Opinion: How Can Causal AI Benefit World Models?

Qiuling Pan[1] **Hong Zhou**[1,*] **Zhouchen Lin**[2,3,4,*]
[1]Huawei Technologies Co., Ltd.
[2]State Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[3]Institute for Artificial Intelligence, Peking University
[4]Pazhou Laboratory (Huangpu), Guangzhou, Guangdong, China
panqiuling@huawei.com, zh.zhouhong@huawei.com, zlin@pku.edu.cn

Abstract

World models are regarded as a key pathway toward achieving general artificial intelligence, yet current modeling approaches suffer from correlation limitations that hinder their ability to capture the intrinsic causal mechanisms. This deficiency results in significant shortcomings in out-of-distribution generalization capabilities, sample efficiency, and deep reasoning abilities of world models. This paper argues that integrating principles from causal science is essential for overcoming these challenges and constructing world models aligned with core objectives. We systematically propose a framework where the three pillars of causal science address these shortcomings. Ultimately, we contend that the shift from a correlation-driven paradigm to a causality-driven paradigm represents not merely a technical refinement, but a necessary leap toward constructing agents that genuinely understand and interact with the real world.

1 Introduction

Researchers currently hold differing views on the definition of world models. Both the compressed representations of environments in the reinforcement learning (RL) community and large-scale generative models trained on massive real-world physical data are considered a type of world model [1–4]. No matter what the definition is, almost all researchers agree that the core functionality of world models must encompass modeling the dynamic changes, causal relationships, and spatiotemporal structures of the physical world [5–7]. There are multiple types of relationships in the real world. Current deep learning excels at handling statistical and correlational relationships, but to truly engage with the real world, it must also master formal, causal, and stochastic relationships. Therefore, beyond connectionism, we should also incorporate symbolism and evolutionism, forming a multi-engine driven world model. As the Turing Award laureate Yann LeCun believes, world models represent one of the key pathways toward achieving general artificial intelligence (AGI) [8].

Current world models have achieved tremendous success in complex 3D simulations [9, 10], gaming [11], and continuous control for robotics [12, 13]. However, a fundamental limitation of mainstream world models lies in their focus on learning correlations rather than causal relationships between variables [14, 15]. This results in current world models failing to meet their core capability requirements, indicated by: (a) Poor generalization: Vulnerable in environments out of the training distribution, easily disrupted by changes in superficial features. (b) Low sample efficiency: Requires massive data to learn relationships among numerous variables in the environment, struggling to extract essential patterns from limited local interactions. (c) Lack of deep understanding: Remains confined to pattern matching, incapable of genuine reflection or imagination beyond observed data.

*Hong Zhou and Zhouchen Lin are the corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Embodied World Models for Decision Making.

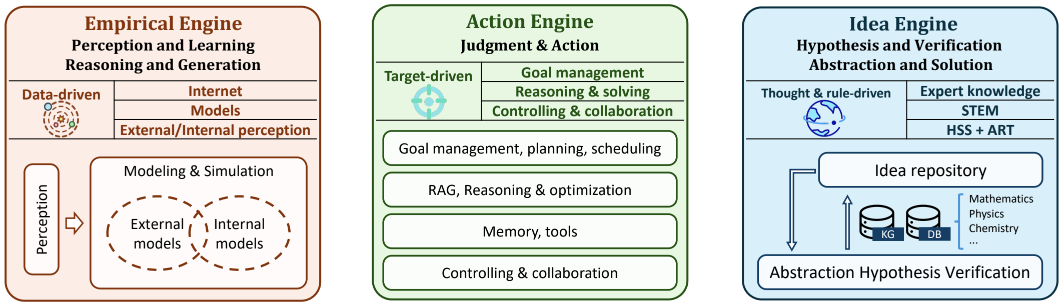

Figure 1: A conceptual framework for an advanced AI designed to build a world model.

As Figure 1 shows, to develop a better world model system, we argue that previous efforts and future directions include three engines. Firstly, the *empirical engine* is the foundation for the perception and learning of the system, covering the most advanced models and techniques such as LLM, MoE, SFT and continual RL. Secondly, the *action engine* is the decision and execution center of the system, requiring goal management, reasoning and collaboration abilities. Lastly, the *idea engine* is the highest cognitive layer of the system, responsible for abstract thinking and innovation.

