Causal Inference as an Inductive Bias for Reinforcement Learning

Ahmad B. Abdallah, Daniel Park, Tuan Minh Pham, Udit Ekansh

04/18/2025

Abstract

Reinforcement Learning (RL) has demonstrated strong performance across diverse domains, yet its generalization ability remains limited, particularly in the presence of distributional shifts. This paper investigates whether integrating causal inductive biases into RL can improve robustness and transferability across related tasks. We conduct a comparative study using model-free baselines and causal RL agents, including variants of the CausalCF algorithm that leverage interventions and counterfactual reasoning. All agents are trained on a robotic *picking* task in the `CausalWorld` environment and evaluated under a suite of domain-shift protocols, as well as on a distinct *pushing* task to assess representation transfer. Results show that causal agents outperform non-causal baselines on several out-of-distribution evaluations and exhibit promising signs of transferability. However, all methods degrade under severe domain perturbations. These findings provide empirical support for the hypothesis that causal structure can enhance generalization in RL, while also highlighting current limitations and motivating future research in scalable causal representation learning.

1 Introduction

Reinforcement Learning (RL) offers a principled approach for sequential decision-making through interaction with an environment. The learning problem is typically framed as a Markov Decision Process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, R the reward function, P the transition dynamics, and γ the discount factor. The agent seeks to learn a policy $\pi_\phi(a_t \mid s_t)$ that maximizes expected cumulative reward.

Although modern RL methods have shown success across a variety of domains, their generalization ability remains limited. In particular, many algorithms overfit to statistical regularities in the training environment, leading to

poor performance under distributional shift or in novel settings. This problem is especially evident in real-world applications such as robotics, where environments may vary in subtle but important ways.

One promising approach to improve generalization is to incorporate *causal inductive biases* into RL. By leveraging causal structures—such as intervention and counterfactual reasoning—agents may learn representations that are more invariant across tasks and environments. Recent work in Causal Reinforcement Learning (CRL) aims to formalize and implement such ideas, though the practical benefits remain challenging to validate empirically.

In this project, we explore the extent to which causal representations learned by a causal RL agent exhibit invariance across related tasks. We train **CausalCF**, a causal RL algorithm that integrates elements of the Pearl Causal Hierarchy, on a robotic *picking* task in the `CausalWorld` environment. We then evaluate the learned policy components on a distinct *pushing* task, without additional training. Our goal is to examine whether the model's learned causal structure is transferable and whether any signs of generalization emerge.

While our work does not aim to definitively quantify causal RL's advantages over non-causal baselines, it provides empirical evidence and visualizations that support the hypothesis of causal invariance. This preliminary investigation contributes to the broader understanding of how causal reasoning may aid RL agents in performing robustly across structurally similar but dynamically different tasks. The remainder of this document is organized as follows: Section 2 provides background on causal inference, inductive bias, and related literature in RL. Section 3 covers our primary objectives and scope of this project. Section 4 describes the experimental setup, including the environment, models, and training configuration. Section 5 presents our empirical findings and visualizations. Section 6 discusses key observations and limitations, and concludes with a summary and potential directions for future work.

2 Literature Review

Causal inference enhances reinforcement learning (RL) by addressing key challenges in sample efficiency, generalization, and interpretability [1]. It provides a framework combining data with structural environmental knowledge, enabling counterfactual reasoning [2], which is crucial for RL agents operating in uncertain, interactive environments

Causal RL (CRL) formalizes the integration of causal knowledge into RL systems through the tuple $(\mathcal{M}, \mathcal{G})$, where $\mathcal{M}$ is an RL model (e.g., MDP) and $\mathcal{G}$ encodes environmental causal structure [3]. Two distinct approaches emerge: (1) utilizing predefined causal information, or (2) learning causal relationships from data. These methods capture how actions influence states and rewards through causal dependencies. The underlying structural equations establish causal links between variables, with the *Causal Markov Assumption* enabling tractable inference by enforcing conditional independence between variables and their non-descendants given parents [1]. This principled abstraction facilitates

causal discovery in complex environments.

