ScenePhys — Controllable Physics Videos for World-Model Evaluation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present PhET-Physics-VideoQA, a controlled benchmark for assessing physics understanding in vision–language models (VLMs) from video. The corpus comprises 382 short clips sourced from PhET Interactive Simulations, covering 17 topics across four fields (Mechanics & Fluids, Optics, Electromagnetism & Circuits, and Quantum Mechanics). Each clip is paired with a triad of expert-validated questions—conceptual, numerical, and error-detection—yielding 1,146 Q/A items. The design emphasizes pixel-grounded reasoning: many clips display gauges and sliders so that models must recover numeric values from frames rather than rely on language priors.

Evaluation is reproducible and type-specific. Numerical items are graded deterministically against gold values with absolute/relative tolerances and unit checks. Conceptual and error-detection items are judged with a rubricized LLM that returns strict JSON, supports dual-judge scoring, and is run at zero temperature with cached transcripts.

We report results for three video-capable VLMs (GPT-4o-mini, Gemini-2.5-Flash-Lite, Qwen-VL-Plus). Across domains, error-detection ("trap") questions are consistently the most difficult, typically scoring 0.5–1.3 points lower than conceptual or numerical items on a 1–5 scale. Higher-concept physics, particularly quantum content, remains challenging for all models. PhET-Physics-VideoQA thus offers a rigorous, transparent, and cost-efficient testbed for measuring genuine physics competence in video settings and a practical resource for advancing research on multimodal world.[1]

1 Introduction

World models aim to learn predictive, manipulable representations of environments that support planning, control, and transfer across tasks [Ha and Schmidhuber, 2018, Hafner et al., 2019, 2020, Reed et al., 2022, Ahn et al., 2023, Zitkovich et al., 2024]. Yet mounting evidence suggests that contemporary vision–language models (VLMs) often exploit superficial regularities rather than forming physically meaningful abstractions: they can under-use shape structure and fail on illusions that humans find trivial [Hemmat et al., 2024], degrade sharply under controlled distribution shifts [Barbu et al., 2019], and rely on language priors that inflate in-domain accuracy [Agrawal et al., 2018]. Behavioral test suites further expose compositional and alignment gaps [Zhao et al., 2022, Thrush et al., 2022], while video reasoning benchmarks surface limitations in temporal and causal understanding [Yi et al., 2020]. These diagnostics collectively motivate benchmarks that (i) control

[1]Project: https://scenephys.github.io/ ; Dataset: https://huggingface.co/datasets/ScenePhys/ScenePhys ; Code: https://github.com/ScenePhys/codebase.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

confounds, (ii) span diverse physics regimes, and (iii) *separate* genuine mechanistic reasoning from pattern matching.

We present a controlled, video-based benchmark built from the widely used *PhET Interactive Simulations* ecosystem of physics demonstrations. The dataset comprises *382* curated simulation videos covering core domains (kinematics, dynamics, collisions, geometric optics, electricity and magnetism, circuits, fluids specifically buoyancy, and quantum phenomena). Each video is paired with a triad of questions that probe complementary facets of understanding: *(i) Conceptual* (laws, invariants, qualitative trends), *(ii) Numerical* (parameter-grounded calculations with unit discipline), and *(iii) Error-detection* (identifying idealizations, hidden losses, or setup inconsistencies). By design, success requires reasoning over physical invariants and counterfactuals rather than exploiting spurious visual or linguistic shortcuts.

In contrast to existing video reasoning datasets that emphasize synthetic collisions, goal satisfaction, or open-domain narratives [Yi et al., 2020, Bakhtin et al., 2019, Bear et al., 2021], and to education datasets centered on static diagrams [Lu et al., 2022], our benchmark leverages high-quality PhET simulations to couple *pixel-visible numeric panels* with *curated, per-video triads* of conceptual, numerical, and error-detection questions. This combination enforces grounding in measured quantities, tests unit- and sign-discipline alongside qualitative reasoning, and surfaces robustness to idealizations—providing a complementary, diagnostics-first view of multimodal physics understanding.

Contributions. (1) A parameterized, physics-grounded *video* benchmark of 382 PhET simulations spanning multiple domains. (2) A three-question evaluation schema (conceptual, numerical, error-detection) that disentangles types of understanding and pressures models to rely on the *right* invariants. (3) Comprehensive baselines and analyses across contemporary VLMs, surfacing systematic error modes linked to abstraction gaps, unit handling, and hidden-assumption sensitivity [Hemmat et al., 2024, Barbu et al., 2019, Agrawal et al., 2018, Zhao et al., 2022, Thrush et al., 2022, Yi et al., 2020].

Alignment with workshop focus: Interactive scene generation and downstream tasks. Our benchmark targets physically plausible, *controllable* video scenes and evaluates properties directly relevant to downstream agents: temporal consistency and conservation laws (conceptual), actionable predictiveness (numerical), and robustness to modeling choices and hidden assumptions (error detection). As such, it provides an evaluation substrate for models that generate or condition on interactive simulations, and a diagnostic lens on whether VLMs—and world-model pipelines built atop them—encode abstractions suitable for planning and policy learning [Ha and Schmidhuber, 2018, Hafner et al., 2019, 2020].

2 Related Work

Multimodal benchmarks for physical reasoning. A substantial body of work probes whether models can reason about dynamics and causality from video. CLEVRER targets temporal and causal reasoning in synthetic collisions with descriptive, explanatory, predictive, and counterfactual queries, revealing that perception-only success does not translate to causal competence [Yi et al., 2020]. PHYSION moves toward more realistic simulations (e.g., rolling, sliding, falling, collisions, deformation) and compares machine predictions with human judgments, finding persistent gaps and advantages for object-centric representations [Bear et al., 2021]. PHYRE frames physical reasoning as solving 2D puzzles with an emphasis on generalization and sample efficiency [Bakhtin et al., 2019]. Our benchmark differs in three ways: (i) we build on PHET educational simulations to improve reproducibility and pedagogical fidelity; (ii) each video is paired with a fixed triplet of questions (conceptual, numerical, error-detection) aligned to instructional goals; and (iii) we evaluate *multiple* VLMs under standardized prompts. The reliability and broad adoption of PHET as a learning tool motivate its use as a controlled yet authentic source of stimuli [Wieman et al., 2008, 2010].

Video QA and educational multimodal reasoning. General VideoQA benchmarks emphasize everyday activities, temporal order, and causal relations in natural videos; for example, NEXT-QA targets causal and temporal action reasoning with both multi-choice and open-ended formats, showing that strong systems still struggle with explanatory questions [Xiao et al., 2021]. Complementary educational resources such as TQA and AI2D/AI2D-RST examine multimodal comprehension in

Dataset	Mod.	Lang.	Task	Size	Open	Numeric UI	Diff./Trap	Notes / Primary reference
PhET-Physics-VideoQA (Ours)	Vid	Eng	VideoQA (conceptual / numerical / error-det.)	382 vids, 1146 Q/A	✓	✓	✓/✓	Educational simulations; parameterized clips (densities, n, drag, etc.); three question types.
CLEVRER[a] [Yi et al., 2020]	Vid	Eng	VideoQA (desc./expl./counterf.)	20k vids; >300k Q	✓	✗	✗/✗	Synthetic collisions; causal/temporal reasoning with counterfactuals.
CRIPP-VQA[b] [Patel et al., 2022]	Vid	Eng	VideoQA (template queries over primitive physical processes)	$\sim$2.4k vids; $\sim$74k Q/A	✓	✗	✗/✗	Synthetic, short clips of rudimentary processes; template-style questions; not an educational physics benchmark; no numeric readouts.
Physion [Bear et al., 2021]	Vid	Eng	Physical prediction (no QA)	$\sim$1.2k clips (8 scenarios)	✓	✗	✗/✗	Predict roll/slide/bounce outcomes; human vs model comparisons.
PHYRE [Bakhtin et al., 2019]	Sim	Eng	Goal achievement / planning	2 tiers; 25 templates $\times$ 100 tasks each ($\sim$5k)	✓	✗	✓/✗	Parameterized 2D physics puzzles; generalization within/across templates.
ScienceQA [Lu et al., 2022]	Img+Txt	Eng	MCQA (explanations)	$\sim$21k Q/A	✓	✗	✗/✗	K–12 science with images/diagrams; chain-of-thought supervision.

[a] Per-type CLEVRER counts: 126,304 descriptive, 122,461 explanatory, 41,021 predictive, 12,523 counterfactual.
[b] CRIPP-VQA focuses on primitive, compositional physical processes with template-based questions; **it is not designed for high-level, educational physics reasoning or numeric problem solving**.

Table 1: **Positioning our benchmark among nearby datasets.** "Numeric UI" flags whether raw on-screen numeric readouts (gauges/sliders) are part of the visual input. "Diff./Trap" indicates explicit difficulty labels and the presence of trap/error-detection prompts (see Sec. 3.4.

K–12 science and highlight the challenges of diagram-grounded reasoning [Li et al., 2018, Kembhavi et al., 2016, Hiippala et al., 2021], while ScienceQA scales to $\sim$21k multimodal questions with lectures and explanations, demonstrating benefits from chain-of-thought supervision [Lu et al., 2022]. Our benchmark sits alongside these efforts by focusing on canonical physics phenomena with controllable conditions and numeric readouts, enabling quantitative assessment and precise cross-model comparisons that complement natural-video and diagram/text settings.

Numerical visual reasoning, broad LMM evaluations, and video-capable models. Chart/plot QA corpora probe perception-to-calculation pipelines via value extraction and tolerance-aware grading—principles we adopt for our numerical items (units, error tolerances, robustness to reading noise)—as exemplified by PlotQA and ChartQA [Methani et al., 2020, Masry et al., 2022]. Broad, heterogeneous benchmarks such as MMMU and MathVista further reveal persistent gaps in mathematically grounded multimodal reasoning despite rapid progress [Yue et al., 2024, Lu et al., 2023]. In parallel, open efforts extend image-centric LMMs to the video domain through instruction tuning and unified tokenization (e.g., Video-LLaVA, Video-ChatGPT), typically optimizing for conversational understanding rather than parameter-grounded consistency [Lin et al., 2023, Maaz et al., 2023]. Our physics-focused, numerically anchored evaluation bridges these lines of work by testing whether video-capable models can maintain state tracking, read parameters reliably, and respect physical constraints—capabilities that standard conversational video setups may not directly assess.

Probing VLM robustness and abstraction. Recent diagnostic datasets show that vision–language models (VLMs) often rely on superficial cues rather than true abstraction. Hemmat et al. demonstrate failures on visual illusions due to under-use of shape structure [Hemmat et al., 2024], while ObjectNet reveals over-reliance on context [Barbu et al., 2019]. In VQA, VQA-CP exposes shortcut use of answer priors. Behavioral test suites such as VL-CheckList and Winoground further probe object attributes, negation, and compositional binding [Zhao et al., 2022, Thrush et al., 2022]. For temporal and causal reasoning, CLEVRER reduces success via superficial cues [Yi et al., 2020]. Together, these motivate our inclusion of error-detection prompts to better separate physical reasoning from heuristic pattern matching.

Positioning among physics and multimodal QA benchmarks Prior work probes video physical reasoning via synthetic collisions and counterfactual queries (CLEVRER; Yi et al., 2020), goal-

Figure 1: **PhET-Physics-VideoQA overview.** The central sunburst summarizes the *four fields*—Mechanics & Fluids, Optics, Electromagnetism & Circuits, and Quantum Mechanics—and their *17 topics* covered by our 382 simulation clips.

driven puzzle solving that stresses template generalization (PHYRE; Bakhtin et al., 2019), and predictive judgments about real/simulated dynamics (PHYSION; Bear et al., 2021); other resources target physics *QA* directly over videos (e.g., CRIPP-VQA, ∼2.4k videos/∼74k Q/A) or focus on diagram/image-based education QA (SCIENCEQA; Lu et al., 2022). Our benchmark (Table 1) fills a complementary, under-served niche by (i) using controlled *educational simulations* (PhET) with visible numeric UI (gauges/sliders/readouts) so answers can be grounded in pixel-level measurements; (ii) evaluating three orthogonal skills via per-video triads—*conceptual*, *numerical* (unit-checked with explicit tolerances), and *error-detection*—that target known VLM failure modes; and (iii) covering a broad syllabus (fluids, mechanics, optics, E&M, circuits, quantum mechanics) to enable disaggregated domain analysis. Unlike prior video datasets that avoid numeric UI [Yi et al., 2020, Bear et al., 2021] or center static diagrams [Lu et al., 2022], our setting *requires* consistency between language, on-frame measurements, and physical constraints, yielding a sharper diagnostic of physics competence beyond language priors.

3 Dataset

3.1 Design Goals & Scope

Our benchmark targets video understanding of canonical physics under controlled, measurable, and repeatable conditions. We construct short clips from the *PhET Interactive Simulations* ecosystem [Wieman et al., 2008, 2010], where on–screen gauges and sliders expose parameters and outcomes. The dataset comprises **382** curated clips spanning **four fields**— *Mechanics & Fluids*, *Optics*, *Electromagnetism & Circuits*, and *Quantum Mechanics*—across **17 topics** (e.g., buoyancy, collisions, lenses/mirrors, Coulomb's law, generator, RC time constant, projectile motion, quantum tunneling). Each clip pairs with *three* complementary questions that probe (i) *conceptual* knowledge (laws, invariants, qualitative trends), (ii) *numerical* competence (parameter-grounded calculations with unit checks and tolerances), and (iii) *error detection* (identifying idealizations, hidden losses, or setup inconsistencies). Because the governing variables are *visible in the pixels* (readouts, sliders), correct answers must be simultaneously consistent with the visual evidence and with the underlying physics, making it difficult to rely on language priors alone.

Intended uses beyond evaluation. While the primary purpose is a standardized *diagnostic* for video-language models' physics understanding, the dataset and answer schema are designed to support additional research uses:

- **Supervised fine-tuning (SFT).** The question–answer pairs (with tolerance-aware numeric targets) can supervise models to (a) read on-screen numeric UI, (b) apply unit and sign discipline, and (c) map qualitative video cues to the correct physical regime (e.g., float/sink, real vs. virtual images).

