DeepPHY: Benchmarking Agentic VLMs on Physical Reasoning

Xinrun Xu[1,2,3], **Pi Bu**[1], **Ye Wang**[4], **Börje F. Karlsson**[5],
Ziming Wang[1], **Tengtao Song**[1], **Qi Zhu**[1], **Jun Song**[1,*], **Zhiming Ding**[2,*], **Bo Zheng**[1]

[1]Taobao & Tmall Group of Alibaba,
[2]Institute of Software, Chinese Academy of Science,
[3]University of Chinese Academy of Sciences,
[4]Renmin University of China,
[5]Informatics Department, PUC-Rio

Abstract

Although Vision Language Models (VLMs) exhibit strong perceptual abilities and impressive visual reasoning, they struggle with attention to detail and precise action planning in complex, dynamic environments, leading to subpar performance. Real-world tasks typically require complex interactions, advanced spatial reasoning, long-term planning, and continuous strategy refinement, usually necessitating understanding the physics rules of the target scenario. However, evaluating these capabilities in real-world scenarios is often prohibitively expensive. To bridge this gap, we introduce DeepPHY, a novel benchmark framework designed to systematically evaluate VLMs' understanding and reasoning about fundamental physical principles through a series of challenging simulated environments. DeepPHY integrates multiple physical reasoning environments of varying difficulty levels and incorporates fine-grained evaluation metrics. Our evaluation finds that even state-of-the-art VLMs struggle to translate descriptive physical knowledge into precise, predictive control.

Code — https://github.com/XinrunXu/DeepPHY

1 Introduction

Vision Language Models (VLMs) have demonstrated remarkable results in both static visual content understanding tasks (Ye et al. 2024; Yao et al. 2024). Building on this success, a significant research frontier has emerged in applying these models to dynamic, interactive visual environments (including games (Wang et al. 2024), GUI (Qin et al. 2025; Gu et al. 2025), and embodied AI (Zitkovich et al. 2023)). However, prevalent benchmarks and environments exhibit significant limitations in simulating the authenticity and complexity of physical interactions. Game environments (Bellemare et al. 2013; Fan et al. 2022; Tan et al. 2025a) typically offer high-level observation/action space and simplified physics, bypassing the need for low-level physical reasoning. GUI environments (Xie et al. 2024; Rawles et al. 2024) are not grounded in real-world physics, featuring discrete, non-continuous actions. And embodied AI environments (Kolve et al. 2017; Cheng et al. 2024; Nasiriany et al. 2024) focus primarily on semantic-level interactions, usually oversimplifying physical dynamics. This insufficient modeling of complex physical phenomena restricts agents' ability to learn deep causal relationships between actions and longer-term physical consequences. To address this gap, we propose DeepPHY, a benchmark emphasizing agents' need to perceive and understand physical consequences of their actions through sustained interaction.

DeepPHY systematically integrates six challenging physics-based simulation environments: PHYRE (Bakhtin et al. 2019), I-PHYRE (Li et al. 2024), Kinetix (Matthews et al. 2025), Pooltool (Kiefl 2024), and games Angry Birds and Cut the Rope[1]. **None** of which have been previously aggregated for benchmarking agentic VLMs. This integrated collection stands in stark contrast to existing LLM physical reasoning benchmarks that primarily evaluate physical reasoning through static question-answering formats or text-based physics problems (Wang et al. 2025a; Shen et al. 2025; Xiang et al. 2025; Xu et al. 2025). DeepPHY instead immerses agents in interactive sandboxes where success hinges on performing actions and understanding their physical consequences over time. By curating these diverse challenges from different environments, we create the first comprehensive benchmark specifically dedicated to evaluating the interactive physical reasoning capabilities of agentic VLMs.

Through this work, we aim to reveal the boundaries and core shortcomings of current VLMs. Our extensive evaluation across the DeepPHY suite sheds light on their limits in complex physical interaction, long-horizon planning, and dynamic adaptation. The key contributions of this paper are:

- We introduce **DeepPHY**, the first comprehensive benchmark suite to systematically evaluate interactive physical reasoning in agentic VLMs.
- We develop a **unified framework and standardized metrics** that transform diverse physics simulators into a rigorous and accessible testbed. This platform evaluates VLMs and collects interaction data useful for training more physically realistic AI agents.
- We conduct an **extensive empirical study** of leading open- and closed-source VLMs, providing clear baselines and revealing their limitations in physical interaction, planning, and adaptation.

[1]Popular physics-based puzzle games by Rovio Entertainment and ZeptoLabs, respectively.