Furthermore, causal models have an equivalent causal directed acyclic graph (DAG) representation $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the tuple elements represent the set of variables and their causal relationships, respectively. The graph is constructed using three building blocks: the chain, fork, and collider. Under this structure, a form of product decomposition utilizing causal relationships can be employed, known as causal factorization, to model complex joint probability distributions $P(\boldsymbol{V}) = \prod_{i=1}^{n} P(V_i|pa_i(V_i))$ [3]

Research has demonstrated the effectiveness of causal RL approaches in various domains. Examples include Barenhoim's unified CRL framework, which leverages the structural invariances within causal inference and sample efficiency of reinforcement learning to produce a new, robust system for RL. This framework has been applied to research areas such as decision-making, learning, and intelligence, as well as applied domains, including robotics and healthcare [4].

These causal mechanisms take on deeper significance when intentionally deployed as inductive biases - a concept formally defined as "imposing constraints on relationships and interactions among entities in a learning process" [5]. The goal of implementing inductive biases is to prioritize certain solutions over another, through which one can discover a desirable generalized solution or improve the overall search without significantly decreasing in performance. In reinforcement learning (RL), these biases manifest as structural or domain-specific constraints that guide the agent's learning process while introducing potential trade-offs between performance and model complexity.

In Reinforcement Learning (RL), *inductive bias* often takes the form of structural or domain-related constraints that guide the agent's learning process. One early and influential example is *reward shaping*, where an auxiliary function $\Phi(s)$ is introduced to modify the original reward R [6]. Formally, the shaped reward R' is given by

$$R'(s_{t+1}, a_t) \;=\; R(s_{t+1}, a_t) \;+\; \gamma\,\Phi(s_{t+1}) \;-\; \Phi(s_t), \tag{1}$$

where $\gamma \in [0, 1]$ is the discount factor. A well-designed shaping function can bias exploration toward promising regions of the state space while preserving policy optimality. Another common bias appears in *hierarchical RL* [7], where tasks are decomposed into subtasks, each with a specialized policy $\{\pi_1, \pi_2, \ldots\}$. These policies can be sequenced to solve complex problems more efficiently. Similarly, *physics-based* biases improve control in continuous domains by embedding known kinematic or dynamic constraints into the agent's model [8, 9].

A growing area of interest is *graph-based* inductive biases, often introduced via Graph Neural Networks (GNNs) [5]. In multi-agent or physically structured environments, representing entities as nodes and interactions as edges can significantly boost sample efficiency and improve the agent's ability to handle complex relationships [10, 11]. Finally, *causal inductive biases* have gained traction for enhancing out-of-distribution generalization and interpretability [12,13]. By leveraging principles such as do-calculus or counterfactual reasoning, agents can move beyond correlation-based policies and instead learn true cause-effect mechanisms in their environment.

Overall, these diverse strategies demonstrate that carefully chosen inductive biases can enhance sample efficiency, promote better exploration, and improve policy robustness—often at the cost of potential model complexity or performance constraints if the assumed bias is not well matched to the environment.

Our initial proposal to "inject causal bias" was refined through literature analysis and feedback from Prof. Xue. Recognizing that all RL methods (especially causal approaches) are inherently biased per Mitchell's definition, "any basis for choosing one generalization over another", we have shifted our focus to a comparative analysis of causal and non-causal reinforcement learning methods using the Causal World Dataset rather than attempting to create an unbiased causal reinforcement learning method. This new direction allows us to explore the strengths and limitations of different approaches in various scenarios, providing a more nuanced understanding of when and how causal methods can outperform traditional reinforcement learning techniques.

3 Problem Statement & Objectives

3.1 Problem Statement

Reinforcement learning models frequently fail to generalize due to their dependence on correlational rather than causal relationships, leading to poor out-of-distribution performance in real-world applications. Current approaches that expand training data or adapt domains remain fundamentally limited by their lack of causal reasoning. Our initial research direction sought to enhance RL generalization by integrating causal inductive biases through structural causal models (SCMs) and counterfactual reasoning. This framework aimed to develop causal agents capable of distinguishing true environmental mechanisms from spurious correlations, thereby improving out-of-distribution robustness and decision reliability.

