COGNPHYS: A BENCHMARK FOR INDUCTIVE PHYSICAL REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing benchmarks for physical reasoning in Vision-Language Models (VLMs) primarily assess rule following—predicting outcomes governed by known Newtonian laws—yet a hallmark of human physical cognition is rule discovery: observing unfamiliar phenomena and inducing the underlying principles from scratch. This capability is essential for embodied agents adapting to novel environments, robot control, and automated scientific discovery. We introduce CognPhys, a diagnostic benchmark comprising multiple physics worlds, each governed by distinct rules that systematically deviates from real-world physics. We evaluate a broad suite of open-source and closed-source VLMs using a structured scoring protocol. Our extensive evaluations reveal a profound deficiency in state-of-the-art VLMs: they are severely bottlenecked at rule induction. Even the most capable models consistently fail to synthesize coherent rules from observed motion, constrained by an over-reliance on memorized physical priors. CognPhys uncovers a fundamental blind spot in current VLM physical reasoning and establishes a challenging testbed for scientific and embodied AI.

1 INTRODUCTION

The rapid progress of large language models (LLMs) and multimodal foundation models has substantially advanced their ability to reason about complex information. As demonstrated by recent frontier models, including GPT-6 Astra, Claude Opus 5.5, and Gemini 3.8 Flash (OpenAI, 2026; Anthropic, 2026; Google DeepMind, 2026), these systems can interpret visual scenes, integrate perception with domain knowledge, and solve challenging scientific problems requiring multi-step reasoning. Dedicated physics evaluations such as PHYSICS (Feng et al., 2025) further show that models can reliably apply established physical laws to solve problems posed under explicit conditions. Yet, this impressive performance largely remains within the realm of deduction under given rules, leaving the more challenging task of rule induction from raw observation largely unexplored. The ability to observe unfamiliar phenomena and infer the underlying laws that govern them is a hallmark of intelligence, essential not only to scientific discovery but also to robot control and embodied agents rapidly adapting to novel environments. For vision-language models (VLMs) to serve as effective reasoning agents in the physical world, they must go beyond memorized physics and demonstrate genuine inductive physical reasoning—the capacity to discover hidden rules from visual observation alone.

Existing physical reasoning benchmarks, however, predominantly test *rule following* (Feng et al., 2025; Yi et al., 2020; Grunde-McLaughlin et al., 2021; Riochet et al., 2018; Xu et al., 2025; Qiu et al., 2025; Xiang et al., 2025; Dai et al., 2025; Wang et al., 2026): under fixed physical assumptions, predict future trajectories or answer questions about object motion. While valuable, this paradigm assumes the governing laws are fixed and known a priori, missing a critical dimension of intelligence—the ability to detect when familiar physics breaks down and to infer the alternative rules at play. Consider a robot entering an unfamiliar room where objects behave unexpectedly, or a scientist observing anomalous planetary motion; both must induce the underlying principles rather than apply pre-learned templates.

To bridge this gap, we introduce **CognPhys** (Cognitive Physical Reasoning Benchmark), a benchmark designed to evaluate VLMs on inductive physical reasoning. CognPhys is generated by a procedural rule engine and comprises 284 worlds each governed by distinct rules. We evaluate a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

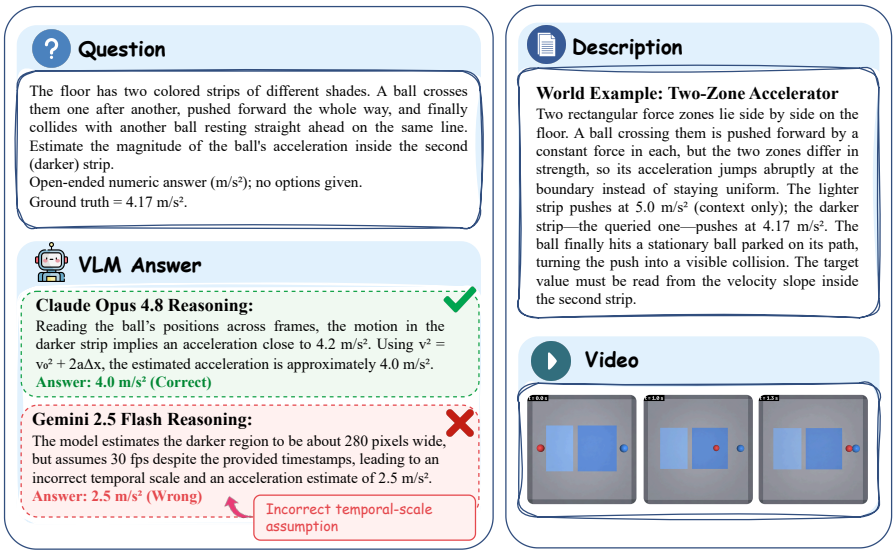


Figure 1: **Overview of CognPhys through a representative instance.** A Two-Zone Accelerator example illustrates the benchmark format, including a procedurally generated physical world, visual observations, an open-ended parameter-estimation question, and contrasting VLM responses.

range of open-source and closed-source VLMs on CognPhys across three tasks at increasing levels of abstraction: judging the laws governing a world, estimating a law's parameter in SI units, and predicting where the body will be given the first half of a clip. Figure 1 illustrates a representative parameter-estimation instance and contrasting VLM responses.

Our main contributions are:

A challenging CognPhys benchmark. We construct a systematic benchmark of counterfactual physics worlds, each with curated video episodes and diagnostic questions, providing a testbed for evaluating inductive physical reasoning in VLMs.

A hierarchical diagnostic protocol. Rather than treating physical reasoning as a monolithic score, we structure our evaluation across three tasks at increasing levels of abstraction: *law identification* (qualitatively discerning the governing axiomatic rule), *parameter estimation* (quantitatively calibrating physical coefficients in SI units), and *trajectory prediction* (forecasting future object states from observed motion).

A detailed analysis of rule reasoning. Experiments across open-source and closed-source VLMs reveal that current models struggle substantially with rule discovery. The results agree on one thing: what the models lack is not perception but a procedure for turning observations into a law.

