BilliardPhys-Bench: Benchmarking Physical Reasoning and Visual Dynamics of Multimodal LLMs

Anonymous ACL submission

Abstract

Current multimodal models handle static image recognition well, but intuitive physical reasoning remains a weakness. Predicting how objects will move and interact from a single image is still difficult for these systems. We present BilliardPhys-Bench, a benchmark for physical reasoning in synthetic billiards environments. Its procedural engine generates randomized scenarios with friction and elastic collisions. The benchmark tests three abilities: (1) predicting ball-to-ball collisions, (2) reasoning about wall bounces, and (3) estimating final ball positions after motion stops. We evaluate recent MLLMs from the GPT, Claude, Gemini, and Qwen families. Performance drops as simulation time increases and scene geometry grows more complex. We also observe a consistent failure mode we call "stasis bias": when the correct physical outcome is harder to infer, models tend to predict no interaction. These findings show where current MLLMs break down on visual dynamics and point toward the need for better physical inductive biases in multimodal architectures.

1 Introduction

Reasoning about the physical world is a prerequisite for general-purpose AI (NVIDIA et al., 2025; Wu et al., 2024; Upadhyay et al., 2026; Zhao et al., 2026; Zhou et al., 2026). Humans do this easily: from a single image of a billiard table, most people can anticipate likely trajectories, collisions, and outcomes. Current Multimodal Large Language Models (MLLMs), however, still struggle to infer accurate physical evolution from a 2D image (Yao et al., 2024; Ye et al., 2024; Zhang et al., 2026a,b).

Billiards is well-suited for evaluation because it requires geometric estimation of collision angles, multi-body tracking, and causal reasoning over dynamics (Zhao et al., 2015; Kiefl, 2024; Alberico et al., 2025). Unlike standard visual question answering, which mainly tests semantic recognition, billiard prediction is sensitive to initial conditions: small errors in perceived position or velocity compound into substantially different outcomes (Wang et al., 2025; Xu et al., 2025b,a; Yu et al., 2026; Mak et al., 2026). This sensitivity makes billiards a test of predictive physical reasoning rather than descriptive scene understanding (Yuan et al., 2025; Jiang et al., 2025; Liu et al., 2025).

Existing benchmarks for physical reasoning fall into three categories. Static QA benchmarks test knowledge of physical laws through text or image queries (Wang et al., 2025; Shen et al., 2025; Xiang et al., 2025; Chung et al., 2025; Zhu et al., 2025; Ye et al., 2025; LSST Dark Energy Science Collaboration et al., 2026)—these measure declarative knowledge but do not require reasoning about evolving environments. Symbolic or code-based benchmarks present physics through abstract interfaces, reducing the need for perceptual grounding (Bakhtin et al., 2019; Matthews et al., 2025; Cherian et al., 2024; Li et al., 2025; Zhang et al., 2025). Visual interactive benchmarks evaluate VLMs as agents that plan and adapt through trial and error (Li et al., 2025; Zhang et al., 2025). Single-frame forward simulation—predicting future states from a static initial condition without iterative interaction—remains less studied.

This gap reflects a separation between perception and reasoning. Models may describe scenes or state physical rules correctly yet fail to produce accurate predictions (Xu et al., 2025; Lee et al., 2025; Puyin et al., 2025; Lin et al., 2025). In billiards, this appears as stasis bias and failures on sequential collisions. Prior work on ball trajectory prediction has mainly used frame-by-frame tracking in video (Yu, 2025; Kienzle et al., 2025; Gomez-Gonzalez et al., 2020; Chiang et al., 2024) rather than prediction from a static state.

We introduce BilliardPhys-Bench with three contributions:

(1) A procedurally generated, multi-tier benchmark. Our engine creates randomized billiard sce-

narios, each providing a high-resolution initial-state image paired with physics-engine ground truth. The benchmark decomposes the problem into three levels: (a) discrete collision prediction (will the cue ball collide, and with which object?), (b) continuous final-state estimation (precise coordinates of all balls after energy dissipates), and (c) complex interaction chains (reasoning through secondary and tertiary collisions).

(2) A diagnostic framework for physical reasoning failures. We evaluate leading proprietary and open-source MLLMs and analyze failure modes—stasis bias, misunderstanding of momentum transfer, sensitivity to visual details—providing specific insight into why current architectures fail at visual dynamical reasoning.

(3) Directions for stronger physical inductive biases. Our results show the limits of next-token prediction for physical simulation and outline pathways through hybrid neural-physical approaches (Achterhold et al., 2023), factor graphs, or differentiable physics engines.

By shifting evaluation from interactive action to one-shot prediction, our benchmark tests whether MLLMs can maintain an internal model of physics.

2 Data Generation

2.1 Physical modeling of constant friction

To describe the motion of a billiard ball under constant friction, we consider a ball of mass m moving on a horizontal surface with a coefficient of friction μ. The kinetic friction force f acting against the direction of motion is:

$$f = \mu N = \mu m g \tag{1}$$

where N is the normal force and g is gravitational acceleration. By Newton's second law ($F = ma$), the constant deceleration a is:

$$a = \frac{-f}{m} = \frac{-\mu m g}{m} = -\mu g \tag{2}$$

For uniformly decelerated linear motion, the relationship between final velocity v, initial velocity v_0, acceleration a, and displacement s is:

$$v^2 - v_0^2 = 2as \tag{3}$$

Substituting $a = -\mu g$ gives velocity as a function of displacement:

$$v^2 = v_0^2 - 2\mu g s, \qquad v(s) = \sqrt{v_0^2 - 2\mu g s} \tag{4}$$

Under this constant friction model, ball velocity decreases non-linearly with distance traveled (Eq. 4).