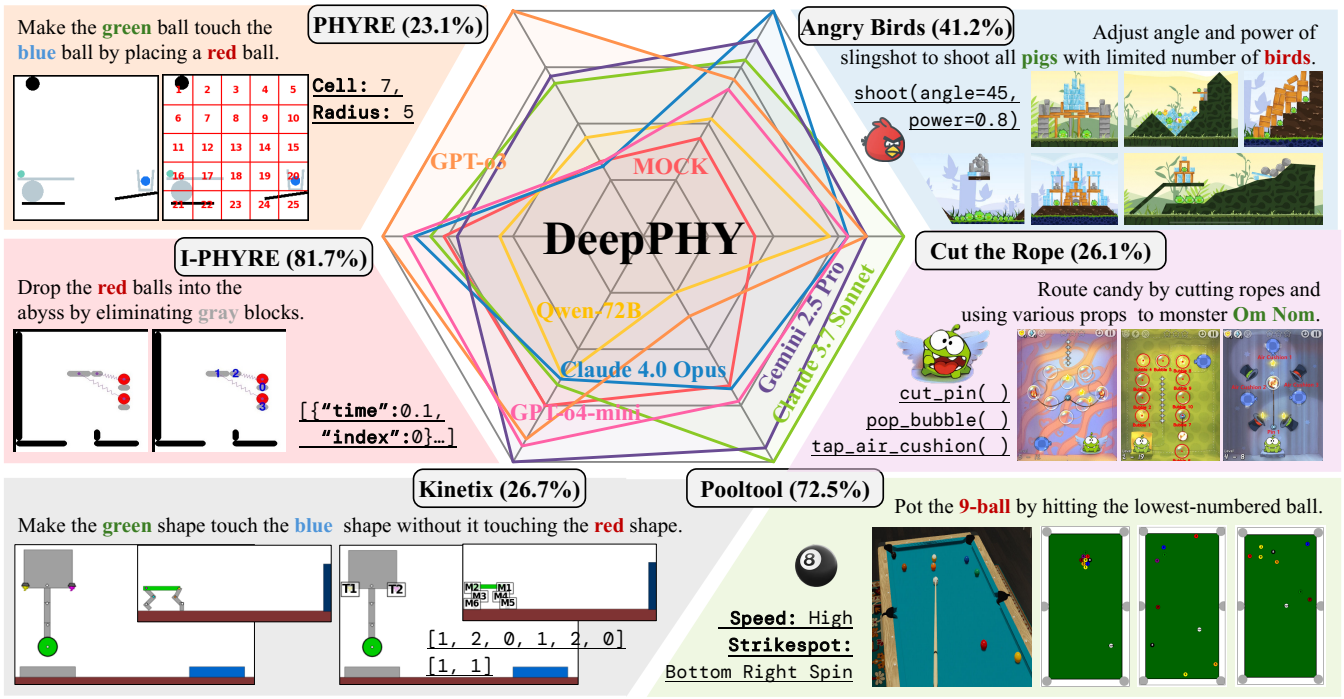

Figure 1: **DeepPHY Benchmark Suite.** Six diverse and challenging environments for evaluating interactive physical reasoning in agentic VLMs. For each environment, a representative task and its goal are displayed. The values in parentheses show the success rate of the best performing VLMs (VLA Prompt Format), and the underlined text provides examples of the structured action spaces. These results indicate that even state-of-the-art models have significant performance gaps to address.

2 Related Work

Physical reasoning capability serves as the cornerstone of world model (Wu et al. 2024; Agarwal et al. 2025) construction and embodied intelligence (Yuan et al. 2025) tasks. However, most evaluations rely on static problem-solving benchmarks. These benchmarks, often presented as large-scale QA about object properties (Wang et al. 2023; Chow et al. 2025) or as text-based physics exams (Wang et al. 2025a; Chung et al. 2025; Zhang et al. 2025), test agents' ability to recall scientific knowledge or deduce logical outcomes from fixed context. While valuable for assessing declarative knowledge, they fundamentally avoid challenges of real-time visual perception and continuous interaction with dynamic worlds. Consequently, such benchmarks are insufficient for holistically evaluating physical intelligence, which necessitates a closed loop of observation, action, and interaction within physics-driven environments.

Another category of research focuses on physical reasoning within simulated environments, but often abstracts away the challenge of perceptual grounding by relying on symbolic inputs. Common approaches provide agents with pre-processed symbolic inputs like object property matrices (Bakhtin et al. 2019; Li et al. 2024; Matthews et al. 2025) or interact with simulators via code generation (Cherian et al. 2024). While these methods are powerful for isolating specific planning, they limit generalizability by bypassing the understanding of raw sensory data. DeepPHY, in contrast, is the first benchmark designed specifically to bridge this gap, evaluating agents' interactive physical reasoning directly from visual input in dynamic settings.

Moreover, research on agents in gaming environments (Xu et al. 2024), while demonstrating impressive capabilities, often operates on a different level of abstraction. Even when visual information is involved, most existing game agents operate in narrative-driven environments — such as GTA (Yang et al. 2024), Escape Room (Wang et al. 2025b), StarCraft II (Shao et al. 2024), Red Dead Redemption II (Tan et al. 2025b), or Civilization VI (Qi et al. 2024) — where progression hinges on scripted storylines or discrete mechanics. This allows agents to learn game-specific rules and heuristics rather than inferring the underlying physical laws governing the world. This focus on game mechanics, rather than fundamental physics, sidesteps the core challenge of building agents that can reason from first-principles based on raw visual observation.

To address these issues, we introduce DeepPHY, an interactive physics simulation benchmark designed to comprehensively evaluate the capabilities of agentic VLMs. By immersing agents in diverse, dynamic, and vision-based environments that require direct interaction, DeepPHY moves beyond static knowledge recall and symbolic reasoning to directly assess an agent's ability to perceive, act, and reason in worlds governed by physical principles.