2 Related Work

2.1 Fixed-physics visual reasoning benchmarks

Early physical reasoning datasets assume a single, fixed Newtonian physics and focus on event understanding and causal reasoning. Yi et al. (2020) introduced CLEVRER, a synthetic 3D environment with collision events and descriptive, causal, and counterfactual questions. Grunde-McLaughlin et al. (2021) constructed AGQA for large-scale video question answering covering temporal and spatial reasoning under real-world physics. Riochet et al. (2018) proposed IntPhys for judging physical plausibility (e.g., object permanence, continuity). More recent benchmarks extend this direction toward vision-language models: SeePhys (Xiang et al., 2025) evaluates vision-based physics reasoning, while CausalPhys (Tang et al., 2026) and MASS-Bench (Wu et al., 2026) investigate causal physical reasoning and motion-aware spatial-temporal grounding, respectively. These

Table 1: Comparison of CognPhys with representative physics reasoning benchmarks.

Benchmark	Multi-world	Rule discovery	Visual input	Specific rule deviation
CLEVRER (Yi et al., 2020)	×	×	✓	×
AGQA (Grunde-McLaughlin et al., 2021)	×	×	✓	×
IntPhys (Riochet et al., 2018)	×	×	✓	×
SeePhys (Xiang et al., 2025)	×	×	✓	×
MASS-Bench (Wu et al., 2026)	×	×	✓	×
CausalPhys (Tang et al., 2026)	×	×	✓	×
PHYSICS (Feng et al., 2025)	×	×	×	×
UGPhysics (Xu et al., 2025)	×	×	×	×
PHYBench (Qiu et al., 2025)	×	×	✓	×
NewtonBench (Zheng et al., 2026)	✓	✓	×	×
LLM-SRBench (Shojaee et al., 2025)	✓	✓	×	×
PhysGym (Chen et al., 2025)	✓	✓	×	×
Gravity-Bench (Koblischke et al., 2025)	✓	✓	×	✓
WorldBench (Upadhyay et al., 2026)	✓	×	✓	×
PhysicsMind (Mak et al., 2026)	×	×	✓	×
NovPhy (Pinto et al., 2024)	×	×	✓	✓
CognPhys (Ours)	✓	✓	✓	✓

benchmarks evaluate reasoning under fixed physics but do not require models to infer governing rules.

2.2 Physics problem-solving benchmarks

Another line of work evaluates models through textbook or exam-style problems under real-world Newtonian physics. Feng et al. (2025) collected over 1,200 PhD-level physics problems across multiple subfields. Xu et al. (2025) systematically covered undergraduate physics courses with an automatic scoring framework. Qiu et al. (2025) further emphasized problem originality and expression-level evaluation. Multimodal extensions such as Xiang et al. (2025), Dai et al. (2025), Wang et al. (2026) incorporate diagrams and tables alongside text. These benchmarks only focus on problem-solving and do not require rule induction.

2.3 Physical law discovery and symbolic regression benchmarks

Recent work has moved toward evaluating scientific law discovery from data. Zheng et al. (2026) constructed law-discovery tasks requiring LLMs to infer mathematical forms from experimental observations. Shojaee et al. (2025) focused on symbolic regression—recovering analytic equations from numerical observations. In interactive settings, Chen et al. (2025) allowed LLM agents to conduct experiments and induce physical laws, while Koblischke et al. (2025) specifically targeted gravitational law discovery under non-standard gravity. However, these benchmarks operate on structured numerical data without visual input.

2.4 Multi-world and open-world physical generalization benchmarks

A growing body of work diagnoses world models across diverse scenario distributions. Upadhyay et al. (2026) proposed concept-disentangled scenes for video prediction evaluation. Pinto et al. (2024) introduced physical novelty (new objects, new interactions). Mak et al. (2026) introduced PhysicsMind, benchmarking physical reasoning and prediction across simulated and real mechanical scenarios for VLMs and world models. However, these works primarily study generalization across novel scenarios, rather than systematic induction of world-specific physical rules.

2.5 Position of CognPhys

As summarized in Table 1, existing benchmarks remain fundamentally partitioned: visual physics benchmarks predominantly assess *rule following* under fixed Newtonian priors, scientific discovery suites evaluate *rule induction* solely on clean numerical tables devoid of visual perception, and

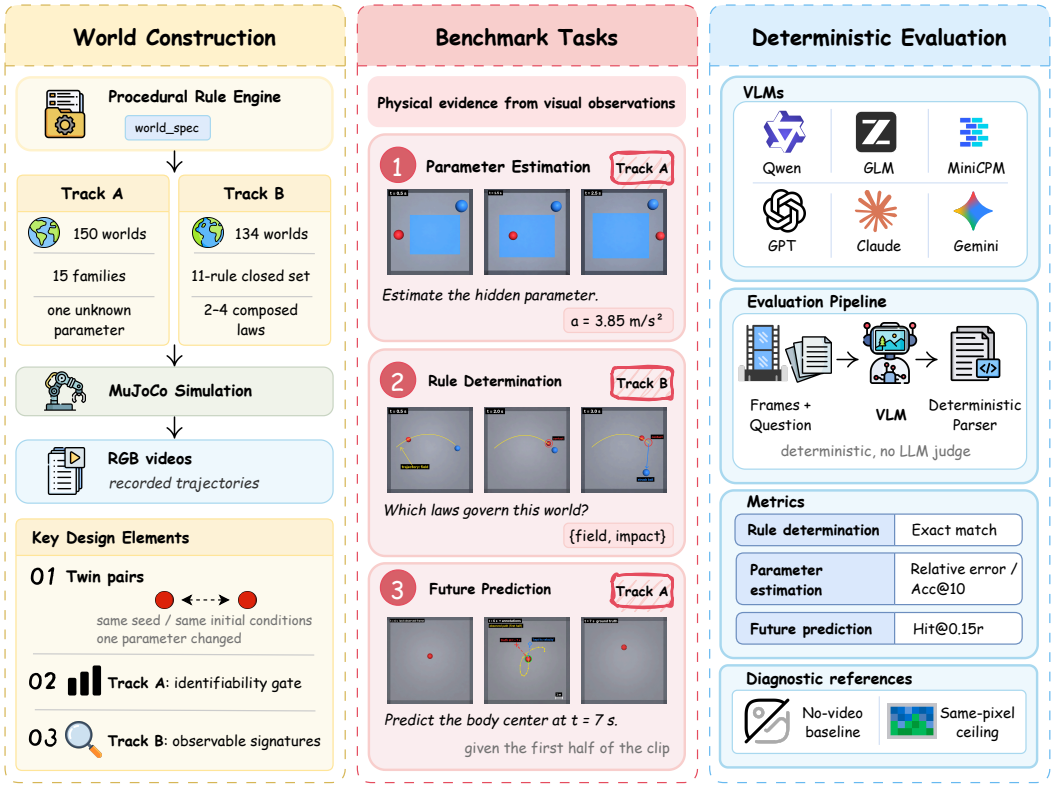


Figure 2: Methodological pipeline of CognPhys, covering procedural world construction, benchmark task formulation, and deterministic evaluation.

