LMC and Red-Pass, where the required reasoning depended on the color of the objects. For these scenarios, we provided four demonstration samples: two scenarios with red-colored objects that violate the true physics, and two scenarios without any red-colored objects and that do not violate any true physics. Since we did not inform them specifically to use color information for reasoning in these scenarios, many subjects answered that there was not enough information to predict the answer.

E.9 Does using all evaluation frames improve performance?

In our main experiments, we provided only the initial frame from each evaluation sample to the LMMs, following the standard procedure for predictive physical reasoning. Thus, we posed physical reasoning as a normative task that requires LMMs to apply some rule to make their predictions. Intuitively, if all the frames from the evaluation sample were provided to the LMM, then any inferred physical rule becomes redundant.

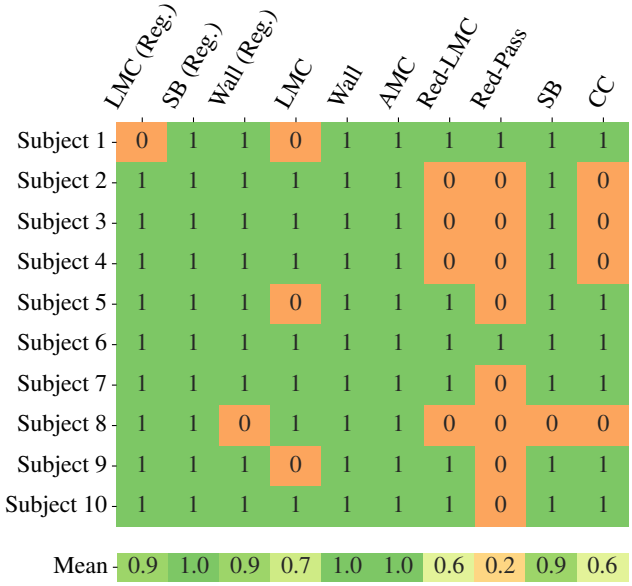


Figure 22: Accuracy of human subjects in various scenarios. "1" indicates correct prediction, and "0" indicates otherwise.

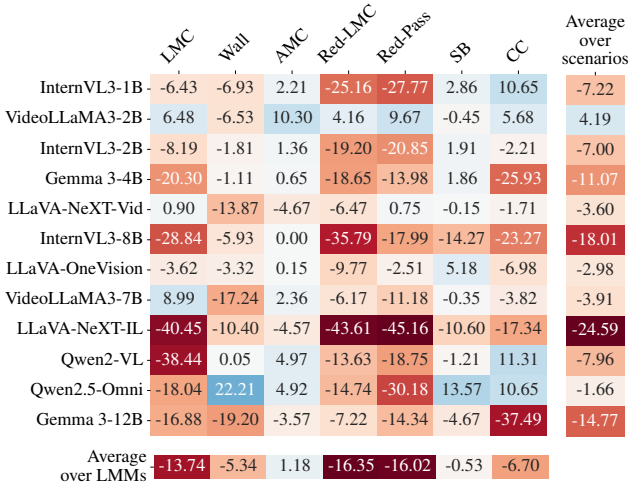


Figure 23: Difference in 2-shot performance when all evaluation frames are provided instead of just the first frame. The results indicate that despite having access to all the frames, the LMMs generally struggle to infer the dynamics from visual inputs, further asserting language bias.

Interestingly, through the following experiments, we find that even when provided with all the frames in the evaluation sample, LMMs fail to perceive object dynamics from the visual inputs.