Category	PHYRE	I-PHYRE	Kinetix	Pooltool	Angry Birds	Cut the Rope
1. Fundamental Physics						
Collision & Stability	☆	✔	✔	☆	☆	✔
Gravity & Friction	✔	✔	✔	☆	✔	✔
Momentum Transfer	✔	✔	☆	☆	✔	✔
2. Dynamics						
Articulated Dynamics	✗	✔	☆	✔	✗	✔
Tension & Oscillation	✗	☆	✗	✗	✗	☆
Rotational Force	✗	✔	☆	☆	✗	✗
3. Action & Control						
Decision Horizon	Single-step	Sequential	Sequential	Sequential	Sequential	Sequential
Planning Strategy	In-advance	In-advance	On-the-fly	On-the-fly	On-the-fly	On-the-fly
Control Complexity	Single	Multi	Multi	Single	Multi	Multi
Timing Criticality	Low	High	Medium	Low	Low	High
4. Reasoning & Strategy						
Causal Chain Reasoning	☆	☆	✔	✔	✔	☆
Tool Use & Affordance	✔	✔	✔	✔	☆	☆
Dynamic Novelty	✗	✗	✔	✗	☆	☆
5. Evaluation Setup						
# Test Instance	1000	40	74	100	34	88
Evaluation Strategy	Env.	Env.	Env.	Env.	Manual	Manual
Max Attempts or Steps	10	10	16	15	# Birds	10

Table 1: Comparative analysis of the DeepPHY benchmark suite environments across five key dimensions. Legend: ☆ - core challenge or implemented with high fidelity; ✔ - present and relevant for solving tasks; and ✗ - not a primary focus or missing.

3 DeepPHY

DeepPHY is a benchmark framework designed to evaluate whether existing VLMs acting as agents possess the ability to understand physical environments and perform sequential action planning tasks. Table 1 provides a comparative analysis of the six environments across key dimensions. This breakdown illustrates how each environment contributes unique challenges to the DeepPHY benchmark, ensuring a comprehensive evaluation of an agent's physical intelligence. In the remainder of this section, we introduce the physics environments employed in the benchmark.

3.1 Problem Formalization

We formalize the physical reasoning challenge in DeepPHY as a trial-based decision process. The underlying environment for any single attempt is a Partially Observable Markov Decision Process (POMDP), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O})$, where $\mathcal{S}$ is the latent physical state, $\mathcal{A}$ is the action space for a single step, and $\mathcal{R}$ is a sparse reward given only for final task success. An agent is allowed up to K attempts to solve a given task, which always starts from the same initial observation o_{initial}. The core challenge is not just to find a solution, but to learn from failed attempts. We model this as follows: Let $k \in \{1, \ldots, K\}$ be the attempt index. A single attempt, or trial, results in a trajectory $\tau^{(k)}$.

- For **in-advance planning** environments (e.g., PHYRE), the agent submits a complete action plan $a^{(k)}$ based on o_{initial} and past history. The resulting trajectory is a tuple capturing the plan and its outcome:

$$\tau^{(k)} = (o^{(k)}, a^{(k)}, o^{(k)}_{\text{final}}, r^{(k)})$$

  where $o^{(k)}_{\text{final}}$ is the final visual observation after the plan is executed, and $r^{(k)}$ is the terminal reward (e.g., 1 for success, 0 for failure).
- For **on-the-fly planning** environments (e.g., Kinetix), the agent interacts sequentially with the environment. The trajectory consists of the full sequence of interactions, culminating in a final reward:

$$\tau^{(k)} = (o^{(k)}_0, a^{(k)}_0, o^{(k)}_1, a^{(k)}_1, \ldots, o^{(k)}_T, r^{(k)})$$

  where $o^{(k)}_0 = o_{\text{initial}}$, and $r^{(k)}$ is the terminal reward received after the final observation $o^{(k)}_T$.

At the beginning of each new attempt k, the agent has access to the history of all previous failed trials, $H^{(k)} = \{\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(k-1)}\}$. To succeed, the agent must use this history to refine its internal world model, f_{phy}, and improve its planning. The policy for generating the plan for the k-th attempt, $\pi^{(k)}$, can be expressed as:

$$\pi^{(k)} = \arg\max_{\pi \in \Pi} V(f_{\text{phy}}(o_{\text{initial}}, H^{(k)}, \pi))$$

where Π is the space of all possible plans (either a single action or a sequence policy). The function $f_{\text{phy}}(\cdot)$ now represents the agent's ability to simulate the outcome of a candidate plan π, conditioned not only on the initial observation

but also on the history of past failures $H^{(k)}$. The value function $V(\cdot)$ estimates the likelihood of success for that simulated outcome. This formulation highlights that DeepPHY fundamentally evaluates an agent's capacity for iterative refinement and learning from failure by updating its predictive model, f_{phy}, across trials.

3.2 Environments

DeepPHY integrates the following environments:

PHYRE (Bakhtin et al. 2019). PHYRE provides a suite of 2D physical reasoning tasks that require agents to achieve specified goals by placing interactive objects in the scene that trigger correct physical chain reactions within limited attempts. Detailed in Appendix B.

I-PHYRE (Li et al. 2024). I-PHYRE is a dynamically evolving interactive physics reasoning benchmark where agents solve puzzles by removing obstacles at precise timings in correct sequences. Detailed in Appendix C.

Kinetix (Matthews et al. 2025). An open-source 2D physics simulation platform designed for agents. It generates vast and diverse physical control tasks, spanning scenarios from robotic locomotion and grasping to classical control problems. Detailed in Appendix D.

Pooltool (Kiefl 2024). A high-fidelity `billiards` simulation benchmark that accurately models multibody collisions, `English spin` effects, and friction-induced trajectory alterations. Detailed in Appendix E.