multi-world benchmarks focus on qualitative visual novelty rather than formal axiomatic variations. **CognPhys** bridges this divide as the first benchmark to unify raw video perception, multi-world diversity, and systematic physical rule discovery. By subjecting models to 284 procedurally generated worlds with mathematically rigorous counterfactual laws, CognPhys prevents reliance on memorized physical priors and diagnoses the complete discovery pipeline—from raw pixel tracking to physical parameter estimation and high-level law induction.

3 METHOD

Figure 2 gives an overview of the full CognPhys pipeline, from procedural world generation through benchmark task design to deterministic scoring.

3.1 DATA CONSTRUCTION

3.1.1 WORLD DESIGN

CognPhys comprises 284 procedurally generated 3D physics worlds, each governed by one or more physical laws that deviate from Newtonian mechanics. A world is specified as a composition of laws rather than as a hand-authored scene: the engine supplies a library of rule templates, each pairing an optional trigger condition with an effect, and several of them can act simultaneously. We organize the worlds into two tracks. **Track A** comprises 150 worlds across 15 families, each governed by a single law whose quantitative parameter is to be estimated. **Track B** comprises 134 worlds in which several laws are composed, drawn from a closed set of eleven.

Worlds are generated in *twin pairs*. The two members of a pair share a generation seed, and therefore identical initial conditions, and differ only in the value of a single rule parameter; a model that reads dynamics rather than recalling a prior must therefore answer the two members differently. To keep

pairs informative we record for every world an *effect size*, defined as the divergence of its trajectory from that of the same world with the parameter nulled and measured in ball diameters (median 4.5, range 0.8–12.1).

Every candidate world must pass an *identifiability gate*: a classical computer-vision estimator, working from the same pixels a model sees, has to recover the rule parameter to within 20%. Of 222 candidate worlds, 72 fail the gate and are discarded, leaving 150; the survivors are recovered by the reference estimator with a median relative error of 2.8%. The gate is what licenses reading a model's failure as a statement about the model rather than about the task. Track A is additionally re-rendered from a 45° oblique viewpoint; the worlds, parameters, questions and ground truth are unchanged, so the second view isolates failures caused by the camera rather than by the physics.

3.1.2 SIMULATION PIPELINE

Worlds are simulated in MuJoCo at 30 FPS and rendered to 800×600 MP4 video at 50 pixels per metre, with the floor drawn from a fixed top-down camera. Unlike the engine's default scenario generator, which drops bodies at random positions where a localised feature is frequently never reached, we place bodies under *controlled initial conditions*: the observed body is launched from a fixed point along $+x$ at a specified speed, so that its interaction with the annotated feature falls inside a clean, observable window. Track A episodes last 12 seconds, extended to 20 seconds for the five families in which a second body must be struck; Track B episodes last 14 seconds.

For every frame we log the complete ground-truth state to a `states.jsonl` file, containing the position, velocity, radius and mass of each object, its hidden properties, and discrete events such as collisions and entries into a localised region. Crucially, the ground-truth parameter is *measured from the recorded trajectory* rather than read from the rule's nominal value: a continuous force rule delivers only half of its nominal acceleration, because the force is applied once per frame and consumed by the first physics substep, so treating the nominal value as ground truth would mark correct answers as wrong.

3.1.3 TASK DESIGN

For each world we pose one of three tasks, ordered by increasing level of abstraction.

Rule determination (Track B) — the model must name *every* law that governs the world, drawn from a closed set of eleven. The prompt lists the entire taxonomy with one line of physics for each entry, so that listing every candidate is a viable strategy; scoring is therefore exact set match as the primary metric, with precision and recall reported alongside it.

Parameter estimation (Track A) — the model must report the quantitative value of the governing law's parameter in SI units, for example an acceleration in $\mathrm{m/s^2}$ or a damping rate in $\mathrm{s^{-1}}$. This task requires an absolute rather than a relative judgement: the model must both measure a rate of change from the video and commit to a physical scale.

Future prediction — the model is shown only the first half of a clip and must report the pixel coordinates of the observed body's centre at a later moment, either one second past the observation window or at 75% of the episode. Errors are normalised by the radius of the body in that world, so that the metric is comparable across the different apparent sizes produced by depth, and the primary metric is the fraction of predictions falling within 0.15 ball radii.

All prompts are stated in fixed natural language and require the final answer on a designated last line. Answers are extracted by regular expression matching and unparseable responses are counted as incorrect.

3.2 EVALUATION PROTOCOL

Let $\mathcal{W}$ be the set of evaluated worlds and, for each $w \in \mathcal{W}$, let y_w be its ground-truth answer and r_w the radius of the observed body in pixels. For evaluation we sample T evenly spaced frames $\{I_1, \ldots, I_T\}$ from the episode and construct the multimodal prompt

Table 2: VLMs evaluated in this work. Open-source models were run from local weights; closed-source models were accessed through a commercial API.