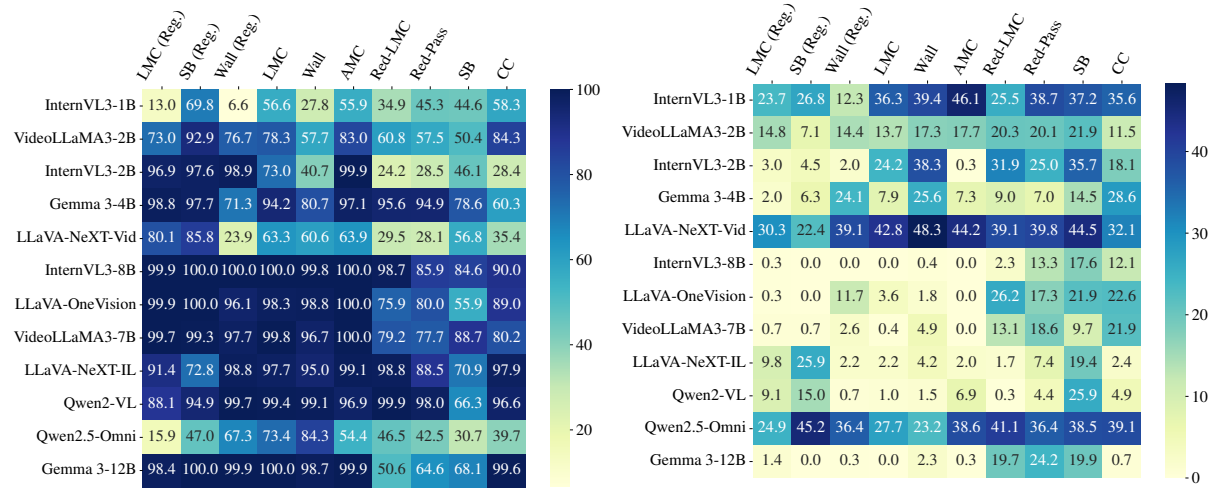
In our current experiment using all evaluation frames, we provide only two demonstration samples since using more frames (hence, more tokens) results in nearly approaching the LMMs' token limit. We compare the performance in the current experiment with that from a similar setup, but with only one evaluation frame. The differences in accuracy are shown in Fig. 23. Compared to the main experiments where only the initial frame was provided, we notice that across scenarios and LMMs, performance either remained roughly the same or decreased significantly. Our results convey two important findings:

1. LMMs struggle to perceive object dynamics from visual samples when the underlying physics contradict the existing parametric knowledge.

2. This behavior is further evidence of language bias, as LMMs were largely able to apply the underlying physical laws in Tab. 8 when these laws were provided as explicit reasoning in the demonstration samples. In § F, we will show that Gemma3-12B spent an order of magnitude less attention on image tokens compared to the text tokens, echoing similar findings in prior works. We believe that the phenomenon we observed in § F, along with the general *attention dilution* due to more tokens, is the primary reason for the accuracy drop observed in Fig. 23.

E.10 Effect of Structural Perturbations to Prompts

This section evaluates the effect of minor prompt changes on the accuracy of LMMs. These minor changes, which we refer to as "structural perturbations," do not affect the semantic meaning of the demonstration samples. We ablate the effects of the following types of structural perturbations: (1) changing object attributes in the demonstration samples, (2) changing order of demonstration samples, (3) minor text rewrites in the question and the options (e.g., "the speed will increase" or "the speed increases"), and (4) option ordering in the evaluation and demonstration queries.



(a) Mean performance (%) (b) Standard deviation

Figure 24: Mean performances of various LMM-scenario combinations and their standard deviations when there are structural perturbations in the prompt.

For this experiment, we create multiple smaller versions of InPhyRe with only the first 100 samples from each scenario. In each version, the structural perturbations are randomly chosen. The results over 10 such versions are shown in Fig. 24. For each LMM and each scenario, we compute the mean performance (Fig. 24a), the standard deviation in performance (Fig. 24b), and the coefficient of variation (CoV) in the performance (Fig. 25). The CoV is measured as the ratio between the mean performance and its standard deviation. Our primary metric will be the CoV, as it captures the relative variation due to the chosen structural perturbations.

The CoV for each LMM-scenario combination is shown in Fig. 25. Most LMM-scenario combinations have very low CoV values, indicating the robustness of the evaluation to structural perturbations. Some LMM-scenario combinations have unacceptable levels of CoV[7]. However, these values could either be due to the LMMs or the scenario setups. The higher CoV values concentrate around smaller LMMs ($<$ 7B parameters), and there is no noticeable pattern where a specific scenario has consistently high CoV. This suggests that higher CoV values are due to the sensitivity of the smaller LMMs, and not due to the benchmark design.

[7]Following standard practice, an acceptable CoV is less than 0.3.

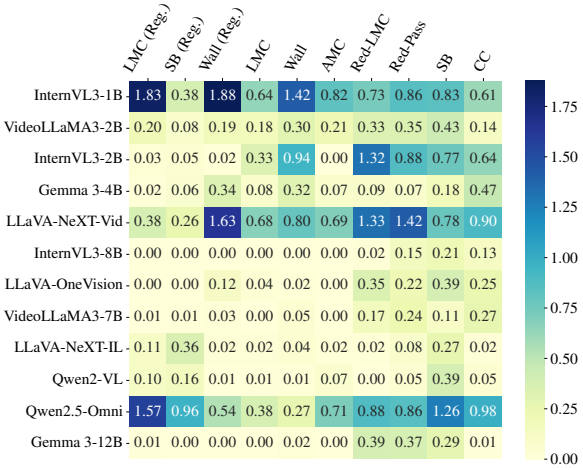


Figure 25: The coefficients of variation (CoV) for LMM-scenario combinations. CoV is computed as the ratio between the corresponding mean (Fig. 24a) and standard deviation (Fig. 24b) values.