To achieve these goals, causal science aiming to uncover genuine causal relationships between variables can make contributions [16, 17]. Its core lies in defining, identifying, and quantifying causality. Within the potential outcome framework [18] and structural causal model framework [19], causal science has established three foundational tasks: causal invariance [20, 21], causal discovery [22, 23], and causal inference [24, 25]. Each addresses the limitations of correlation from distinct perspectives. Previous applications in deep learning demonstrate that causal science can help build more robust [26, 27], efficient [28], interpretable [29, 30], and generalizable AI models [31, 32].

As Figure 1 illustrates, this position paper argues that integrating techniques from causal invariance, causal discovery, and causal inference into world model learning is a key pathway to overcoming the aforementioned limitations. In the following sections, we will first explore how the principle of causal invariance necessitates a more active world model learning paradigms. Subsequently, this paper will elaborate on how leveraging the concept of causal discovery enables the construction of an efficient, more fine-grained dynamic model, by treating an agent's actions as causal interventions. Finally, we will further demonstrate that endowing world models with causal inference capabilities, particularly treatment effect estimation and counterfactual reasoning, elevates them from mere "observation" to higher levels of "reflection" and "imagination", enabling true world modeling.

2 Causal Invariance Call for Active Environment Exploration

2.1 Vulnerability of Associative Models: When Surface Features Change

The vulnerability of current world models originates from their nature as associative models, trained to minimize prediction errors on a given dataset. For instance, a world model might erroneously associate a specific floor texture with the property of being slippery, rather than learning the more fundamental physical principle of the friction coefficient. An associative model is defined as a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ learning via empirical risk minimization [33], where $\mathcal{X}$ is the input space (containing surface features like color, texture, lighting), $\mathcal{Y}$ is the output space, ℓ is a loss function, and $\{(x_i, y_i)\}_{i=1}^n$ is training data from a source environment $\mathcal{E}_s$.

When $\mathcal{E}_s$ shifts to a target $\mathcal{E}_t$ with altered surface features, the joint distribution $P_{\mathcal{E}_t}(x, y) \neq P_{\mathcal{E}_s}(x, y)$. For an associative model, the risk on $\mathcal{E}_t$ is: $\mathcal{R}_{\mathcal{E}_t}(f) = \mathbb{E}_{(x,y)\sim P_{\mathcal{E}_t}} [\ell(f(x), y)]$, and due to $P_{\mathcal{E}_t}(x, y)$ divergence from $P_{\mathcal{E}_s}(x, y)$, $\mathcal{R}_{\mathcal{E}_t}(f) \gg \mathcal{R}_{\mathcal{E}_s}(f)$, showcasing vulnerability.

2.2 Causal Invariance: Learning the Laws of Physics

Define a causal dynamic model as $M : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}'$, where $\mathcal{S}$ is the state space (encoding intrinsic physical properties), $\mathcal{A}$ is the action space, and $\mathcal{S}'$ is the next-state space. The model satisfies causal invariance if for any two environments $\mathcal{E}_1, \mathcal{E}_2$ with different surface feature distributions

$P_{\mathcal{E}_1}(x) \neq P_{\mathcal{E}_2}(x)$ (where x is a surface-feature-augmented state $x = (s, c)$, $c \in \mathcal{C}$ for surface features like color), the conditional distribution of next state given state and action is invariant:

$$P_{\mathcal{E}_1}(s' \mid s, a) = P_{\mathcal{E}_2}(s' \mid s, a) = P(s' \mid s, a).$$

Training such a model minimizes a risk over a collection of environments $\{\mathcal{E}_k\}_{k=1}^K$:

$$M = \arg\min_{m \in \mathcal{M}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{(s,a,s') \sim P_{\mathcal{E}_k}} \left[d\left(m(s,a), s'\right)\right],$$

where d is a distance metric (e.g., mean squared error). This enforces M to capture invariant physical laws $P(s' \mid s, a)$, stable across $\{\mathcal{E}_k\}_{k=1}^K$.

2.3 From Passive Data Collection to Active "Natural Experiments"

Let $\mathcal{D}_{\text{passive}} = \{(s_i, a_i, s'_i)\}_{i=1}^N$ be data collected passively from a single environment. In contrast, active data collection involves a model M selecting environments $\mathcal{E} \sim \pi(\mathcal{E} \mid M)$ (where π is a policy for environment selection) to maximize the invariance-aware information gain:

$$\mathcal{I}(M) = \mathbb{E}_{\mathcal{E} \sim \pi(\mathcal{E}|M)} \left[\mathcal{D}_{\text{KL}}\left(P(s' \mid s, a, \mathcal{E}) \parallel P(s' \mid s, a)\right)\right],$$

where $\mathcal{D}_{\text{KL}}$ is the Kullback-Leibler divergence. The goal is to solve:

$$\pi^* = \arg\max_{\pi} \mathcal{I}(M_\pi),$$

where M_π is trained on data from environments selected by π. This turns data collection into an active search for "natural experiments" (environments $\mathcal{E}$) that best reveal invariant causal structures.

