Language Agents Meet Causality – Bridging LLMs and Causal World Models

John Gkountouras[1], Matthias Lindemann[2], Phillip Lippe[3], Efstratios Gavves[3], Ivan Titov[2,1]
[1]Institute for Logic, Language and Computation (ILLC), University of Amsterdam
[2]Institute for Language, Cognition and Computation (ILCC), University of Edinburgh
[3]QUVA Lab, University of Amsterdam

Introduction

Why Do We Need Causal Understanding in LLMs?

"Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality." - Judea Pearl

- Current LLM limitations:
  - **Incomplete** or **incorrect** causal knowledge of the world.
  - **Poor generalization** to novel scenarios.
  - **Limited understanding** of action consequences.

Our Solution:

- Integrate Causal Representation Learning (CRL) with LLMs.
- Learn structured causal world models.
- Enable robust planning and reasoning.

Causal Representation Learning

Why Causal Representation Learning?

Goal:

- Learn disentangled representations of causal variables
- Discover underlying causal structure from observations
- Enable reliable prediction of intervention effects

Modeling Setup:

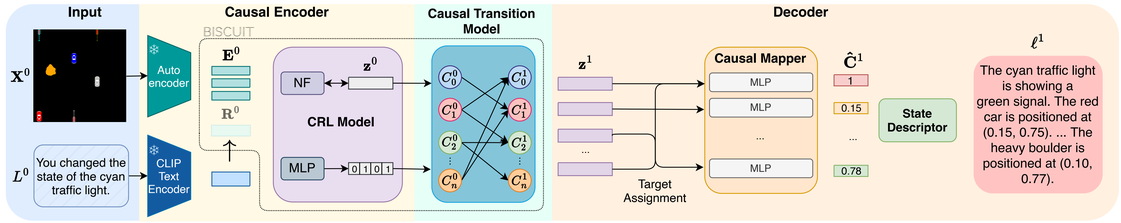
Input: High-dimensional observations $\mathbf{X}^t$ and actions $\mathbf{R}^t$

Learn mapping: $f : \mathbf{X}^t \to \mathbf{z}^t$

Temporal dynamics: $p(\mathbf{z}^{t+1} \mid \mathbf{z}^t, \mathbf{R}^t)$

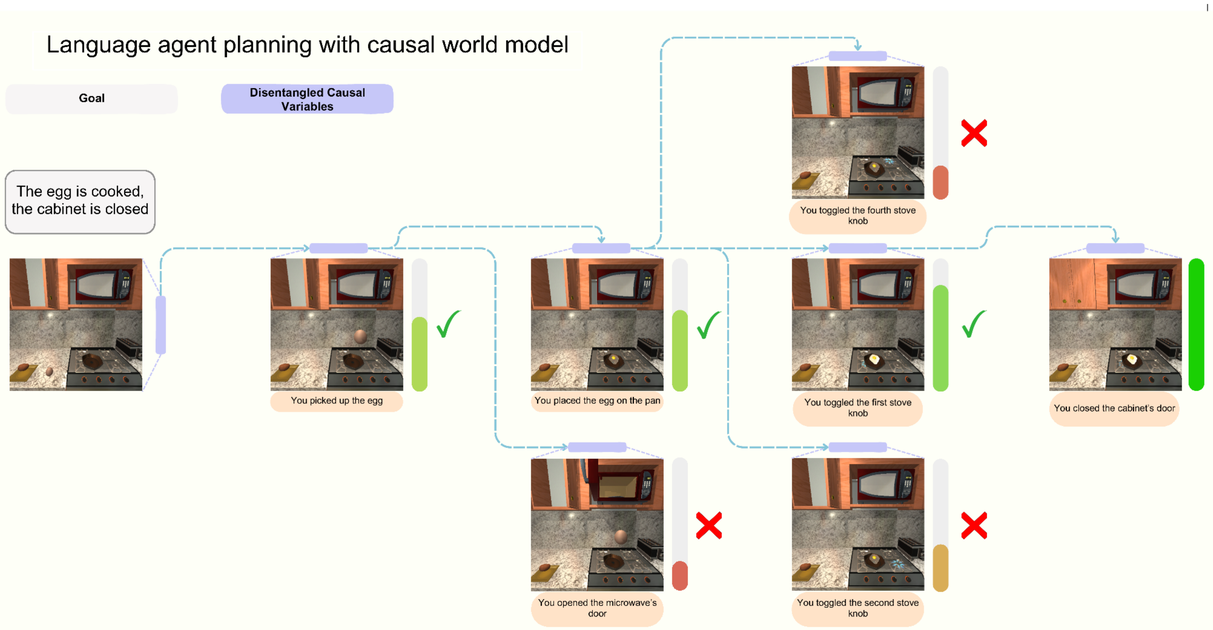
Key equation: $p(\{\mathbf{X}^t\} \mid \{\mathbf{R}^t\}) = \int p(\mathbf{z}^0) \prod_{t=1}^{T} p_\omega(\mathbf{z}^t \mid \mathbf{z}^{t-1}, \mathbf{R}^t)\, p_g(\mathbf{X}^t \mid \mathbf{z}^t)\, \mathrm{d}\mathbf{z}$