Angry Birds. A physics-based puzzle game where agents strategically launch `birds` to dismantle complex structures (e.g., wood, stone, ice), aiming to eliminate designated targets through precise force and angle calculations, and trigger large-scale chain-reaction collapses for efficient puzzle resolution. Detailed in Appendix F.

Cut the Rope. A physics-based puzzle game where agents manipulate a dynamic system by cutting ropes and touching other props (e.g., bubbles and cushions) at precise moments to guide candy to a little green monster named *Om Nom*. Detailed in Appendix G.

3.3 Observation Space

DeepPHY's primary aim is to evaluate the physical reasoning of agents, rather than mere perceptual localization. To this end, we refactor and augment each environments' observation space, providing clear annotated-image renderings of interactive objects' locations and identities, thus minimizing the burden of object detection. This design directs agents' to understanding physical dynamics and planning manipulations, enabling a more targeted assessment of their physical reasoning capabilities. This augmentation is realized in distinct ways across the benchmark suite:

Gridded and Labeled Overlays: Visual scenes are optionally overlaid with grids or numerical IDs. In PHYRE, a 5x5 grid is superimposed on the scene to discretize object placement locations. In I-PHYRE, Kinetix and Cut the Rope, interactive elements are annotated with numerical labels.

Dimensionality Reduction: For Pooltool, the native 3D environment is converted into a more VLM-friendly 2D top-down view, simplifying the visual information and making spatial relationships clearer.

Direct Visual Input: In Angry Birds, observation space consists of a raw screenshot of the game—showing structures and pigs without explicit labels—and textual information of available birds.

By providing these modified observations, DeepPHY ensures that the core challenge remains centered on understanding and predicting physical dynamics.

3.4 Action Space

A core challenge for current VLMs not specifically pretrained or fine-tuned for agentic control is their poor performance in generating actions within a continuous space. Specifying precise coordinates, forces, or angles in a free-form text response is often unreliable. To address this, a common principle across all DeepPHY environments is the transformation of continuous or complex action spaces into discrete and structured formats. This conversion is tailored to each environment to preserve its core physical challenges, while making interaction more feasible for VLMs. This strategy is vital as existing games are typically designed for humans, who can provide timely and precise analog feedback. For VLMs, especially in a zero-shot setting, these discrete and structured action spaces make the tasks tractable.

Our approach differs across the suite environments, demonstrating various methods of discretization:

Discretized Parameter Space: In PHYRE, the continuous action of placing a ball at any (x, y) coordinates with any radius is converted into selecting one of 25 grid cells and one of 5 radius levels. Similarly, in Pooltool, the agent chooses from a small, predefined set of named options for shot power (3 selections) and spin type (9 selections), abstracting away the need to specify continuous force and offset values.

Integer-to-Command Mapping: In Kinetix, the agent outputs a simple integer vector. Each integer maps to a specific action for a corresponding motor or thruster. This allows for coordinated control of multiple components through simple, structured output.

Structured Command Language: For games like Angry Birds and Cut the Rope, which involve a variety of interaction types, we have designed a predefined Python code. In Angry Birds, the agent must generate a launch command with angle and power parameters - `shoot(angle, power)`. In Cut the Rope, the agent generates actions as `cut_pin(Index)`, `pop_bubble(Index)`, etc. This provides the agent with a powerful yet constrained set to interact with the environment.

Structured Data Format: In I-PHYRE, the agent outputs a JSON list where each object specifies the index of a block and the precise time of its removal, enabling complex timed sequences of actions.

Table 2 provides a summary of the observation and action space conversions for each environment in the DeepPHY benchmark suite.

4 Evaluation Protocol

In this section, we describe our protocols for evaluating various current state-of-the-art VLMs on DeepPHY.

Environments	Observation Space	Action Space
PHYRE	Initial scene image and identical image overlaid with a 5x5 grid.	**Discretized selection** of grid `Cell` (1-25) and `Radius` level (1-5).
I-PHYRE	Scene image with interactive blocks annotated.	**JSON array** specifying a sequence of block `index` and `time` (in seconds) for each elimination.
Kinetix	High-clarity rendered image, plus an annotated version of controllable motors and thrusters (e.g., 'M0', 'T0').	**JSON array** of integers where each integer corresponds to a discrete action (e.g., 0: *off*, 1: *forward*, 2: *reverse*).
Pooltool	A 3D to 2D top-down rendered view of the pool table.	**Discretized selection** from a predefined list of `Speed` options ("Low", "Medium", "High") and `Strikespot` options (e.g., "Top Spin", "Bottom Left Spin").
Angry Birds	Screenshot of the current game state, including structures, pigs, and available birds.	**Python code** specifying shoot `angle` (0-90) and `power` (0.0-1.0).
Cut the Rope	Annotated screenshot with IDs on all interactive elements (pins, cushions, etc.). Bubbles are labeled in real-time.	**Python code**, predefined action like `cut_pin(id)`, `pop _bubble(id)`, `tap_air_cushion(id, times)`, etc.

Table 2: Summary of observation & action space conversions in the DeepPHY benchmark.

4.1 Evaluation Setting

We aim to keep the evaluation setting simple and consistent across environments. During each timestep of interaction, agents are prompted to output their next action, conditioned on their past interaction history with the environment. To perform successfully in DeepPHY, models must demonstrate robust instruction-following capabilities, including reading visual observation scenes and interpreting environment rules, understanding the action space, inferring physical principles, and producing valid actions to complete tasks effectively.