3 Causal Discovery for More Fine-grained Dynamic Modeling

Beyond learning robust dynamics, an effective world model must also be efficient and adaptable. Monolithic models that attempt to represent the entire world's dynamics with a single, massive function are notoriously sample-hungry and difficult to adapt when the environment changes. Causal discovery offers a path to a more fine-grained, modular, and efficient representation by treating an agent's actions as targeted experiments that reveal the world's sparse causal structure.

3.1 Actions of an Agent are Causal Interventions

Let a causal graph be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ is the set of variables (representing environmental states and action-related factors) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of directed edges (causal relationships). An agent's action $A \in \mathcal{A}$ is a causal intervention, denoted as $do(A = a)$, which modifies the causal graph by fixing A to value a. The post-intervention distribution $P_{\text{post}}(V \mid do(A = a))$ is related to the pre-intervention distribution $P_{\text{pre}}(V)$ via the causal mechanism:

$$P_{\text{post}}(V \mid do(A = a)) = \prod_{V_i \in \mathcal{V} \setminus \{A\}} P(V_i \mid \text{Pa}(V_i)),$$

where $\text{Pa}(V_i)$ is the set of parent nodes of V_i in $\mathcal{G}$. Observing the changes in variables, the agent infers their causal links between variables by analyzing conditional independences under interventions.

3.2 The Architecture of the Causal World Model (CWM)

A CWM should consist of two components:

Encoder: A function $E : \mathcal{O} \to \mathcal{Z}$, where $\mathcal{O}$ is the high-dimensional observation space (e.g., pixel space) and $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_m\}$ is the space of decoupled, object-centered latent variables. Formally, $Z = E(O)$, and the latent variables satisfy $P(Z_i \mid Z_{\neg i}) = \prod_{j=1}^m P(Z_j)$ (statistical independence, indicating decoupling), where $Z_{\neg i}$ is $\mathcal{Z}$ without Z_i.

Causal Dynamics Model: A causal graph $\mathcal{G}_d = (\mathcal{Z} \cup \mathcal{A}, \mathcal{E}_d)$ defining dynamics. The next-state of each latent variable Z'_i is a function only of its direct causal parents $\text{Pa}(Z_i)$ in $\mathcal{G}_d$ and action A:

$$Z'_i = f_i(\text{Pa}(Z_i), A),$$

where f_i is the local causal function for Z_i, in contrast to a global state transition function $S' = F(S, A)$ (where S is a monolithic global state).

3.3 Advantages of the Factorized Architecture

Sample Efficiency: Let the parameter complexity of a monolithic model be $\Theta(D^k)$ for D-dimensional state space and degree k, while for the factorized CWM, each local function f_i has parameter complexity $\Theta(d_i^l)$ with $d_i \ll D$ (dimension of $\text{Pa}(Z_i)$) and l a small degree. The total parameter complexity is $\sum_{i=1}^{m} \Theta(d_i^l) \ll \Theta(D^k)$. The sample complexity $\mathcal{S}$ scales with parameter complexity, so $\mathcal{S}_{\text{CWM}} \ll \mathcal{S}_{\text{monolithic}}$, meaning fewer samples suffice for learning.

Generalization and Transfer: Suppose a local mechanism f_k changes (due to environmental shift). The loss function for re-learning is $\mathcal{L}_{\text{local}} = \mathbb{E}\left[d(Z'_k, f'_k(\text{Pa}(Z_k), A))\right]$, focusing only on Z_k. Other latent variables $Z_j (j \neq k)$ use unchanged functions f_j, so their causal knowledge $P(Z'_j \mid \text{Pa}(Z_j), A) = f_j(\text{Pa}(Z_j), A)$ remains valid and can be transferred to new tasks, as new tasks still rely on these invariant local causal mechanisms.