However, upon further investigation and valuable feedback from Professor Xue, we have revised our problem definition to address a more fundamental question in the field of causal reinforcement learning. Given the inherent bias in any causal RL method, as highlighted by Mitchell's (1980) definition of bias as "any basis for choosing one generalization over another, other than strict consistency with the observed training instances," we aim to conduct a comparative analysis of causal and non-causal RL methods using the Causal World Dataset.

We seek to understand the specific scenarios and conditions under which causal RL methods may outperform traditional RL approaches, despite their inherent biases. Our research will investigate the trade-offs between different types of biases present in causal and non-causal RL methods, focusing on their impact on sample efficiency, generalization, and robustness to domain shifts. We aim to identify which causal methods are best suited for particular applications and explore the role of hyperparameters in causal RL performance. Through systematic experimentation, we will test the hypothesis that learned causal structures remain invariant across different environments, potentially of-

fering advantages in transfer learning and out-of-distribution generalization.

This revised problem definition shifts our focus from developing a novel causal RL method to a more comprehensive evaluation of existing approaches. By doing so, we aim to contribute valuable insights to the ongoing discourse on the role of causality in reinforcement learning and its potential to enhance the capabilities of AI systems in complex, real-world scenarios.

3.2 Objectives

The primary aim of this project is to investigate the generalization capabilities of causal RL methods in comparison to traditional, non-causal approaches. Building on recent advancements in causal inference and inductive bias in machine learning, our study focuses on evaluating how well these methods perform in structurally diverse and distributionally shifted environments. Specifically, we address the following objectives:

1. **Conduct a systematic comparison of causal and non-causal RL methods using the `CausalWorld` dataset.**

   We evaluate several variants of reinforcement learning agents—including traditional model-free algorithms like Soft Actor-Critic (SAC), as well as causal agents based on the CausalCF framework. Agents are trained on a robotic manipulation task within the `CausalWorld` environment and evaluated on a range of benchmark protocols designed to introduce controlled distributional shifts. We assess performance using fractional success, emphasizing scenarios where causal methods demonstrate clear advantages over traditional baselines, with particular attention to environment complexity and variable interactions.

2. **Investigate the invariance and transferability of learned causal structures.** A central hypothesis in causal RL is that learned causal representations are more invariant and thus transferable across tasks. To evaluate this, we examine the extent to which causal representations learned during training on the *picking* task can be reused for a distinct but structurally related task (*pushing*). We compare the performance of transferred causal agents with those trained directly on the new task, analyzing whether causal structure facilitates improved generalization under domain shift.

These objectives aim to advance our understanding of when and how causal reasoning can support more robust and generalizable reinforcement learning, particularly in robotic settings where real-world variability poses significant challenges.

4 Experimental Setup

This section outlines the methodology used to evaluate causal and non-causal reinforcement learning agents in a robotic manipulation environment. All experiments were conducted in the `CausalWorld` simulator, a benchmark suite designed to study causal reasoning and generalization in goal-conditioned robotic tasks. Our study primarily focused on training agents on a *picking* task and evaluating their performance on a related but distinct *pushing* task. We rely on the official implementation of the CausalCF algorithm, which incorporates counterfactual reasoning into the RL pipeline, to examine how causal representations influence transfer and robustness.

We also attempted to run experiments on the *stacking* task, which is among the more complex manipulation scenarios available in the CausalWorld environment. However, we encountered an observation space mismatch error during initialization, which we were ultimately unable to resolve. As a result, no training or evaluation was conducted for stacking, and we leave this task as part of future scope for more comprehensive multi-task evaluation.

The following subsections detail the codebase used, task design, evaluation protocols, and training configurations.

4.1 Codebase and Environment

Our implementation is based entirely on the official GitHub repository provided by the authors of CausalCF[1]. This repository includes training scripts, model definitions, and environment wrappers for the `CausalWorld` simulator. We preserved the architecture and training protocol as closely as possible to ensure reproducibility and maintain consistency with the original design.