Planning Strategy. We categorize the environments into two distinct interaction paradigms based on their decision-making process.

- **In-advance Planning**, where an agent is required to devise a complete solution plan from the initial state and output it as a single action (or a full sequence of actions). If the plan fails, the agent receives feedback on the failure and must generate a new complete plan in the next attempt. This setup tests the agent's ability for comprehensive causal reasoning and foresight.
- **On-the-fly Planning**, requiring sequential, turn-by-turn interaction with the environment. At each step, the agent outputs a single action, observes immediate physical consequences, and decides on the next action based on the new state. This mode evaluates the agent's capacity for continuous observation, dynamic adaptation, and reactive control.

Prompt Format. We use two prompting strategies.

- **Vision-Language-Action (VLA)**, where the model receives environment rules, current visual observation, and history of failed attempts, then directly outputs an action.
- **World Model (WM)**, which builds on the VLA prompt, also requiring predicting the environmental changes that will result from the model's chosen action.

Metrics. We use three primary metrics.

- **Success Rate:** The fraction of tasks solved successfully.
- **Pass@K:** The proportion of tasks solved within a maximum of K attempts.
- **Average Attempts:** The mean number of interactive attempts taken to solve a task, averaged over successful trials only.

4.2 Models

We evaluate 17 popular open- and closed-source models, as detailed in Table 7. Open-source models include Qwen2.5-VL-3B/7B/32B/72B-Instruct (Qwen Team 2025). Closed-source models include Claude 3.5/3.7/4.0 Sonnet and Claude 4.0 Opus (Anthropic 2024, 2025); Gemini-2.0-Flash, Gemini-2.5-Pro-06-17, and Gemini-2.5-Flash-06-17 (Mallick and Kilpatrick 2025; Comanici and et al. 2025); and GPT-4-Vision-Preview, GPT-4o-mini-0718, GPT-4o-0806, GPT-o3-0416, and GPT-o4-mini-0416 (Achiam et al. 2023; OpenAI 2024, 2025). Furthermore, a *MOCK* model, which performs random actions, is included as a baseline.

5 Experimental Results

Here, we evaluate all VLMs on DeepPHY, to establish their **zero-shot** baseline performance. We report Success Rate, Pass@K, and Average Attempts over 3 runs under the different physical environments, with a temperature set to 0.1.

5.1 Overall Performance Analysis

The experimental results reveal that current VLMs face significant challenges in interactive physical reasoning tasks. As shown in Tables 3, 4, and 5, most models (especially open-source ones) still cannot surpass *MOCK* results, even with the simplified action spaces designed for VLMs. This indicates a lack of deep understanding of underlying physical principles and for zero-shot planning ability. Among

Model	PHYRE (Success Rate % / Avg. Step)						I-PHYRE (Success Rate % / Avg. Step)					
	VLA			WM			VLA			WM		
	Att. 1 ↑	Att. 10 ↑	Avg. ↓	Att. 1 ↑	Att. 10 ↑	Avg. ↓	Att. 1 ↑	Att. 10 ↑	Avg. ↓	Att. 1 ↑	Att. 10 ↑	Avg. ↓
MOCK	1.12	7.64	5.00	1.12	7.64	5.00	20.50	62.50	3.67	20.50	62.50	3.67
Qwen-7B	0.60	**9.62**	5.10	0.60	6.38	**4.70**	12.50	20.00	**1.75**	**27.50**	**32.50**	2.00
Qwen-72B	1.38	**10.10**	4.48	**1.86**	9.18	**4.12**	12.50	**45.00**	2.00	12.50	29.17	**1.60**
Claude 4S	0.59	2.45	**3.25**	**1.93**	**12.86**	4.88	29.17	**67.50**	2.64	18.33	55.00	**2.08**
Claude 4O	1.13	7.16	4.32	**3.35**	**13.33**	4.32	**33.33**	**71.67**	2.79	14.29	51.26	**2.18**
Gemini 2.5P	1.50	**16.37**	5.02	**1.95**	15.92	**4.53**	25.83	58.33	**2.41**	**27.50**	**67.50**	2.48
Gemini 2.5F	1.06	9.26	4.96	**1.45**	**10.40**	**4.41**	16.67	45.00	**2.44**	**21.67**	**50.83**	2.49
GPT-4o	**1.80**	**11.88**	**4.21**	1.10	8.40	5.14	**22.50**	**53.33**	2.61	15.00	45.83	**2.11**
GPT-o3	**2.87**	**23.06**	4.58	2.70	21.40	**4.53**	**39.17**	**81.67**	2.84	35.83	76.67	**2.79**
GPT-o4m	0.92	7.52	**5.13**	**1.22**	**9.72**	5.20	**34.17**	**75.00**	2.87	25.00	70.83	**2.80**

Table 3: Quantitative results on the **PHYRE** and **I-PHYRE** benchmarks.