Causal World Model Architecture



- Causal Encoder: Maps high-dimensional states to causal variables $\mathbf{z}^t = \mathcal{E}(\mathbf{X}^t) = (f_\phi \circ e_\psi)(\mathbf{X}^t)$
- Transition Model: Predicts next states given actions using learned causal structure $\mathbf{z}^{t+1} \sim p_\omega\big(\mathbf{z}^{t+1} \mid \mathbf{z}^t, \mathcal{L}_e(L^t)\big)$
- Decoder: Decodes the disentangled causal representation to text descriptions $\ell^{t+1} = \mathcal{G}(\mathbf{z}^{t+1}) = (s \circ m_\theta)(\mathbf{z}^{t+1})$

Planning rollout



1. LLM proposes actions based on current state

2. Causal world model simulates outcomes

3. State evaluation and action selection

4. Iterative refinement through MCTS

Results

Planning

Environment	Steps	Causal World Model			RAP Baseline		
		Success Rate ↑	Avg. Steps (Success) ↓	Avg. Steps (Failure) ↓	Success Rate ↑	Avg. Steps (Success) ↓	Avg. Steps (Failure) ↓
iTHOR	2	**0.58**	**1.78**	**3.36**	0.25	4.00	4.00
	4	**0.44**	**2.14**	**5.43**	0.11	6.00	6.00
GridWorld	2	**0.95**	**1.92**	3.20	0.20	2.00	**3.19**
	4	**0.73**	**2.71**	**5.27**	0.11	3.27	5.48
	6	**0.46**	**3.65**	**7.72**	0.08	5.50	7.72
	8	**0.42**	**4.62**	**9.76**	0.06	7.00	9.93

Planning: Finding sequence of actions to reach desired goals. Requires understanding action consequences and long-term effects.

- Our approach shows superior success rates across environments
- Particularly effective for complex, multi-step tasks

Causal Inference

Steps	iTHOR 1	iTHOR 2	iTHOR 4	GridWorld 1	GridWorld 2	GridWorld 4	GridWorld 6	GridWorld 8
Causal Model	**0.824**	**0.680**	**0.630**	**0.954**	**0.922**	**0.829**	**0.797**	**0.758**
RAP World Model	0.482	0.285	0.110	0.391	0.220	0.085	0.045	0.005

Causal Inference: Predicting the effects of actions on the environment.

- Metric: Accuracy
- Our model maintains high accuracy even for long-term predictions
- Performance gap widens with prediction horizon

Conclusion

➔ First framework integrating CRL with language models for planning

➔ Demonstrated benefits of structured causal understanding in planning

➔ Introduced effective text-based interface for causal world models

Key findings:

➔ CRL enables more robust and reliable AI planning systems

➔ Bridges the gap between language understanding and causal reasoning

➔ Provides foundation for scaling to real-world applications

Future Directions:

➔ Extension to more complex environments

➔ Integration with advanced search algorithms

➔ Applications in robotics and embodied AI

References

[1] Schölkopf, Bernhard, et al. "Toward causal representation learning." *Proceedings of the IEEE* 109.5 (2021): 612-634.

[2] Lippe, Phillip, et al. "Biscuit: Causal representation learning from binary interactions." *Uncertainty in Artificial Intelligence*. PMLR, 2023.

[3] Richens, Jonathan, and Tom Everitt. "Robust agents learn causal world models." *arXiv preprint arXiv:2402.10877* (2024).

[4] Hao, Shibo, et al. "Reasoning with language model is planning with world model." *arXiv preprint arXiv:2305.14992* (2023).

[5] Gupta, Tarun, et al. "The essential role of causality in foundation world models for embodied ai." *arXiv preprint arXiv:2402.06665* (2024).