Open-source		
Model	Parameters	Base LLM
Qwen3 VL 32B	32B	Qwen3
GLM 4.6V Flash	9B	GLM-4
MiniCPM-V 4.5	8B	Qwen2.5
Qwen2.5 VL 7B	7B	Qwen2.5
Qwen3 VL 2B	2B	Qwen3
Closed-source		
Provider	Models	
OpenAI	GPT-6 Astra; GPT-5.6 sol, terra, luna; GPT-5.5; GPT-5.4; GPT-5.3 codex; GPT-5.2; GPT-5.1; GPT-5 mini; GPT-4.1; GPT-4o	
Anthropic	Claude Opus 4.8; Claude Sonnet 5; Claude Haiku 4.5	
Google DeepMind	Gemini 3.8 Flash; Gemini 3.7 Flash; Gemini 3.1 Pro	

$$\mathrm{prompt}(w, \mathbf{I}) = [\text{instruction}] \oplus [I_1, \ldots, I_T] \oplus [\text{question}], \tag{1}$$

where $\oplus$ denotes interleaved image–text concatenation. The instruction states that the world's physics may differ from Newtonian mechanics and that the answer must be grounded in visual evidence.

The model generates a free-text response, from which a deterministic parser extracts the answer $\hat{y}_w$ by regular expression matching. The three tasks are scored respectively by relative error, exact set match, and a radius-normalised hit rate,

$$\mathrm{Err}(w) = \frac{|\hat{y}_w - y_w|}{|y_w|}, \qquad \mathrm{Exact}(w) = [\hat{y}_w = y_w], \qquad \mathrm{Hit}(w) = [\|\hat{y}_w - y_w\| \le 0.15\, r_w], \tag{2}$$

and each is aggregated by the mean over $\mathcal{W}$. For rule determination, $\hat{y}_w$ and y_w are sets of rule names and Exact denotes set equality; for parameter estimation, y_w is a scalar in SI units and we additionally report the fraction of predictions within 5% and 10% relative error; for future prediction, y_w is a pair of pixel coordinates. The entire pipeline is deterministic and requires no human judge or LLM-as-judge.

4 Experiments

4.1 Setup

We evaluate eighteen closed-source models from three families (OpenAI, Anthropic, Google DeepMind) through a commercial API, and five open-source models run from local weights. Table 2 lists them. All models see the same input: frames sampled at 2 fps, at most sixteen frames per item, with the long side downscaled to 512 pixels, which matches the effective resolution of the downsampled image path used by the API models. Scoring follows the protocol in Section 3.2 and is fully deterministic, with no human judge and no model-as-judge.

Each task is reported against a single reference point, measured on the same worlds: what a classical method recovers from the same pixels. Parameter estimation is compared against a hand-written computer-vision estimator, and future prediction against a replay of the world's own rules, including walls and contacts. Rule determination is a multi-label choice over the closed vocabulary, for which no classical estimator yields a comparable number, so that column carries no reference.

4.2 Main results

Table 3 reports all three tasks.

Table 3: Main results (%). **Rule Determination**: exact set match. **Parameter Estimation**: accuracy at 10% relative error on the SI parameter. **Future Prediction**: fraction of predictions within 0.15 ball radii at $+1$ s. Best per column in bold.

	Rule Determination	**Parameter Estimation**	**Future Prediction**
	exact set	acc@10	$\leq 0.15r$
Closed-source (API)			
GPT-6 Astra	**20.1**	**15.3**	**33.3**
GPT-5.6 sol	15.7	7.3	30.0
Gemini 3.8 Flash	15.7	7.3	3.3
GPT-5.6 terra	12.7	4.7	16.7
Gemini 3.7 Flash	11.9	6.0	6.7
Gemini 3.1 Pro	10.4	6.7	3.6
GPT-5.5	9.7	8.0	6.7
GPT-4.1	7.5	6.7	0.0
Claude Sonnet 5	6.7	4.0	0.0
GPT-5.3 codex	6.7	3.3	3.6
GPT-5.6 luna	6.0	5.3	3.3
Claude Opus 4.8	5.2	4.0	0.0
GPT-5.4	4.5	6.7	0.0
GPT-5 mini	4.5	4.0	3.3
Claude Haiku 4.5	3.0	5.7	0.0
GPT-5.1	2.2	5.3	0.0
GPT-5.2	0.7	5.3	0.0
GPT-4o	0.0	4.1	4.5
Open-source (local weights)			
GLM 4.6V Flash	2.2	5.0	0.0
MiniCPM-V 4.5	0.0	5.0	0.0
Qwen3 VL 32B	0.0	5.0	0.0
Qwen2.5 VL 7B	0.0	2.5	0.0
Qwen3 VL 2B	0.0	0.0	0.0
Same-pixel ceiling			
classical estimator	—	90.7	87

Rule determination: naming which laws govern a world. The best model returns the exact set of governing laws for 20.1% of worlds, against 1.6% for choosing a rule subset at random. Models are therefore well above chance, but they do not demonstrate an adequate ability to judge *which kind* of law is acting: the scores are carried by the two rules whose signature is a discrete event, and collapse on the rules defined by the shape of the motion (Figure 3a).

Parameter estimation. The best model reaches 15.3% within 10% of the true value, while a classical estimator working from the same pixels reaches 90.7%. The shortfall is not a matter of degree: seven of the eighteen closed-source models score below 5%, and every open-source model sits at or below 5%.

Future prediction. The best model places 33.3% of its predictions within 0.15 ball radii at $+1$ s, while replaying the world's own rules from ground-truth parameters reaches 87%. The task is solvable from the information in the clip; the shortfall is the model's.

Across all three tasks the open-source models collapse to a trivial predictor: 0% on rule determination for four of the five, 0% within the grasp tolerance on future prediction for all five, and at or below 5% on parameter estimation.

5 Analysis

Section 4 shows that the models score poorly. This section asks where the failure lies, and the answer is narrow: the models recognise the scene and can compare two of them, but they do not *induce the law* that governs the motion. What they supply instead are the prototypes they already have, and accuracy falls away as soon as a law's parameter has to be pinned down.