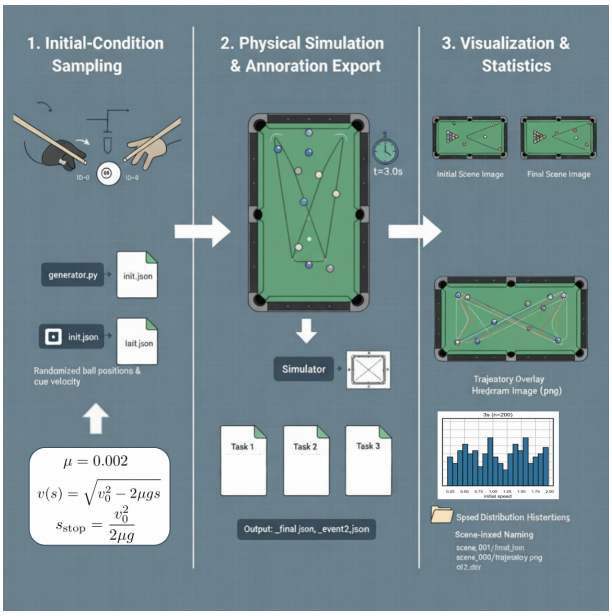

Figure 1: BilliardPhys-Bench data generation pipeline. The pipeline has three stages: (1) initial-condition sampling, which randomizes ball positions and cue velocities; (2) physical simulation, in which a high-fidelity engine computes elastic collisions and friction-based deceleration while exporting structured annotations; and (3) visualization and statistics, which generate the rendered images and dataset distribution reports.

2.2 Dataset generation

The dataset was synthesized using a custom generator. Each example is created through three stages: initial-condition sampling, physical simulation and annotation export, and visualization (see Figure 1).

First, an initial scene is sampled and saved as `init.json`: cue ball (ID=0) velocity (direction and magnitude) and positions of all other balls are randomly sampled according to configured ranges and seed. Friction is modeled as speed decrease proportional to path length (Eq. 4), so a ball stops after traveling:

$$s_{\text{stop}} = \frac{v_0^2}{2\mu g}. \tag{5}$$

Here $g = 9.8\,\mathrm{m/s^2}$ and $\mu = 0.002$. We chose this friction coefficient so that collisions occur in roughly 50% of generated scenarios given our initial velocity intervals. All collisions (ball-ball and ball-wall) are perfectly elastic. If a ball is pocketed during simulation, its final position is recorded as `null`. Second, the simulator advances to the target time ($t \in \{1, 2, 3, 4, 5\}$ s) and writes `_final.json` with:

- **Task 1 (Q1):** Per-ball collision labels saved

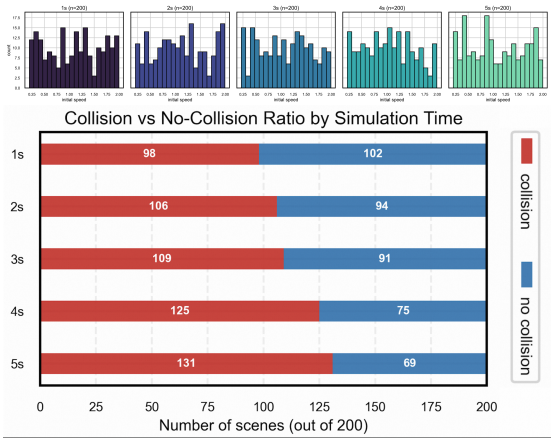

Figure 2: Distribution of the generated benchmark samples. The top histograms show sampled cue-ball initial speeds for simulation windows from 1 s to 5 s. The bottom bar chart shows the number of collision and no-collision cases among the 200 sampled scenes at each time window.

as a list of objects in the form `{"id": i, "answer": "T" | "F"}`, with one entry per non-cue ball.

- **Task 2 (Q2):** Per-wall-collision answers stored as an answer list (e.g., `{"wall": "TOP", "answer": "T" | "F"}`) representing collision or no collision.
- **Task 3 (Predictions):** Final positions saved under the key `predictions` as a list of `{"id": <int>, "pos": [x, y]}`; pocketed balls use `pos = null`. Numeric coordinates are recorded to at most four decimal places.

Third, the generator produces visual outputs: initial and final scene images, a trajectory overlay rendered with the same canvas size and ball radii as the frame renderer, and a histogram of sampled cue-ball initial speeds (Figure 2). All files for a scene are stored under a scene-indexed naming convention for automated evaluation and reproducibility. The BilliardPhys-Bench dataset and code will be released under the MIT License.

2.3 Model I/O and evaluation pipeline

The pipeline (Figure 3) constructs a structured prompt with an optional scene image, submits it to a chat-style LLM API, extracts JSON, and judges correctness against ground truth.

Inputs sent to the model

- **Image:** Scene image showing ball positions, IDs, and cue velocity arrow (Figure 4), with axes $x \in [0, 2], y \in [0, 1]$.
- **Prompt:** States friction ($v(s) = \sqrt{v_0^2 - 2\mu g s}$, $\mu = 0.002$, $g = 9.8$), stopping distance ($s_{\text{stop}} = v_0^2/(2\mu g)$), three task definitions, and JSON output format.

The three tasks:

1. **Task 1 (Q1):** For each non-cue ball, predict whether the cue ball collides with it.
2. **Task 2 (Q2):** For each wall, predict whether the cue ball collides with it.
3. **Task 3 (Q3):** Predict the 2D position of every ball at the target time as `{"id": <int>, "pos": [x, y]}`; pocketed balls use `pos = null`.

API call strategy The implementation uses separate model calls for Tasks 1, 2, and 3 to reduce output-format ambiguity and allow validation rules tailored to each task. All calls use the configured chat-completions endpoint with JSON-formatted messages. Chain-of-thought is enabled for Q1/Q2 but disabled for Q3, where free-form reasoning interferes with the required JSON output. Runtime parameters include `-max-tokens`, `-timeout`, and `-enable-cot`.

Validation and retry After each reply, the client extracts JSON by locating the first balanced `{...}` substring. Failed or truncated responses trigger retries with escalation strategies (fresh session or increased timeout).