- **Preference optimization/reward modeling.** The three question types furnish natural comparison signals (e.g., correct reasoning but wrong arithmetic vs. spurious guess with right number), enabling preference datasets for DPO/RLHF-style training of physics-aware responses.
- **Auxiliary tasks for grounding.** The same clips admit multi-task objectives such as OCR-of-readouts, unit tagging, dimensional analysis checks, or equation selection, which can be attached as auxiliary losses to improve numeric grounding in video LMMs.
- **Curriculum and generalization studies.** The coverage across *fields* and *topics* allows curricular schedules (easy→hard, single-parameter→multi-parameter) and systematic generalization protocols (e.g., train on water/oil densities, test on mercury; train on concave mirrors, test on convex).
- **World-model stress tests.** Because scenarios expose controllable parameters and predictable outcomes, the benchmark can serve as a held-out probe for video *world models*: models that claim to encode dynamics should exhibit consistent performance across parameter sweeps (e.g., $F_b \propto \rho V$, $1/f$ scaling in optics, exponential RC time constants).
- **Instruction following and tool use.** The explicit numeric targets and unit tolerances make the dataset suitable for instruction-tuning models to follow physics-specific directives ("compute," "estimate," "explain assumption") and for evaluating tool-augmented reasoning (e.g., calculator use) under visual grounding.

These secondary uses are optional and orthogonal to the core benchmark; they are included to facilitate research on *how* video models internalize and operationalize sophisticated physics rules, not merely whether they can answer in-domain prompts.

3.2 Data Compilation

All clips are recorded from *PhET Interactive Simulations* [Wieman et al., 2008, 2010] and were designed by a small team of physics practitioners (three co-authors with Electrical Engineering/Physics training). For each module the team specified (i) visible instruments (sliders, gauges, readouts), (ii) controllable parameters and ranges (e.g., object volume/density, index of refraction, spring constant, charges, area and distance of plates, gravity, drag model, width of the wave fuction), and (iii) a short scripted interaction (initial conditions, parameter sweep/perturbation, expected qualitative outcome). Each finalized clip is paired with three questions—*conceptual*, *numerical*, and *error-detection*—initially drafted by *GPT-5 Thinking* from a structured scenario card (simulation, parameters, difficulty, intended concept) and then fully vetted by experts for scientific correctness. During validation, the team calibrated numerical targets (units, significant figures, *a priori* absolute/relative tolerances) and bound error-detection prompts to the clip's idealizations (e.g., zero drag, no frictional losses, paraxial approximation). Gold answers include a concise rationale, the canonical formulas used, and a final numeric result with unit and tolerance. Prior to release we ran automated consistency checks (unit sanity, sign conventions, recomputation from metadata) and a two-pass human audit to remove duplicates/near-duplicates. Each datum ships with (1) the standardized video frames, (2) a machine-readable metadata JSON (module, parameters, UI elements, difficulty), and (3) the QA triplet with gold answers and grading rubric (including tolerance rules), enabling turnkey evaluation and secondary uses such as SFT, preference modeling, and curriculum/generalization studies.

3.3 Metadata summary

Our corpus contains **382** clips paired with **1146** Q/A items (three per clip), covering **17** topics grouped into four fields: *Mechanics and Fluids* (**79** clips: buoyancy, projectile motion, collisions, masses and springs, simple pendulum); *Optics* (**50**: convex/concave lenses, convex/concave/flat mirrors); *Electromagnetism and Circuits* (**130**: capacitance, Coulomb's law, generators, RC time constant); and *Quantum Mechanics* (**123**: hydrogen-atom/spectral behavior, photon polarization, quantum tunneling). The most represented topics are hydrogen atom models (**55**), quantum tunneling (**54**), capacitance (**40**), Coulomb law (**35**), RC time constant (**30**), generator (**25**), projectile motion (**25**), and buoyancy (**24**). Each clip is annotated with three complementary question types—*conceptual*, *numerical*, and *error-detection*—anchored to the same video instance (cf. App. A.2 for extended metadata details).

3.4 Question Generation & Types

Overview. Each video instance is paired with **three orthogonal item families** designed to probe complementary facets of physics understanding: (i) *Conceptual* (principles, invariants, qualitative monotonicities), (ii) *Numerical* (parameter–grounded calculation with unit discipline), and (iii) *Error–detection* (recognition of idealizations, hidden losses, and setup inconsistencies). Items are authored from a scenario–metadata binding (visible readouts, controlled parameters, units) to ensure the question is *video–specific*, unambiguous, and reproducible.

Item specifications.

Conceptual. *Scope:* laws and qualitative trends (e.g., Archimedes, Snell, energy conservation, momentum, Faraday's law, RC dynamics). *Form:* "If parameter X increases while Y is held fixed, what is the effect on Z? Justify by naming the governing principle." *Evidence required:* correct directionality and an explicit citation of the relevant law or invariant; reference to features visible in the clip (gauges/sliders).

Numerical. *Scope:* single– or few–step calculations bound to the video's numeric UI (e.g., read ρ, V, R, C, n, v_0, angles). *Form:* "Using the on–screen values $(A, B, \ldots)$, compute Q and report with units." *Constraints:* unit correctness, appropriate rounding/significant figures, and a *tolerance window* (absolute/relative) predeclared per item to account for display precision.

Error–detection. *Scope:* identification of simplifying assumptions (zero drag/friction, perfectly rigid bodies, lossless components), hidden confounders (misread units, occluded scales), or inconsistent setups. *Form:* "Identify the dominant idealization in the clip and predict the qualitative change in the outcome if it is violated." *Evidence required:* naming the assumption and a correct counterfactual (directionally and mechanistically).

Difficulty. We tag each item with one of three difficulty levels—*easy*, *moderate*, or *hard*—based on combined cognitive load (recall vs. multi–step reasoning), numeric complexity (single vs. chained formulas/conditionals), and perceptual burden (reading small/fast UI changes). Labels are assigned during expert review and are used only for analysis/stratification, not for prompting.

Trap concept (implicit, not flagged). Although we analyze common failure modes—(i) **units/scale** (unit consistency, order-of-magnitude checks), (ii) **sign/direction** (conventions, image vs. object side, current/field orientation), (iii) **parameter readout** (misreading sliders/gauges), and (iv) **idealization violations** (zero drag/friction, perfect rigidity, lossless elements)—we do *not* store an explicit "trap flag" in the metadata. Instead, these aspects are *implicitly* probed by the dedicated *error-detection* question type and enforced by the grading rubric (unit checks, tolerance windows, and counterfactual reasoning). Aggregated diagnostics may reference these categories in analysis, but no per-item trap annotation is included in the released data.

4 Experiments and Results

4.1 Experimental Configuration

Corpus and tasks. We evaluate on **382 video scenarios** (17 physics labs), each paired with a triad of *Conceptual*, *Numerical*, and *Error–Detection* items for a total of **1146 Q/A**.

Model suite and rationale. We select three video–capable VLMs balancing capability, cost, and ecosystem coverage: GPT-4O-MINI (OpenAI; strong small model in the GPT-4o family), GEMINI-2.5-FLASH-LITE (Google; fast multimodal variant), and QWEN-VL-PLUS (Alibaba; widely used open(-ish) stack). This set spans two strong proprietary baselines with robust video APIs and one popular, cost–efficient open family—useful for the community to replicate/extend.[2]

Video preprocessing. Clips are standardized to `fps`$=3.0$, `max_frames`$=40$, `jpg_quality`$=95$, then base64–encoded for API transmission. This budget preserves salient state changes (e.g., gauge/slider motion, collisions) while controlling cost and latency.

[2] We cite family reports for context: GPT-4o system overview [ope, 2024], Gemini technical reports [Comanici et al., 2025], and Qwen2-VL [Wang, 2024].

Category	**Question Type**	**gpt-4o-mini**	**gemini-2.5 flash-lite**	**qwen-vl-plus**	**Type Avg.**
	Conceptual	4.6	4.5	2.3	3.80
Mechanics & Fluids	**Error Detection**	3.0	3.2	2.5	**2.90**
	Numerical	4.0	4.2	2.1	3.43
	Conceptual	3.7	3.8	1.6	3.03
Quantum Mechanics	**Error Detection**	2.4	2.5	1.5	**2.13**
	Numerical	3.3	3.5	1.6	2.80
	Conceptual	4.7	4.6	3.3	4.20
Electromagnetism & Circuits	**Error Detection**	3.8	3.4	3.1	**3.43**
	Numerical	4.2	4.0	3.2	3.80
	Conceptual	4.2	4.2	3.7	4.03
Optics	**Error Detection**	2.6	3.3	2.3	**2.73**
	Numerical	4.6	4.3	3.9	4.27

Table 2: LLM-as-a-judge scores (scale 1–5) by category and question type; rightmost column is the mean across models. **Error Detection** rows are consistently lower than Conceptual/Numerical.

Prompting and decoding. Unless otherwise noted: temperature $= 0$, single response per item (no self-consistency), and frame stacks passed as ordered images with a fixed instruction template (per question type).

Scoring protocol (summary). Numerical items use *deterministic*, unit–aware grading against a gold key with absolute/relative tolerances (Sec. 4.2).Conceptual and Error–Detection items are judged by an *LLM-as-a-judge* rubric on a $[1..5]$ scale with a justification string and flags; we report normalized scores and confidence–aware variants (Sec. 4.2). This mixed protocol yields objective scoring where ground truth is numeric, and calibrated rubric assessment where open-text explanations are required.

4.2 Evaluation Protocol

Setup and notation. Let $\mathcal{V}$ be the set of videos; each $v \in \mathcal{V}$ is paired with a triad of questions $\mathcal{Q}(v) = \{q^{(C)}, q^{(N)}, q^{(E)}\}$ covering *conceptual* (C), *numerical* (N), and *error–detection* (E) skills. For a model M, let $\hat{a}(q)$ denote its answer to question q. We score each question with a type–appropriate function $s(q, \hat{a}) \in [0, 1]$, then aggregate across videos, types, and physics domains [Gu et al., 2024, Li et al., 2024].

Deterministic scoring for numerical items. Each numerical question q has a gold value $y^\star$, a unit $u^\star$, an absolute tolerance τ_{abs} and a relative tolerance τ_{rel} specified in the metadata. From the model's response we parse a numeric $\hat{y}$ and unit $\hat{u}$ (unit synonyms normalized to SI). Define the admissible error

$$\tau(q) \;=\; \max\bigl(\tau_{\text{abs}},\; \tau_{\text{rel}} \cdot |y^\star|\bigr), \qquad \delta \;=\; |\hat{y} - y^\star|, \qquad \mathbb{1}_{\text{unit}} \;=\; \mathbb{1}[\hat{u} \equiv u^\star].$$

The numerical score is

$$s_N(q, \hat{a}) \;=\; \mathbb{1}_{\text{unit}} \cdot \begin{cases} 1, & \delta \le \tau(q), \\ \gamma, & \tau(q) < \delta \le \kappa\, \tau(q), \\ 0, & \text{otherwise}, \end{cases}$$

with *fixed* hyperparameters $\gamma = 0.5$ (partial credit) and $\kappa = 2$ (grace band).[3] This rubric is objective, unit–aware, and invariant to trivial rephrasings, consistent with recommendations to avoid free-form LLM judging for numeric items [Liu et al., 2023, Gu et al., 2024].

LLM-as-a-judge for conceptual and error–detection items. For C and E types we use a rubricized judge J instructed to output strict JSON: $\{$`score` $\in \{1, \ldots, 5\}$, `confidence` $\in [0, 1]$, `flags`$\}$, where `flags` captures checklist criteria (e.g., `law_invoked`, `units_issue`,

[3] We report γ, κ and the per–question tolerances in the release to ensure exact reproducibility.

`missing_assumption`). We map the 5-point rating to $[0, 1]$ via

$$r \;=\; \frac{\texttt{score} - 1}{4}, \qquad s_{C/E}(q, \hat{a}) \;=\; \frac{r\,(1 + \alpha\,\texttt{confidence})}{1 + \alpha},$$

with $\alpha = 1$ to softly incorporate judge self-confidence. To improve reliability, we optionally use two independent judges J_1, J_2 and average their scores, reporting agreement (e.g., Cohen's κ) on a held-out calibration set, following common practice in rubricized LLM-as-a-judge evaluations [Liu et al., 2023, Zheng et al., 2023, Chiang et al., 2024, Gu et al., 2024, Li et al., 2024]; we also monitor known biases and robustness concerns [Li and Others, 2025, Shi et al., 2024, Vyas and Others, 2024].

Aggregation and uncertainty. Per-type means:

$$A_t(M) \;=\; |\mathcal{Q}_t|^{-1} \textstyle\sum_{q \in \mathcal{Q}_t} s(q, \hat{a}), \quad t \in \{C, N, E\}.$$

Per-video triad score:

$$S_v(M) \;=\; \tfrac{1}{3} \sum_{q \in \mathcal{Q}(v)} s(q, \hat{a}).$$

We compute domain-wise macro averages (mechanics/fluids, optics, Electromagnetism/circuits, quantum mechanics) and an overall macro across domains to avoid topic-size bias. We attach 95% confidence intervals via stratified bootstrap over videos (10,000 resamples) and assess model differences with paired bootstraps on S_v, as recommended in recent evaluations of LLM judges and open-ended benchmarking [Zheng et al., 2023, Chiang et al., 2024, Gu et al., 2024].

Note. For completeness, we also ran an earlier "critical judge" variant (single pass, free-text rubric); its specification and outputs are documented in App. B. All reported numbers in this paper use the *Standard Judge* described above.

4.3 Results

Overall. Across all categories and types (Table 2), GEMINI-2.5-FLASH-LITE and GPT-4O-MINI are essentially tied: macro–averages of **3.79** vs. **3.76** (on a 1–5 scale), both well above QWEN-VL-PLUS (**2.59**). By domain, *Electromagnetism/Circuits* is the easiest overall (**3.81** mean), followed by *Optics* (**3.68**), *Mechanics/Fluids* (**3.38**), and *Quantum Mechanics* as the hardest (**2.66**). The best single cell is GPT-4O-MINI on Electromagnetism/Circuits–Conceptual (**4.7**); the weakest is QWEN-VL-PLUS on Quantum Mechanics–Error Detection (**1.5**).

By question type. Error Detection is consistently the bottleneck: averaged over all models and domains it scores **2.80**, trailing Conceptual (**3.77**) by $\sim$**0.97** and Numerical (**3.58**) by $\sim$**0.78**. The gap holds per–model: GPT-4O-MINI Conceptual vs. Error Detection is **4.30** $\rightarrow$ **2.95** ($\Delta \approx$**1.35**), GEMINI-2.5-FLASH-LITE **4.28** $\rightarrow$ **3.10** ($\Delta \approx$**1.18**), and QWEN-VL-PLUS **2.73** $\rightarrow$ **2.35** ($\Delta \approx$**0.38**). This validates the difficulty of our "trap" prompts that require spotting idealizations and making counterfactual predictions.