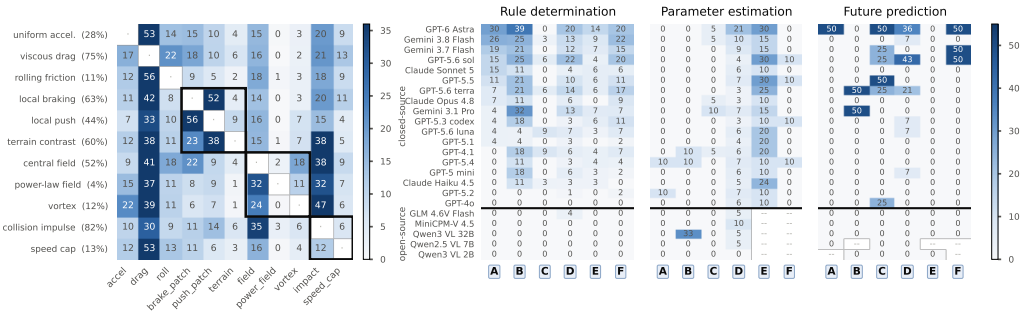


Figure 3: **(a)** When a rule is missed, what is named instead: the cell is the share of that rule's misses, over the closed-source models — rows and columns grouped by kind of mechanics, the boxed blocks being the kinds with more than one rule. **(b)** Every model against the same six kinds — **A** constant force, **B** velocity-dependent drag, **C** ground friction, **D** coloured patches, **E** fields, **F** events — one panel per task, one row per model, the score printed in each cell.

5.1 Naming a prototype, not seeking the law

Varying only the question, accuracy falls two orders of magnitude: forced choice — which of two counterfactual twins was pushed or slowed — is at 97.5 against a ceiling of 100, and paired ranking far above chance ($p < 0.001$). Asking for the parameter collapses every model to single digits against 90.7%; asking it in image pixels does not help either, so the failure is not unit conversion but the step where a law has to be named and given a value.

Rule determination is the one task where the models are clearly above chance, so there the errors are readable (Figure 3a). Identification spans 3.9% to 82.4%, and two rules are the most common substitute in nine of the eleven rows — drag wherever the speed changes, a collision wherever it changes abruptly. Naming *some* laws is evidently easy: the most generous model recalls 95.4% of the rules in force. Committing to the set that is actually in force is not — the best precision is 61.4%, and only 20.1% of worlds get the exact set.

What the models name instead is a prototype. Of the 24 vortex worlds with no collision, 52.7% of the misses name a collision and 47.5% viscous drag, the two prototypes they reach for. Of the 486 push-patch cells, 36.0% answer a brake patch, reading the region's effect backwards. Shape-defined rules are missed most (vortex 11.7%, power-law field 3.9%) while the discrete, abrupt ones do best (collision impulse 82.4%). The information needed is in the pixels: a detector whose thresholds are fitted per rule on one half of the worlds and applied to the other reaches 74.1% and 15.2% on those two. So the failure is induction, not evidence.

5.2 Accuracy at every tolerance

Figure 4 sweeps the tolerance. The models are order-of- magnitude right and nothing more: within a factor of two of the true parameter the closed-source models run from 47% to 98% with a median of 85%, but at the 10% the benchmark scores the best is at 15.3% against the classical estimator's 90.7%, and at 2% at 6.0%. Position prediction is worse at every tolerance: the best model is at 33.3% at the 0.15 ball radii the task scores, against the rule replay's 86.7%, and the median model is at 18.3% even within a whole ball radius.

Panel (b) of Figure 3 says where the shortfall sits, and no kind of mechanics is close. On parameter estimation the best model reaches 30% on fields and 0% on constant force and velocity drag; on future prediction it runs from 0% to 50%, with the median model at 0% on every kind; on rule determination its exact-set rate runs from 0% to 39%. The kind a model handles best changes from task to task, and 39% is the ceiling anywhere on the row. (Future prediction rests on 30 worlds, and the open-source models were run on a subset that omits two families.)

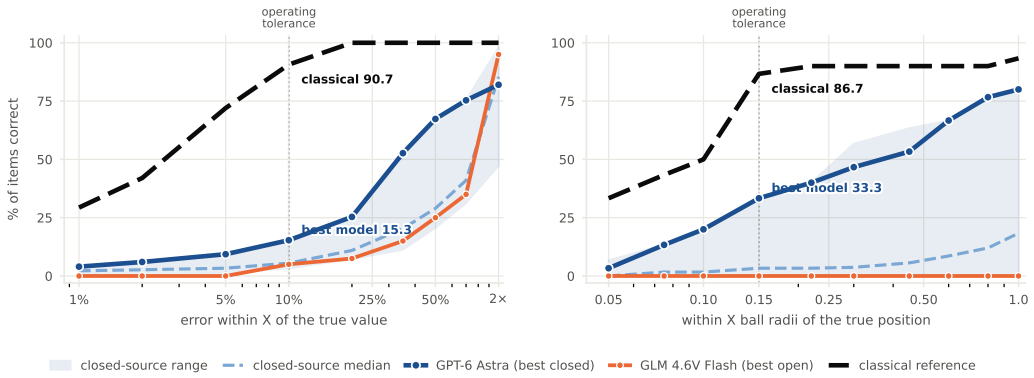


Figure 4: Accuracy against how much precision is demanded, with the classical method that solves the task on the same axes; the dotted line is the tolerance each task is scored at.

5.3 Weak rule induction

The measurements agree on one thing: what the models lack is not perception but a procedure for turning observations into a law. They see the scene — forced choice is at 97.5 against a ceiling of 100 — they can compare two scenes, they can name a great many plausible laws, and they can place an answer within a factor of two. What they never do is derive the law from the clip: asked which laws govern a world, they answer with the prototype the scene reminds them of (drag when the speed changes, a collision when it changes abruptly), and when the prototype is wrong the answer is wrong by an entire kind of mechanics rather than by a neighbouring law — 15.9% land in the same kind, against 13.6% for names drawn at random. Asked for the parameter, they supply the order of magnitude and lose the value: at the tolerance the benchmark scores, the best of them is at 15.3% where a classical estimator on the same pixels is at 90.7%.

This is a habit rather than a limit of the input. The information needed is in the frames throughout, and a hand-written estimator recovers it; the models have not been trained to look for it, and they answer from the nearest familiar mechanism instead. That habit is what matters away from a benchmark. A body encountering ground it has not walked on, an instrument whose reading drifts in an unaccustomed way, a simulator whose resistance law is not the textbook one: each is a case where the governing law has to be inferred from a few seconds of observation, and where answering with the most similar familiar mechanism is confidently wrong. These tasks are small and synthetic, and no model was trained for them — which is precisely what makes the result worth reporting. What they measure is not how much physics the models know, but whether they can find the physics they do not know.

