Position: The Physics-Physical Reasoning Interplay is Key for Future Embodied World Models

Terry Jingchen Zhang[*2], **Kun Xiang**[*1], **Yinya Huang**[*2,3],
Jixi He[1], **Zirong Liu**[1], **Yueling Tang**[1], **Ruizhe Zhou**[1], **Chengyu Yu**[4],
Xiaodan Liang[1†]

[1]Sun Yat-sen University [2]ETH Zurich [3]ETH AI Center [4]Università della Svizzera italiana

Abstract

World modeling represents a critical frontier towards autonomous AI agents capable of capturing environmental dynamics for intelligent decision-making. While current progress scales on massive data corpora, we argue that future world models require combining physics reasoning (understanding the fundamental laws of nature) with physical reasoning (applying these laws to predict observable behaviors and outcomes of intervention). In this paper, we outline the promise and reality as to how frontier models perform in both physics and physical reasoning, then propose a new pathway for future world models and embodied intelligence to internalize the laws of physics to internalize the mechanism of surrounding physical world just like how humans learn to interact with the world in a universally generalizable manner, thereby establishing the foundation for autonomous, efficient and reliable embodied intelligence.

1 Introduction

The pursuit of artificial general intelligence (AGI) calls for world models that can capture, perceive and internalize the physical world [21], thereby enabling autonomous planning and effective action. While frontier models increasingly excel on scientific reasoning benchmarks with accelerating pace by scaling and leveraging multi-agent structure [72], a critical bottleneck stands in the way as agents powered by language models often showed limited generalizability in understanding highly diverse physical scenarios and take actions accordingly [36]. This gap stems from the current paradigm relying on massive data scaling without grounding models using the laws of physics that govern our physical world [29]. While this paradigm has delivered striking advances in pattern recognition and linguistic reasoning, it fails in scenarios requiring causal understanding rather than memorized correlations. This gap becomes particularly dire when embodied agents must take irreversible physical actions under safety-critical and highly diverse conditions. We argue that future embodied intelligence powered by world models must combine physics reasoning to understand the laws of physics with physical reasoning that applies such laws for predicting surrounding phenomena and take actions accordingly in a safe and effective manner.

2 Physics Reasoning: Internalizing The Laws of Nature

Current Progress and Limitations of AI Physics Reasoning Modern AI systems, especially large language models, have shown a growing ability to apply known physical laws in solving standard textbook-level problems and answering conceptual questions. However, this competence

[*]These authors contributed equally to this work.
[†]Corresponding author. Email: `xdliang328@gmail.com`

is largely confined to low-complexity tasks and masks deeper persistent limitations. For instance, models struggle to maintain logical integrity across extended reasoning chains, with accuracy on benchmarks like PhysReason [73] collapsing when problems require more than a dozen inference steps. Similarly, the integration of visual data with formal reasoning remains a critical weakness; even top-tier models achieve less than 60% accuracy on vision-dependent problems that require interpreting diagrams not just as illustrations, but as sources of essential data [61]. This reveals a deeper challenge to apply implicit physical commonsense, a limitation highlighted by models that generate physically implausible outcomes unless explicitly guided by prompts that describe the correct physical state [41]. At the expert level, this manifests as a failure to correctly model complex scenarios, with models often introducing false assumptions or misapplying advanced principles when confronted with PhD-qualifying exam problems [17]. These widespread challenges suggest that current systems still heavily rely on pattern recognition over robust reasoning.

Physical Laws as a Cure for Generalizability Gap and Data Shortage for Scaling The inherent limitations of current models present a fundamental obstacle to developing robust world models capable of reliable prediction and planning. To address these challenges, a range of advanced methodologies has emerged, which can be broadly categorized into four pillars. The first, tool-augmented reasoning, outsources tasks like precise calculation to deterministic code interpreters, allowing the model to focus on high-level planning [18, 69]. A second pillar is process-supervised rewards, which shifts the training paradigm from rewarding only correct final answers to rewarding correct intermediate reasoning steps, thereby improving the model's logical fidelity [28, 37, 12]. Building on this, the third pillar focuses on internalized self-correction, training the model to spontaneously identify and rectify its own errors within a single generative pass [74, 66, 19]. Finally, for problems of the highest complexity, multi-agent collaboration leverages collective intelligence by assigning different roles—such as generator, verifier, and refiner—to multiple LLM agents that work together to explore and validate diverse solutions [43, 68, 70]. Collectively, these methodologies enhance the efficiency and reliability of scientific reasoning, forming a crucial foundation for more capable and trustworthy AI systems.

