Seeing is not Believing: Robust Reinforcement Learning against Spurious Correlation

Wenhao Ding [*1] **Laixi Shi** [*1] **Yuejie Chi** [1] **Ding Zhao** [1]

Abstract

In this work, we consider one critical type of robustness against spurious correlation, where different portions of the state do not have causality but have correlations induced by unobserved confounders. These spurious correlations are ubiquitous in real-world tasks, for instance, a self-driving car usually observes heavy traffic in the daytime and light traffic at night due to unobservable human activity. A model that learns such useless or even harmful correlation could catastrophically fail when the confounder in the test case deviates from the training one. Although motivated, enabling robustness against spurious correlation poses significant challenges since the uncertainty set, shaped by the unobserved confounder and sequential structure of RL, is difficult to characterize and identify. To solve this issue, we propose *Robust State-Confounded Markov Decision Processes* (RSC-MDPs) and theoretically demonstrate its superiority in breaking spurious correlations compared with other robust RL. We also design an empirical algorithm to learn the robust optimal policy for RSC-MDPs, which outperforms all baselines in eight realistic self-driving and manipulation tasks.

1. Introduction

Although standard RL has achieved remarkable success in simulated environments, a growing trend in RL is to address another critical concern – **robustness** – with the hope that the learned policy still performs well when the deployed (test) environment deviates from the nominal one used for training (Ding et al., 2022). Robustness is highly desirable since the performance of the learned policy could significantly deteriorate due to the uncertainty and variations of the test environment induced by random perturbation, rare events, or even malicious attacks (Mahmood et al., 2018; Zhang et al., 2021a).

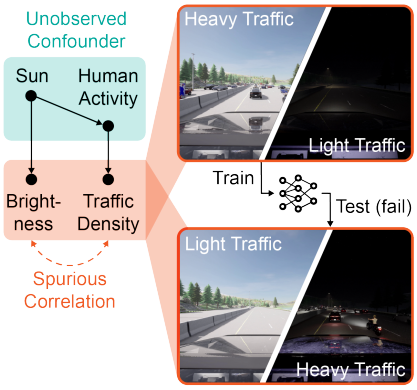

Figure 1. A model trained only with heavy traffic in the daytime learns the spurious correlation between brightness and traffic density and could fail to drive in light traffic in the daytime.

Despite various types of uncertainty have been investigated in RL, this work focuses on the uncertainty of environment with semantic meanings resulting from some unobserved underlying variables. Such environment uncertainty, denoted as **semantic uncertainty**, is motivated by innumerable real-world applications but still receives little attention in sequential decision-making tasks (De Haan et al., 2019). To specify the phenomenon of semantic uncertainty, let us consider a concrete example (illustrated in Figure 1) in a driving scenario, where a shift between training and test environments caused by an unobserved confounder can potentially lead to a severe safety issue. Specifically, the observations *brightness* and *traffic density* do not have cause and effect on each other but are controlled by a confounder (i.e. *sun* and *human activity*) that is usually unobserved to the agent. During training, the agent could memorize the **spurious correlation** between *brightness* and *traffic density*, that is, traffic is heavy during the daytime but light at night. However, such correlation could be problematic during testing when the value of the confounder deviates from the training one, e.g., the traffic becomes heavy at night due to special events (*human activity* changes), as shown in Figure 1. Consequently, the policy dominated by the spurious correlation

[*]Equal contribution [1]Carnegie Mellon University, Location, Country. Correspondence to: Wenhao Ding <wenhaod@andrew.cmu.edu>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

in training fails on out-of-distribution samples (observations of heavy traffic at night) in the test scenarios.

The failure of the driving example in Figure 1 is attributed to the widespread and harmful spurious correlation, namely, the learned policy is not robust to the semantic uncertainty of the test environment. However, ensuring robustness to semantic uncertainty is challenging since the targeted uncertain region – the semantic uncertainty set of the environment – is carved by the unknown causal effect of the unobserved confounder, and thus hard to characterize. In contrast, prior works concerning robustness in RL (Moos et al., 2022) usually consider a homogeneous and structure-agnostic uncertainty set around the state (Zhang et al., 2020b; 2021a; Han et al., 2022), action (Tessler et al., 2019; Tan et al., 2020), or the training environment (Iyengar, 2005; Yang et al., 2022; Shi & Chi, 2022) measured by some heuristic functions (Zhang et al., 2020b; Shi & Chi, 2022; Moos et al., 2022) to account for unstructured random noise or small perturbations. Consequently, these prior works could not cope with the semantic uncertainty since their uncertainty set is different from and cannot tightly cover the desired semantic uncertainty set, which could be heterogeneous and allows for potentially large deviations between the training and test environments.

In this work, to address the semantic uncertainty, we first propose a general RL formulation called State-confounded Markov decision processes (SC-MDPs), which model the possible causal effect of the unobserved confounder in an RL task from a causal perspective. SC-MDPs better explain the reason for semantic shifts in the state space than traditional MDPs. Then, we formulate the problem of seeking robustness to semantic uncertainty as solving Robust SC-MDPs (RSC-MDPs), which optimizes the worst performance when the distribution of the unobserved confounder lies in some uncertainty set. The key contributions of this work are summarized as follows.