- **Task 1/2:** Parsed against expected schema; repeated if malformed or incomplete.
- **Task 3:** Must contain a `predictions` list covering IDs 0–6 with at least one non-null numeric coordinate and range checks. When `force_numeric` is enabled, the prompt instructs numeric estimates rather than `null`.

Successful outputs are saved as three JSON files, one per task.

Correctness judgments A separate utility evaluates predictions against ground truth:

- **Q1 (collisions):** T/F compared per ball against GT labels.
- **Q2 (walls):** Correct only if the predicted wall set exactly matches GT.

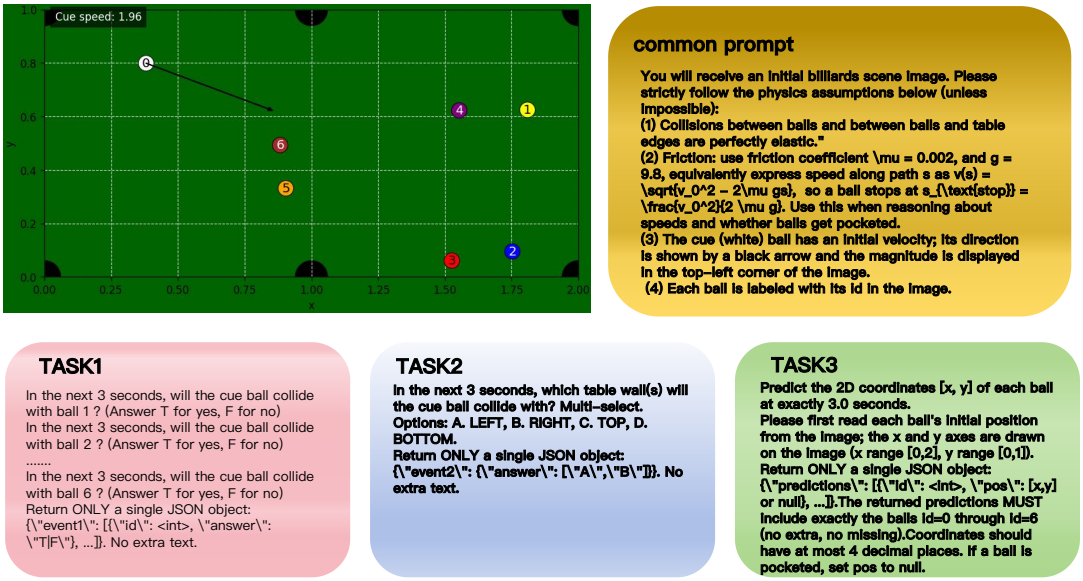

Figure 3: Illustration of the prompt engineering and task definitions. The evaluation pipeline provides models with a "common prompt" detailing physical constants (friction, elasticity) and the input image. The three diagnostic tasks, collision detection (Task 1), wall interaction (Task 2), and coordinate prediction (Task 3), require the model to return structured JSON outputs for automated scoring. This is an example for the 3s scenes.

- **Q3 (positions):** Correct if Euclidean distance $\leq$ ball radius; both-null counts as correct.

Practical notes The prompt requires JSON-only outputs with explicit examples. We set temperature to 0 and impose no maximum token limit for all models. In base vs. "Pro" GPT comparisons, prompt format and client settings are fixed; observed token-usage differences reflect model-side behavior.

3 Results

Our evaluation on `BilliardPhys-Bench` reveals clear differences in how MLLMs handle physical reasoning. We analyze performance across three diagnostic tasks at temporal horizons from 1s to 5s.

3.1 Task 1: collision event prediction

Task 1 separates models with event-level reasoning from those that only produce plausible final states (Figure 5, Figure 6).

- **Top tier:** *GPT-5.4-Pro* and *GPT-5.5* form the first tier with mean accuracies of 78.02% and 74.43%, both remaining above 60% even at 5 s. *Qwen3.6-Plus* forms a competitive second tier, showing that strong event reasoning is no longer limited to one model family.

- **Structural weakness:** *GPT-5.4* and *Claude-Opus-4.7* achieve only 15.66% and 20.01% mean accuracy—weak from 1 s onward, indicating a basic deficit in discrete event identification rather than a long-horizon failure.

3.2 Task 2: wall interaction reasoning

Task 2 is the most polarized: some models sustain wall-collision reasoning across all horizons while others never develop a stable model of boundary reflection.

- **Stable boundary reasoning:** *GPT-5.5* leads with 90.34% mean accuracy, *Qwen3.6-Plus* follows at 87.39%; both remain strong across 1–5 s rather than only at short windows. *GPT-5.4-Pro* also remains in the top tier, showing that strong Task 2 performance is shared by several of the best overall models.

- **Long-horizon specialization:** *Claude-Opus-4.6* attains the best 5 s score (89.54%) despite not ranking near the top overall. *GPT-5.4* and *Claude-Opus-4.7* remain weak on Task 2, confirming that terminal-state prediction alone does not translate into high overall rank.

3.3 Task 3: precise coordinate estimation

Task 3 requires exact $[x, y]$ coordinates for all balls, measuring whether a model can maintain numerically grounded predictions rather than only classify events.