Model	Level: Small (S)				Level: Medium (M)				Level: Large (L)			
	VLA		WM		VLA		WM		VLA		WM	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
MOCK	30.00	35.00	30.00	35.00	31.25	40.28	31.25	40.28	10.83	6.67	10.83	6.67
Qwen-7B	30.00	20.00	50.00	20.00	25.00	25.00	22.22	12.50	7.50	5.00	2.50	3.93
Qwen-72B	30.00	32.00	40.00	50.00	32.50	28.33	20.83	28.13	3.61	2.50	3.33	4.38
Claude 4S	56.67	50.00	43.33	42.50	28.47	29.17	29.17	27.78	5.50	5.00	2.50	7.50
Claude 4O	54.29	68.33	30.00	30.00	23.61	34.72	29.17	16.67	3.50	5.00	7.50	2.50
Gemini-2.5P	60.00	53.33	55.00	65.00	39.58	37.50	38.89	34.17	10.62	8.75	10.83	8.33
Gemini-2.5F	47.50	55.00	43.33	57.50	33.33	43.75	37.50	42.71	10.00	8.75	2.50	8.33
GPT-4o	46.00	56.00	36.67	52.50	36.67	27.50	31.94	27.78	5.00	3.00	2.50	2.50
GPT-o3	50.00	63.33	53.33	60.00	36.46	40.28	50.00	36.11	10.00	10.00	7.50	14.17
GPT-o4m	47.50	56.67	56.67	55.00	37.50	43.75	37.50	30.56	11.50	9.17	8.12	4.17

Table 4: Detailed performance breakdown (Success Rate %) on **Kinetix** benchmark across three difficulty levels (Small, Medium, Large). The table compares two prompt formats (VLA vs. WM) and the effect of visual annotations (w/o vs. w/).

all tested models, the latest flagship closed-source models (e.g., GPT-o3, Gemini-2.5-Pro, and Claude 4.0 Opus) generally demonstrate superior performance to `MOCK`, which suggests that model scale, data quality, and more advanced architectures benefit physical reasoning capabilities. However, all models behave differently across environments, and crucially, a stark performance gap exists when compared to humans. Models' success rates fall considerably shorter than desirable, underscoring the long road ahead in bridging the gap to achieve physical intelligence.

5.2 Detailed Analysis by Environment

PHYRE: Even the most advanced models generally achieve Success Rate at 1st attempt below 4% (Tables 3, 9 & 10 in Appendix B). While successes increase with additional attempts, the improvement is slow, suggesting that models struggle to learn effectively from failed attempts and to revise their strategies accordingly (best at only 23.1%).

Model	Att. 15	SuccRate	SuccRate
Mock	48.00%	17.65%	11.36%
Open-Sourced Model			
Qwen-72B	18.00%	29.41%	13.64%
Close-Sourced Model			
Claude 4.0 Opus	49.00%	32.35%	26.14%
Gemini-2.5-Pro	68.00%	35.29%	22.73%
GPT-o3	25.67%	35.29%	18.18%
Human	100%	64.71%	41.36%

Table 5: Model performance summary across Pooltool, Angry Birds, and Cut the Rope.

Model	PHYRE (Pass@5)		I-PHYRE (Pass@3)	
	VLA	WM	VLA	WM
MOCK	22.7	22.7	82.5	82.5
Qwen-7B	**9.7**	6.8	20.0	**32.5**
Qwen-72B	**13.6**	12.9	**50.0**	40.0
Claude 4S	4.5	**24.3**	**82.5**	65.0
Claude 4O	11.2	**25.8**	**77.5**	60.0
Gemini-2.5P	27.0	**27.6**	70.0	**82.5**
Gemini-2.5F	18.5	**20.1**	62.5	**67.5**
GPT-4o	**18.2**	17.4	**60.0**	55.0
GPT-o3	**33.5**	30.2	87.5	87.5
GPT-o4m	15.5	**20.6**	**85.0**	77.5

Table 6: Comparison of Pass@k metrics across different prompt formats. We report **Pass@5** for the PHYRE benchmark and **Pass@3** for the I-PHYRE benchmark. Bold numbers indicate the better performing prompt format (VLA vs. WM) for each model.

I-PHYRE: Leading models such as GPT-o3 achieve a relatively high Success Rate in this setting, reaching 81.67% at Attempt 10 (Tables 3, 13 and 14 in Appendix C), demonstrating that these models are capable of effective temporal planning and causal reasoning. However, open-source models still perform poorly in this setting, resulting in lower success rates than the *MOCK* baseline.

Kinetix: Model success rates decrease significantly as task difficulty increases (Tables 4, 17, 18 in Appendix D). The effect of visual annotation is mixed. In simple S-level tasks, annotations help models identify controllable components and improve performance. However, on more difficult M- and L-level tasks, this benefit vanishes or even harms performance, especially with the WM prompt. This suggests on harder tasks, the extra labels become a cognitive distraction. Furthermore, the WM prompt also fails to help, often lowering success rates. This reveals a VLM's inability to form an accurate world model from visual input.

Pooltool: As shown in Table 5 (for extra details, see Tables 19 & 20), top models, e.g., GPT-4o-mini, achieve high success rates, sometimes even outperforming the human player in average steps. However, a closer look reveals this "efficiency" stems not from strategy, but from a "brute-force heuristic". Models consistently use maximum power to shoot the target ball along the most direct path into a pocket. This strategy works on simple layouts, but ignores the core skill: cue ball control. Our analysis shows that no tested model uses complex physics for planning. For example, they rarely use side spin to alter angles or top/bottom spin to control the cue ball for the next shot. The 100% success rate of GPT-4o-mini is thus misleading, a product of deterministic output and a non-random environment. The model simply repeats the same brute-force solution that happened to work.