Towards Native Physical Understanding with Physics-Native Models Achieving a genuine, physics-grounded understanding requires a paradigm shift from models learning *about* physics to those learning *through* physics, where physical laws form the intrinsic computational fabric. One primary direction involves re-engineering models to embed known physical principles, whether by integrating governing equations into the optimization process (PINNs) [46, 42, 14], designing the core generative mechanism to simulate a physical process (Flow Matching, PFGM/PFGM++, and GANs) [31, 4, 62, 63, 67], or encoding fundamental symmetries directly into the network architecture (Equivariant Networks) [24]. A parallel, more ambitious direction pursues automated scientific discovery, with systems like the "AI Physicist" and AI Feynman using symbolic regression to distill observational data into intelligible physical formulas [60, 53, 52]. The convergence of these two directions may be enabled by novel architectures like Kolmogorov-Arnold Networks (KANs), whose structure is inherently better suited for representing the compositional nature of physical laws, thereby accelerating scientific discovery [34, 32]. Ultimately, the synthesis of these physics-native approaches—embedding known laws while fostering the discovery of new ones—is crucial for developing the next generation of robust world models [8, 33].

3 Physical Reasoning: Understand the Physical World by the Physical Laws

Current Progress and Gaps in Physical Understanding Physical reasoning represents the ability to apply governing principles to understand real-world phenomena and predict outcomes of intervention, bridging mathematical formalism with observable reality [13, 8]. Benchmarks like IntPhys 2 [7] and GRASP [48] show that while multimodal models exhibit basic visual grounding for simple attributes, their performance heavily depends on prompting and falls short in intuitive reasoning, such as whether the AI understands whether pushing an object will cause it to move or fall over. Their performance is at chance level, while human accuracy is near-perfect. Physics-oriented benchmarks like ContPhy [77] and PhysBench [58] demonstrate that even state-of-the-art models struggle to infer latent properties like mass, density, and friction from object dynamics, particularly for soft bodies and fluids. Video generation benchmarks such as VideoPhy [25] and PhyGenBench [16] often find AI generating physically implausible scenarios violating basic governing principles despite visual coherence. More realistically, various benchmarks inspired by video games simulating real-world physical scenarios like PhysGame[11], Phy-Q [64], PHYRE [3], I-PHYRE [50], and PhyBlock [38], systematically demonstrate that models, unlike humans, do not construct a generalizable, intuitive physical model of concepts such as gravity, stability, and causality through active interaction with their environment.

Bridging Symbolic Reasoning and Physical Understanding Sim-to-Real transfer [57, 65, 10] and PINNs [46, 14, 42], by assigning abstract rules learned in idealized physical environments or embed physical laws into AI architectures, have successfully transformed symbolic physics reasoning into physical reasoning for prediction and interaction in complex realities. A typical example is when predicting the flow field around an object, PINN helps the neural network reason with only a few scattered sensor data by injecting the laws of fluid mechanics as a powerful constraint. This physical constraint allows the network to no longer make blind guesses but to reasonably infer the complete physical state of the area where the data is missing. While physics reasoning establishes mathematical foundations, physical reasoning enables systems to apply these governing principles to novel situations, predicting intervention effects and understanding causal structures underlying observable phenomena. Consciousness-inspired designs [6] demonstrate how physical reasoning connects abstract governing principles with embodied experience through probabilistic world models that predict action consequences and infer latent properties. By grounding abstract physical concepts within perception-action loops, physical reasoning allows AI to develop commonsense understanding required for navigation, object manipulation, and adaptation beyond training distributions.