- We propose a new type of robustness with respect to semantic uncertainty to address spurious correlation in RL and provide a formal formulation called RSC-MDPs, which are well-motivated by ubiquitous real-world applications.
- We theoretically justify the advantage of the proposed RSC-MDP framework against semantic uncertainty over the prior robust RL without semantic information.
- We implement an empirical algorithm to solve RSC-MDPs and show that it outperforms the baselines on eight real-world tasks in manipulation and self-driving.

2. Preliminaries and Limitations of Robust RL

Standard Markov decision processes (MDPs). A discounted infinite-horizon standard MDP is represented by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, T, r, P\}$, where $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathbb{R}^{d_A}$ are the state and action spaces, respectively, with n/d_A being the dimension of state/action. T is the length of the horizon; $P = \{P_t\}_{1\le t\le T}$, where $P_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the probability transition kernel at time step t, for all $1 \le t \le T$; and $r = \{r_t\}_{1\le t\le T}$ denotes the reward function, where $r_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the deterministic immediate reward function. A policy (action selection rule) is denoted by $\pi = \{\pi_t\}_{1\le t\le T}$, namely, the policy at time step t is $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$ based on the current state s_t as $\pi_t(\cdot \mid s_t)$.

Lack of semantic information in the robustness of RL. In spite of the rich literature on robustness in RL, prior works usually hedge against the uncertainty induced by unstructured random noise or small perturbations, specified as a small and homogeneous uncertainty set around the nominal one. However, the unknown uncertainty in the real world could have a complicated and semantic structure that cannot be well-covered by a homogeneous ball regardless of the choice of the uncertainty level, leading to either over conservative policy (when the uncertainty level is large) or insufficient robustness (when the uncertainty level is small). Altogether, we obtain the natural motivation of this work: *How to formulate such semantic uncertainty and ensure robustness against it?*

3. Robust RL against Semantic Uncertainty from a Causal Perspective

To describe semantic uncertainty, we choose to study MDPs from a causal perspective with a basic concept called the structural causal model (SCM), shall be specified in Appendix. Armed with the concept, we formulate State-confounded MDPs – a broader set of MDPs in the face of the unobserved confounder in the state space.

Structural causal model. We denote a structural causal model (SCM) (Pearl, 2009) by a tuple $\{X, Y, F, P^x\}$, where X is the set of exogenous (unobserved) variables, Y is the set of endogenous (observed) variables, and P^x is the distribution of all the exogenous variables. Here, F is the set of structural functions capturing the causal relations between X and Y such that for each variable $y_i \in Y$, $f_i \in F$ is defined as $y_i \leftarrow f_i\big(\mathsf{PA}(y_i), x_i\big)$, where $x_i \subseteq X$ and $\mathsf{PA}(y_i) \subseteq Y \setminus y_i$ denotes the parents of the node y_i.

3.1. State-confounded MDPs (SC-MDPs)

We now present state-confounded MDPs (SC-MDPs), whose probabilistic graph is illustrated in Figure 2(a) with a comparison to standard MDPs in Figure 2(b). Besides the components in standard MDPs $\mathcal{M}$, we introduce a set of unobserved confounder $C_{\mathsf{s}} = \{c_t\}_{1\le t\le T}$, where $c_t \in \mathcal{C}$ denotes the confounder that is generated from some unknown but fixed distribution P_t^c at time step t, i.e., $c_t \sim P_t^c \in \Delta(\mathcal{C})$.

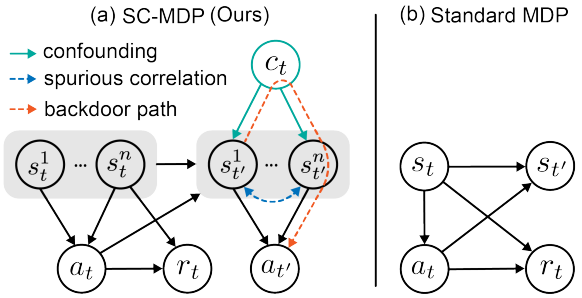

Figure 2. The probabilistic graphs of our formulation (SC-MDP) and standard MDP. s_t^1 means the first dimension of s_t. $s_{t'}$ is a shorthand for s_{t+1}. In SC-MDP, the orange line denotes the backdoor path from $s_{t'}^1$ to $a_{t'}$ opened by the confounder c_t.

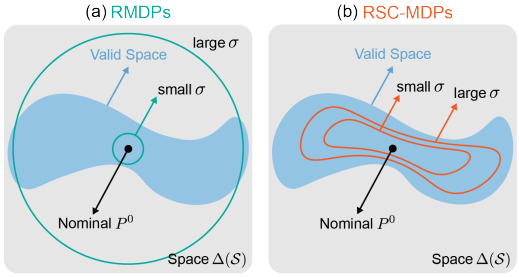

Figure 3. (a) RMDPs add homogeneous noise to states, while (b) RSC-MDPs perturb the confounder to influence states, resulting in a subset of the valid space.