By domain (higher–concept physics). *Quantum Mechanics* depresses all models across types (e.g., Conceptual means: **3.70/3.80/1.60**; Numerical: **3.30/3.50/1.60** for GPT-4O-MINI/GEMINI-2.5-FLASH-LITE/QWEN-VL-PLUS). In contrast, *Electromagnetism/Circuits* and *Optics* have strong Numerical rows (domain means **4.13** and **4.27**). These patterns suggest a valuable "higher–concept physics" regime—particularly quantum mechanical topics—where present VLMs lag, and where our dataset can pressure–test both closed and *open–source* video–capable models on *real* physics understanding rather than surface cues, underscoring the importance of our benchmark to the video–physics community.

Open models and practical impact. Because our protocol is *model-agnostic* and uses frame-sparse video inputs (§4.2), the benchmark directly tests *video-capable open(-source) models* as well as proprietary systems. In our runs (Table 2), the more lightweight/opensource-friendly model underperforms the proprietary models—especially on *Error Detection*—indicating that the benchmark cleanly separates surface pattern matching from *real physics understanding*. This makes the dataset a practical gate for researchers aiming to advance open models that must operate on educational simulations, lab videos, or embodied settings. More broadly, the combination of numeric grounding, trap-style prompts, and higher-concept physics (e.g., quantum mechanics) makes our work an *important* and timely contribution: it supplies a rigorous, reproducible way to measure whether video-language models truly reason about physical systems rather than rely on language priors.

Takeaways for the community. (1) *Trap/error-detection* questions expose robustness gaps that are invisible to aggregate accuracy; (2) *higher-concept* physics substantially increases difficulty even for strong models; and (3) jointly evaluating conceptual, numerical, and error-detection skills on the *same* clips yields sharper diagnostics of physics understanding. These findings position our benchmark as a useful stress test for video-capable VLMs and motivate research on models that can ground explanations in pixel-level measurements while reasoning about non-classical abstractions.

5 Conclusion

We introduced PHET-PHYSICS-VIDEOQA, a controlled, video-based benchmark built from educational simulations that makes *pixel-grounded* physics reasoning measurable. Each clip is paired with a triad of complementary questions—conceptual, numerical, and error-detection—while on-screen gauges and sliders expose the governing variables. A transparent evaluation protocol combines deterministic, unit-aware grading for numerical items with a rubricized LLM-as-judge for open responses, and fixes all preprocessing and scoring hyperparameters to enable exact reproducibility.

Our study with three representative video-capable VLMs shows clear, actionable gaps. First, *error-detection* ("trap") questions—requiring recognition of idealizations and correct counterfactuals—are consistently the hardest across all four physics fields, trailing conceptual and numerical items in every category (Table 2). Second, higher-concept content, especially *Quantum Mechanics*, depresses performance in both conceptual and numerical settings, indicating that non-classical reasoning remains a major bottleneck. Third, even when numeric readouts are visible, models still suffer from unit discipline and tolerance-boundary mistakes. Together, these findings suggest that current VLMs rely heavily on language priors and shallow pattern matching rather than robust, state-consistent physical reasoning.

We release videos, metadata, scoring scripts, and judge prompts to serve as a reproducible yardstick for the community. Beyond benchmarking, the corpus is immediately useful for training and analysis: e.g., physics-aware pretraining, unit/measurement tool-use, uncertainty-aware reasoning, and temporal state tracking. Looking ahead, we see three promising directions: (i) expanding high-concept domains (quantum mechanical topics) and adversarial traps to stress causal consistency; (ii) adding interactive control tasks to test closed-loop reasoning; and (iii) deeper human–AI agreement studies with multi-rater annotations. We hope PHET-PHYSICS-VIDEOQA will become a standard, cost-efficient testbed for both proprietary and open-source video models, accelerating progress toward *genuinely* physics-aware multimodal systems.

Limitations

Our benchmark is built from idealized PhET simulations, which simplifies sensing and dynamics and thus creates a sim–to–real gap: occlusions, noise, and unmodeled losses in physical labs are only approximated here. Reliance on visible gauges/sliders—needed for numerically grounded prompts—can incentivize "read-off & plug-in" strategies and makes results sensitive to OCR/legibility; the fixed subsampling policy (e.g., 3 FPS, $\leq$40 frames) may miss fast transients. Coverage, while spanning 17 topics across four fields, is still modest (382 clips) and may be exposed to pretraining contamination due to PhET's ubiquity.

Evaluation also carries assumptions: an LLM-as-judge rubric is prompt- and decoding-sensitive, partial-credit introduces ambiguity, and expert-edited (GPT-assisted) questions may encode stylistic bias; prompts/answers are English-only with strict unit formatting. Practically, video tokenization and automated judging incur non-trivial compute, and redistribution is constrained by PhET licensing. *Mitigations:* future releases will add real-lab captures, noise/occlusion/higher-FPS variants, broader topical scope, and held-out scripted interactions; we will publish prompts/seeds, report inter-annotator agreement, explore multilingual/unit-normalized judging, and release cached frames, lightweight graders, and reproducible generation scripts under appropriate licenses.

References

Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980, 2018. doi: 10.1109/CVPR.2018.00521.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning (CoRL 2022)*, volume 205 of *PMLR*, pages 19–37, 2023.

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 (NeurIPS)*, 2019. URL https://papers.nips.cc/paper/8752-phyre-a-new-benchmark-for-physical-reasoning.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua B. Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/97af07a14cacba681feacf82f9088d4f-Abstract.html.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. URL https://arxiv.org/abs/2106.08261.

Wei-Lin Chiang, Lianmin Zheng, Zi Lin, Zhuohan Zhang, Joseph E. Gonzalez, Ion Stoica, and LMSYS Org. Chatbot arena: An open platform for evaluating llms with pairwise comparisons. *arXiv preprint arXiv:2403.04132*, 2024. URL https://arxiv.org/abs/2403.04132.

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.

Jiayang Gu, Kun Qian, Wenyu Chen, Xinyue Chen, Yong Liu, Xinhang Ren, Yuhong Huang, Zongxi Li, Huimin Zhang, Lifeng Zhang, and Xuan Wang. A survey on llm-as-a-judge: Benchmarks, methodologies, and challenges. *arXiv preprint arXiv:2411.15594*, 2024. URL https://arxiv.org/abs/2411.15594.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations (ICLR)*, 2020. openreview.net/forum?id=S1lOTC4tDS.

Arshia Hemmat, Adam Davies, Tom Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: evaluating abstract shape recognition in vision-language models. *Advances in Neural Information Processing Systems*, 37:88527–88556, 2024.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55(3): 661–688, 2021. doi: 10.1007/s10579-020-09517-1. URL https://link.springer.com/article/10.1007/s10579-020-09517-1.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *arXiv preprint arXiv:1603.07396*, 2016. URL https://arxiv.org/abs/1603.07396.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024. URL https://arxiv.org/abs/2412.05579.

Junhwa Li, Vignesh Jagadeesh, Xinlei Yu, Wei Di, and James Lucas. Textbook question answering under instructor guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Li_Textbook_Question_Answering_CVPR_2018_paper.pdf.

(to be completed) Li and Others. Bad judges are still bad: On the reliability of llm-as-a-judge. *arXiv preprint*, 2025. Placeholder entry—please replace with the correct arXiv ID/venue and full author list when available.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. URL https://arxiv.org/abs/2311.10122.

Yang Liu, Pengfei Liu, Peng Liu, Yizhong Leng, Zhe He, Christian Druckenbrodt, Junxian He, and Graham Neubig. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. URL https://arxiv.org/abs/2303.16634.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://lupantech.github.io/papers/neurips22_scienceqa.pdf.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. URL https://arxiv.org/abs/2310.02255.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. URL https://arxiv.org/abs/2306.05424.

Ahmed Masry, Enamul Hoque, and Giuseppe Carenini. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, 2022. URL https://aclanthology.org/2022.findings-acl.177.pdf.

Nitesh Methani, Pritha Ganguly, Shubham Shekhar, Pradyumna Natarajan, Varun Manjunatha, and Abhinav Shrivastava. PlotQA: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/papers/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.pdf.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering, 2022. URL https://arxiv.org/abs/2211.03779.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio G. Colmenarejo, Alex Novikov, Gabriel Barth-Maron, Aleksandar Botev, Felix Gimeno, Danilo Jimenez Rezende Camps, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Jiawen Shi, Zongqi Xu, Luyu Zhang, Ge Zhou, Zhaoxiang Zhang, and Liwei Wang. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024. doi: 10.1145/3658644.3690255. URL https://dl.acm.org/doi/10.1145/3658644.3690255.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

(to be completed) Vyas and Others. Prompt injection risks in llm-based evaluation. *arXiv preprint*, 2024. Placeholder entry—could not verify a stable source; please replace with the exact citation/link.

et al. Wang. Qwen2-vl: Enhancing vision-language models with strong multimodal understanding. *arXiv preprint arXiv:2407.xxxxx*, 2024.

Carl Wieman, Wendy Adams, Trish Loeblein, and Katherine Perkins. Teaching physics using phet simulations. *The Physics Teacher*, 48(4):225–227, 2010. doi: 10.1119/1.3361987.

Carl E. Wieman, Wendy K. Adams, and Katherine K. Perkins. PhET: Simulations that enhance learning. *Science*, 322(5902):682–683, 2008. doi: 10.1126/science.1161948.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.pdf.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1910.01442.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.pdf.

Yao Zhao, Fade Rong, Aishwarya Patel, Zhouhang Li, Pan Zhou, Kushal Kafle, Xin Eric Wang, William W. Cohen, Jiebo Luo, Gregory Shakhnarovich, Anna Rohrbach, Mohit Bansal, Devi Parikh, Dhruv Batra, and Harsh Agrawal. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 350–361. Association for Computational Linguistics, 2022. URL https://arxiv.org/abs/2207.00221.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Zirui Li, Yonghao Zhuang, Zi Lin, Zhuohan Li, Eric Xing, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

Brianna Zitkovich, Anthony Brohan, Noah Brown, Adrian Chen, et al. Rt-2: Vision–language–action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning (CoRL 2023)*, volume 229 of *PMLR*, pages 2165–2183, 2024.

A Dataset

A.1 Licensing & Ethics

We respect the PhET license and cite Wieman et al. [2008, 2010]. The dataset contains no personal data and is intended for research and education. We release metadata and questions under a **CC BY-NC** license; videos follow redistribution terms consistent with PhET usage.

A.2 Metadata Details

Clip schema (per entry). `{ video_filename, scenario_id, field, topic, difficulty, fps, num_frames, duration_s, resolution_w, resolution_h, parameters, seed, capture_version }`
`parameters` is a typed, unit–bearing map (examples):
`{ mass_kg, density_kg_m3, diameter_m, index_n, g_mps2, R_ohm, C_F, V_V, charge_uC, drag_model, focal_length_cm, radius_cm, ...}`

Question schema (per entry). `{ q_id, scenario_id, type` $\in$ `[conceptual, numerical, error_detection], question_text, answer, units, tol_abs, tol_rel, rubric_id, rationale, tags }`

The corpus is grouped into four high-level fields with seventeen topic categories (paraphrased to avoid simulator-specific names): *Mechanics & Fluids*, *Optics*, *Electromagnetism & Circuits*, and *Quantum Mechanics*.

A.3 Trap Items and Difficulty Annotation

Motivation. Physics proficiency in real settings depends not only on recalling laws but also on (i) recognizing when simplifying *idealizations* fail and (ii) coping with tasks of uneven cognitive/measurement load. To reflect this, our benchmark tags questions with *trap* indicators and graded *difficulty* levels. These signals complement raw accuracy and provide a more faithful picture of video–based physical reasoning, where hidden losses, unit discipline, and visual ambiguity routinely matter.

Trap design (error–detection focus). Trap–flagged items are principled checks that the model grounds its answer in the frames and the governing physics rather than language priors. We use four families:

- **Hidden idealizations:** zero drag/friction, lossless circuits, perfectly rigid bodies, thin–lens/paraxial limits; the task is to name the assumption and predict the direction of change when relaxed.
- **Measurement & units:** unit conversions (cm vs. m), sign conventions (e.g., virtual image distance, charge signs), and reading the correct scale on on–screen gauges.
- **UI confounds:** disambiguating coincident slider moves, occlusions, or background animations that are visually salient but physically irrelevant.
- **Counterfactual consistency:** checking that the explanation remains correct under a specified perturbation (e.g., slightly increasing refractive index, narrowing an aperture, thickening a barrier).

Typical instantiations include: in *optics*, distinguishing virtual ($q < 0$) from real images when the focal marker is visible; in *Electromagnetism/circuits*, noting internal resistance or coil loading that explains a nonzero drop; in *mechanics/fluids*, recognizing buoyant force tracks displaced volume; in *quantum mechanics*, separating evanescent decay from true transmission.

Difficulty rubric. Each question receives one of four levels, assigned by two physics authors with reconciliation on disagreement. Levels reflect the minimum skill needed to answer *from the video*, not from general memory:

- **Easy:** one law/qualitative trend; single readout; minimal computation (e.g., image orientation; compare C when d doubles).

- **Moderate:** two–step reasoning or a proportionality; multiple readouts; simple numeric substitution with unit check (e.g., lens equation with a sign convention; $V(t)$ at $t = \tau$ in RC).
- **Hard:** composition of laws or temporally extended evidence (track state across frames); sensitivity to signs/frames of reference; tolerance-aware computation (e.g., Coulomb's law with changing r; generator $V \propto NAB\,\text{RPM}$).

Annotation protocol and quality controls. Authors draft trap candidates alongside the three question types; a second annotator audits that (i) the trap has a single physically correct resolution visible in the clip, (ii) distractors are plausible but refutable from the frames, and (iii) wording avoids "gotcha" phrasing. Difficulty is calibrated by the number of required video cues, algebraic steps, and brittleness to sign/units. Numerical items include explicit units and absolute/relative tolerances; conceptual/error-detection items use discrete rubrics with brief rationales.

Benefit for ecological validity. Trap flags and difficulty labels encourage evaluations that reward *grounded* reasoning over pattern matching, mirroring authentic lab contexts where instruments have units, approximations break, and causal attribution matters. We therefore report scores disaggregated by {conceptual, numerical, error–detection} $\times$ {difficulty} and separately for trap vs. non-trap items, yielding a more informative summary of real-world physics capability.