`CausalWorld` is a physics-based robotic manipulation environment built on PyBullet, simulating a three-fingered robotic hand (TriFinger) tasked with arranging colored blocks into target configurations. It offers structured observations, goal-conditioned rewards, and multiple control modes. For all experiments, we used structured observations and controlled the robot via joint positions, as recommended in the original CausalCF paper.

Due to the complexity of the CausalCF pipeline, including its counterfactual training phase, we ran all experiments on Purdue's high-performance computing clusters (Gilbreth and Anvil), which provided necessary GPU resources. Environment setup was managed through Conda and PyTorch, with minor modifications made only to logging and evaluation routines for our analysis.

4.2 Task and Evaluation Protocol

The CausalWorld environment includes a built-in evaluation pipeline comprising 12 protocols (P0–P11), each of which systematically alters one or more environment variables. These variables include block pose, block mass, block

[1] `https://github.com/Tom1042roboai/CausalCF`

size, goal pose, and floor friction. The purpose of these protocols is to test the generalization ability of RL agents across increasing levels of distributional shift.

Each protocol is associated with either `Space A` or `Space B`, which define different sampling ranges for the affected variables. `Space A` represents the training distribution, while `Space B` introduces out-of-distribution variation. All our models were trained using samples from `Space A`, making protocols involving `Space B` especially useful for measuring robustness.

Protocol	Space	Variables Modified
P0	A	None
P1	A	Block mass
P2	B	Block mass
P3	A	Block size
P4	A	Block pose
P5	A	Goal pose
P6	B	Block pose, Goal pose, Block mass
P7	A	Block pose, Goal pose, Block mass
P8	B	Block pose, Goal pose, Block mass
P9	B	Block pose, Goal pose, Block mass, Floor friction
P10	A	All variables
P11	B	All variables

Table 1: Evaluation protocols from CausalWorld as defined in the CausalCF paper [14].

CausalWorld utilizes a success/reward metric called *fractional success* which is defined as "the fractional volumetric overlap of the blocks with the goal shape, which ranges between 0 (no overlap) and 1 (complete overlap)" [14]. This is calculated as follows

$$\text{fractional success} = \frac{\sum_{i=1}^{n} intersection(O\_vol_i, G\_vol_i)}{\sum_{i=1}^{n} G\_vol_i} \tag{2}$$

Where n represents the number of objects, i represents the i^{th} object, O_vol_i is the current location of the i^{th} object, and G_vol_i is the goal location for the same object. As stated above, a higher fractional success represents more ideal performance by the model. We utilize this metric as a means of comparing the model performance.

Each configuration is evaluated using the *fractional success* metric, which essentially measures the overlap between the goal configuration and the achieved block configuration. A value of 1 indicates perfect alignment, while 0 indicates no overlap. We use three variants of this metric to capture different aspects of agent performance:

- **Last Fractional Success:** The success achieved at the final timestep of the evaluation trajectory.

- **Full Integrated Fractional Success:** The average success accumulated across all timesteps.
- **Last Integrated Fractional Success:** A cumulative score weighted toward the final portion of the trajectory.

4.3 Training Details and Hyperparameters

For our baseline models, we have trained a Soft-Actor Critic (SAC) model using the following hyperparameters.

Parameter	Value
γ	0.95
τ	1×10^{-3}
`ent_coef`	1×10^{-3}
`target_entropy`	auto
`learning_rate`	1×10^{-4}
`buffer_size`	1000000
`learning_starts`	100
`batch_size`	256

Table 2: Parameter settings for SAC

Parameter	Value
Total time steps	7000000
Training time	$\approx$ 30 hours
Episode length	840
Number of episodes	???
Skipframe	3
Space	A
Checkpoint frequency	50000

Table 3: Training parameters used for all models

When choosing our hyperparameter settings for SAC models, we matched the hyperparameter settings used in [**?**] as they have been shown to be effective in similar reinforcement learning tasks within CausalWorld. These settings provided a strong baseline for our experiments and allowed us to focus model evaluation. Similarly, when training the CausalCF models, we matched the settings used in [14] since they have demonstrated strong performance in comparable environments and tasks. Given the resource constraints we faced, particularly limited GPU access, we were unable to perform an extensive hyperparameter tuning process of our own. By adopting the hyperparameter settings from these prior works, we leveraged their proven effectiveness to ensure that our

models could perform optimally without the need for additional computational resources.