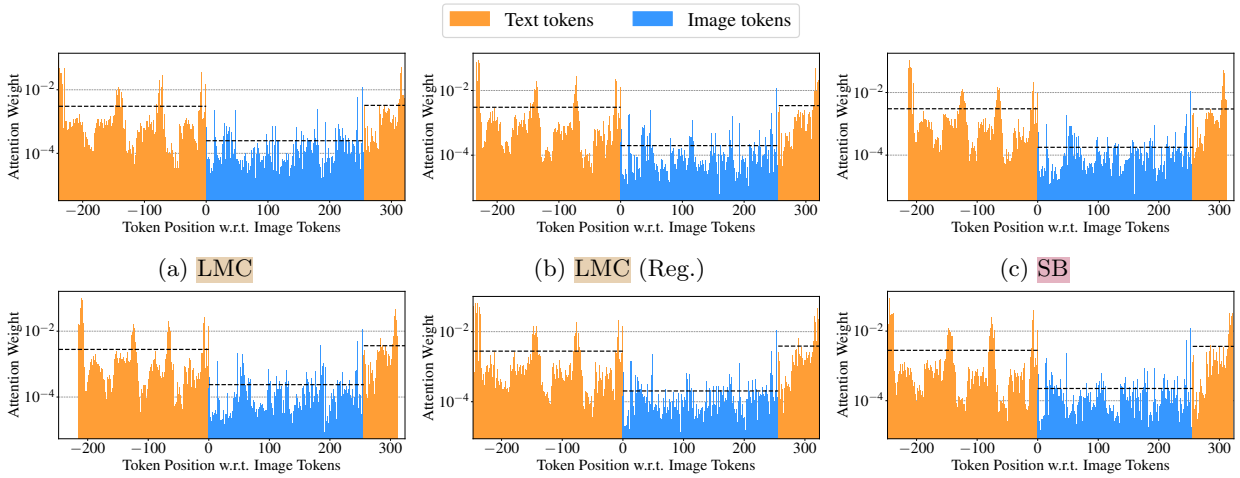
F Why Do LMMs Perform Poorly in Irregular Scenarios

We observed that LMMs generally struggled in irregular scenarios, although they performed well in regular scenarios. To develop methods to improve inductive physical reasoning in LMMs, we must first understand its causes. However, there are several practical challenges in investigating the causes of poor inductive physical reasoning in LMMs. The foremost challenge is the difficulty in interpreting the outputs of an LMM. Common approaches inspect the hidden states and the attention maps between tokens. In this section, we will use attention maps and linear probes on hidden states to gain insights into why LMMs fail in irregular scenarios. Specifically, we will (1) evaluate the attention values over image tokens and text tokens in Gemma3-12B, and (2) compare the hidden states of a pre-trained InternVL3-1B with those of a fine-tuned InternVL3-1B that demonstrated considerable inductive physical reasoning in § E.3 on regular and irregular scenarios. Although our findings do not fundamentally explain the lack of inductive physical reasoning in LMMs, we believe this discussion will foster future efforts.

F.1 Analysis of Attention Maps in Gemma3-12B

We plot the normalized attention values over the tokens in the final layer of Gemma3-12B in Fig. 26. To obtain these plots, we first choose three corresponding regular-irregular scenario pairs – LMC and LMC (Reg.), SB and SB (Reg.), and Wall and Wall (Reg.). In each pair, the scenarios only vary in the underlying physical law. Note that the exact visual attributes may vary in the selected samples. The plots are obtained by summing the attention values from 20 different samples from these regular-irregular scenario pairs. Since each sample may have a different number of text tokens, we use the image tokens as the "hinge" around which we arrange the pre-image and the post-image text tokens. The x-axis shows the position of these tokens w.r.t. the image tokens. Orange- and blue-colored regions denote text and image tokens, respectively. The black dashed lines show the average attention values for each colored segment.

Gemma3-12B spent an order of magnitude less attention on image tokens, compared to text tokens. The lower attention values over image tokens are a possible reason for the observed language bias. It may also explain why LMMs failed to perceive the underlying physical laws even when all frames (including those that show the outcome of the collision event) from the evaluation sample were provided. Similar findings were reported in (Chen et al., 2024a), albeit in the context of natural images.