6 Conclusion

We introduced **CognPhys**, a benchmark for evaluating inductive physical reasoning in vision-language models. Unlike existing benchmarks that assume fixed Newtonian laws, CognPhys requires models to infer unfamiliar physical rules from visual observations through rule identification, parameter estimation, and future prediction.

Our experiments show that current VLMs remain strongly dependent on memorized physical priors. Although some models can make coarse comparisons and identify certain rules, they perform poorly when required to estimate absolute quantities or predict trajectories under non-standard dynamics. The results indicate that the main challenge lies not in basic visual perception, but in converting observed motion into calibrated physical rules.

The limitations are those of a controlled, synthetic study. The worlds are generated rather than filmed, so what the benchmark measures is whether a model can induce a rule inside a designed space of physics. Extending the same tasks to filmed scenes is the natural next step.

REFERENCES

Anthropic. Claude Opus 5.5 System Card. https://www.anthropic.com/claude-opus-5-5-system-card, 2026. Accessed: 2026-09-25.

Yimeng Chen, Piotr Piękos, Mateusz Ostaszewski, Firas Laakom, and Jürgen Schmidhuber. PhysGym: Benchmarking LLMs in interactive physics discovery with controlled priors. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1556. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/42abcf9dffa48ca44fea1c497539a914-Paper-Datasets_and_Benchmarks_Track.pdf.

Song Dai, Yibo Yan, Jiamin Su, Zihao Dongfang, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and Xuming Hu. PhysicsArena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17290–17316, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.937. URL https://aclanthology.org/2025.findings-emnlp.937/.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. PHYSICS: Benchmarking foundation models on university-level physics problem solving. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11717–11743. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.610. URL https://aclanthology.org/2025.findings-acl.610/.

Google DeepMind. Gemini 3.8 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-8-Flash-Model-Card.pdf, 2026. Accessed: 2026-09-25.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11287–11297, June 2021.

Nolan Koblischke, Hyunseok Jang, Kristen Menou, and Mohamad Ali-Dib. Gravity-Bench-v1: A benchmark on gravitational physics discovery for agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31084–31104. PMLR, 2025. URL https://proceedings.mlr.press/v267/koblischke25a.html.

Chak-Wing Mak, Guanyu Zhu, Boyi Zhang, Hongji Li, Xiaowei Chi, Kevin Zhang, Yichen Wu, Yangfan He, Chun-Kai Fan, Wentao Lu, Kuangzhi Ge, Xinyu Fang, Hongyang He, Kuan Lu, Tianxiang Xu, Li Zhang, Yongxin Ni, Youhua Li, and Shanghang Zhang. PhysicsMind: Sim and real mechanics benchmarking for physical reasoning and prediction in foundational VLMs and world models, 2026. URL https://arxiv.org/abs/2601.16007.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, 2026. Accessed: 2026-09-25.

Vimukthini Pinto, Chathura Gamage, Cheng Xue, Peng Zhang, Ekaterina Nikonova, Matthew Stephenson, and Jochen Renz. NovPhy: A physical reasoning benchmark for open-world AI systems. *Artificial Intelligence*, 336:104198, 2024. doi: 10.1016/j.artint.2024.104198. URL https://www.sciencedirect.com/science/article/pii/S0004370224001346.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao,

Qing-Hong Cao, Ming xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. PHYBench: Holistic evaluation of physical perception and reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-0044. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/01e793f8cf5689aa0ea46a1b01071bea-Paper-Datasets_and_Benchmarks_Track.pdf.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics reasoning. *CoRR*, abs/1803.07616, 2018. URL https://arxiv.org/abs/1803.07616.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K. Reddy. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55325–55359. PMLR, 2025. URL https://proceedings.mlr.press/v267/shojaee25a.html.

Tianyi Tang, Zhuoyi Lin, Zeyu Feng, Tianyi Ma, Yew-Soon Ong, Ivor Tsang, and Haiyan Yin. Causal scaffolding for physical reasoning: A benchmark for causally-informed physical world understanding in VLMs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9848–9859, New York, NY, USA, 2026. Association for Computing Machinery. doi: 10.1145/3770855.3817582. URL https://doi.org/10.1145/3770855.3817582.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. WorldBench: Benchmarking physical understanding of world models by isolating physics concepts, 2026. URL https://arxiv.org/abs/2601.21282.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. PhysUniBench: A multi-modal physics reasoning benchmark at undergraduate level, 2026. URL https://arxiv.org/abs/2506.17667.

Xiyang Wu, Zongxia Li, Jihui Jin, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. MASS: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9433–9443, June 2026.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. SeePhys: Does seeing help thinking? – benchmarking vision-based physics reasoning. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-3130. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/86a472735333e2a9195b9c04eb324dea-Paper-Datasets_and_Benchmarks_Track.pdf.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69849–69877. PMLR, 2025. URL https://proceedings.mlr.press/v267/xu25ai.html.

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. URL https://openreview.net/forum?id=HkxYzANYDB.

Tianshi Zheng, Kelvin Kiu Wai Tam, Newt Nguyen Kim Hue Nam, Baixuan Xu, Zhaowei Wang, Cheng Jiayang, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. NewtonBench: Benchmarking generalizable scientific law discovery in LLM agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Gk6umqW74m.

A Appendix

A.1 The laws the worlds are built from

Worlds are assembled from rule templates rather than hand-authored. Track A draws one law per world from the families below; Track B composes several laws drawn from a closed vocabulary of eleven. Both are listed here so that the task descriptions in the main text can be read against concrete physics.

The two vocabularies come from the same engine and overlap, but neither is a subset of the other. `vortex` and `speed_cap` appear only in Track B; uniform acceleration is realised in Track A as a global gravity field and in Track B as a constant body force; and the remaining labels in the Track B vocabulary each correspond to a Track A family testing the same law.