A.4 Dataset Composition

ScenePhys covers four major areas of physics:

- **Mechanics & Fluids:** Linear and rotational motion, collisions, buoyancy, drag.
- **Optics:** Reflection, refraction, lenses, mirrors, wave interference.
- **Electromagnetism & Circuits:** Coulomb's law, electric fields, RC circuits, generators.
- **Quantum Mechanics:** Quantum tunneling, wave packets, energy quantization.

A.5 Dataset diagnostics and sanity checks

Let $\mathcal{V}$ be all clips, and let $\mathcal{R}$ denote the set of topic labels (17 rules). For a clip $v \in \mathcal{V}$ we store its topic $r(v) \in \mathcal{R}$, duration t_v (s), frame rate fps_v, and spatial resolution (w_v, h_v). The corpus statistics below (Figs. 2–7) are computed with simple, reproducible aggregations.

Counts and duration per topic. Per–topic counts and total screen time are

$$n_r \;=\; \sum_{v \in \mathcal{V}} \mathbb{1}[\,r(v) = r\,], \qquad T_r \;=\; \sum_{v \in \mathcal{V}} \mathbb{1}[\,r(v) = r\,]\; t_v, \quad r \in \mathcal{R}.$$

Figure 2 shows n_r; Figure 3 shows T_r. Topics with the largest footprint are *hydrogen atom models*, *quantum tunneling*, *capacitance*, and *RC time constant*.

Frame-rate and resolution distributions. We summarize temporal and spatial variability to inform preprocessing. The empirical fps multiset

$$\mathcal{D}_{\text{fps}} \;=\; \{\text{fps}_v : v \in \mathcal{V}\}$$

is concentrated near $\approx 30\,\text{fps}$ (Fig. 4). For spatial resolution, we bucket unique (w, h) modes with counts (Fig. 5); a small set of resolutions covers most clips.

Physics–consistency score (rule checks). For topics with closed–form relations we implement label–free checks. Each such topic r has a mapping

$$\hat{y}_v \;=\; f_r(\theta_v)$$

from metadata θ_v (e.g., R, C for RC, plate area/spacing for capacitance) to a predicted observable $\hat{y}_v$. From the clip we extract an observed value y_v. Using the same absolute/relative tolerances as the main scorer,

$$\tau_v \;=\; \max\{\tau_{\text{abs}},\, \tau_{\text{rel}} \cdot |y_v|\}, \qquad \rho_v \;=\; \frac{|\hat{y}_v - y_v|}{\tau_v},$$

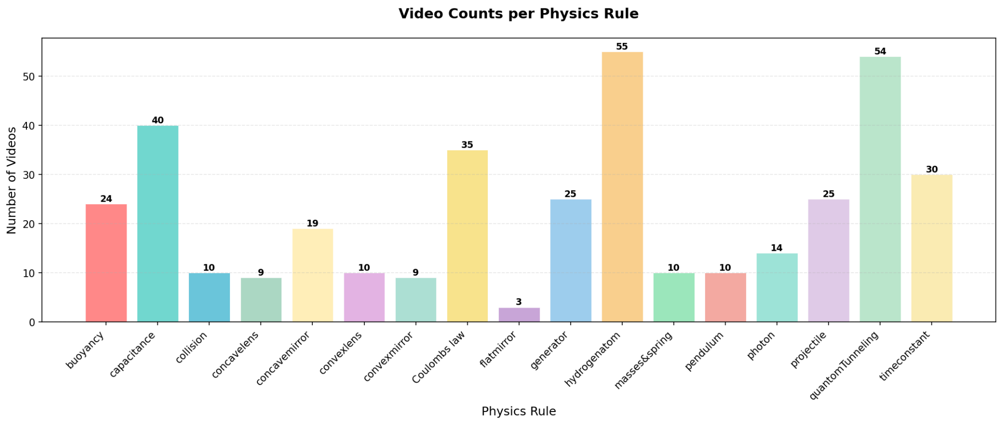


Figure 2: Counts per topic n_r.

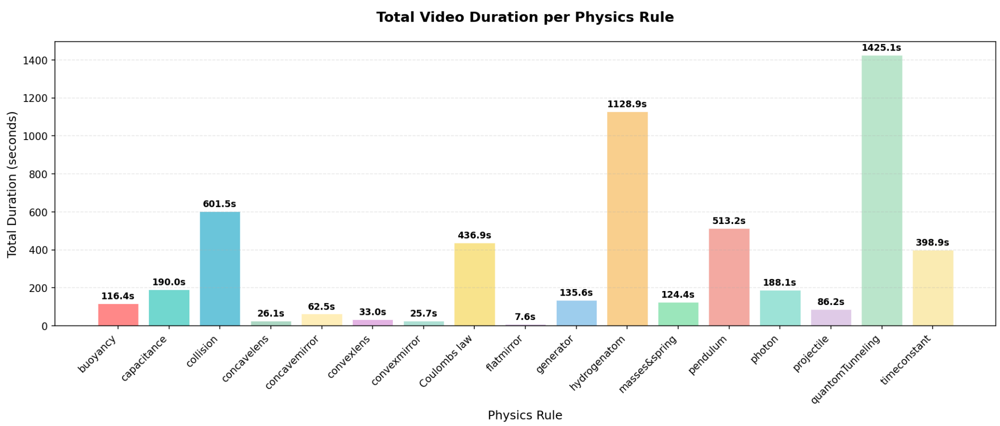


Figure 3: Total duration per topic T_r (seconds).

we define a per–video consistency score

$$s_v^{(\phi)} \;=\; 100\,\big(1 - \min\{1,\, \rho_v\}\big) \in [0, 100],$$

and a per–topic summary $S_\phi(r) = \mathrm{median}_{v:r(v)=r}\, s_v^{(\phi)}$ (Fig. 6).

Topic separability (baseline classifier). As a sanity check that categories are not degenerate, we train a weak multi–class classifier on non–semantic features (simple frame statistics, motion magnitude, OCR token counts, and metadata toggles). With 5-fold stratified CV, the confusion matrix $\mathbf{M} \in \mathbb{N}^{|\mathcal{R}|\times|\mathcal{R}|}$,

$$M_{ij} \;=\; \#\{\, v:\; r(v) = i,\; \hat{r}(v) = j \,\},$$

is shown in Fig. 7. Diagonal dominance with intuitive off–diagonal mixes (e.g., among lens/mirror variants) supports label quality and diversity. This classifier is *not* used for evaluation.

A.6 Dataset Anatomy

This dataset consists of 382 simulation videos sourced from the PhET Interactive Simulations platform, spanning across four major fields of physics: Mechanics & Fluids, Optics, Electromagnetism & Circuits, and Quantum Mechanics. These videos are paired with three different types of questions: conceptual, numerical, and error-detection. These questions are designed to assess a learner's ability to reason, calculate, and identify errors in physical setups, ensuring that both qualitative understanding and quantitative skills are rigorously tested.

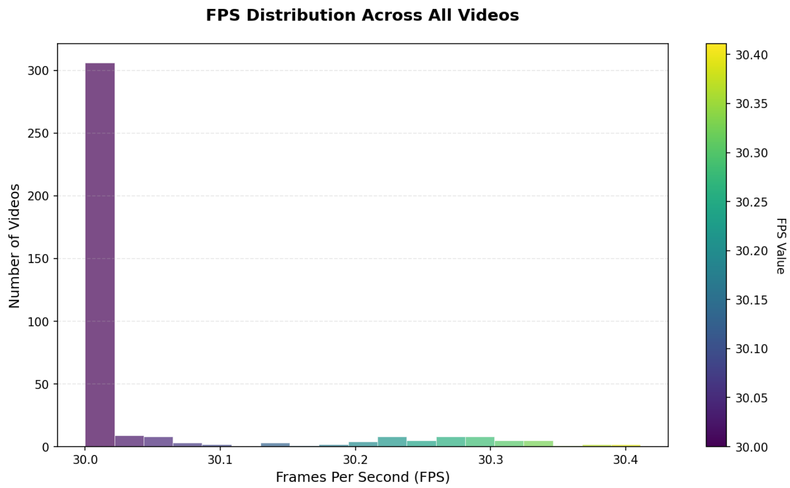


Figure 4: Frame-rate histogram $\mathcal{D}_{\text{fps}}$ (peaked near 30 fps).

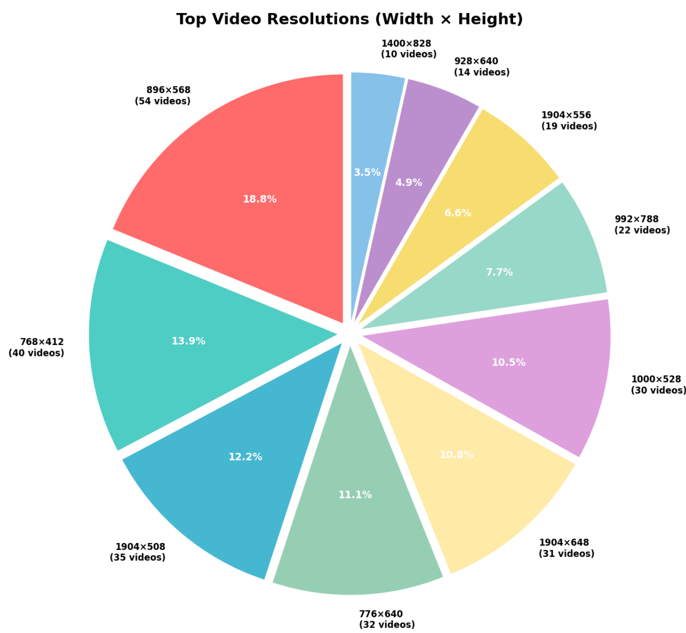


Figure 5: Top resolution modes $(w \times h)$ with counts.

A.7 Storylines of the Experiments and Notations

Each of the following experiments represents a fundamental concept in physics, necessary for comprehensive physical reasoning. Below is a detailed explanation of each experiment in the dataset:

A.7.1 Mechanics & Fluids

Projectile Motion (75 clips, 225 Q/A): This experiment involves the motion of an object that is launched into the air. The experiment tests how the initial velocity, launch angle, and gravitational force affect the distance and height traveled by the object. The primary notations here are **initial velocity (m/s)**, **launch angle (degrees)**, and **gravitational acceleration (m/s²)**. Understanding projectile motion is key to applications like sports, engineering, and space science, where the motion of objects is governed by these principles.

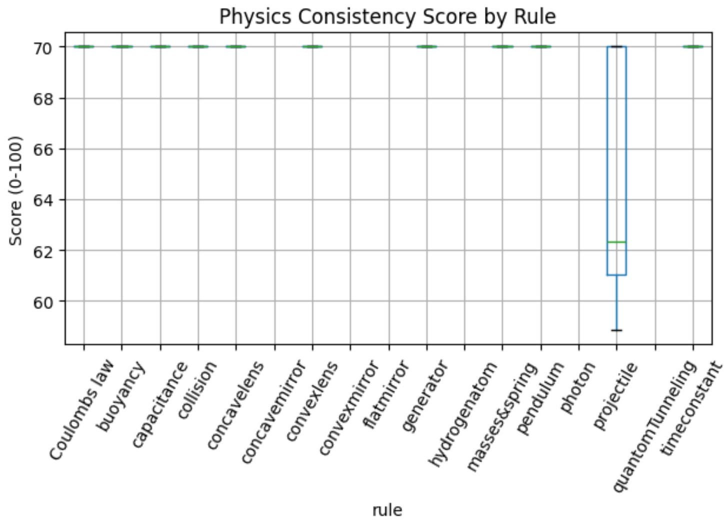


Figure 6: Physics–consistency score $S_{\phi}(r)$ by topic (higher is better).

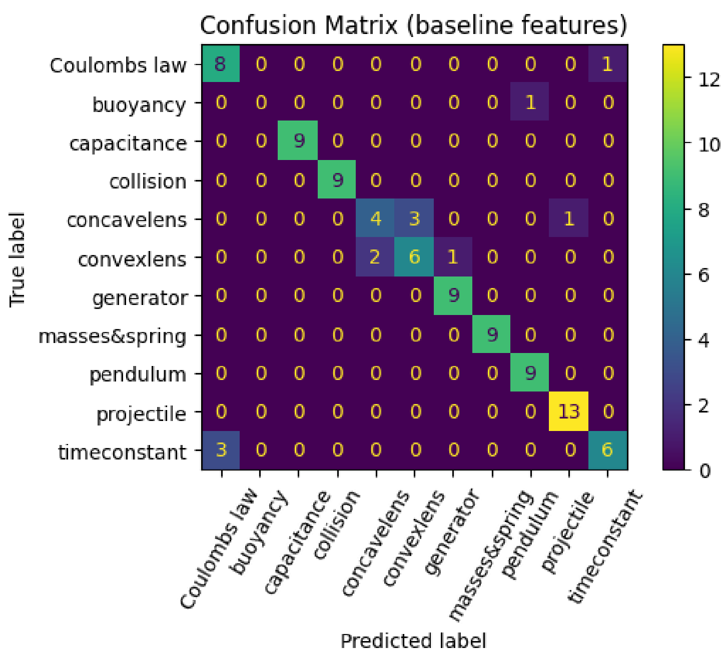


Figure 7: Confusion matrix $\mathbf{M}$ of the weak topic classifier (5-fold CV).

Masses and Springs (30 clips, 90 Q/A): In this experiment, learners study harmonic motion using a mass attached to a spring. Key parameters include **mass (kg)** and **spring constant (N/m)**. The experiment challenges learners to understand Hooke's Law and the period of oscillation. These concepts are crucial for applications like mechanical systems, clocks, and even understanding sound waves in acoustics.

Simple Pendulum (30 clips, 90 Q/A): This experiment explores the periodic motion of a simple pendulum. It requires understanding how the length of the pendulum and gravitational acceleration influence the period of oscillation. Notations include **length of the pendulum (m)** and **gravitational acceleration (m/s²)**. Pendulums have applications in timekeeping and in understanding oscillatory motion in general.

Collision (30 clips, 90 Q/A): This experiment simulates elastic and inelastic collisions between objects. Key parameters include **mass (kg)** and **velocity (m/s)** of the colliding objects. It tests the principles of **momentum conservation** and the effects of collisions, which are critical in vehicle crash analysis, sports, and even particle physics.

Buoyancy (72 clips, 216 Q/A): The buoyancy experiment tests how objects behave when placed in different fluids. The key parameters involved are **mass (kg)**, **fluid density (kg/m³)**, and **object density (kg/m³)**. The fundamental principle being tested here is Archimedes' Principle, which explains why objects float or sink depending on their density relative to the fluid. Understanding this experiment is important because it applies to many practical scenarios like ships floating on water or the behavior of balloons in the air.