4 Causal Inference for Reflection and Imagination

4.1 Beyond Prediction: Climbing the Causal Ladder

Pearl [19] defines the causal ladder with three levels:

Level 1 (Association): Models $P(Y \mid X)$, capturing statistical correlations.

Level 2 (Intervention): Models $P(Y \mid do(X = x))$, describing effects of actions.

Level 3 (Counterfactual): Models $P(Y \mid do(X = x), \text{obs}(X = x', Y = y'))$, reasoning about "what if" scenarios.

The causal ladder provides a powerful framework for understanding different levels of reasoning. True intelligence requires M_{causal} operating on Level 2 and 3, unlike M_{pred} stuck at Level 1.

4.2 Second-level: Precise Planning Based on Intervention

A world model trained only for prediction learns M_{wrong} such that $M_{\text{wrong}}(X) = \mathbb{E}[Y \mid X]$, a wrong mathematical object as it fails to capture $P(Y \mid do(X = x))$.

For causal planning, let π be a policy (action sequence $A_1, A_2, \dots, A_T$). The causal model computes $P(S_T \mid do(A_1), do(A_2), \dots, do(A_T))$ via the causal dynamics:

$$P(S_t \mid do(A_1), \dots, do(A_t), S_{t-1}) = P(S_t \mid \text{Pa}(S_t), A_t),$$

where $\text{Pa}(S_t)$ are causal parents of S_t. This lets the agent find $\pi^* = \arg\max_\pi P(\text{Success} \mid \pi)$ for precise planning.

4.3 Third-level: In-depth Reflection and Creation Based on Counterfactuals

After a complex plan fails, it is often difficult to pinpoint the exact misstep. A Causal World Model can address this through counterfactuals. After a failed plan $\pi_{\text{fail}} = (A_1^{\text{fail}}, \dots, A_T^{\text{fail}})$ with outcome $O = \text{Fail}$, counterfactual inference asks:

$$P(\text{Success} \mid do(A_t = a'_t), \pi_{\text{fail} \setminus \{A_t\}}),$$

for $t \in \{1, \dots, T\}$, to assign credit by finding $t^* = \arg\max_t P(\text{Success} \mid do(A_t = a'_t), \pi_{\text{fail} \setminus \{A_t\}})$.

Counterfactuals are the basis of imagination. A CWM can explore hypothetical worlds by altering the rules of its learned model. For example, $P(S \mid do(C = c))$ where C is a causal factor (e.g., $C = \text{gravity}$, $c = \text{weaker gravity}$, $S = \text{human can fly}$):

$$P(S \mid do(C = c)) = \sum_{S_{\text{pre}}} P(S_{\text{pre}}) P(S \mid do(C = c), S_{\text{pre}}),$$

enabling exploration of hypothetical worlds.

5 Conclusion

World models are critical for advancing toward general artificial intelligence, yet their correlation-driven design limits generalization, sample efficiency, and deep reasoning. This paper argues that integrating causal science is necessary to fix these drawbacks. Specifically, causal invariance drives active, multi-environment exploration to learn robust physical laws and mathematic functions, such as in IMO 2025 contest, Celia earned silver medal (34 points); causal discovery treats agent actions as interventions, building efficient and refined dynamic models; causal inference enables reflection and imagination beyond observation. Therefore, future work can further integrate causal principles with the world model and build relevant evaluation benchmarks.

Acknowledgments

Z. Lin was supported by the NSF China (No. 62276004) and the State Key Laboratory of General Artificial Intelligence.

References

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

[2] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

[3] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.

[4] Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7346–7355, 2024.

[5] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[6] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[7] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

[8] Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104011, 2024.

[9] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

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

[11] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[12] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

[13] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[15] Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

[16] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[17] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. 2022.

[18] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

[19] Judea Pearl. *Causality*. Cambridge university press, 2009.

[20] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[21] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022.

[22] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[23] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, page 3. Springer, 2016.

[24] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[25] Peng Ding and Fan Li. Causal inference. *Statistical Science*, 33(2):214–237, 2018.

[26] Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33:289–301, 2020.

[27] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17081–17092, 2022.

[28] Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22905–22918, 2021.

[29] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[30] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. Causality learning: A new perspective for interpretable machine learning. *arXiv preprint arXiv:2006.16789*, 2020.

[31] Paras Sheth, Raha Moraffah, K Selçuk Candan, Adrienne Raglin, and Huan Liu. Domain generalization–a causal perspective. *arXiv preprint arXiv:2209.15177*, 2022.

[32] Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025.

[33] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.