Given our models were all trained on the *picking* task, we initially assess their performance on picking to determine how well the model training process is working. Subsequently, to answer our question of model generalization using causal information, we transfer the learned causal representation of the picking and evaluate on the *pushing* task.

The models we evaluate span the three layers of the Pearl Causal Hierarchy, allowing us to distinguish the benefit of incorporating each additional layer

1. **Layer 1 (Association):** Our baseline SAC model
2. **Layer 2 (Intervention):** The *Intervene* model which incorporates interventions into the SAC baseline model
3. **Layer 3 (Counterfactuals)**:
   (a) The *Counterfactual + Intervene* model, which utilizes both counterfactual and intervention information atop the SAC baseline, but operates on a static causal representation without performing Counterfactual training
   (b) The *CausalCF* model, which also utilizes both counterfactual and intervention information, and alternates between learning the causal representation and agent training

Finally, the *TransferCausalRepIntervene* model which utilizes interventions and the causal representation learned for the picking task is evaluated on the pushing task. This is compared to the pushing task trained intervention model to gauge how well the learned causal representation transfers to adjacent tasks.

5 Results

We evaluate four variants of Soft Actor-Critic (SAC) agents on a suite of 12 robotic manipulation tasks (P0–P11) from the CausalWorld benchmark. The focus of this evaluation is to understand how the use of causal representations, counterfactual reasoning, and intervention-based training affects the agent's generalization performance.

The four configurations evaluated are:

1. **No Intervention SAC:** Standard model-free SAC agent trained on `picking`, with no causal reasoning.
2. **CausalCF Iterative:** SAC trained on `picking` using the causal model without explicit interventions.
3. **CausalCF + Intervene:** SAC trained on the `picking` task using both interventions and counterfactual updates.

4. **Transfer-CausalRep:** Agent trained on `picking` with causal representation, then fine-tuned on the `pushing` task.

In the sections that follow, we provide a detailed breakdown of each configuration, presenting bar plots and radar plots for all three success variants and interpreting the patterns observed in the context of structural generalization and causal transferability.

5.1 No Intervention SAC (Picking Task)

This configuration represents a standard model-free SAC agent trained on the `picking` task without any causal representation, counterfactuals, or interventions. It serves as a baseline to assess the extent to which causal structure contributes to generalization performance.

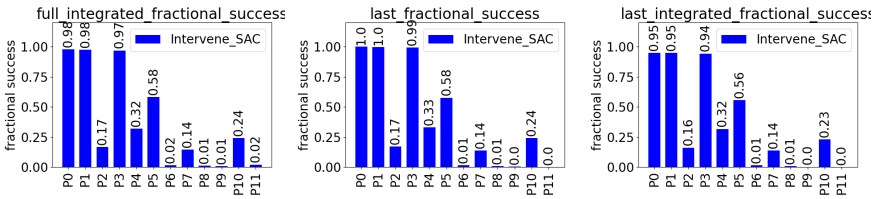

Figure 1: No Intervention SAC (Picking): Bar plots showing full integrated, last, and last integrated fractional success.

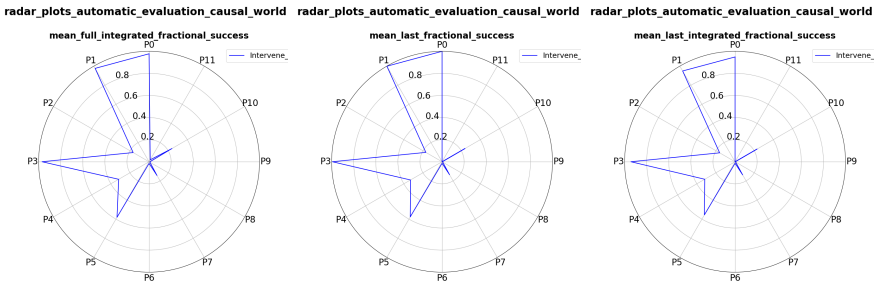

Figure 2: No Intervention SAC (Picking): Radar plots corresponding to the same success variants.