A.1.1 Track A: the families

Within a family the worlds form twin pairs that differ only in the value of the parameter being estimated. The two members of a pair are compared through the effect size rather than the nominal parameter, so the ranges quoted below are measured ground-truth values across the family.

Table 4: The Track A families. The *hit* variants instantiate the same law as their base family but place a second body in the path of the observed one, so that the world ends in a collision.

Family	Law	Parameter	Range
`gravity_g`	uniform acceleration in a fixed direction	g	3.3–10.9 m/s^2
`drag_b`	resistance proportional to speed	b	0.33–1.37 s^{-1}
`roll_resist`	constant deceleration from the floor	deceleration	0.92–2.73 m/s^2
`friction_zone`	stronger braking inside a coloured patch	deceleration	2.23–11.45 m/s^2
`push_zone`	sustained push inside a coloured patch	acceleration	3.33–10.89 m/s^2
`two_zone`	two push patches in a row	acceleration, far patch	2.06–8.62 m/s^2
`push_brake_zone`	a push patch followed by a braking patch	deceleration, far patch	2.17–7.94 m/s^2
`field_strength`	attraction to a point, $\propto 1/d$	acceleration at $d = 2$ m	2.01–7.25 m/s^2
`field_exponent`	attraction to a point, $\propto 1/d^p$	exponent p	1.0–2.0
`impact_dv`	velocity injected by a collision	Δv	4.58–8.88 m/s
`_hit`	the families `roll_resist`, `friction_zone`, `push_zone`, `two_zone` and `push_brake_zone` each also appear with a second body standing in the path of the observed one, so that the world ends in a collision		

A.1.2 Track B: the closed rule vocabulary

The prompt lists the whole vocabulary below, and the model must return every entry that governs the world. Naming the full list is therefore always available as a strategy, which is why exact set match is reported alongside precision and recall.

A.2 Zero-information references

The main text reports one reference per column — what a classical method recovers from the same pixels. Three references that ignore the video entirely are reported here, because an absolute score is not interpretable on its own and because the size of a model's shortfall depends on which of them one has in mind.

The three columns say different things about the models, and it is worth keeping them apart. On parameter estimation every closed-source model is below the reference: the best reaches 15.3% at 10% relative error where the family median — a constant per family, which never looks at the clip

Table 5: The closed vocabulary used by Track B. Each world composes several of these laws.

Label	Definition given to the model
accel	uniform acceleration under a constant force
drag	resistance proportional to speed; the body slows exponentially
roll	continuous deceleration from the ground, with no marked region
brake_patch	a coloured region of ground where the body decelerates faster
push_patch	a coloured region of ground that pushes the body forward
terrain	two coloured regions with different effects
field	attraction to a fixed point, falling off as $1/d$
power_field	the same, but falling off as a power law with an exponent other than one
vortex	a tangential force about a fixed point, so the body orbits it
impact	the struck body suddenly gains speed in a collision
speed_cap	speed saturates at a terminal value

Table 6: What is achievable without looking at the video, on the same worlds as the corresponding column of Table 3.

Task	Reference	Score	Best model
Rule determination	a uniformly drawn rule subset	1.6%	20.1%
Parameter estimation	the family median (no video)	24.0%	15.3%
Future prediction	keeping the last velocity	43%	33.3%

— reaches 24.0% (16.7% at 5%). On future prediction the best model is also below the reference, but narrowly: 33.3% within 0.15 ball radii against 43% for extrapolating the last observed velocity. On rule determination the reference is the chance level of drawing a rule subset at random, 1.6%, and the models are well above it (Table 3), which is why that task is analysed separately in Section 5.

A.3 The failure is not an artefact of the viewpoint

Track A and Track A+ render the identical worlds from a second, $45°$ oblique viewpoint: the same physics, parameters, questions and ground truth, so any change in score is attributable to the camera. Figure 5 pairs the two runs per model. The magnitude of the failure, and every model's position relative to the no-video reference, are unchanged by the viewpoint: every model sits below that reference in both views, and no model moves from one regime to the other. The benchmark therefore measures something the choice of camera does not decide.

A.4 What accounts for a wrong answer

The main table is indexed by model and by task. It is worth asking how much either of those, and the physics itself, actually explains. Over the same 2,818 (model, world) observations we decompose the variance of the log relative error into the factors available (Figure 6A).

Which *law* is being tested explains 19.7% of the variance; which *model* is being asked explains 3.7%. Knowing the exact world — its family, its parameter, its initial conditions — bounds any world-level predictor at 39.7%, and the paired-camera runs cap the reproducible part of that at roughly 28%. The reading that matters here is the first one: *the benchmark separates laws far better than it separates models*. The frontier models are close enough to interchangeable on this task that the identity of the model is a weaker predictor of the error than the identity of the law.

The third candidate is the one the benchmark was designed around. Worlds were admitted only if a classical estimator could recover the parameter from the same pixels, and the *effect size* — the divergence of the world's trajectory from that of the same world with the parameter nulled, in ball diameters — was recorded as the difficulty axis. Within the admitted range it explains **0.4%** of the variance: across a fifteen-fold span in signal strength, pooled model accuracy moves between 3.6% and 8.5% with no monotone trend, while the classical estimator stays between 83% and 96% (Figure 6B). A world is not hard because its effect is subtle. The gate did its job — every admitted world is solvable — and what the models fail on is orthogonal to how much signal that world contains.

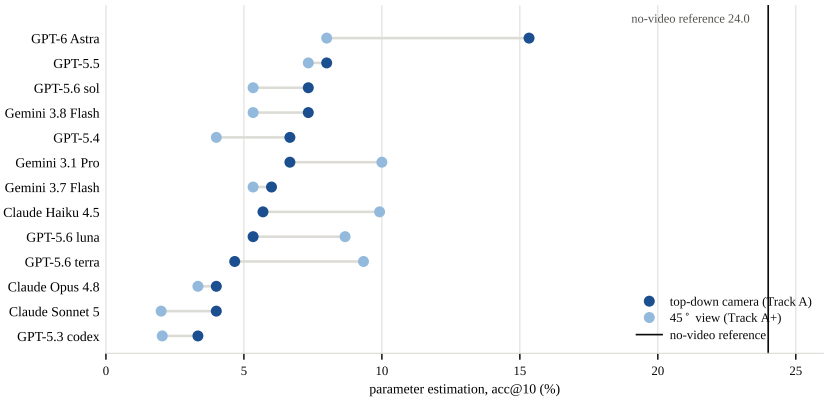


Figure 5: Same worlds, two cameras. Each model is one dot per camera, joined; the horizontal position of the pair is what matters. The median change between views is 2.0 points and Spearman $\rho = 0.39$, and all 13 models stay below the no-video reference line in both views — the classical ceiling (90.7) is off scale to the right.