Towards Causal Understanding of the Physical World Physical reasoning fundamentally requires predicting consequences of actions based on understanding underlying physical processes rather than pattern associations. Developing robust reasoning capabilities through benchmarks like NovPhy [45], VideoPhy-2 [5], and Morpheus [71] establishes foundations for systems that reliably predict responses across diverse temporal and spatial contexts. Future AI systems must continuously update their physical understanding through experiential learning that maintain consistency while enabling robust generalization through principled causal mechanisms governed by physical laws rather than pattern matching approaches. These systems should seamlessly combine forward prediction (what will happen) with inverse reasoning (what have caused this outcome) and counterfactual analysis ("what if" scenario, e.g., if gravity were reversed, apples will not fall from the tree but ascend towards the sky), enabling comprehensive understanding of underlying causal structure behind physical interaction that supports safe and effective real-world interaction.

4 World Models: Perceive, Deduce and Internalize

The Promise and Reality of World Models World models calls for perception and internal representations of environmental dynamics to support prediction, planning, and decision-making. Current systems like DreamerV3 [22] and S4WM [15] achieve impressive performance on sequential decision-making problems, demonstrating sophisticated temporal prediction and superior long-term memory across simulated environments. However, comprehensive evaluation against physical benchmarks reveals fundamental limitations where these systems excel at data memorization but fall short when genuine physical understanding becomes necessary. A true world model should be capable of achieving a wider range of tasks, including counterfactual reasoning, grasping novel objects, and adapting to unseen physical constraints. While language-based world models like Dynalang [30] similarly demonstrate impressive linguistic reasoning, they, along with trajectory diffusion approaches like PolyGRAD [47] and video generation models like MineWorld [20], all collapse when confronted with scenarios requiring principled physical understanding rather than pattern matching.

Physical Principles Enable More Robust Understanding to Unseen Scenarios Grounding world models in physical principles establishes computational systems that understand why environmental dynamics occur rather than merely memorizing past occurrences or performing pattern matching on massive datasets, as demonstrated by models like OccWorld [76] and GAIA-1 [26]. This approach is essential because physical environments diversify heavily beyond specific training scenarios and foundational structure is key towards robust generalization. Neural physics approaches [39] demonstrate how incorporating physical understanding enables world models like FusionForce [1] to maintain consistency across different scales, materials, and environmental conditions. The synthesis of physics reasoning and physical reasoning within world models creates systems capable of bridging the gap between symbolic knowledge and embodied experience that characterizes human-level environmental understanding.

Towards Physical World Models Through Neural Physics The pathway toward physics-grounded world models involves combining physical computation with neural learning. This approach builds on neural physics frameworks [39] and specific implementations like MoSim, a neural motion simulator [23], SAIN, a hybrid physical-neural dynamics model [2], and PhysORD, a neuro-symbolic approach for physics-infused prediction [75], enabling world models to develop intuitive physical understanding through embodied interaction while maintaining end-to-end learning capabilities. Future world models must demonstrate robust sim-to-real transfer by grounding representations in physical law, establishing computational foundations that naturally generalize across different contexts.

5 Embodied Intelligence: Observe, Predict and Act

Current Progress and Critical Limitations in Embodied AI Embodied intelligence refers to AI systems that possess physical bodies and can interact with the world through sensorimotor experiences, learning from the consequences of their actions in real physical environments. Recent advances highlight growing capabilities in navigation and manipulation tasks. Contemporary benchmarking efforts, such as BEHAVIOR [51], Meta-World [40], and RoboTwin [44], have established standardized environments for evaluating embodied agents across diverse manipulation and navigation tasks. Complementary to these, simulation and asset-generation tools like RoboScape [49] and EmbodiedGen [56] enhance realism by supporting video-based dynamics learning, depth prediction, and scalable creation of physically plausible 3D assets. Together, these advances demonstrate measurable progress in reducing sim-to-real gaps through improved physical simulation and world model grounding [35, 59]. However, current embodied agents still struggle with persistent sim-to-real gaps when transferred to novel environments or unseen object configurations, and often rely on a narrow set of primitive actions rather than rich, compositional behaviors. As a result, these agents remain confined to narrow operational contexts and fall short when confronted with physical scenarios that demand principled reasoning about mass, friction, elasticity, and other fundamental properties.