Angry Birds & **Cut the Rope**: As shown in Table 5 (for extra details, see Tables 21 & 22), a vast performance gap separates all models from the human player in these games. This exposes a core weakness of current VLMs: they struggle with complex, multi-stage physics tasks that require precise timing. Our design gives clear visual labels to all interactive elements, like slingshots, ropes, and bubbles. This shifts the challenge from low-level perception to high-level reasoning. Although models can see what is interactive, they still cannot create reliable action sequences to reach goals. This highlights their fundamental limitations in spatiotemporal reasoning for dynamic physical processes, with most failures stemming from wrong timing or sequencing. For example, in Cut the Rope, a model might cut a rope too early, stopping the candy from gaining enough momentum. In Angry Birds, models find it hard to plan attack sequences that cause chain reactions, failing to understand how one bird's impact changes the structure for the next attack. These systematic failures show that current models cannot build a coherent predictive internal world model for multi-step decisions in dynamic environments.

5.3 Case Study of Prompt Format

Our comparative analysis of VLA v.s. WM prompt formats (Tables 3, 4, and 6) exposes the intrinsic limitations of current agentic VLMs. Across all environments, we find that the WM approach currently offers only limited benefits, primarily in simpler tasks (e.g., in PHYRE/I-PHYRE). Forcing a prediction might help the model avoid purely random exploration in initial, "zero-history" situations, but this advantage is fragile and quickly diminishes as complexity increases. In fact, WM often becomes a liability in more complex tasks.

We theorize that when the intrinsic difficulty of a physical planning task already pushes a model to its limits, the additional demand of generating an accurate dynamic prediction (the WM task) imposes excessive overhead. Moreover, the world modeling capabilities of current models remain underdeveloped. Revealingly, even if models can generate textually correct predictions (as shown in Figure 7), offering accurate descriptions of the desired physical outcome, they still fail to translate this descriptive knowledge into a precise, executable control signal to realize that outcome. This highlights models' **physical understanding** as largely descriptive, rather than possessing true predictive and procedural control capabilities.

6 Conclusion

In this paper, we introduce DeepPHY, the first comprehensive benchmark to evaluate VLMs' interactive physical reasoning. Our systematic evaluation reveals that even state-of-the-art models struggle with precise, multi-step planning in dynamic environments. We discover a fundamental disconnect between a model's ability to describe physical phenomena and its ability to use that knowledge to predict and control outcomes. We release DeepPHY as a rigorous testbed to benchmark such limitations and facilitate the development of more physically grounded AI agents.

Acknowledgments

This work was supported by the **National Key R&D Program of China** under Grant 2022YFF0503900, and the Research Project of **Institute of Software, Chinese Academy of Sciences** (ISCAS-ZD-202401, ISCAS-ZD-202403).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; and et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; et al. 2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Accessed: 2025-07-01.

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. Accessed: 2025-07-01.

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

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47: 253–279.

Cheng, Z.; Wang, Z.; Hu, J.; Hu, S.; Liu, A.; Tu, Y.; Li, P.; Shi, L.; Liu, Z.; and Sun, M. 2024. LEGENT: Open Platform for Embodied Agents. In Cao, Y.; Feng, Y.; and Xiong, D., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*.

Cherian, A.; Corcodel, R.; Jain, S.; and Romeres, D. 2024. LLMPhy: Complex Physical Reasoning Using Large Language Models and World Models. arXiv:2411.08027.

Chow, W.; Mao, J.; Li, B.; Seita, D.; Guizilini, V.; and Wang, Y. 2025. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. arXiv:2501.16411.

Chung, D. J. H.; Gao, Z.; Kvasiuk, Y.; Li, T.; Münchmeyer, M.; Rudolph, M.; Sala, F.; and Tadepalli, S. C. 2025. Theoretical Physics Benchmark (TPBench) – a Dataset and Study of AI Reasoning Capabilities in Theoretical Physics. arXiv:2502.15815.

Comanici, G.; and et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Fan, L.; Wang, G.; Jiang, Y.; Mandlekar, A.; Yang, Y.; Zhu, H.; Tang, A.; Huang, D.-A.; Zhu, Y.; and Anandkumar, A. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Gu, J.; Ai, Q.; Wang, Y.; Bu, P.; Xing, J.; Zhu, Z.; Jiang, W.; Wang, Z.; Zhao, Y.; Zhang, M.-L.; et al. 2025. Mobile-R1: Towards Interactive Reinforcement Learning for VLM-Based Mobile Agent via Task-Level Rewards. *arXiv preprint arXiv:2506.20332*.

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

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Li, S.; Wu, K.; Zhang, C.; and Zhu, Y. 2024. I-PHYRE: Interactive Physical Reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Mallick, S. B.; and Kilpatrick, L. 2025. Gemini 2.0: Flash, Flash-Lite and Pro. Google for Developers Blog. Accessed: 2024-08-08.

Matthews, M. T.; Beukman, M.; Lu, C.; and Foerster, J. N. 2025. Kinetix: Investigating the Training of General Agents through Open-Ended Physics-Based Control Tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Nasiriany, S.; Maddukuri, A.; Zhang, L.; Parikh, A.; Lo, A.; Joshi, A.; Mandlekar, A.; and Zhu, Y. 2024. RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. arXiv:2406.02523.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2025. OpenAI o3 and o4-mini System Card.

Qi, S.; Chen, S.; Li, Y.; Kong, X.; Wang, J.; Yang, B.; Wong, P.; Zhong, Y.; Zhang, X.; Zhang, Z.; Liu, N.; Wang, W.; Yang, Y.; and Zhu, S.-C. 2024. CivRealm: A Learning and Reasoning Odyssey in Civilization for Decision-Making Agents. arXiv:2401.10568.