As expected, the model performs very well on the in-distribution task P0, with near-perfect success. It also shows competence on low-variance tasks like P1 (mass variation) and P3 (block size variation). However, the absence of any structural prior leads to steep performance degradation on even moderate task perturbations. Performance is particularly poor on protocols involving pose and friction shifts (P5–P9), and nearly zero in the complex multi-variable settings P10–P11.

The radar plots show this trend clearly — the model exhibits a narrow band of high performance centered on familiar conditions, but is unable to extend its learning to scenarios with overlapping changes or latent causal dependencies. This illustrates the inherent limitations of model-free RL agents when generalization is not scaffolded by inductive structure or causal reasoning.

5.2 CausalCF Iterative (Picking Task)

This variant of the causal model does not use explicit interventions or counterfactuals, relying instead on causal representation learning through iterative training. The agent is trained and evaluated on the `picking` task.

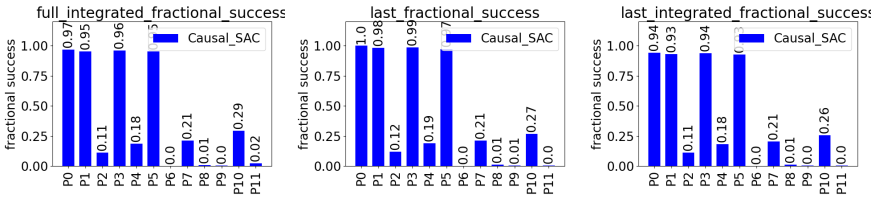

Figure 3: CausalCF Iterative (Picking): Bar plots showing full integrated, last, and last integrated fractional success.

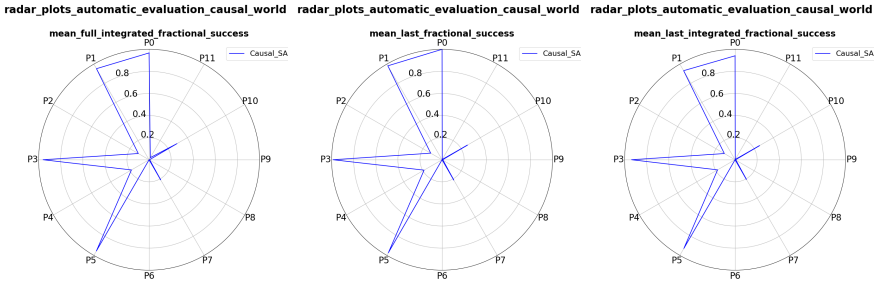

Figure 4: CausalCF Iterative (Picking): Radar plots corresponding to the same success variants.

Compared to the full CausalCF model, this variant shows relatively stable performance on the simpler tasks (P0–P2) and some mid-complexity settings (P3–P5). However, the absence of interventions appears to limit its robustness in generalizing to more complex variations. Notably, success rates in P6 and beyond are consistently lower than in the previous configuration.

This highlights the potential benefits of explicit intervention data in reinforcing structural understanding. While the causal representation alone does offer improvements over purely reactive baselines, it is less effective in environments where structural changes require active disentanglement of variables.

5.3 CausalCF + Intervene (Picking Task)

This configuration represents the full CausalCF pipeline with both counterfactual reasoning and interventions. The model is trained solely on the picking task and evaluated across all 12 benchmark protocols.

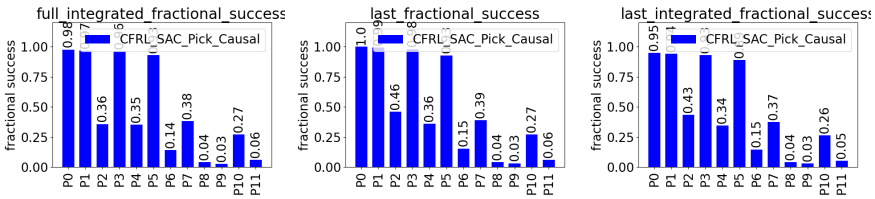

Figure 5: CausalCF + Intervene (Picking): Bar plots showing full integrated, last, and last integrated fractional success.