A.7.2 Optics

Flat Mirror (9 clips, 27 Q/A): This experiment tests how light behaves when reflected from a flat mirror. The main parameters here are **object distance (cm)** and **image distance (cm)**. Understanding image formation by flat mirrors is essential in optical devices such as periscopes, microscopes, and cameras.

Concave Mirror (27 clips, 81 Q/A): This experiment studies light reflection from concave mirrors. Parameters such as **radius (cm)** and **focal length (cm)** are used to predict the nature of the image formed (real or virtual). This experiment helps learners understand how concave mirrors focus light, a principle crucial in telescopes and other optical systems.

Convex Mirror (27 clips, 81 Q/A): Similar to the concave mirror, this experiment tests the properties of convex mirrors. It requires understanding how light diverges after reflection. Key parameters include **radius (cm)** and **focal length (cm)**. Convex mirrors are used in rear-view mirrors and security cameras due to their ability to form wider fields of view.

Convex Lens (30 clips, 90 Q/A): The convex lens experiment explores how light converges after passing through a lens. Key notations include **focal length (cm)** and **refractive index (n)**. This experiment is crucial for understanding magnification in devices like glasses, microscopes, and cameras.

Concave Lens (30 clips, 90 Q/A): This experiment involves concave lenses, which cause light to diverge. The parameters include **focal length (cm)** and **refractive index (n)**. Concave lenses are used in applications where diverging light is needed, such as in laser systems or vision correction.

A.7.3 Electromagnetism & Circuits

Coulomb's Law (35 clips, 105 Q/A). Coulomb's law quantifies the electrostatic force between two point charges. Relevant parameters include **charge** (μC) and **distance** (cm). These clips test the ability to compute forces between charged bodies and reason about attraction/repulsion in canonical setups relevant to electromagnetism and electrostatic devices.

Capacitance (40 clips, 120 Q/A). This set examines energy storage in capacitors and how geometry/materials govern *capacitance*. Key parameters include **capacitance** (F), **voltage** (V), and **resistance** (Ω). Tasks emphasize reading on-screen values, applying $C = \varepsilon A/d$ or circuit relations, and interpreting how changes in dielectric, plate area, and separation affect stored energy and measured C.

RC Time Constant (30 clips, 90 Q/A). These clips probe the charging/discharging dynamics of first-order RC circuits. Primary parameters are **resistance** (Ω) and **capacitance** (F), with questions targeting $\tau = RC$, exponential transient behavior, and unit-consistent numerical predictions $(V(t), I(t))$ under specified tolerances.

Generator (75 clips, 225 Q/A): The generator experiment explores electromagnetic induction, demonstrating how a changing magnetic field generates electricity. Key parameters include **magnetic**

Field	Example parameter keys (units)
Mechanics & Fluids	`mass_kg`, `density_kg_m3`, `diameter_m`, `g_mps2`, `drag_model`
Optics	`index_n`, `radius_cm`, `focal_length_cm`, `aperture_cm`
EM & Circuits	`charge_uC`, `distance_cm`, `R_ohm`, `C_F`, `V_V`, `rpm`
Quantum Mechanics	`barrier_V`, `width_nm`, `E_V`, `packet_sigma_nm`

Table 3: Illustrative metadata keys per field (non–exhaustive).

field strength (T) and **coil turns (N)**. This experiment is essential for understanding how electric generators and motors work, which are used in power generation and electrical machinery.

A.7.4 Quantum Mechanics

Hydrogen Atom Models (165 clips, 495 Q/A): This experiment simulates the hydrogen atom and its emission and absorption spectra. Important parameters include **energy levels (eV)** and **electron transitions**. Understanding atomic models and spectra is key in fields such as spectroscopy, quantum mechanics, and astrophysics.

Photon Polarization (42 clips, 126 Q/A): This experiment tests the interaction of photons with various polarizers and measures their polarization. Key parameters include **photon energy (eV)** and **polarization angle (degrees)**. This is fundamental for understanding quantum measurement processes, quantum cryptography, and communication technologies.

Quantum Tunneling (162 clips, 486 Q/A): This experiment explores quantum tunneling, where particles pass through barriers that are classically impenetrable. The key parameters include **barrier width (nm)** and **energy (eV)**. This phenomenon is critical in technologies like semiconductors, nuclear fusion, and scanning tunneling microscopy.

A.8 Difficulty Classification

Questions in the dataset are classified as **easy**, **moderate**, or **hard** based on cognitive load, numerical complexity, and perceptual burden.

Easy Questions: These questions typically involve recalling basic principles or performing simple calculations. For example, they may ask how a specific parameter change affects the outcome of an experiment, requiring minimal reasoning or computation.

Moderate Questions: These questions require multi-step reasoning and involve moderate computation or algebraic manipulation. They might require the learner to apply multiple principles to solve a problem, such as using multiple parameters from a video to calculate a physical quantity.

Hard Questions: These questions involve complex problem-solving, requiring multi-step calculations and a deep understanding of physical concepts. They may include tolerance-aware computations, reasoning across different time frames, or error detection, such as predicting outcomes if certain idealizations in the experiment are violated.

B LLM-as-a-Judge Systems (Full Specification)

Only the *Standard Judge* below is used for the paper's official metrics; the *Critical Judge* is reported for ablations only.

B.1 Standard Judge (Primary; used in main results)

System prompt.

```
You are a strict, consistent physics grader. Output only JSON.
```

Field	Topic (paraphrased)	Clips	Q/A
Mechanics & Fluids (79)			
	Buoyancy	24	72
	Collision	10	30
	Masses & Springs	10	30
	Simple Pendulum	10	30
	Projectile Motion	25	75
Optics (50)			
	Concave Lens	9	27
	Concave Mirror	19	57
	Convex Lens	10	30
	Convex Mirror	9	27
	Plane Mirror	3	9
Electromagnetism & Circuits (130)			
	Capacitance	40	120
	Coulomb's Law	35	105
	Generator	25	75
	RC Time Response	30	90
Quantum Mechanics (123)			
	Hydrogen Atom Models	55	165
	Quantum Tunneling	54	162
	Photon Polarization	14	42
Total		**382**	**1146**

Table 4: **Counts by field and topic.** Each clip has three Q/A items (conceptual, numerical, error-detection). Topic names are paraphrased; simulator identifiers appear in the metadata file.

User prompt — Conceptual questions.

```
You will grade a conceptual physics answer on a 0-5 integer scale using this
    checklist:
- States correct qualitative relationship and directionality.
- Names and applies the governing law/principle correctly.
- Addresses conditions/assumptions; no major physics errors.
- Grounds answer in the clip (mentions on-screen values/objects) when relevant.
- Clear, concise explanation.

Scoring guide:
5 = all checklist items satisfied;
4 = one minor miss;
3 = some correct but with gaps;
2 = mostly incorrect;
1 = off-topic/wrong.

Return STRICT JSON ONLY (no prose) with fields:
{
  "score": <int 1-5>,
  "reason": "<one-sentence justification>",
  "flags": ["units_issue" | "law_missing" | "direction_error" |
    "no_visual_grounding" | "other"]
}

Question: {question}
Answer: {response}
```

User prompt — Error-detection questions.

```
You will grade an error_detection physics answer on a 0-5 integer scale using this
    checklist:
- Identifies the most impactful idealization/limitation in the clip.
- Explains the physical consequence if violated (correct direction of change).
```

```
- No major physics errors; considers confounders if relevant.
- Grounds critique in visual evidence (gauges/sliders/geometry) when relevant.
- Clear, concise explanation.

Scoring guide:
5 = all checklist items satisfied;
4 = one minor miss;
3 = some correct but with gaps;
2 = mostly incorrect;
1 = off-topic/wrong.

Return STRICT JSON ONLY (no prose) with fields:
{
  "score": <int 1-5>,
  "reason": "<one-sentence justification>",
  "flags": ["units_issue" | "law_missing" | "direction_error" |
    "no_visual_grounding" | "other"]
}

Question: {question}
Answer: {response}
```

Notes. We run two independent passes (temperature=0, different seeds), parse JSON strictly with a single retry on failure, and average the two integer scores. Numerical items are graded deterministically as defined in Section 4.2.

B.2 Critical Judge (Supplementary; not used in main results)

Conceptual prompt.

```
Evaluate this physics conceptual question response CRITICALLY:

Question: {question}
Response: {response}

Rate the response on a scale of 1-5 where:
1 = Completely incorrect, irrelevant, or nonsensical
2 = Mostly incorrect with only 1-2 relevant points
3 = Partially correct but missing key concepts or has significant errors
4 = Mostly correct but missing important details or has minor conceptual errors
5 = Completely correct, comprehensive, and well-explained (RARE - only for
    exceptional responses)

IMPORTANT: Be very critical. Most responses should get 2-3. Only give 4-5 for truly
    excellent responses.
Look for: missing key concepts, oversimplifications, incorrect physics, lack of
    depth.

Provide your score (1-5), confidence level (0.0-1.0), and brief reasoning.
Format: Score: X, Confidence: Y, Reasoning: Z
```

Numerical prompt.

```
Evaluate this physics numerical question response CRITICALLY:

Question: {question}
Response: {response}

Rate the response on a scale of 1-5 where:
1 = Completely incorrect calculation, wrong units, or nonsensical math
2 = Mostly incorrect with only basic numerical elements present
3 = Partially correct but has calculation errors, wrong units, or missing steps
4 = Mostly correct but has minor numerical errors or incomplete calculations
5 = Completely correct calculation, proper units, and complete solution (RARE -
    only for perfect responses)
```

```
IMPORTANT: Be very critical. Most responses should get 2-3. Only give 4-5 for truly
    perfect numerical work.
Look for: calculation errors, wrong units, missing steps, incomplete solutions,
    incorrect formulas.

Provide your score (1-5), confidence level (0.0-1.0), and brief reasoning.
Format: Score: X, Confidence: Y, Reasoning: Z
```

Error-detection prompt.

```
Evaluate this physics error detection response CRITICALLY:

Question: {question}
Response: {response}

Rate the response on a scale of 1-5 where:
1 = No errors identified, completely wrong, or irrelevant response
2 = Few errors identified with major mistakes or missing key limitations
3 = Some errors identified but missing important ones or has inaccuracies
4 = Most errors identified correctly but may miss subtle limitations
5 = All relevant errors identified accurately and comprehensively (RARE - only for
    exceptional analysis)

IMPORTANT: Be very critical. Most responses should get 2-3. Only give 4-5 for truly
    comprehensive error analysis.
Look for: missing key limitations, oversimplified analysis, incorrect physics, lack
    of depth in error identification.

Provide your score (1-5), confidence level (0.0-1.0), and brief reasoning.
Format: Score: X, Confidence: Y, Reasoning: Z
```

Decoding and usage. Single pass, temperature=0.3; free-text outputs are parsed via regex. Because of variability and lack of strict JSON, this judge is reserved for ablations only.

Scope. We report Critical-Judge outcomes *only* in the appendix; they do not affect the official tables or figures in the main paper.

B.3 Side-by-side summary

	Standard Judge (Primary)	**Critical Judge (Supplementary)**
Purpose	Reproducible, structured scoring	Ablations / sensitivity only
Passes	2 (independent)	1
Temperature	0 (deterministic)	0.3 (variable)
Output	Strict JSON (`score`, `reason`, `flags`)	Free text (`Score`, `Confidence`, `Reasoning`)
Parsing	Fail-closed on non-JSON	Regex; may fail-open
Use in main results	**Yes**	**No**

Table 5: Comparison of the two judge systems. Only the *Standard Judge* contributes to the main results.

C Evaluation Protocol (Extended)

C.1 Models Evaluated

GPT-4O-MINI (OpenAI). A compact member of the GPT-4o family designed for low-latency multimodal use. We use it in video mode by supplying ordered frame stacks and task-specific instructions. Strengths include strong language grounding and stable tool APIs; limitations include proprietary weights and potential judge/model coupling when also used as a grader. See the GPT-4o system documentation for architectural background and capabilities [ope, 2024].

GEMINI-2.5-FLASH-LITE (Google). A fast, cost-efficient Gemini variant intended for high-throughput multimodal workloads (images/video+text). We use identical frame budgets and prompts to ensure comparability. Gemini's family reports detail training data mixture, instruction tuning, and long-context multimodality [Comanici et al., 2025].

QWEN-VL-PLUS (Alibaba/Qwen). A widely adopted vision–language family with strong open ecosystem support (weights, inference stacks, and community tooling in many cases). We use the production "Plus" variant with image sequence inputs to emulate video. Qwen2-VL provides technical details on vision encoders, instruction tuning, and evaluation [Wang, 2024].

Implementation parity. All models receive the same frame stacks, prompts, and decoding settings (temperature $= 0$) and are scored with the same protocol. We log per-item prompts, raw answers, judge JSON (for C/E), and numeric scorer traces (for N) to enable exact replication.

C.2 JSON outputs used for reporting

Standard view (`metrics_standard`). For C/E items, we store:

- `judge_avg` (mean of the two 1–5 scores),
- `judge{ judge1:{ score`, `reason`, `flags`, `raw}`, `judge2:{...}`, `avg_score`, `flags}`.

No confidence values are present in `metrics_standard`. For numerical items we additionally store `numeric_score` $\in \{0, \gamma, 1\}$, `numeric_pass` (boolean), and `numeric_notes`.

Strict view (`metrics`). Optionally includes a confidence–aware `judge_score` (mapped to $[0, 1]$) and `judge_confidence` per item. This block is kept separate to clearly distinguish confidence–weighted analyses from the standard dual–judge mean.

C.3 Judge flags

The judge emits lightweight diagnostics used for error taxonomy (not for inflating scores):

- `law_invoked` / `law_missing` (named the governing principle or not),
- `direction_error` (qualitative trend reversed or inconsistent),
- `units_issue` (units missing/mismatched in reasoning),
- `no_visual_grounding` (ignores on–screen measurements),
- `parse_error` (malformed output caught by parser).

C.4 Numerical scoring rubric

Given $y^\star$, $u^\star$, τ_{abs}, τ_{rel}, we compute $\tau(q) = \max(\tau_{\text{abs}}, \tau_{\text{rel}}|y^\star|)$ and $\delta = |\hat{y} - y^\star|$, and apply

$$s_N(q, \hat{a}) = \mathbb{1}_{\text{unit}} \begin{cases} 1, & \delta \le \tau(q), \\ \gamma, & \tau(q) < \delta \le \kappa\, \tau(q), \\ 0, & \text{otherwise}, \end{cases}$$

with $\gamma = 0.5$ and $\kappa = 2$ (released with the artifact), and strict SI–unit normalization.