Physics-Physical Reasoning for Consequence-Aware Real-World Actions Embodied intelligence represents the ultimate test for physics-grounded world models, where agents must make irreversible physical actions with potential safety implications in dynamic, unpredictable environments. Beyond perception and instruction following, recent Vision–Language–Action (VLA) models integrate multimodal understanding with action generation [9, 27], yet benchmarks such as PhysBench [58] reveal that state-of-the-art systems still struggle with physical reasoning. Physics–physical reasoning addresses this limitation by enabling agents to systematically evaluate potential outcomes, and select behaviors whose effects are both desirable and safe. This synthesis proves essential because agents must not only comprehend governing mathematical principles but also predict how they manifest in observable behaviors across diverse materials and environmental conditions. Recent research emphasizes combining high-fidelity physical simulators with learned world models for grounding abstract cognitive functions in real-world interactions [35].

Towards Embodied Intelligence Powered by Physics-Empowered World Models Embodied intelligence relies on world models to predict the consequences of actions and guide decision-making in complex physical environments. Current approaches, such as Dreamer [54] and GENIE [55], demonstrate complementary strengths: Dreamer excels at learning latent dynamics for planning and policy optimization, while GENIE emphasizes causal and physically informed reasoning for robust adaptation. Future work should integrate these paradigms with explicit physics-grounded priors, combining forward and inverse dynamics, counterfactual reasoning, and hierarchical multi-scale physical models. Such synthesis would enable embodied agents to anticipate outcomes across temporal and spatial scales, refine their physical understanding through experience, and execute actions that are both effective and safe.

6 Conclusion

Our vision encompasses a paradigm shift from current pattern-matching approaches to physics-grounded world models embodying genuine physical understanding , ultimately enabling robust, adaptive, and safe autonomous agents representing the true promise of artificial intelligence in physical environments.

References

[1] Ruslan Agishev and Karel Zimmermann. Fusionforce: End-to-end differentiable neural-symbolic layer for trajectory prediction. *arXiv preprint arXiv:2502.10156*, 2025.

[2] Anurag Ajay, Maria Bauza, Jiajun Wu, Nima Fazeli, Joshua B Tenenbaum, Alberto Rodriguez, and Leslie P Kaelbling. Combining physical simulators and object-based networks for control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3217–3223. IEEE, 2019.

[3] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross B. Girshick. PHYRE: A new benchmark for physical reasoning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5083–5094, 2019.

[4] Giacomo Baldan, Qiang Liu, Alberto Guardone, and Nils Thuerey. Flow matching meets pdes: A unified framework for physics-constrained generation, 2025.

[5] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025.

[6] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[7] Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

[8] Oualid Bougzime, Samir Jabbar, Christophe Cruz, and Frédéric Demoly. Unlocking the potential of generative ai through neuro-symbolic architectures: Benefits and limitations, 2025.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

[10] Arunkumar Byravan, Jan Humplik, Leonard Hasenclever, Arthur Brussee, Francesco Nori, Tuomas Haarnoja, Ben Moran, Steven Bohez, Fereshteh Sadeghi, Bojan Vujatovic, and Nicolas Heess. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields, 2022.

[11] Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. Physgame: Uncovering physical commonsense violations in gameplay videos, 2024.

[12] Wenxiang Chen, Xiangyun Xiao, Dihong Gong, Tian-Yuan He, Ling-Hao Chen, Yizhou Wang, Zhifang Sui, and Zixuan Li. Better process supervision with bi-directional rewarding signals, 2025.

[13] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models, 2024.

[14] Junwoo Cho, Seungtae Nam, Hyunmo Yang, Seok-Bae Yun, Youngjoon Hong, and Eunbyung Park. Separable physics-informed neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23761–23788. Curran Associates, Inc., 2023.

[15] Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: Rnns, transformers, and s4. *Advances in Neural Information Processing Systems*, 36:72904–72930, 2023.

[16] Meng Fanqing, Liao Jiaqi, Tan Xinyu, Shao Wenqi, Lu Quanfeng, Zhang Kaipeng, Cheng Yu, Li Dianqi, Qiao Yu, and Luo Ping. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

[17] Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11717–11743. Association for Computational Linguistics, 2025.