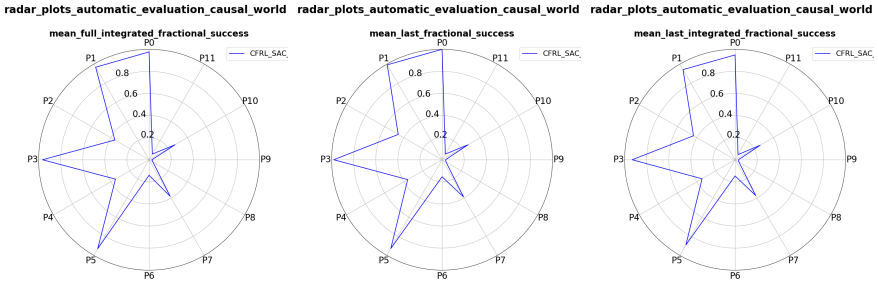

Figure 6: CausalCF + Intervene (Picking): Radar plots corresponding to the same success variants.

The bar plots indicate high performance on in-distribution protocols (P0–P1) with fractional success close to 1.0. Performance remains strong on moderate-shift tasks such as P3 (block size variation) and P5 (goal pose change), suggesting the model is able to generalize across modest structural changes. However, a significant drop is observed in out-of-distribution protocols (P6–P11) compared to P0-P5, especially those involving combinations of variation in mass, pose, and friction. The radar plots reinforce this: while performance is evenly distributed across lower-index protocols, success is fragmented and inconsistent on the more complex ones.

This result supports the hypothesis that counterfactual and interventional training encourages the learning of more robust, transferable representations within related domains. However, it also highlights the limits of this approach when faced with more severe distributional shifts that the causal model was not trained to explicitly capture.

5.4 Transfer-CausalRep (Pick → Push)

This configuration evaluates the model's ability to transfer causal representations learned from the `picking` task to a different but related task: `pushing`. The policy is fine-tuned on the pushing task while reusing the causal modules learned during the initial training.

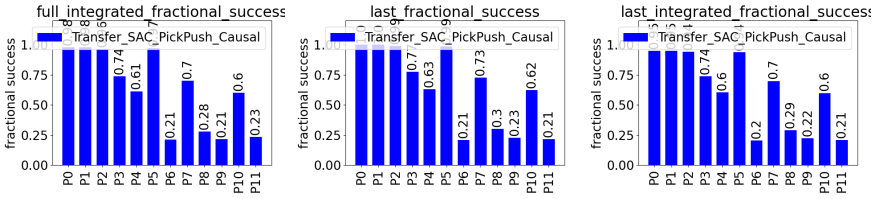

Figure 7: Transfer-CausalRep (Pick → Push): Bar plots showing full integrated, last, and last integrated fractional success.

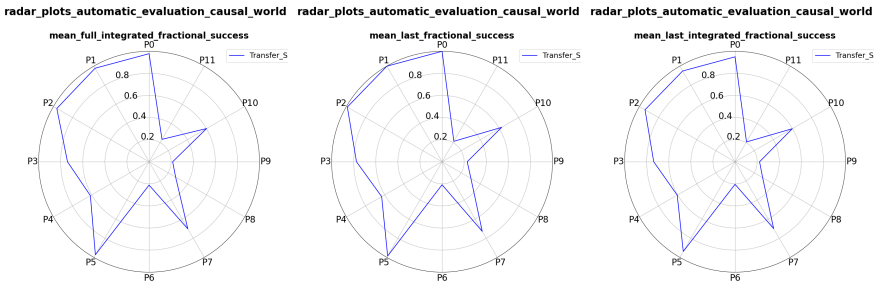

Figure 8: Transfer-CausalRep (Pick → Push): Radar plots corresponding to the same success variants.