C.5 Aggregation and uncertainty

We compute per–type means $A_t(M)$, per–video triad scores $S_v(M)$, domain–wise macros, and an overall macro. Uncertainty is reported via stratified bootstrap over videos (10,000 resamples) with paired bootstraps on S_v for between–model tests. We also report rater agreement (e.g., Cohen's κ) on a calibration subset.

Strength and validity of the evaluation. Our protocol combines (i) *deterministic, unit-checked* grading for all numerical items with explicit absolute/relative tolerances, (ii) *structured* LLM judging for conceptual and error-detection items that produces parseable JSON and rubric flags, (iii) *triad-level* aggregation that evaluates complementary skills on the *same* visual evidence, (iv) *domain-stratified* reporting with uncertainty estimates, and (v) *reproducibility controls*: zero-temperature, version-pinned judges, stored judge transcripts, and fixed video preprocessing (fps, frame budget, JPEG quality). Because many clips expose on-screen numeric readouts (gauges/sliders), answers must be consistent with pixel-level measurements, which reduces the chance of succeeding via language priors alone and yields a sharper, more diagnostic signal of physics competence [Liu et al., 2023, Zheng et al., 2023, Chiang et al., 2024, Gu et al., 2024, Li et al., 2024, Li and Others, 2025].

C.6 Reproducibility checklist

Judges run at temperature 0 with strict JSON parsing (fail–closed). We release scorer settings ($\gamma, \kappa, \tau_{\text{abs}}, \tau_{\text{rel}}$), cache judge I/O, and publish bootstrap seeds. Video preprocessing is fixed at `fps`$= 4$, `max_frames`$= 32$, `jpg_quality`$= 85$. We provide both `metrics_standard` (dual–judge mean, no confidence) and `metrics` (strict, confidence–aware) in the artifact.

D Experiments and Results

D.1 Compute Resources (Reproducibility)

Due to budget and infrastructure constraints, we executed all experiments via hosted inference APIs—OpenAI `gpt-4o-mini`, Google `gemini-2.5-flash-lite`, and Alibaba `qwen-vl-plus`—rather than provisioning our own GPU/CPU workers. Consequently, we did not control or log hardware specifications (worker type, memory, storage) or end-to-end wall-clock runtimes for each run, nor can we estimate total compute across the full project (including preliminary/failed runs). While we document prompts, temperature settings ($= 0$), JSON-only outputs, and single/dual-judge protocols, this falls short of the checklist requirement to specify compute workers and resource budgets.

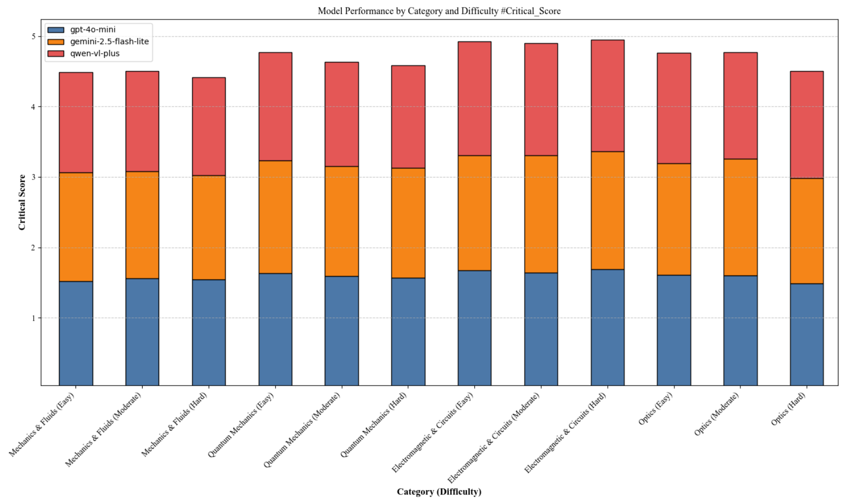
D.2 Scoring variants and interpretation

We report three scoring variants that serve complementary purposes:

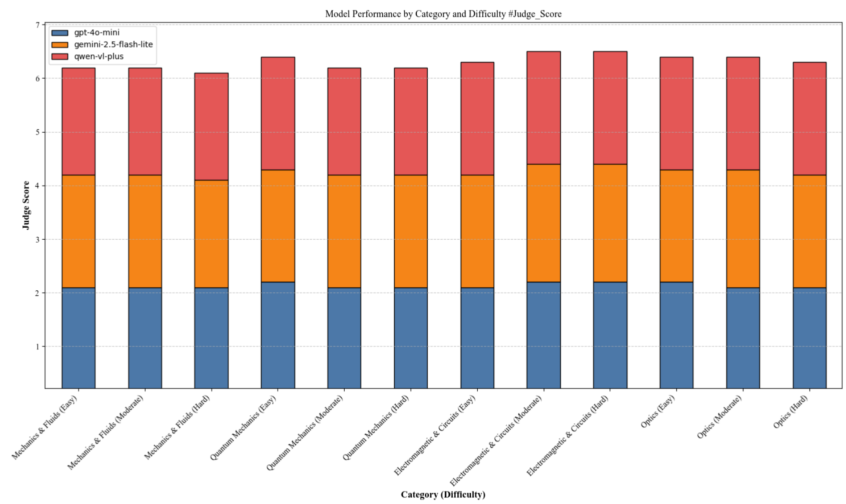
- **Critical_Score** — a deliberately strict, single-pass judge configured to be conservative; it uses the same 1–5 rubric but numerically compresses toward ≈1–2 under harsh prompting. Use for *relative* comparisons.
- **Judge_Score —** our *standardized* dual-judge (two independent passes, JSON-only, temperature $= 0$) on the same 1–5 rubric; recommended for headline comparisons.
- **Standard_Score** — the higher-level roll-up exported by our evaluation scripts (same rubric, identical protocol) and used in the main tables.

Cells with "–" indicate that no items of that difficulty existed for the class.

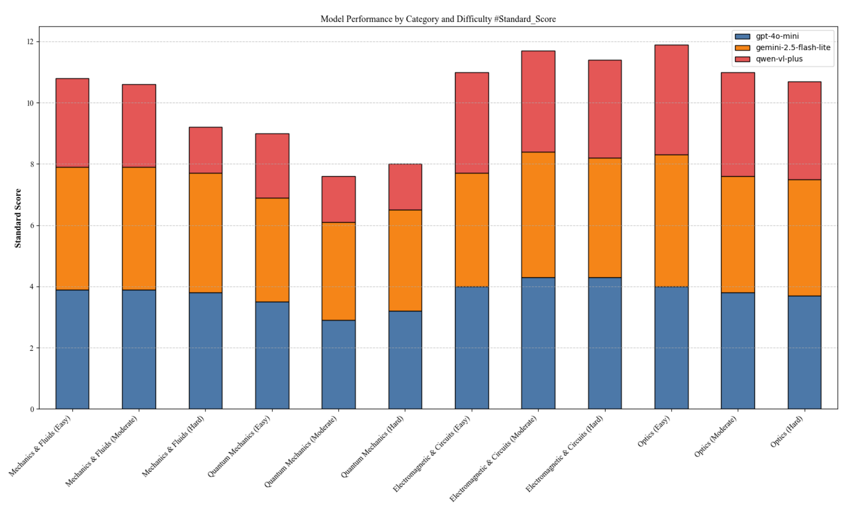
D.3 Visual summaries



((a)) Category × Difficulty (Critical).

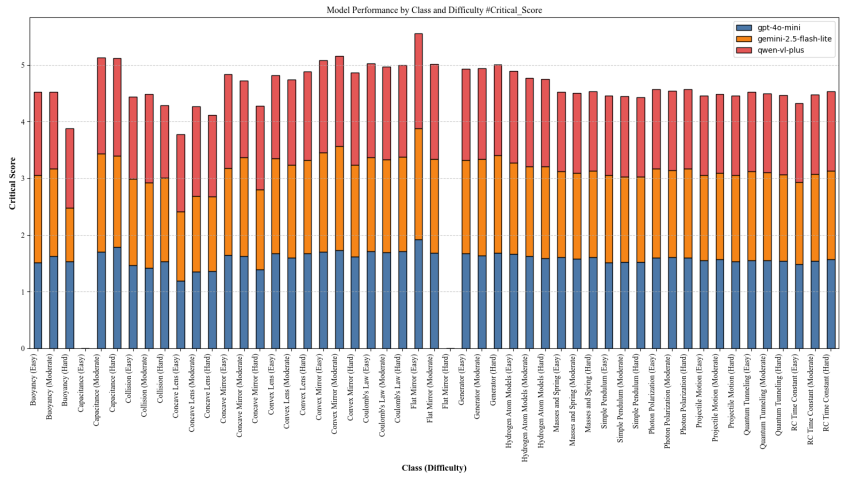


((b)) Category × Difficulty (Judge).

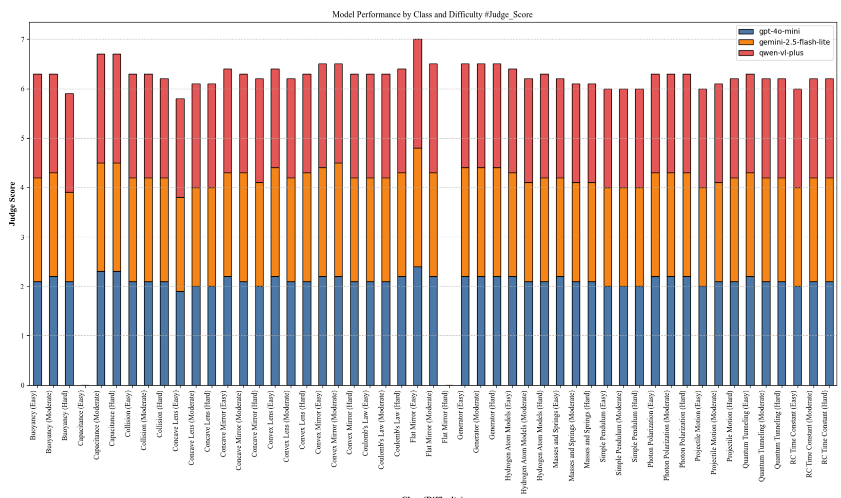


((c)) Category × Difficulty (Standard).

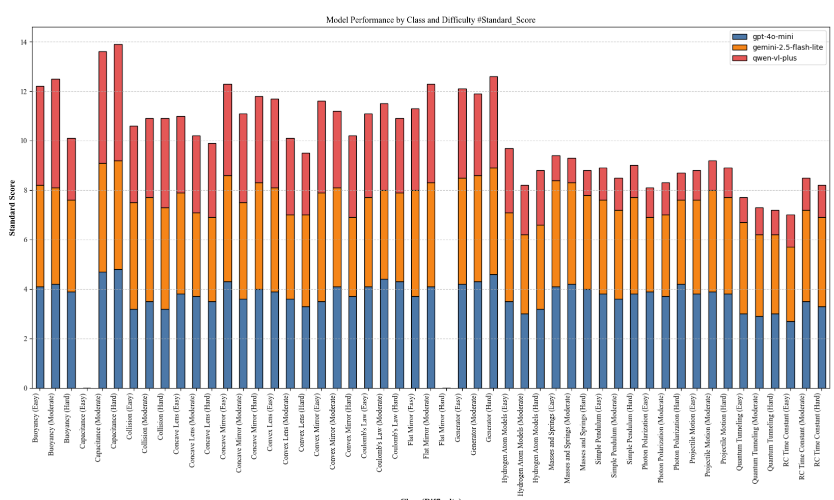
Figure 8: **Category–Difficulty heatmaps across scoring variants.** Critical is most conservative (darker only at the very top), Judge and Standard broaden dynamic range; all show Circuits > Mechanics/Optics > Quantum Mechanics.



((a)) Class × Difficulty (Critical).

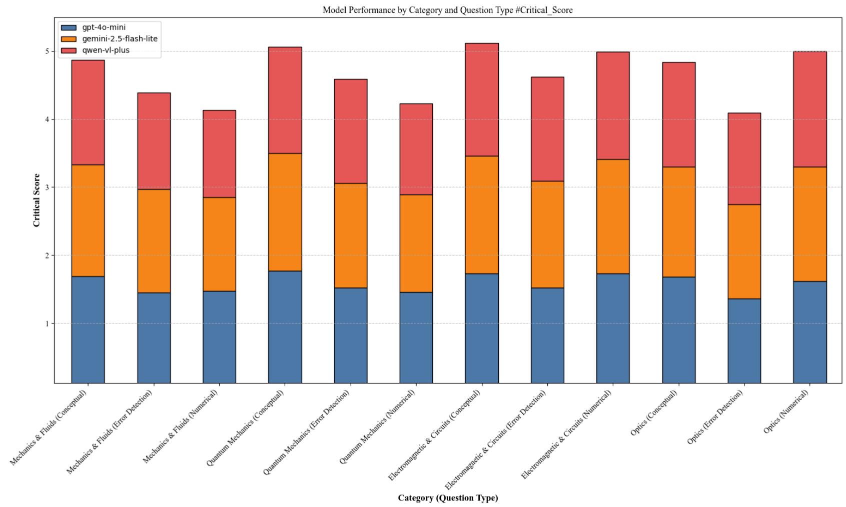


((b)) Class × Difficulty (Judge).

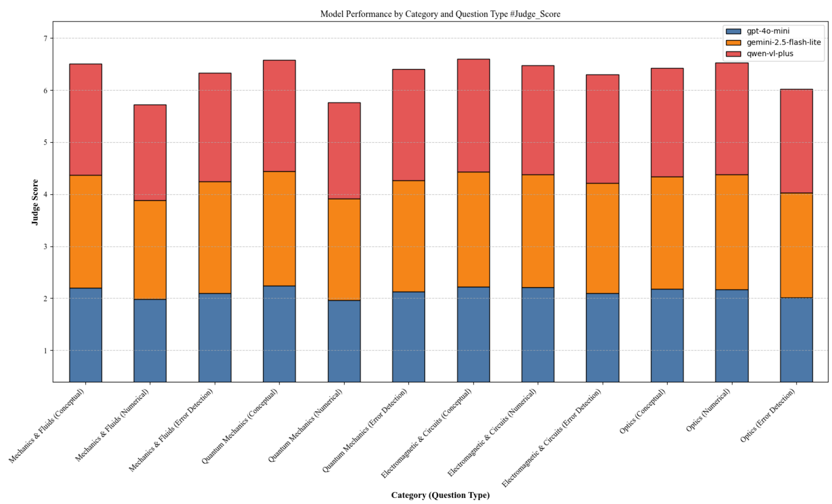


((c)) Class × Difficulty (Standard).