[18] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Graham Neubig, and Karthik Narasimhan. Pal: Program-aided language models, 2023.

[19] Yong-Xiang Gao, Hong-Yu Lin, and Xian-He Sun. Embedding self-correction as an inherent ability in large language models for enhanced mathematical reasoning, 2025.

[20] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

[21] David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018.

[22] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[23] Chenjie Hao, Weyl Lu, Yifan Xu, and Yubei Chen. Neural motion simulator pushing the limit of world models in reinforcement learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27608–27617, 2025.

[24] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d, 2022.

[25] Bansal Hritik, Lin Zongyu, Xie Tianyi, Zong Zeshun, Yarom Michal, Bitton Yonatan, Jiang Chenfanfu, Sun Yizhou, Chang Kai-Wei, and Grover Aditya. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

[26] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[27] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024.

[28] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[29] Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

[30] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. *arXiv preprint arXiv:2308.01399*, 2023.

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022.

[32] Ziming Liu, Pingchuan Ma, Yixuan Wang, Wojciech Matusik, and Max Tegmark. Kan 2.0: Kolmogorov-arnold networks meet science, 2024.

[33] Ziming Liu, Eric J Michaud, and Max Tegmark. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *arXiv preprint arXiv:2305.08746*, 2023.

[34] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks, 2025.

[35] Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, Wei Li, Wei Yin, Yao Yao, Jia Pan, Qiu Shen, Ruigang Yang, Xun Cao, and Qionghai Dai. A survey: Learning embodied intelligence from physical simulators and world models, 2025.

[36] Sirui Lu, Zhijing Jin, Terry Jingchen Zhang, Pavel Kos, J. Ignacio Cirac, and Bernhard Schölkopf. Can theoretical physics research benefit from language agents?, 2025.

[37] Liangchen Luo, Peiyi Wang, Jiaqi Chen, Yuxuan Liu, Yeye Ruan, Han An, Zichao Yang, Chen-Yu Lee, Huisheng Wang, Ru-Yuan Zhang, Jing-Jun Liu, and Tomas Pfister. Improve mathematical reasoning in language models by automated process supervision, 2024.

[38] Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, Mingfei Han, Meng Cao, Bokui Chen, Ivan Laptev, and Xiaodan Liang. Phyblock: A progressive benchmark for physical understanding and planning via 3d block assembly, 2025.

[39] Pingchuan Ma. *Building World Models with Neural Physics*. Phd thesis, Massachusetts Institute of Technology, 2025.

[40] Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, K. R. Zentner, Ryan Julian, J K Terry, Isaac Woungang, Nariman Farsad, and Pablo Samuel Castro. Meta-world+: An improved, standardized, rl benchmark, 2025.

[41] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024.

[42] Sarvin Moradi, Burak Duran, Saeed Eftekhar Azam, and Massood Mofid. Novel physics-informed artificial neural network architectures for system and input identification of structural dynamics pdes. *Buildings*, 13(3), 2023.

[43] Sumeet Motwani, Megha Srivastava, Nikolaos Pappas, and Xifeng Yan. Malt: Improving reasoning with multi-agent llm training, 2025.

[44] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version), 2025.

[45] Vimukthini Pinto, Chathura Gamage, Cheng Xue, Peng Zhang, Ekaterina Nikonova, Matthew Stephenson, and Jochen Renz. Novphy: A physical reasoning benchmark for open-world ai systems. *Artificial Intelligence*, 335:104179, 2024.

[46] Zhiyuan Ren, Shijie Zhou, Dong Liu, and Qihe Liu. Physics-informed neural networks: A review of methodological evolution, theoretical foundations, and interdisciplinary frontiers toward next-generation scientific computing. *Applied Sciences*, 15(14), 2025.

[47] Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion. *arXiv preprint arXiv:2312.08533*, 2023.

[48] Jassim Serwan, Holubar Mario, Richter Annika, Wolff Cornelius, Ohmer Xenia, and Bruni Elia. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023.

[49] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model, 2025.

[50] Li Shiqian, Wu Kewen, Zhang Chi, and Zhu Yixin. I-phyre: Interactive physical reasoning. *arXiv preprint arXiv:2312.03009*, 2023.