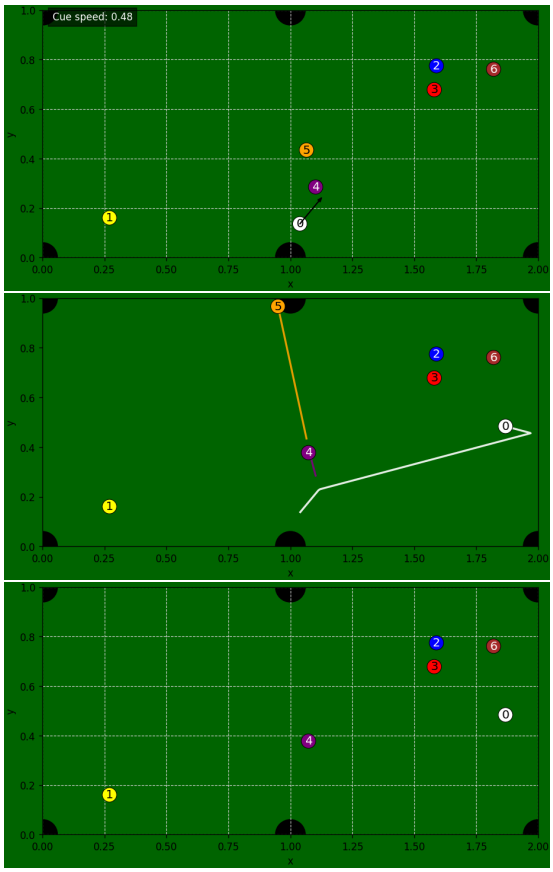

Figure 4: Example from the billiard reasoning benchmark. (Top) The input image, with the initial ball positions and cue-ball velocity vector. (Middle) The ground-truth trajectory overlay, showing wall reflections and ball-to-ball collisions. (Bottom) The board state at the target prediction time.

- **Top performance:** *GPT-5.4-Pro* and *GPT-5.5* reach 87.93% and 86.93% at 1 s and remain above 60% at 5 s. Precise final-state prediction is substantially stronger in these models than in earlier baselines.

- **Different axis from event reasoning:** *GPT-5.4* and *Claude-Opus-4.7* are strong on Task 3 despite weak event reasoning, whereas *Qwen3.6-Plus* shows the reverse. The benchmark thus distinguishes predicting a plausible final state from identifying the events that causally produce it.

3.4 Temporal dynamics and biases

Longer horizons do not affect all models equally. Extending from 1 s to 5 s introduces longer interaction chains, greater friction effects, and more elastic reflections. Balanced models (*GPT-5.5*, *GPT-5.4-Pro*) degrade gradually; uneven models mainly expose existing weaknesses rather than creating new ones.

Failures are non-uniform: *GPT-5.4* and *Claude-Opus-4.7* miss event-level questions even when final-state estimates remain strong, while *Qwen3.6-Plus* loses accuracy on terminal-state prediction. This is consistent with "stasis bias": under growing uncertainty, some models default to conservative event judgments even when their terminal-state predictions remain plausible. The benchmark becomes more diagnostic as the horizon increases, not merely harder.

3.5 Overall physical reasoning performance

The total score (Figure 6, bottom-right) combines the three tasks:

$$S_{total} = 0.3 \cdot Acc_{T1} + 0.3 \cdot Acc_{T2} + 0.4 \cdot Acc_{T3} \quad (6)$$

Task 3 receives the highest weight (0.4) because it requires precise coordinate prediction and is the most demanding.

The leaderboard rewards models that avoid a sharp trade-off between event reasoning and terminal-state prediction. *GPT-5.5* ranks first (73.58%), with *GPT-5.4-Pro* close behind (72.29%); their profiles differ slightly (*GPT-5.5* strongest on Task 2, *GPT-5.4-Pro* more balanced). *GPT-5.2-Pro* is the strongest earlier baseline (62.34%). Below this group, *Gemini-3-Flash-Preview* and *Qwen3.6-Plus* remain competitive—the former balanced, the latter carried by event reasoning. *GPT-5.4* and *Claude-Opus-4.7* confirm that strong terminal-state prediction alone is insufficient when event reasoning is weak.

4 Discussion

4.1 The GPT family: model progression and test-time compute

The GPT family contains both cross-generation changes (*GPT-5.2* → *GPT-5.4* → *GPT-5.5*) and within-generation reasoning-depth variants (base vs. Pro). The "Pro" variants are deeper-reasoning configurations rather than separate base models.

Additional reasoning depth helps most on event inference. This is visible in the *GPT-5.2*/*GPT-5.2-Pro* comparison, where Task 1 and Task 2 improve far more than terminal-state prediction. It becomes clearer in the *GPT-5.4*/*GPT-5.4-Pro* pair: under the lighter setting, *GPT-5.4* is "terminal-state strong, event-level weak"; under deeper reasoning, *GPT-5.4-Pro* becomes one of the most balanced models. This suggests that event-level weakness in base