(a) LMC (b) LMC (Reg.) (c) SB

(d) SB (Reg.) (e) Wall (f) Wall (Reg.)

Figure 26: Attention values over image tokens are around an order of magnitude less compared to those over text tokens.

F.2 Analysis of Hidden States in InternVL3-1B

To understand why an LMM showed poor performance in irregular scenarios, we compare its hidden states to a similar LMM that performed well in irregular scenarios. To that end, we compare a pre-trained (PT) InternVL3-1B with a fine-tuned (FT) InternVL3-1B. The results for the FT model were shown in § E.3. We compare the hidden states of the two models by training linear probes on them. We consider two target tasks for the linear probes: (T1) predict whether the given scenario is regular or irregular, and (T2) predict the attributes (color and shape) of the objects in demonstration and evaluation samples. Through (T1), we understand whether the LMMs learn different hidden states for regular and irregular scenarios. Through (T2), we check for any discrepancy in the amount of information carried by the hidden states from demonstration and evaluation samples.

(T1) Can we predict the underlying scenario from the hidden states? The task here is to predict whether the hidden states correspond to samples from a regular scenario or an irregular scenario. We evaluate PT and FT LMMs on this task using the three corresponding pairs of regular-irregular scenarios that we used in our previous experiment – LMC, SB, and Wall. We also vary the number of demonstration samples from one to three. To train the linear probe, we collect hidden states from all samples in each of the scenarios. The results are shown in Fig. 27.

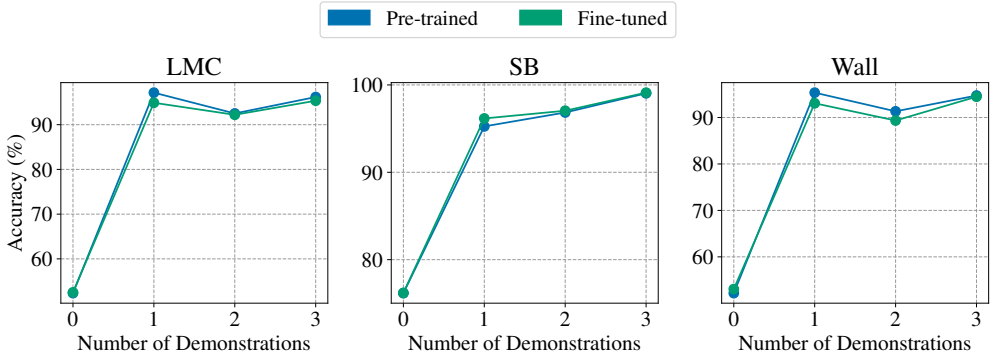
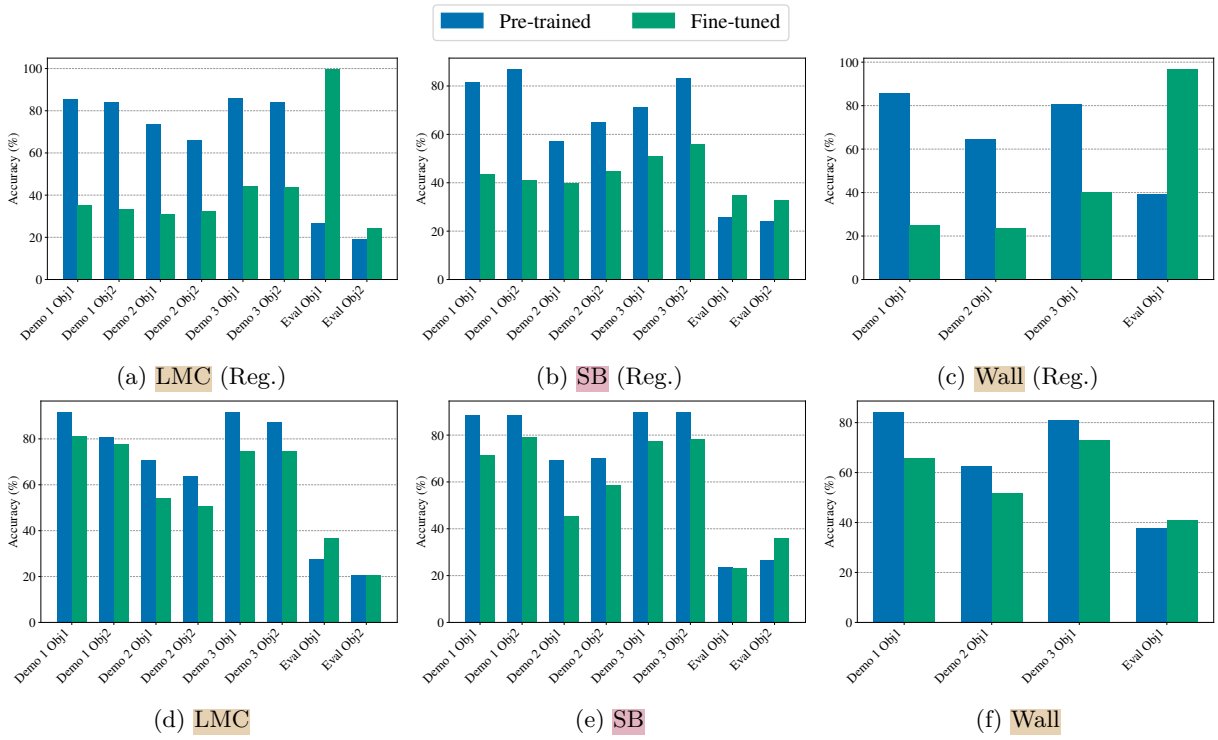