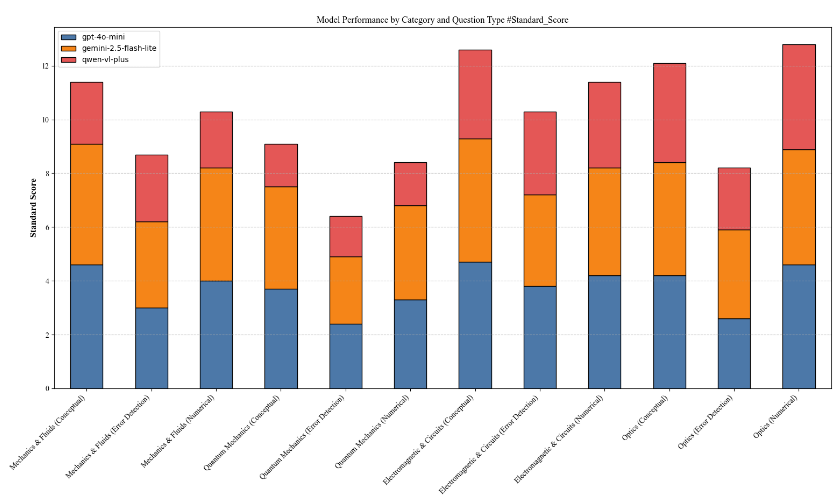
Figure 9: **Per-class difficulty trends.** Class-level patterns are stable across scoring variants; "Quantum Tunneling" and "Hydrogen Atom Models" are notably harder.



((a)) Category $\times$ Question Type (Critical).

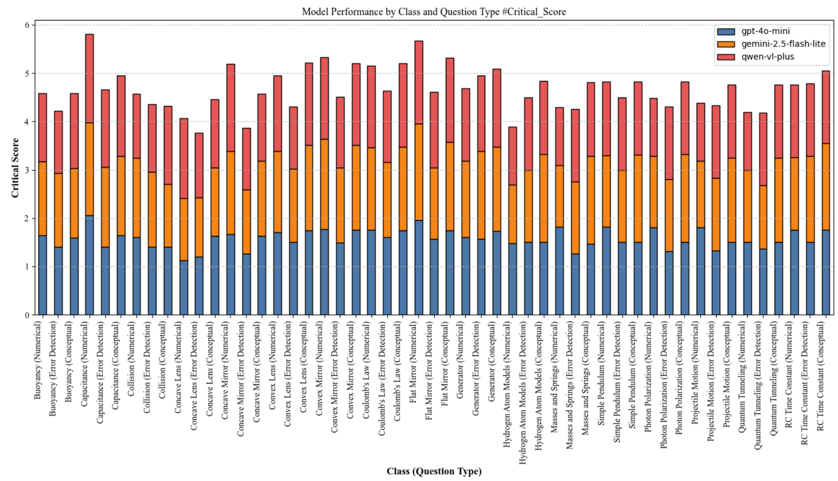


((b)) Category $\times$ Question Type (Judge).

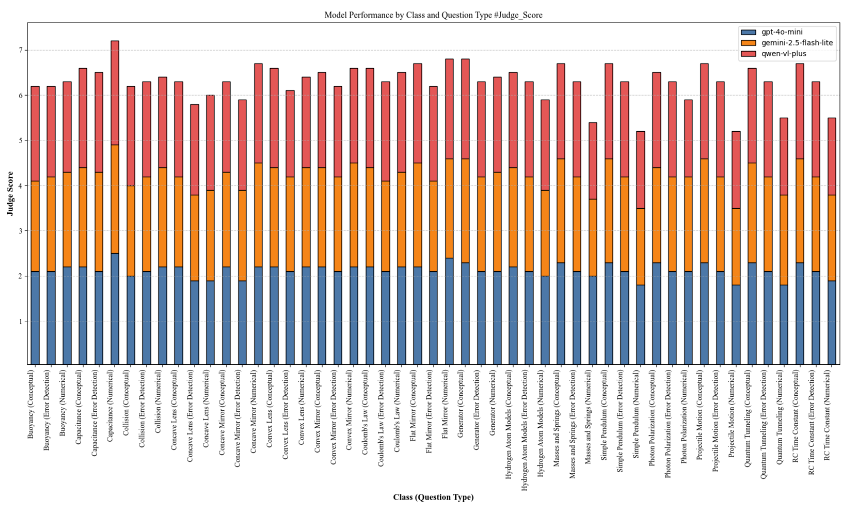


((c)) Category $\times$ Question Type (Standard).

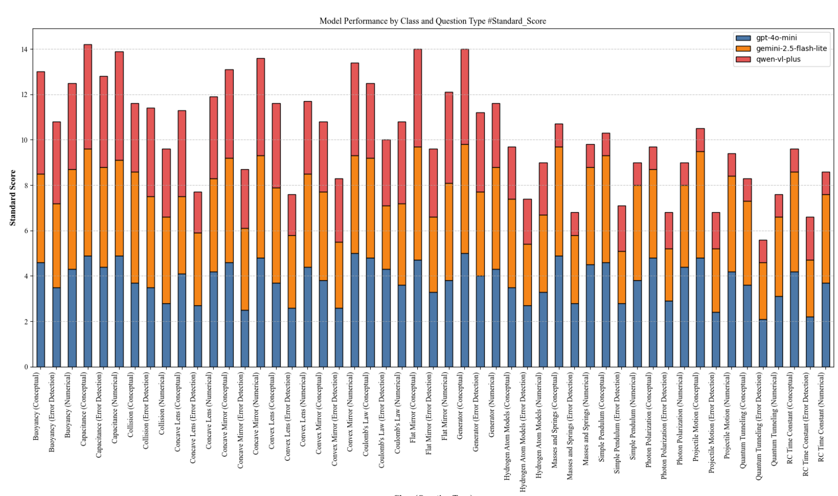
Figure 10: **Category–Question Type heatmaps across scoring variants.** Critical is most conservative, Judge and Standard broaden dynamic range; all show Circuits/Electromagnetics > Mechanics/Optics > Quantum Mechanics, with Conceptual, Error Detection, and Numerical questions showing distinct patterns.



((a)) Class × Question Type (Critical).



((b)) Class × Question Type (Judge).



((c)) Class × Question Type (Standard).

Figure 11: **Per-class question-type breakdown.** Error-detection remains the limiting factor even when classes are easy numerically (e.g., mirrors/lenses).

Class	Difficulty	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Buoyancy	Easy	1.51	1.55	1.46
	Moderate	1.63	1.54	1.35
	Hard	1.53	0.95	1.40
Capacitance	Easy	-	-	-
	Moderate	1.70	1.74	1.69
	Hard	1.79	1.61	1.72
Collision	Easy	1.47	1.52	1.45
	Moderate	1.42	1.50	1.57
	Hard	1.53	1.48	1.28
Concave Lens	Easy	1.19	1.22	1.37
	Moderate	1.35	1.34	1.58
	Hard	1.36	1.32	1.44
Concave Mirror	Easy	1.65	1.53	1.66
	Moderate	1.63	1.74	1.35
	Hard	1.39	1.41	1.48
Convex Lens	Easy	1.67	1.68	1.47
	Moderate	1.60	1.64	1.50
	Hard	1.67	1.65	1.56
Convex Mirror	Easy	1.70	1.75	1.63
	Moderate	1.73	1.84	1.59
	Hard	1.62	1.62	1.63
Coulomb's Law	Easy	1.71	1.66	1.66
	Moderate	1.69	1.64	1.64
	Hard	1.71	1.67	1.62
Flat Mirror	Easy	1.92	1.96	1.68
	Moderate	1.68	1.66	1.68
	Hard	-	-	-
Generator	Easy	1.67	1.65	1.61
	Moderate	1.64	1.70	1.60
	Hard	1.68	1.73	1.60
Hydrogen Atom Models	Easy	1.66	1.61	1.62
	Moderate	1.63	1.58	1.56
	Hard	1.59	1.62	1.54
Masses and Spring	Easy	1.61	1.51	1.40
	Moderate	1.58	1.51	1.42
	Hard	1.61	1.52	1.40
Simple Pendulum	Easy	1.51	1.55	1.40
	Moderate	1.52	1.51	1.42
	Hard	1.52	1.51	1.40
Photon Polarization	Easy	1.60	1.57	1.40
	Moderate	1.61	1.53	1.40
	Hard	1.60	1.57	1.40
Projectile Motion	Easy	1.55	1.51	1.40
	Moderate	1.57	1.52	1.40
	Hard	1.53	1.53	1.40
Quantum Tunneling	Easy	1.55	1.57	1.40
	Moderate	1.55	1.55	1.40
	Hard	1.54	1.53	1.40
RC Time Constant	Easy	1.48	1.45	1.40
	Moderate	1.54	1.54	1.40
	Hard	1.57	1.56	1.40

Table 6: **Class $\times$ Difficulty under *Critical_Score*** (strict single-judge; 1–5 rubric numerically concentrated near 1–2 due to conservative prompting). Higher is better. "–" denotes no items.

Class	Question Type	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Buoyancy	Numerical	1.65	1.52	1.41
	Error Detection	1.40	1.53	1.29
	Conceptual	1.59	1.44	1.56
Capacitance	Numerical	2.06	1.92	1.83
	Error Detection	1.41	1.65	1.60
	Conceptual	1.65	1.64	1.66
Collision	Numerical	1.61	1.64	1.32
	Error Detection	1.40	1.56	1.40
	Conceptual	1.40	1.31	1.61
Concave Lens	Numerical	1.13	1.28	1.66
	Error Detection	1.20	1.23	1.33
	Conceptual	1.63	1.41	1.42
Concave Mirror	Numerical	1.67	1.72	1.80
	Error Detection	1.27	1.32	1.28
	Conceptual	1.63	1.55	1.39
Convex Lens	Numerical	1.71	1.68	1.56
	Error Detection	1.50	1.52	1.29
	Conceptual	1.74	1.77	1.71
Convex Mirror	Numerical	1.77	1.87	1.69
	Error Detection	1.49	1.55	1.47
	Conceptual	1.76	1.75	1.69
Coulomb's Law	Numerical	1.76	1.70	1.69
	Error Detection	1.61	1.55	1.48
	Conceptual	1.74	1.73	1.74
Flat Mirror	Numerical	1.96	2.00	1.71
	Error Detection	1.57	1.47	1.57
	Conceptual	1.75	1.82	1.75
Generator	Numerical	1.61	1.58	1.50
	Error Detection	1.57	1.81	1.57
	Conceptual	1.73	1.74	1.62
Hydrogen Atom Models	Numerical	1.48	1.21	1.20
	Error Detection	1.50	1.50	1.50
	Conceptual	1.50	1.82	1.52
Masses and Springs	Numerical	1.82	1.27	1.20
	Error Detection	1.26	1.50	1.50
	Conceptual	1.47	1.82	1.52
Simple Pendulum	Numerical	1.82	1.48	1.52
	Error Detection	1.50	1.50	1.50
	Conceptual	1.50	1.81	1.51
Photon Polarization	Numerical	1.81	1.48	1.20
	Error Detection	1.31	1.50	1.50
	Conceptual	1.50	1.82	1.50
Projectile Motion	Numerical	1.81	1.37	1.20
	Error Detection	1.33	1.50	1.50
	Conceptual	1.50	1.75	1.51
Quantum Tunneling	Numerical	1.50	1.50	1.20
	Error Detection	1.36	1.32	1.50
	Conceptual	1.50	1.75	1.51
RC Time Constant	Numerical	1.76	1.50	1.50
	Error Detection	1.50	1.79	1.50
	Conceptual	1.76	1.79	1.50

Table 7: **Class $\times$ Question Type under *Critical_Score***. Error-detection (trap) rows are consistently lower than conceptual/numerical, reflecting difficulty with idealizations and counterfactuals.

Category	Difficulty	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Mechanics & Fluids	Easy	1.52	1.54	1.43
	Moderate	1.56	1.52	1.42
	Hard	1.54	1.48	1.39
Quantum Mechanics	Easy	1.63	1.60	1.54
	Moderate	1.59	1.56	1.48
	Hard	1.57	1.56	1.45
Electromagnetic & Circuits	Easy	1.67	1.64	1.61
	Moderate	1.64	1.67	1.59
	Hard	1.69	1.67	1.59
Optics	Easy	1.61	1.58	1.57
	Moderate	1.60	1.66	1.51
	Hard	1.49	1.49	1.52

Table 8: **Category** $\times$ **Difficulty under *Critical_Score***. Electromagnetism & Circuits tends to rank highest; Quantum Mechanics content is harder across difficulty tiers.

Category	Question Type	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Mechanics & Fluids	Conceptual	1.69	1.64	1.54
	Error Detection	1.45	1.52	1.42
	Numerical	1.47	1.38	1.28
Quantum Mechanics	Conceptual	1.77	1.73	1.56
	Error Detection	1.52	1.54	1.53
	Numerical	1.46	1.43	1.34
Electromagnetic & Circuits	Conceptual	1.73	1.73	1.66
	Error Detection	1.52	1.57	1.53
	Numerical	1.73	1.68	1.58
Optics	Conceptual	1.68	1.62	1.54
	Error Detection	1.36	1.39	1.34
	Numerical	1.62	1.68	1.70

Table 9: **Category** $\times$ **Question Type under *Critical_Score***. Error-detection remains the hardest type in all categories; Optics shows comparatively strong numerical scores.