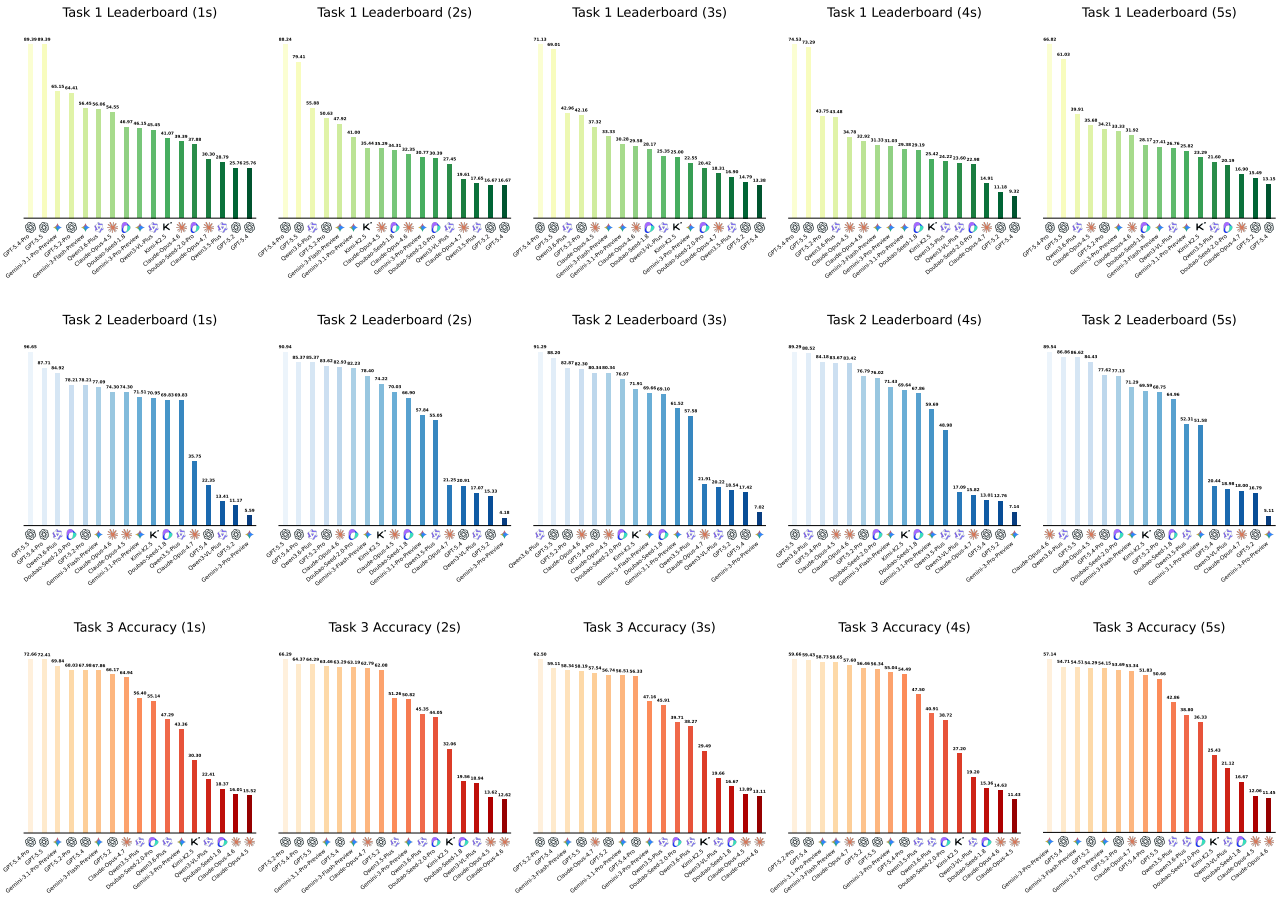

Figure 5: Performance of leading MLLMs across the three tasks. The bar charts compare model accuracy as the simulation time increases from 1 s to 5 s. Accuracy decreases across all three tasks as the prediction window lengthens.

configurations can be mitigated by more test-time reasoning.

Cross-generation change is not monotonic. Moving from *GPT-5.2* to *GPT-5.4* does not produce uniform gain; instead the model becomes more polarized, with stronger terminal-state prediction but persistently weak event reasoning. *GPT-5.5* then improves on the high-reasoning regime represented by *GPT-5.4-Pro*, achieving the best overall score and Task 2 performance, but its advantage is incremental rather than a restructuring of capability.

A cost trade-off: *GPT-5.4-Pro* and *GPT-5.5* consume roughly $8\times$ more tokens than other models including *GPT-5.4*, with most of the extra budget going to reasoning tokens. Because prompt and client settings are held fixed, this gap reflects model-side default reasoning behavior rather than something introduced by our evaluation setup. We cannot inspect hidden reasoning traces directly, so this conclusion is based on usage metadata. This does not diminish the strongest GPT results, but accuracy and inference efficiency are not aligned in this comparison.

4.2 The Gemini family

Gemini-3-Flash-Preview (58.06%) and *Gemini-3.1-Pro-Preview* (53.84%) offer balanced profiles, remaining competitive on terminal-state prediction but clearly behind the top GPT models on event reasoning and aggregate score. The Gemini family shows useful consistency across tasks, yet consistency alone is no longer enough to secure top ranks once stronger event-reasoning models are included.

4.3 The Qwen series

Qwen3.6-Plus (57.36%) far exceeds *Qwen3.5-Plus* (43.09%) and *Qwen3-VL-Plus* (22.23%). Its profile is distinctive:

- **Strong event reasoning:** Task 2 at 87.39% (stable between 84.92% and 91.29% across horizons), competitive on Task 1 at 47.66%.
- **Weaker terminal-state prediction:** Task 3 drops from 57.00% at 1 s to 44.71% at 5 s, suggesting better local event judgments than long-horizon physical rollout.

