Toward Causal Foundation World Models: From Representation to Decision-Making

Mengyue Yang
School of Engineering Mathematics and Technology
University of Bristol
mengyue.yang@bristol.ac.uk

World models have emerged as a unifying paradigm, aiming to capture the dynamics of an environment so that agents can reason, predict, and plan without direct interaction. Recent advances in generative modeling and reinforcement learning have produced increasingly powerful world models across perception and control, yet most remain correlation-based and assume stationary dynamics. This restricts their ability to generalize, adapt, and be reliable in domains such as robotics, healthcare, and social systems that are inherently *dynamic, non-stationary, and open-ended*.

To address these challenges, we outline the vision of *causal foundation world models*: scalable and general-purpose models that unify representation, reasoning, and decision-making. By grounding world models in causality, the goal is to move beyond pattern recognition toward systems that can explain, predict, and intervene in complex environments. This talk will review recent lines of research that illustrate progress toward this direction.

- **Causal representation learning.** Disentangled causal representation and generalization in perceptual settings (CVPR 2021; CIKM 2021; KDD 2023; NeurIPS 2023 Spotlight (Yang et al. 2021b,a, 2023a,b; Sun et al. 2025)).
- **General decision-making systems.** Single/Multi-agent, and a foundation decision-making model. (ICLR 2025; ICML 2025; NeurIPS 2025, NeurIPS 2023 (Yan et al. 2024; Jin et al. 2025; Mi et al. 2025; Feng et al. 2023)).
- **Causality in foundation models.** Causal principles for reasoning, explanation, and planning with foundation models (NeurIPS 2025; ICML 2025; ICLR 2025 (Yu et al. 2025; Jin et al. 2025; Zhu et al. 2025)).
- **Causal world modeling.** Curiosity-driven exploration and intervention to reconstruct the dynamic causal diagrams under the hood in single-agent and multi-agent (NeurIPS 2025 (Zhao et al. 2025; Liu et al. 2025)).

Future Works. We will discuss directions on (i) dynamic and non-stationary causal mechanisms: theoretical foundation to practical challenges, (ii) causal multi-agent games, and (iii) open-ended exploration, toward scalable and trustworthy world models for real-world applications.

References

Feng, X.; Luo, Y.; Wang, Z.; Tang, H.; Yang, M.; Shao, K.; Mguni, D.; Du, Y.; and Wang, J. 2023. Chessgpt: Bridging policy learning and language modeling. *Advances in Neural Information Processing Systems*, 36: 7216–7262.

Jin, J.; Wu, Y.; Li, H.; He, X.; Zhang, W.; Yang, Y.; Yu, Y.; Wang, J.; and Yang, M. 2025. Large Language Models are Demonstration Pre-Selectors for Themselves. In *ICML*.

Liu, A.; Wang, J.; Kaski, S.; Wang, J.; and Yang, M. 2025. A Principle of Pre-Strategy Intervention for Multi-Agent Reinforcement Learning. In *NeurIPS*.

Mi, Q.; Yang, M.; Yu, X.; Zhao, Z.; Deng, C.; An, B.; Zhang, H.; Chen, X.; and Wang, J. 2025. Mf-llm: Simulating collective decision dynamics via a mean-field large language model framework. *arXiv e-prints*, NeurIPS 2025.

Sun, Y.; Kong, L.; Chen, G.; Li, L.; Luo, G.; Li, Z.; Zhang, Y.; Zheng, Y.; Yang, M.; Stojanov, P.; et al. 2025. Causal Representation Learning from Multimodal Biomedical Observations. In *ICML 2025*.

Yan, X.; Song, Y.; Feng, X.; Yang, M.; Zhang, H.; Ammar, H. B.; and Wang, J. 2024. Efficient Reinforcement Learning with Large Language Model Priors. In *ICLR*.

Yang, M.; Cai, X.; Liu, F.; Zhang, W.; and Wang, J. 2023a. Specify Robust Causal Representation from Mixed Observations. KDD '23.

Yang, M.; Dai, Q.; Dong, Z.; Chen, X.; He, X.; and Wang, J. 2021a. Top-N Recommendation with Counterfactual User Preference Simulation. CIKM '21, 2342–2351.

Yang, M.; Fang, Z.; Zhang, Y.; Du, Y.; Liu, F.; Ton, J.-F.; Wang, J.; and Wang, J. 2023b. NeurIPS 2023. In *Advances in Neural Information Processing Systems*, 79832–79857.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021b. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *CVPR 2021*.

Yu, X.; Wang, Z.; Yang, L.; Li, H.; Liu, A.; Xue, X.; Wang, J.; and Yang, M. 2025. Causal Sufficiency and Necessity Improves Chain-of-Thought Reasoning. *NeurIPS 2025*.

Zhao, Z.; Li, H.; Zhang, H.; Wang, J.; Faccio, F.; Schmidhuber, J.; and Yang, M. 2025. Curious Causality-Seeking Agents Learn Meta Causal World. *arXiv preprint arXiv:2506.23068. NeurIPS 2025*.

Zhu, Y.; de Souza, D. A.; Shi, Z.; Yang, M.; Minervini, P.; D'Amour, A.; and Kusner, M. J. 2025. When Can Proxies Improve the Sample Complexity of Preference Learning? *ICML 2025*.