Qin, Y.; Ye, Y.; Fang, J.; Wang, H.; Liang, S.; Tian, S.; Zhang, J.; Li, J.; Li, Y.; Huang, S.; et al. 2025. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. *arXiv preprint arXiv:2501.12326*.

Qwen Team. 2025. Qwen2.5-VL.

Rawles, C.; Clinckemaillie, S.; Chang, Y.; Waltz, J.; Lau, G.; Fair, M.; Li, A.; Bishop, W.; Li, W.; Campbell-Ajala, F.; Toyama, D.; Berry, R.; Tyamagundlu, D.; Lillicrap, T.; and Riva, O. 2024. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. arXiv:2405.14573.

Shao, X.; Jiang, W.; Zuo, F.; and Liu, M. 2024. SwarmBrain: Embodied agent for real-time strategy game StarCraft II via large language models. arXiv:2401.17749.

Shen, H.; Wu, T.; Han, Q.; Hsieh, Y.; Wang, J.; Zhang, Y.; Cheng, Y.; Hao, Z.; Ni, Y.; Wang, X.; et al. 2025. PhyX: Does Your Model Have the" Wits" for Physical Reasoning? *arXiv preprint arXiv:2505.15929*.

Tan, W.; Jiang, C.; Duan, Y.; Lei, M.; Li, J.; Hong, Y.; Wang, X.; and An, B. 2025a. StarDojo: Benchmarking Open-Ended Behaviors of Agentic Multimodal LLMs in Production-Living Simulations with Stardew Valley. *arXiv preprint arXiv:2507.07445*.

Tan, W.; Zhang, W.; Xu, X.; Xia, H.; Ding, Z.; Li, B.; Zhou, B.; Yue, J.; Jiang, J.; Li, Y.; An, R.; Qin, M.; Zong, C.; Zheng, L.; Wu, Y.; Chai, X.; Bi, Y.; Xie, T.; Gu, P.; Li, X.; Zhang, C.; Tian, L.; Wang, C.; Wang, X.; Karlsson, B. F.; An, B.; Yan, S.; and Lu, Z. 2025b. Cradle: Empowering Foundation Agents towards General Computer Control. In *Forty-second International Conference on Machine Learning (ICML)*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Trans. Mach. Learn. Res.*, 2024.

Wang, L.; Su, E.; Liu, J.; Li, P.; Xia, P.; Xiao, J.; Zhang, W.; Dai, X.; Chen, X.; Meng, Y.; Ding, M.; Bai, L.; Ouyang, W.; Tang, S.; Wang, A.; and Ma, X. 2025a. PhysUniBench: An Undergraduate-Level Physics Reasoning Benchmark for Multimodal Models. arXiv:2506.17667.

Wang, Y. R.; Duan, J.; Fox, D.; and Srinivasa, S. 2023. NEWTON: Are Large Language Models Capable of Physical Reasoning? arXiv:2310.07018.

Wang, Z.; Dong, Y.; Luo, F.; Ruan, M.; Cheng, Z.; Chen, C.; Li, P.; and Liu, Y. 2025b. EscapeCraft: A 3D Room Escape Environment for Benchmarking Complex Multimodal Reasoning Ability. arXiv:2503.10042.

Wu, J.; Yin, S.; Feng, N.; He, X.; Li, D.; Hao, J.; and Long, M. 2024. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37: 68082–68119.

Xiang, K.; Li, H.; Zhang, T. J.; Huang, Y.; Liu, Z.; Qu, P.; He, J.; Chen, J.; Yuan, Y.-J.; Han, J.; et al. 2025. SeePhys: Does Seeing Help Thinking?–Benchmarking Vision-Based Physics Reasoning. *arXiv preprint arXiv:2505.19099*.

Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T. J.; Cheng, Z.; Shin, D.; Lei, F.; et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37: 52040–52094.

Xu, X.; Wang, Y.; Xu, C.; Ding, Z.; Jiang, J.; Ding, Z.; and Karlsson, B. F. 2024. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*.

Xu, X.; Xu, Q.; Xiao, T.; Chen, T.; Yan, Y.; Zhang, J.; Diao, S.; Yang, C.; and Wang, Y. 2025. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*.

Yang, J.; Dong, Y.; Liu, S.; Li, B.; Wang, Z.; Jiang, C.; Tan, H.; Kang, J.; Zhang, Y.; Zhou, K.; and Liu, Z. 2024. Octopus: Embodied Vision-Language Programmer from Environmental Feedback. arXiv:2310.08588.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. arXiv:2408.04840.

Yuan, H.; Bai, Y.; Fu, Y.; Zhou, B.; Feng, Y.; Xu, X.; Zhan, Y.; Karlsson, B. F.; and Lu, Z. 2025. Being-0: A Humanoid Robotic Agent with Vision-Language Models and Modular Skills. arXiv:2503.12533.

Zhang, Y.; Ma, Y.; Gu, Y.; Yang, Z.; Zhuang, Y.; Wang, F.; Huang, Z.; Wang, Y.; Huang, C.; Song, B.; Lin, C.; and Zhao, J. 2025. ABench-Physics: Benchmarking Physical Reasoning in LLMs via High-Difficulty and Dynamic Physics Problems. arXiv:2507.04766.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2165–2183. PMLR.