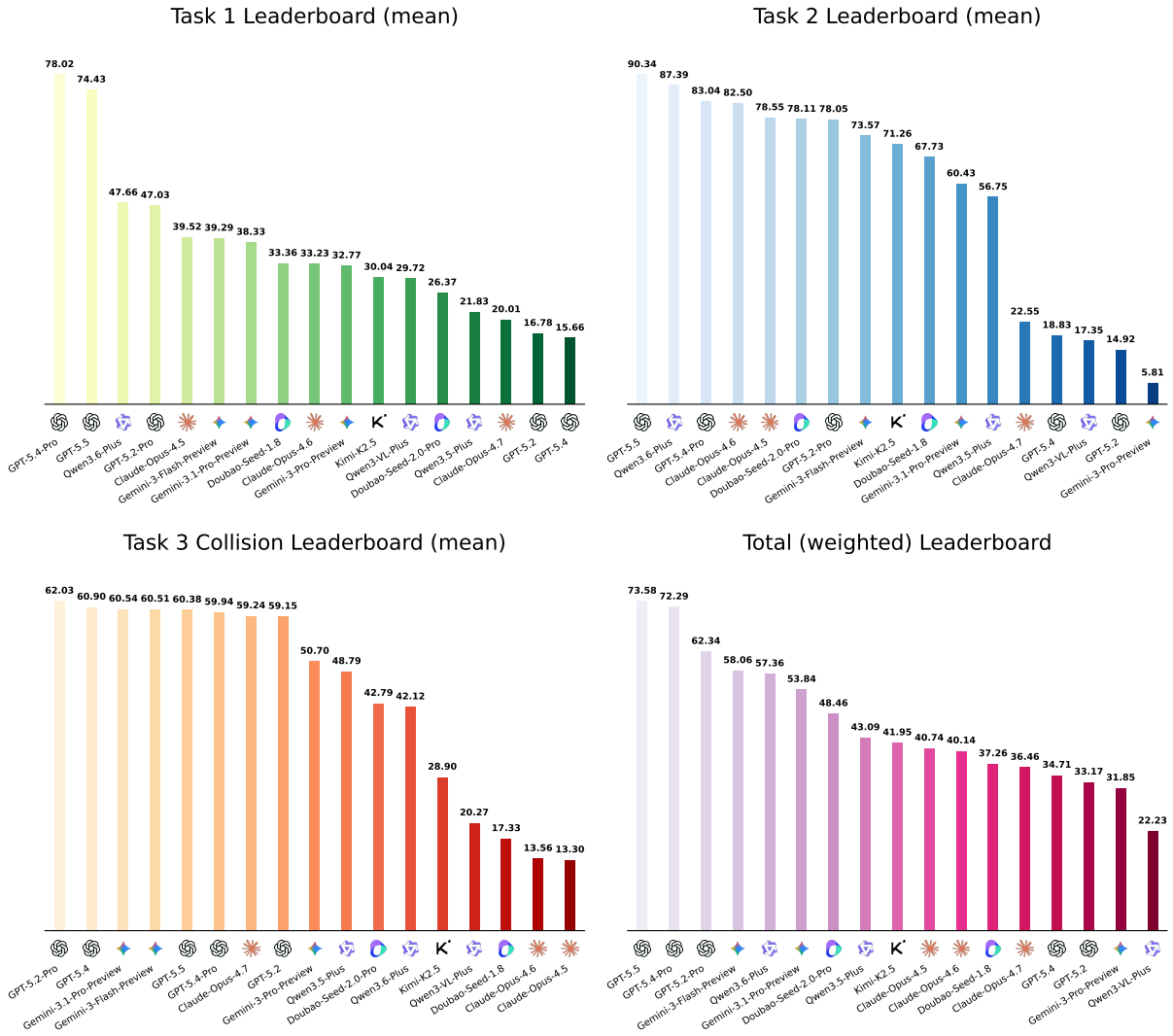

Figure 6: **Aggregate performance leaderboards across the diagnostic tasks.** The bar charts show mean accuracy averaged over all temporal horizons (1 s to 5 s). (Top-left) **Task 1 leaderboard**: mean accuracy for discrete collision prediction. (Top-right) **Task 2 leaderboard**: mean accuracy for wall-interaction reasoning. (Bottom-left) **Task 3 leaderboard**: mean accuracy for precise 2D coordinate estimation. (Bottom-right) **Total (weighted) leaderboard**: overall ranking based on a weighted combination of the three tasks.

Although it does not match the best GPT models overall, it clearly belongs to the second tier and represents a different route to strong performance: robust event reasoning without equally strong final-state prediction.

4.4 The Claude family: an internal split

The Claude family's progression is non-monotonic. *Claude-Opus-4.5* and *Claude-Opus-4.6* are strongest on wall-interaction reasoning (*4.6*: 82.50% Task 2, best 5 s score at 89.54%). They retain meaningful boundary-reflection strength even when overall scores are mid-range.

Claude-Opus-4.7 inverts this: strong on Task 3 (77.93% at 1 s, 62.27% at 5 s) but weak on event reasoning (20.01% Task 1, 22.55% Task 2). Like *GPT-5.4*, it is "terminal-state strong, event-level weak." This model also has a small number of missing samples. The Claude line does not show steady improvement; later variants shift the balance between reflection reasoning and terminal-state prediction rather than uniformly improving.

4.5 Mid-range models: Kimi and Doubao

Doubao-Seed-2.0-Pro (48.46%) gains over *Doubao-Seed-1.8* (37.26%) mainly through better Task 2 (67.73% → 78.11%) and Task 3 (21.07% → 49.03%), while Task 1 drops (33.36% → 26.37%). The confusion statistics are consistent with a more conservative prediction style: fewer

false positives but more missed true collisions. Its stronger overall result comes from wall-interaction reasoning and terminal-state prediction, not from collision-event detection.

Kimi-K2.5 (41.95%) occupies a different middle position with a more even profile: 30.04% Task 1, 71.26% Task 2, 33.53% Task 3. It does not match the Task 2/3 performance of *Doubao-Seed-2.0-Pro* but avoids the severe weaknesses of *Doubao-Seed-1.8*, which is limited by poor `final.json` completeness beyond 3 s. Within this group, *Doubao-Seed-2.0-Pro* is the most competitive and *Kimi-K2.5* the most balanced.

4.6 Implications for physical world modeling

Physical reasoning is not a single ability. Different model families show different trade-offs between terminal-state prediction and event reasoning. Strong performance requires both reliable spatial grounding and a mechanism for reasoning about causal interactions over time. Models strong in only one dimension—*GPT-5.4* and *Claude-Opus-4.7* on spatial grounding, *Qwen3.6-Plus* on event reasoning—do not achieve top overall scores even when individual task performance is high.

4.7 Modeling vs. calculation

Some models approximate where objects end up (*GPT-5.4*: 80.86% Task 3 at 1 s) without identifying the collision events that produce that outcome (25.76% Task 1). *Qwen3.6-Plus* shows the reverse: strong on Task 2 across all horizons but weaker on Task 3.

Task 3 can sometimes be approached as terminal-state extrapolation from the observed scene. Task 1 requires identifying state transitions—collisions that redirect trajectories. The strongest models connect event-level reasoning with long-horizon prediction rather than excelling at only one. This gap suggests some models are better at producing spatially plausible outputs than at representing the causal structure of the physical process. Improving this connection is important for physical reasoning in multimodal systems.