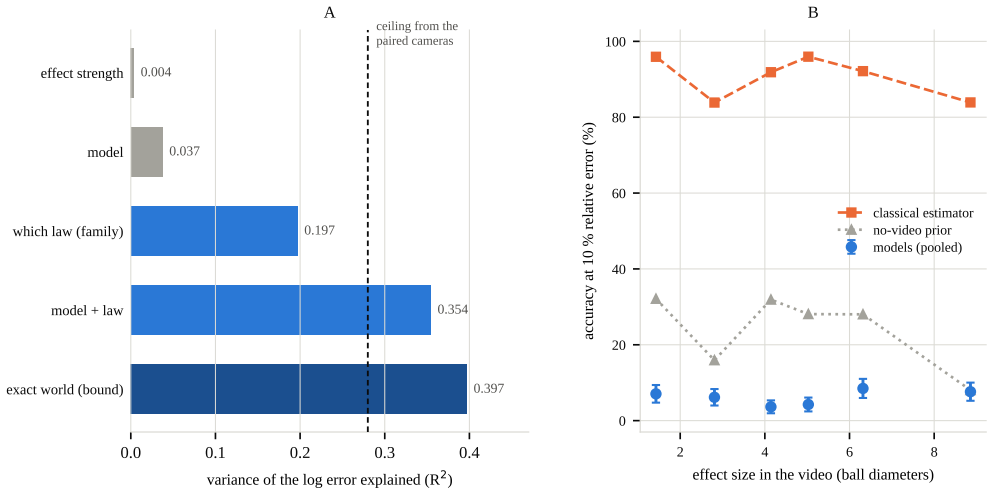


Figure 6: What explains a wrong answer on Track A **(A)**, and accuracy against the signal in the video **(B)**. **(A)**: the share of the log relative error's variance accounted for by each factor, over 2,818 (model, world) observations. **(B)**: accuracy against the effect size the benchmark was designed around, in ball diameters; bars are 95% binomial intervals. Signal strength explains the least — and it is the property the worlds were built around.

Figure 7 gives the same picture per family, with the classical estimator's own error carried as a reference row, so that "hard for the model" can be told apart from "hard for the pixels".

A.5 Reading the coordinates versus reasoning from them

Poor prediction accuracy does not by itself show that the models reason badly about dynamics, since a model that misreads the ball's position will extrapolate badly even with perfect mechanics. The scaffolded condition separates the two: the model is required to write down the ball's coordinates frame by frame before extrapolating, so its reported coordinates can be taken as its own input and extrapolated mechanically.

For GPT-5.5 on the $+1$ s horizon, the coordinates it reports agree with the classical tracker to a median of 3.7 px, or 0.19 ball radii. Yet extrapolating *those same coordinates* mechanically, by a

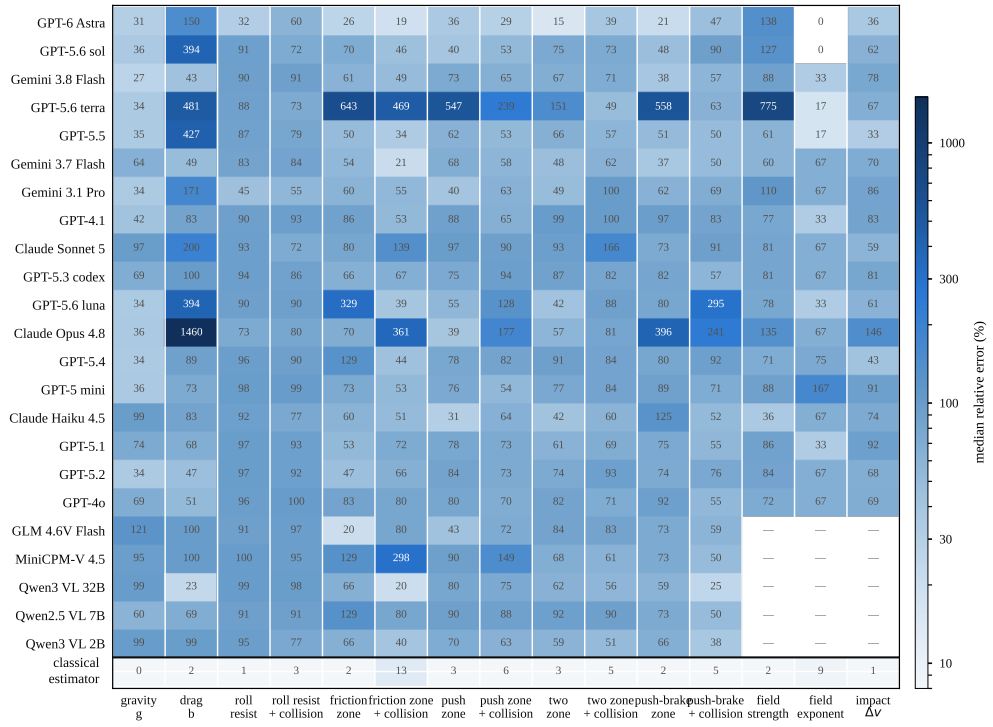


Figure 7: Median relative error (%), by family. Closed-source $n = 150$, open-source $n = 40$; the colour scale is logarithmic over 8–1500%; a dash means the family was not in the open-source subsample. The bottom row is the classical estimator, so a column that is dark for the models but light for the estimator is a failure of the model rather than of the pixels.

quadratic fit, is far worse (median 7.79 radii) than the model's own answer (0.49 radii). The model's reasoning is therefore doing real work on this task; the limit is not that it declines to reason about the motion it has described.