The results from this configuration demonstrate promising signs of transferability. Although the model was originally trained on a different task, it achieves competitive success rates on several pushing protocols—particularly those with minimal frictional and pose changes (P1–P5). This indicates that the causal modules learned on the picking task are not only reusable but provide a structural prior that supports learning in related environments.

Compared to the SAC baseline and even the purely picking-focused causal models, Transfer-CausalRep achieves better performance on higher-index protocols such as P6–P9. This highlights the benefit of transferring causal abstractions, especially in environments where relational or structural knowledge is more important than raw low-level similarity.

However, performance on the most complex protocols (P10 and P11) remains low. This aligns with the broader trend observed in previous models, indicating that even transferred causal representations may struggle when faced with multiple simultaneous variations that deviate heavily from the source training

distribution.

Overall, this configuration provides empirical support for the hypothesis that causal representations are partially transferable across related robotic tasks, especially when task dynamics share structural elements.

6 Discussion & Conclusion

This study explored the potential of causal representations to enhance generalization in reinforcement learning through a comparative analysis of causal and non-causal agents in the `CausalWorld` robotic manipulation environment. Our investigation focused on evaluating whether causal inductive biases, informed by elements of the Pearl Causal Hierarchy, enable reinforcement learning agents to generalize across task variants and transfer knowledge between structurally related tasks.

Empirical results demonstrate that causal models—particularly those incorporating both intervention and counterfactual reasoning—consistently outperform standard model-free baselines in scenarios involving distributional shifts. Notably, the full CausalCF model exhibited superior robustness across evaluation protocols with moderate environment perturbations. Furthermore, our transfer learning experiments suggest that causal representations learned on one task (e.g., picking) retain partial utility when adapted to related tasks (e.g., pushing), validating the hypothesis that structural knowledge contributes to policy reuse and generalization.

However, the study also revealed limitations. All models, including those leveraging causal reasoning, struggled under extreme domain randomization (e.g., P10–P11), highlighting that current causal RL approaches may not fully capture or adapt to complex multi-factor distributional shifts. Additionally, our inability to evaluate on the stacking task limits the breadth of conclusions drawn across manipulation complexities.

In summary, our findings support the promise of causal reinforcement learning as a means to improve generalization and transfer in robotic agents. Future work should expand this investigation to include more diverse tasks, explore adaptive causal discovery during transfer, and evaluate causal RL under richer forms of environment shift. Such directions will help determine the scalability and practicality of causal inductive biases in real-world autonomous systems.

References

[1] Z. Deng, J. Jiang, G. Long, and C. Zhang, "Causal reinforcement learning: A survey," 2023.

[2] P. L. Harris, T. German, and P. Mills, "Children's use of counterfactual thinking in causal reasoning," *Cognition*, vol. 61, no. 3, pp. 233–259, 1996.

[3] Y. Zeng, R. Cai, F. Sun, L. Huang, and Z. Hao, "A survey on causal reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[4] E. Bareinboim, "Causal reinforcement learning (CRL)." `https://crl.causalai.net/`, Dec 2024.

[5] P. Battaglia, J. B. C. Hamrick, V. Bapst, A. Sanchez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. Allen, C. Nash, V. J. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," *arXiv*, 2018.

[6] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999.

[7] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," in *Discrete Event Dynamic Systems*, pp. 41–77, 2003.

[8] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," in *International Symposium on Experimental Robotics (ISER)*, 2016.

[9] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian neural networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 15379–15389, 2019.

[10] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, and A. et, "Learning to simulate complex physics with graph networks," *International Conference on Learning Representations (ICLR)*, 2020.

[11] H. Wang, J. Zhang, and X. Liu, "Graph neural networks in reinforcement learning: A survey," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4822–4828, 2020.

[12] E. Bareinboim, "Causal reinforcement learning: Invited talk," *International Conference on Machine Learning (ICML)*, vol. 119, 2020.

[13] K. Zhang and E. Bareinboim, "Learning causal state representations of partially observed systems," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14346–14356, 2020.

[14] T. He, J. Gajcin, and I. Dusparic, "Causal counterfactuals for improving the robustness of reinforcement learning," *arXiv preprint arXiv:2211.05551*, 2022.