5 Conclusion

We introduced `BilliardPhys-Bench`, a benchmark for evaluating physical reasoning in MLLMs through procedurally generated billiard scenarios. The results reveal a consistent gap between recognizing a scene and reasoning about how it will evolve.

GPT-5.5 and **GPT-5.4-Pro** form the leading tier, with **GPT-5.5** achieving the strongest overall score and **GPT-5.4-Pro** the most balanced profile. **Qwen3.6-Plus** forms a strong second tier on event reasoning. **GPT-5.4** and **Claude-Opus-4.7** show that terminal-state prediction alone is insufficient without reliable event detection.

Performance declines as simulation time increases from 1 s to 5 s: small errors in early trajectory estimation accumulate, and longer interaction chains become harder to predict. Current MLLMs still lack a stable model of physical dynamics over time. `BilliardPhys-Bench` provides a direct way to evaluate these limitations and measure progress toward models with stronger physical inductive biases.

Limitations

Our benchmark uses a constant-friction model with perfectly elastic collisions in two dimensions. Real billiards involve spin, non-uniform friction, imperfect elasticity, and 3D dynamics such as jump shots. The simplified setting isolates core reasoning abilities but does not capture all of real-world billiard physics.

All scenarios are billiard-table scenes. Generalization to other physical domains (fluid dynamics, deformable objects, articulated bodies) cannot be assumed from performance on this benchmark alone.

The benchmark evaluates prediction from a single initial-state image. Models receive no video sequences or multiple frames, so the benchmark tests forward simulation from a static input rather than dynamic state estimation.

We did not collect human performance data. A human baseline would help contextualize model results and separate tasks that are inherently difficult from those that expose model-specific weaknesses.

Most evaluated models are accessed via commercial APIs whose architectures, training data, and parameter counts are not publicly disclosed. This limits reproducibility and detailed failure analysis.

All scenes contain exactly seven balls (IDs 0–6) on a standard table. The benchmark does not vary the number of objects, table geometry, or obstacle configurations.

References

Jan Achterhold, Philip Tobuschat, Hao Ma, Dieter Buechler, Michael Muehlebach, and Joerg Stueckler. 2023. Black-box vs. gray-box: A case study on learning table tennis ball trajectory prediction with spin and impacts. *Preprint*, arXiv:2305.15189.

Ivan Alberico, Marco Cannici, Giovanni Cioffi, and Davide Scaramuzza. 2025. Egocentric Event-Based Vision for Ping Pong Ball Trajectory Prediction. *arXiv e-prints*, arXiv:2506.07860.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. Phyre: A new benchmark for physical reasoning. *Preprint*, arXiv:1908.05656.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2024. Llmphy: Complex physical reasoning using large language models and world models. *Preprint*, arXiv:2411.08027.

Hung-Tse Chiang, Bor-Yao Tseng, Jiann-Liang Chen, and Han-Chuan Hsieh. 2024. Trajectory Analysis in UKF: Predicting Table Tennis Ball Flight Parameters. *IT Professional*, 26(3):65–72.

Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. 2025. Theoretical physics benchmark (TPBench)—a dataset and study of AI reasoning capabilities in theoretical physics. *Mach. Learn. Sci. Tech.*, 6(3):030505.

Sebastian Gomez-Gonzalez, Sergey Prokudin, Bernhard Scholkopf, and Jan Peters. 2020. Real time trajectory prediction using deep conditional generative models. *Preprint*, arXiv:1909.03895.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. 2025. WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control. *arXiv e-prints*, arXiv:2512.11047.

E. Kiefl. 2024. Pooltool: A python package for realistic billiards simulation. *Journal of Open Source Software*, 9(101):7301.

Daniel Kienzle, Katja Ludwig, Julian Lorenz, Shin'ichi Satoh, and Rainer Lienhart. 2025. Uplifting table tennis: A robust, real-world application for 3d trajectory and spin estimation. *Preprint*, arXiv:2511.20250.

Yoonpyo Lee, Kazuma Kobayashi, Sai Puppala, Sajedul Talukder, Seid Koric, Souvik Chakraborty, and Syed Bahauddin Alam. 2025. Agentic Physical AI toward a Domain-Specific Foundation Model for Nuclear Reactor Control. *arXiv e-prints*, arXiv:2512.23292.

Xinyi Li, Zaishuo Xia, Weyl Lu, Chenjie Hao, and Yubei Chen. 2025. SmallWorlds: Assessing Dynamics Understanding of World Models in Isolated Environments. *arXiv e-prints*, arXiv:2511.23465.

Jieru Lin, Zhiwei Yu, and Börje F. Karlsson. 2025. SWITCH: Benchmarking Modeling and Handling of Tangible Interfaces in Long-horizon Embodied Scenarios. *arXiv e-prints*, arXiv:2511.17649.

Guiliang Liu, Bo Yue, Yi Jin Kim, and Kui Jia. 2025. Embracing Evolution: A Call for Body-Control Co-Design in Embodied Humanoid Robot. *arXiv e-prints*, arXiv:2510.03081.

LSST Dark Energy Science Collaboration, Eric Aubourg, Camille Avestruz, Matthew R. Becker, Biswajit Biswas, Rahul Biswas, Boris Bolliet, Adam S. Bolton, and etc Bom. 2026. Opportunities in AI/ML for the Rubin LSST Dark Energy Science Collaboration. *arXiv e-prints*, arXiv:2601.14235.

Chak-Wing Mak, Guanyu Zhu, Boyi Zhang, Hongji Li, Xiaowei Chi, Kevin Zhang, Yichen Wu, Yangfan He, Chun-Kai Fan, Wentao Lu, Kuangzhi Ge, Xinyu Fang, Hongyang He, Kuan Lu, Tianxiang Xu, Li Zhang, Yongxin Ni, Youhua Li, and Shanghang Zhang. 2026. PhysicsMind: Sim and Real Mechanics Benchmarking for Physical Reasoning and Prediction in Foundational VLMs and World Models. *arXiv e-prints*, arXiv:2601.16007.