[51] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, C. Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments, 2021.

[52] Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, and Max Tegmark. Ai feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In *Advances in Neural Information Processing Systems*, volume 33, pages 4040–4054, 2020.

[53] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

[54] Boyuan Wang, Xinpan Meng, Xiaofeng Wang, Zheng Zhu, Angen Ye, Yang Wang, Zhiqin Yang, Chaojun Ni, Guan Huang, and Xingang Wang. Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling, 2025.

[55] Jiaming Wang, Diwen Liu, Jizhuo Chen, Jiaxuan Da, Nuowen Qian, Tram Minh Man, and Harold Soh. Genie: A generalizable navigation system for in-the-wild environments, 2025.

[56] Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. Embodiedgen: Towards a generative 3d world engine for embodied intelligence, 2025.

[57] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Sim-to-real transfer via 3d feature fields for vision-and-language navigation, 2024.

[58] Chow Wei, Mao Jiageng, Li Boyi, Seita Daniel, Guizilini Vitor, and Wang Yue. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

[59] Lik Hang Kenny Wong, Xueyang Kang, Kaixin Bai, and Jianwei Zhang. A survey of robotic navigation and manipulation with physics simulators in the era of embodied ai, 2025.

[60] Tailin Wu and Max Tegmark. Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, 100(3):033311, 2019.

[61] Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. Seephys: Does seeing help thinking? - benchmarking vision-based physics reasoning. *CoRR*, abs/2505.19099, 2025.

[62] Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. *Advances in Neural Information Processing Systems*, 35:16782–16795, 2022.

[63] Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. Pfgm++: Unlocking the potential of physics-inspired generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 38566–38591. PMLR, 2023.

[64] Cheng Xue, Vimukthini Pinto, Chathura Nagoda Gamage, Ekaterina Nikonova, Peng Zhang, and Jochen Renz. Phy-q as a measure for physical reasoning intelligence. *Nat. Mac. Intell.*, 5(1):83–93, 2023.

[65] Mengyuan Yan, Iuri Frosio, Stephen Tyree, and Jan Kautz. Sim-to-real transfer of accurate grasping with eye-in-hand observations and continuous control, 2017.

[66] Zhen-Fu Yan, Yi-Chen Zhang, Dihong Gong, and Yuan-Hai-Tao. Spontaneous step-level self-correction makes large language models better mathematical reasoners, 2025.

[67] Liu Yang, Dongkun Zhang, and George Em Karniadakis. Physics-informed generative adversarial networks for stochastic differential equations. *SIAM Journal on Scientific Computing*, 42(1):A292–A317, 2020.

[68] Yifei Yang, Zian Tang, Chi-Min Chan, and Hong-Ning Wang. Multi-llm collaborative search for complex problem solving, 2025.

[69] Bohan Yao and Vikas Yadav. A toolbox, not a hammer – multi-tag: Scaling math reasoning with multi-tool aggregation, 2025.

[70] Rui Yuan and Yijie Xie. Reinforce llm reasoning through multi-agent reflection, 2025.

[71] Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments, 2025.

[72] Terry Jingchen Zhang, Yongjin Yang, Yinya Huang, Sirui Lu, Bernhard Schölkopf, and Zhijing Jin. Collective intelligence: On the promise and reality of multi-agent systems for ai-driven scientific discovery. *Preprints*, August 2025.

[73] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16593–16615. Association for Computational Linguistics, 2025.

[74] Zhenghao Zhao, Zhaoye Fei, Ansong Ni, Linyi Li, Hai-Tao Zheng, Jian-Guang Lou, and Tat-Seng Chua. Boosting llm reasoning via spontaneous self-correction, 2025.

[75] Zhipeng Zhao, Bowen Li, Yi Du, Taimeng Fu, and Chen Wang. Physord: a neuro-symbolic approach for physics-infused motion prediction in off-road driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11670–11677. IEEE, 2024.

[76] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024.

[77] Zheng Zhicheng, Yan Xin, Chen Zhenfang, Wang Jingzhou, Zhi Qin, Joshua Tenenbaum, and Gan Chuang. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024.