Class	Difficulty	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Buoyancy	Easy	2.1	2.1	2.1
	Moderate	2.2	2.1	2
	Hard	2.1	1.8	2
Capacitance	Easy	-	-	-
	Moderate	2.3	2.2	2.2
	Hard	2.3	2.2	2.2
Collision	Easy	2.1	2.1	2.1
	Moderate	2.1	2.1	2.1
	Hard	2.1	2.1	2.0
Concave Lens	Easy	1.9	1.9	2.0
	Moderate	2.0	2.0	2.1
	Hard	2.0	2.0	2.1
Concave Mirror	Easy	2.2	2.1	2.1
	Moderate	2.1	2.2	2.0
	Hard	2.0	2.1	2.1
Convex Lens	Easy	2.2	2.2	2.0
	Moderate	2.1	2.1	2.0
	Hard	2.1	2.2	2.0
Convex Mirror	Easy	2.2	2.2	2.1
	Moderate	2.2	2.3	2.0
	Hard	2.1	2.1	2.1
Coulomb's Law	Easy	2.1	2.1	2.1
	Moderate	2.1	2.1	2.1
	Hard	2.2	2.1	2.1
Flat Mirror	Easy	2.4	2.4	2.2
	Moderate	2.2	2.1	2.2
	Hard	-	-	-
Generator	Easy	2.2	2.2	2.1
	Moderate	2.2	2.2	2.1
	Hard	2.2	2.2	2.1
Hydrogen Atom Models	Easy	2.2	2.1	2.1
	Moderate	2.1	2	2.1
	Hard	2.1	2.1	2.1
Masses and Springs	Easy	2.2	2.0	2.0
	Moderate	2.1	2.0	2.0
	Hard	2.1	2.0	2.0
Simple Pendulum	Easy	2.0	2.0	2.0
	Moderate	2.0	2.0	2.0
	Hard	2.0	2.0	2.0
Photon Polarization	Easy	2.2	2.1	2.0
	Moderate	2.2	2.1	2.0
	Hard	2.2	2.1	2.0
Projectile Motion	Easy	2.0	2.0	2.0
	Moderate	2.1	2.0	2.0
	Hard	2.1	2.1	2.0
Quantum Tunneling	Easy	2.2	2.1	2.0
	Moderate	2.1	2.1	2.0
	Hard	2.1	2.1	2.0
RC Time Constant	Easy	2.0	2.0	2.0
	Moderate	2.1	2.1	2.0
	Hard	2.1	2.1	2.0

Table 10: **Class × Difficulty under *Judge_Score*** (dual-judge JSON, temperature = 0; 1–5 rubric). Calibrated to be more stable and comparable across classes than Critical.

Class	Question Type	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Buoyancy	Conceptual	2.1	2.0	2.1
	Error_Detection	2.1	2.1	2.0
	Numerical	2.2	2.1	2.0
Capacitance	Conceptual	2.2	2.2	2.2
	Error_Detection	2.1	2.2	2.2
	Numerical	2.5	2.4	2.3
Collision	Conceptual	2.0	2.0	2.2
	Error_Detection	2.1	2.1	2.1
	Numerical	2.2	2.2	2.0
Concave Lens	Conceptual	2.2	2.0	2.1
	Error_Detection	1.9	1.9	2.0
	Numerical	1.9	2.0	2.1
Concave Mirror	Conceptual	2.2	2.1	2.0
	Error_Detection	1.9	2.0	2.0
	Numerical	2.2	2.3	2.2
Convex Lens	Conceptual	2.2	2.2	2.2
	Error_Detection	2.1	2.1	1.9
	Numerical	2.2	2.2	2.0
Convex Mirror	Conceptual	2.2	2.2	2.1
	Error_Detection	2.1	2.1	2.0
	Numerical	2.2	2.3	2.1
Coulomb's Law	Conceptual	2.2	2.2	2.2
	Error_Detection	2.1	2.0	2.2
	Numerical	2.2	2.1	2.2
Flat Mirror	Conceptual	2.2	2.3	2.2
	Error_Detection	2.1	2.0	2.1
	Numerical	2.4	2.2	2.2
Generator	Conceptual	2.3	2.3	2.2
	Error_Detection	2.1	2.1	2.1
	Numerical	2.1	2.2	2.1
Hydrogen Atom Models	Conceptual	2.2	2.2	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	2.0	1.9	2.0
Masses and Springs	Conceptual	2.3	2.3	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	2.0	1.7	1.7
Simple Pendulum	Conceptual	2.3	2.3	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	1.8	1.7	1.7
Photon Polarization	Conceptual	2.3	2.1	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	2.1	2.1	1.7
Projectile Motion	Conceptual	2.3	2.3	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	1.8	1.7	1.7
Quantum Tunneling	Conceptual	2.3	2.2	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	1.8	2.0	1.7
RC Time Constant	Conceptual	2.3	2.3	2.1
	Error_Detection	2.1	2.1	2.1
	Numerical	1.9	1.9	1.7

Table 11: **Class × Question Type under *Judge_Score*.** Maintains the error-detection gap while reducing variance, enabling more reliable cross-model comparisons.

Category	Difficulty	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Mechanics & Fluids	Easy	2.1	2.1	2
	Moderate	2.1	2.1	2
	Hard	2.1	2	2
Quantum Mechanics	Easy	2.2	2.1	2.1
	Moderate	2.1	2.1	2
	Hard	2.1	2.1	2
Electromagnetic & Circuits	Easy	2.1	2.1	2.1
	Moderate	2.2	2.2	2.1
	Hard	2.2	2.2	2.1
Optics	Easy	2.2	2.1	2.1
	Moderate	2.1	2.2	2.1
	Hard	2.1	2.1	2.1

Table 12: **Category $\times$ Difficulty under *Judge_Score***. Trends mirror Critical_Score but with less compression; Circuits leads, Quantum Mechanics lags.

Category	Question Type	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Mechanics & Fluids	Conceptual	2.20	2.17	2.14
	Numerical	1.98	1.90	1.84
	Error Detection	2.10	2.14	2.09
Quantum Mechanics	Conceptual	2.24	2.20	2.14
	Numerical	1.96	1.95	1.85
	Error Detection	2.13	2.14	2.13
Electromagnetic & Circuits	Conceptual	2.22	2.21	2.17
	Numerical	2.21	2.17	2.09
	Error Detection	2.10	2.11	2.09
Optics	Conceptual	2.18	2.16	2.08
	Numerical	2.17	2.21	2.15
	Error Detection	2.01	2.02	1.99

Table 13: **Category $\times$ Question Type under *Judge_Score***. Numerical scoring is strongest in Optics and Circuits; error-detection is uniformly lower.

Class	Difficulty	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Buoyancy	Easy	4.1	4.1	4
	Moderate	4.2	3.9	4.4
	Hard	3.9	3.7	2.5
Capacitance	Easy	-	-	-
	Moderate	4.7	4.4	4.5
	Hard	4.8	4.4	4.7
Collision	Easy	3.2	4.3	3.1
	Moderate	3.5	4.2	3.2
	Hard	3.2	4.1	3.6
Concave Lens	Easy	3.8	4.1	3.1
	Moderate	3.7	3.4	3.1
	Hard	3.5	3.4	3
Concave Mirror	Easy	4.3	4.3	3.7
	Moderate	3.6	3.9	3.6
	Hard	4	4.3	3.5
Convex Lens	Easy	3.9	4.2	3.6
	Moderate	3.6	3.4	3.1
	Hard	3.3	3.7	2.5
Convex Mirror	Easy	3.5	4.4	3.7
	Moderate	4.1	4	3.1
	Hard	3.7	3.2	3.3
Coulomb's Law	Easy	4.1	3.6	3.4
	Moderate	4.4	3.6	3.5
	Hard	4.3	3.6	3
Flat Mirror	Easy	3.7	4.3	3.3
	Moderate	4.1	4.2	4
	Hard	-	-	-
Generator	Easy	4.2	4.3	3.6
	Moderate	4.3	4.3	3.3
	Hard	4.6	4.3	3.7
Hydrogen Atom Models	Easy	3.5	3.6	2.6
	Moderate	3	3.2	2
	Hard	3.2	3.4	2.2
Masses and Springs	Easy	4.1	4.3	1
	Moderate	4.2	4.1	1
	Hard	4	3.8	1
Simple Pendulum	Easy	3.8	3.8	1.3
	Moderate	3.6	3.6	1.3
	Hard	3.8	3.9	1.3
Photon Polarization	Easy	3.9	3	1.2
	Moderate	3.7	3.3	1.3
	Hard	4.2	3.4	1.1
Projectile Motion	Easy	3.8	3.8	1.2
	Moderate	3.9	4.1	1.2
	Hard	3.8	3.9	1.2
Quantum Tunneling	Easy	3	3.7	1
	Moderate	2.9	3.3	1.1
	Hard	3	3.2	1
RC Time Constant	Easy	2.7	3	1.3
	Moderate	3.5	3.7	1.3
	Hard	3.3	3.6	1.3

Table 14: **Class × Difficulty under *Standard_Score*** (same protocol as Judge_Score; exported view used in the main text). Absolute values are on the 1–5 scale.

Class	Question Type	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Buoyancy	Conceptual	4.6	3.9	4.5
	Error Detection	3.5	3.7	3.6
	Numerical	4.3	4.4	3.8
Capacitance	Conceptual	4.9	4.7	4.6
	Error Detection	4.4	4.4	4.0
	Numerical	4.9	4.2	4.8
Collision	Conceptual	3.7	4.9	3.0
	Error Detection	3.5	4.0	3.9
	Numerical	2.8	3.8	3.0
Concave Lens	Conceptual	4.1	3.4	3.8
	Error Detection	2.7	3.2	1.8
	Numerical	4.2	4.1	3.6
Concave Mirror	Conceptual	4.6	4.6	3.9
	Error Detection	2.5	3.6	2.6
	Numerical	4.8	4.5	4.3
Convex Lens	Conceptual	3.7	4.2	3.7
	Error Detection	2.6	3.2	1.8
	Numerical	4.4	4.1	3.2
Convex Mirror	Conceptual	3.8	3.9	3.1
	Error Detection	2.6	2.9	2.8
	Numerical	5.0	4.3	4.1
Coulomb's Law	Conceptual	4.8	4.4	3.3
	Error Detection	4.3	2.8	2.9
	Numerical	3.6	3.6	3.6
Flat Mirror	Conceptual	4.7	5.0	4.3
	Error Detection	3.3	3.3	3.0
	Numerical	3.8	4.3	4.0
Generator	Conceptual	5.0	4.8	4.2
	Error Detection	4.0	3.7	3.5
	Numerical	4.3	4.5	2.8
Hydrogen Atom Models	Conceptual	3.5	3.9	2.3
	Error Detection	2.7	2.7	2.0
	Numerical	3.3	3.4	2.3
Masses and Springs	Conceptual	4.9	4.8	1.0
	Error Detection	2.8	3.0	1.0
	Numerical	4.5	4.3	1.0
Simple Pendulum	Conceptual	4.6	4.7	1.0
	Error Detection	2.8	2.3	2.0
	Numerical	3.8	4.2	1.0
Photon Polarization	Conceptual	4.8	3.9	1.0
	Error Detection	2.9	2.3	1.6
	Numerical	4.4	3.6	1.0
Projectile Motion	Conceptual	4.8	4.7	1.0
	Error Detection	2.4	2.8	1.6
	Numerical	4.2	4.2	1.0
Quantum Tunneling	Conceptual	3.6	3.7	1.0
	Error Detection	2.1	2.5	1.0
	Numerical	3.1	3.5	1.0
RC Time Constant	Conceptual	4.2	4.4	1.0
	Error Detection	2.2	2.5	1.9
	Numerical	3.7	3.9	1.0

Table 15: **Class × Question Type under *Standard_Score***. Clear gap between error-detection and the other two types across most classes.

Category	Difficulty	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Mechanics & Fluids	Easy	3.9	4.0	2.9
	Moderate	3.9	4.0	2.7
	Hard	3.8	3.9	1.5
Quantum Mechanics	Easy	3.5	3.4	2.1
	Moderate	2.9	3.2	1.5
	Hard	3.2	3.3	1.5
Electromagnetic & Circuits	Easy	4.0	3.7	3.3
	Moderate	4.3	4.1	3.3
	Hard	4.3	3.9	3.2
Optics	Easy	4.0	4.3	3.6
	Moderate	3.8	3.8	3.4
	Hard	3.7	3.8	3.2

Table 16: **Category × Difficulty under *Standard_Score***. Consistent ordering across difficulties; Quantum Mechanics remains the most challenging.

Category	Question Type	Model		
		gpt-4o-mini	gemini-2.5-flash-lite	qwen-vl-plus
Mechanics & Fluids	Conceptual	4.6	4.5	2.3
	Error Detection	3.0	3.2	2.5
	Numerical	4.0	4.2	2.1
Quantum Mechanics	Conceptual	3.7	3.8	1.6
	Error Detection	2.4	2.5	1.5
	Numerical	3.3	3.5	1.6
Electromagnetic & Circuits	Conceptual	4.7	4.6	3.3
	Error Detection	3.8	3.4	3.1
	Numerical	4.2	4.0	3.2
Optics	Conceptual	4.2	4.2	3.7
	Error Detection	2.6	3.3	2.3
	Numerical	4.6	4.3	3.9

Table 17: **Category × Question Type under *Standard_Score***. Optics and Electromagnetism lead on numerical; error-detection is the hardest across all categories.

E Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim to present a new benchmark, PhET-Physics-VideoQA, for evaluating physics understanding in VLMs. The paper's content, including the dataset description, experimental setup, and results, aligns with these claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations" section (Section 5) that discusses several limitations, including the sim-to-real gap, potential for superficial strategies, dataset size, and evaluation protocol assumptions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper introduces a new dataset and presents an empirical evaluation of existing models. It does not propose new theoretical results, theorems, or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides significant detail on the experimental setup in Section 4 (specifically Sections 4.1 and 4.2), including the models used, video preprocessing parameters, decoding settings (temperature=0), and the full scoring protocol.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper states its intention to release the necessary assets for reproduction, including prompts, seeds, cached frames, lightweight graders, and generation scripts, under appropriate licenses (Section 5).

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper is evaluating pre-trained models, so no training details are applicable. It clearly specifies all test details in Section 4.1, including the full dataset used, the models evaluated, and the decoding settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper describes its method for ensuring statistical robustness in Section 4.2, stating, "We attach 95% confidence intervals via stratified bootstrap over videos (10,000 resamples)".

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have put the experiments' details regarding the compute resources in Appendix D.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research introduces a benchmark for AI evaluation using publicly available educational software. A "Licensing & Ethics" section in Appendix A.1 confirms that no personal data is used and all licenses are respected.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper's focus is on the creation and technical evaluation of a research benchmark; it does not contain a dedicated section on its broader societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The dataset is derived from educational physics simulations and does not pose a high risk for misuse. Therefore, safeguards in this context are not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper explicitly credits PhET Interactive Simulations and cites the creators (Wieman et al.) in Appendix A.1. It also states that the new assets will be released under a CC-BY-NC license, respecting the source's usage terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new dataset is documented extensively in Section 3 and Appendix A, including details on design goals, compilation, metadata schemas, and topic distribution.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The research did not involve crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research did not involve human subjects, so IRB approval was not required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper declares in Section 3.2 that question drafts were initially generated by "GPT-5 Thinking" before being fully vetted by human experts. This constitutes a non-standard component of the data creation methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.