Michael Matthews, Michael Beukman, Chris Lu, and Jakob Foerster. 2025. Kinetix: Investigating the training of general agents through open-ended physics-based control tasks. *Preprint*, arXiv:2410.23208.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, and etc Fenzi. 2025. Cosmos World Foundation Model Platform for Physical AI. *arXiv e-prints*, arXiv:2501.03575.

Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-fei, and Ehsan Adeli. 2025. QuantiPhy: A Quantitative Benchmark Evaluating Physical Reasoning Abilities of Vision-Language Models. *arXiv e-prints*, arXiv:2512.19526.

Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. 2025. Phyx: Does your model have the "wits" for physical reasoning? *Preprint*, arXiv:2505.15929.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. 2026. WorldBench: Disambiguating Physics for Diagnostic Evaluation of World Models. *arXiv e-prints*, arXiv:2601.21282.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, and 1 others. 2025. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models. *arXiv preprint arXiv:2506.17667*.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. 2024. ivideogpt: interactive videogpts are scalable world models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, and 1 others. 2025. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *arXiv preprint arXiv:2505.19099*.

Wanghan Xu, Yuhao Zhou, Yifan Zhou, Qinglong Cao, Shuo Li, Jia Bu, Bo Liu, Yixin Chen, and etc He. 2025a. Probing Scientific General Intelligence of LLMs with Scientist-Aligned Workflows. *arXiv e-prints*, arXiv:2512.16969.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025b. UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models. *arXiv e-prints*, arXiv:2502.00334.

Xinrun Xu, Pi Bu, Ye Wang, Börje F. Karlsson, Ziming Wang, Tengtao Song, Qi Zhu, Jun Song, Zhiming Ding, and Bo Zheng. 2025. Deepphy: Benchmarking agentic vlms on physical reasoning. *Preprint*, arXiv:2508.05405.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv e-prints*, arXiv:2408.01800.

Christine Ye, Sihan Yuan, Suchetha Cooray, Steven Dillmann, Ian L. V. Roque, Dalya Baron, Philipp Frank, Sergio Martin-Alvarez, Nolan Koblischke, Frank J Qu, Diyi Yang, Risa Wechsler, and Ioana Ciuca. 2025. ReplicationBench: Can AI Agents Replicate Astrophysics Research Papers? *arXiv e-prints*, arXiv:2510.24591.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. *arXiv e-prints*, arXiv:2408.04840.

Fangxu Yu, Xingang Guo, Lingzhi Yuan, Haoqiang Kang, Hongyu Zhao, Lianhui Qin, Furong Huang, Bin Hu, and Tianyi Zhou. 2026. TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models. *arXiv e-prints*, arXiv:2601.18744.

Y. Yu. 2025. Fast detection, localization, and trajectory prediction of table tennis based on shuffle yolov5s model. *Discover Applied Sciences*, 7:1362.

Haoqi Yuan, Yu Bai, Yuhui Fu, Bohan Zhou, Yicheng Feng, Xinrun Xu, Yi Zhan, Börje F Karlsson, and Zongqing Lu. 2025. Being-0: A humanoid robotic agent with vision-language models and modular skills. *arXiv preprint arXiv:2503.12533*.

Yibo Zhang, Liang Lin, Kaiwen Luo, Shilinlu Yan, Jin Wang, Yaoqi Guo, Yitian Chen, Yalan Qin, Zhenhong Zhou, Kun Wang, and Li Sun. 2026a. RSA-Bench: Benchmarking Audio Large Models in Real-World Acoustic Scenarios. *arXiv e-prints*, arXiv:2601.10384.

Yiming Zhang, Yingfan Ma, Yanmei Gu, Zhengkai Yang, Yihong Zhuang, Feng Wang, Zenan Huang, Yuanyuan Wang, Chao Huang, Bowen Song, Cheng Lin, and Junbo Zhao. 2025. ABench-Physics: Benchmarking Physical Reasoning in LLMs via High-Difficulty and Dynamic Physics Problems. *arXiv e-prints*, arXiv:2507.04766.

Yuanhe Zhang, Jiayu Tian, Yibo Zhang, Shilinlu Yan, Liang Lin, Zhenhong Zhou, Li Sun, and Sen Su. 2026b. SEE: Signal Embedding Energy for Quantifying Noise Interference in Large Audio Language Models. *arXiv e-prints*, arXiv:2601.07331.

Guosheng Zhao, Yaozeng Wang, Xiaofeng Wang, Zheng Zhu, Tingdong Yu, Guan Huang, Yongchen Zai, Ji Jiao, Changliang Xue, Xiaole Wang, Zhen Yang, Futang Zhu, and Xingang Wang. 2026. UniDriveDreamer: A Single-Stage Multimodal World Model for Autonomous Driving. *arXiv e-prints*, arXiv:2602.02002.

Yongsheng Zhao, Yifeng Zhang, Rong Xiong, and Jianguo Wang. 2015. Optimal state estimation of spinning ping-pong ball using continuous motion model. *IEEE Transactions on Instrumentation and Measurement*, 64(8):2208–2216.

Rong Zhou, Dongping Chen, Zihan Jia, Yao Su, Yixin Liu, Yiwen Lu, Dongwei Shi, Yue Huang, Tianyang Xu, Yi Pan, Xinliang Li, Yohannes Abate, Qingyu Chen, Zhengzhong Tu, Yu Yang, Yu Zhang, Qingsong Wen, Gengchen Mai, Sunyang Fu, and 8 others. 2026. Digital Twin AI: Opportunities and Challenges from Large Language Models to World Models. *arXiv e-prints*, arXiv:2601.01321.

Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Lifan Yuan, Penghao Zhu, Eli Chertkov, Shengyan Liu, and etc Du. 2025. Probing the Critical Point (CritPt) of AI Reasoning: a Frontier Physics Research Benchmark. *arXiv e-prints*, arXiv:2509.26574.