Figure 27: Accuracy of linear probes in predicting the underlying scenario from hidden states of pre-trained and fine-tuned models.

We find that the hidden states from both PT and FT LMMs have sufficient information to classify the underlying scenario with over 90% accuracy. The classification accuracy is not affected by the number of demonstration samples. We also compare these classification results against the case where no demonstration samples are provided. Note that without demonstration samples, the evaluation frame and the accompanying question-answer pair do not provide any discriminative information to suggest whether the underlying scenario is regular or not. Therefore, the classification accuracy in the case with no demonstration samples only acts as a baseline that validates our observations with ≥ 1 demonstration samples. In all scenario pairs, the linear probes achieve significantly less classification accuracy without demonstration samples. In LMC and Wall, they only match random chance performance. In the next task, we will understand how PT and FT LMMs understand the object attributes from the prompt differently.



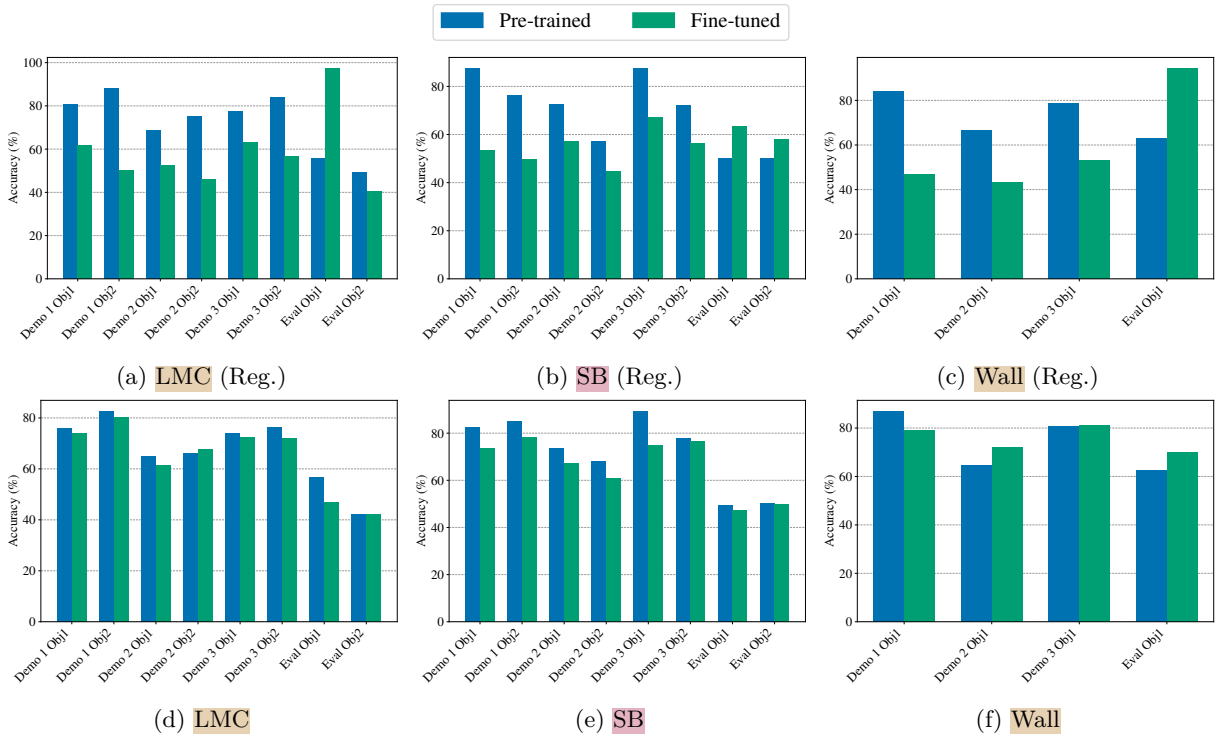
(a) LMC (Reg.) (b) SB (Reg.) (c) Wall (Reg.)

(d) LMC (e) SB (f) Wall

Figure 28: Accuracy of linear probes on hidden states from pre-trained and fine-tuned InternVL3-1B models in classifying the colors of the objects in demonstration and evaluation samples.

(T2) Do PT and FT LMMs perceive object attributes differently? Through task (T1), we saw that the hidden states from both PT and FT InternVL3-1B models carried sufficient information from the demonstration samples to classify the underlying scenario. It is not possible to semantically evaluate these scenario-specific differences in the hidden states exhaustively. Therefore, as a first step, in this task, we will quantify the object attribute information in the hidden states. Similar to task (T1), we will use linear probes on the hidden states to classify the shapes and the colors of the objects in demonstration and evaluation samples. In total, the objects in InPhyRe were composed from three shapes and ten colors. The frames in LMC and SB showed two objects, while those in Wall showed only one. We will compare the attribute classification accuracy across scenarios and between PT and FT InternVL3-1B models.

Figs. 28 and 29 show the attribute classification accuracies on color and shape, respectively, of linear probes on the hidden states from PT and FT models in various scenarios. In each figure, the top row shows regular scenarios, while the bottom row shows irregular scenarios. By using accuracy as a proxy for the amount of information in the hidden states about the predicted attributes, we make the following observations:



(a) LMC (Reg.) (b) SB (Reg.) (c) Wall (Reg.)

(d) LMC (e) SB (f) Wall

Figure 29: Accuracy of linear probes on hidden states from pre-trained and fine-tuned InternVL3-1B models in classifying the shapes of the objects in demonstration and evaluation samples.

(O1) PT models contain more attribute-specific information about the objects in the demonstration samples than the objects in the evaluation sample. This pattern is evident across various scenarios and on both color and shape attributes.

(O2) **FT models adapt their hidden states to the underlying scenario**. In irregular scenarios, their hidden states contain more information from the demonstration samples, and in regular scenarios, their hidden states contain more information from the evaluation sample. Indeed, in regular scenarios, the demonstration samples are not required since the parametric knowledge is sufficient. However, in irregular scenarios, demonstration samples are key in physical reasoning. Thus, this adaptation aligns with the expected behavior of an LMM in irregular scenarios.

(O1) is a surprising result, since we expect PT models to have less information about the demonstration samples due to their poor performance in irregular scenarios. Moreover, it is not clear from the linear probe whether this information came from the frames or the question-answer pairs in the prompt. (O2) suggests that fine-tuning introduces adaptive behavior when the fine-tuning dataset potentially contradicts parametric knowledge. While (O2) is an interesting finding, it does not explain whether information in the hidden states of PT and FT models are of different natures.

Conclusions from our analysis: Although our results indicate that LMMs indeed understand regular and irregular scenarios differently, they do not shed light on what causes them to be different. Understanding the cause of this difference is a prerequisite for building solutions to improve inductive physical reasoning. Our findings through task (T2) – that LMMs adapted their visual attribute reasoning after fine-tuning – are encouraging as the first steps in understanding inductive physical reasoning. Our analysis was also limited by the lack of reliable tools to interpret the outputs and the hidden states of LMMs. Therefore, future analysis would require better interpretability tools for LMMs, specifically regarding the visual inputs.

G Use of LLMs and Generative AI

Grammarly[8] and Writefull[9] (embedded in Overleaf) were used in correcting grammatical errors and spellcheck. LLMs were used to obtain feedback on the writing style. They were not used in generating sentences or summarizing paragraphs. Some figures in the paper used clip-art style images generated using Gemini[10] – the cartoon image of a man holding a torch in the title, robot head, electric bulb, and book in Fig. 1, and palette in Fig. 3. Generative AI was not used in dataset creation.

[8] https://app.grammarly.com/
[9] https://www.writefull.com/
[10] https://gemini.google.com/app