To characterize the causal effect of the confounder C_s on the state dynamic, we resort to an SCM, where C_s is the set of exogenous (unobserved) confounder and endogenous variables include all dimensions of states $\{s_t^i\}_{1\le i\le n, 1\le t\le T}$, and actions $\{a_t\}_{1\le t\le T}$. Specifically, the structural function F is considered as $\{\mathcal{P}_t^i\}_{1\le i\le n, 1\le t\le T}$ – the transition from the current state s_t, action a_t and the confounder c_t to each dimension of the next state s_{t+1}^i for all time steps, i.e., $s_{t+1}^i \sim \mathcal{P}_t^i(\cdot \mid s_t, a_t, c_t)$. Notably, the specified SCM does not confound the reward, i.e., $r_t(s_t, a_t)$ does not depend on the confounder c_t.

Armed with the above SCM, denoting $P^c := \{P_t^c\}$, we can introduce state-confounded MDPs (SC-MDPs) represented by $\mathcal{M}_{\text{sc}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, P^c\}$ (Figure 2(a)). A policy is denoted as $\pi = \{\pi_t\}$, where each π_t results in an intervention (possibly stochastic) that set $a_t \sim \pi_t(\cdot \mid s_t)$ at time step t regardless of the value of confounder.

State-confounded value function and optimal policy. Given s_t, the causal effect of a_t on the next state s_{t+1} plays an important role in characterizing value function/Q-function. To ensure the identifiability of the causal effect, the confounder c_t are assumed to obey the backdoor criterion (Pearl, 2009; Peters et al., 2017), leading to the following *state-confounded value function* (SC-value function) (Wang et al., 2021) (*state-confounded Q-function* (SC-Q function) can be specified similarly):

$$\widetilde{V}_t^{\pi,P^c}(s) = \mathbb{E}_{\pi,P^c}\Big[\sum_{k=t}^{T} r_k(s_k,a_k) \mid s_t = s; \\ c_k \sim P_k^c, s_{k+1}^i \sim \mathcal{P}_k^i(\cdot \mid s_k, a_k, c_k)\Big]. \quad (1)$$

Remark 1. Note that the proposed SC-MDPs serve as a general formulation for a broad family of RL problems that include standard MDPs as a special case. Specifically, any standard MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, T, r\}$ can be equivalently represented by at least one SC-MDP $\mathcal{M}_{\text{sc}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, P^c\}$ as long as $\mathbb{E}_{c_t\sim P_t^c}\left[\mathcal{P}_t^i(\cdot \mid s_t, a_t, c_t)\right] = \left[P(\cdot \mid s_t, a_t)\right]_i$ for all $1 \le i \le n, 1 \le t \le T$.

3.2. Robust state-confounded MDPs (RSC-MDPs)

In this work, we consider robust state-confounded MDPs (RSC-MDPs) – a variant of SC-MDPs promoting the robustness to the uncertainty of the unobserved confounder distribution P^c, denoted by $\mathcal{M}_{\text{sc-rob}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, \mathcal{U}^\sigma(P^c)\}$. Here, the perturbed distribution of the unobserved confounder is assumed in an uncertainty set $\mathcal{U}^\sigma(P^c)$ centered around the nominal distribution P^c with radius σ measured by some 'distance' function $\rho : \Delta(\mathcal{C}) \times \Delta(\mathcal{C}) \to \mathbb{R}^+$, i.e.,

$$\mathcal{U}^\sigma(P^c) := \otimes\, \mathcal{U}^\sigma(P_t^c), \\ \mathcal{U}^\sigma(P_t^c) := \{P \in \Delta(\mathcal{C}) : \rho(P, P_t^c) \le \sigma\}. \quad (2)$$

Consequently, the corresponding *robust SC-value function* is defined as

$$\widetilde{V}_t^{\pi,\sigma}(s) := \inf_{P\in\mathcal{U}^\sigma(P^c)} \widetilde{V}_t^{\pi,P}(s), \quad (3)$$

representing the worst-case cumulative rewards when the confounder distribution lies in the uncertainty set $\mathcal{U}^\sigma(P^c)$.

RSC-MDPs possess benign properties similar to the standard MDPs: there exists at least one optimal policy that maximizes the robust SC-value function $\widetilde{V}_t^{\pi,\sigma}$ for any RSC-MDP which shall be verified in Theorem 2 in Appendix D.3.

Goal. Based on all the definitions and analysis above, this work aims to find an optimal policy for RSC-MDPs that maximizes the robust SC-value function in (3), yielding optimal performance in the worst case when the unobserved confounder distribution falls into an uncertainty set $\mathcal{U}^\sigma(P^c)$.

3.3. Advantages of RSC-MDPs over traditional robust works in RL

The most relevant robust RL formulation to ours is RMDPs (with the details in Appendix A). Here, we provide a comparison between RMDPs and our RSC-MDPs by an illustration and also theoretical justifications and leave the comparisons and connections to other related formulations in Figure 4 and Appendix B.1 due to space limits.

As an illustration, imagining the true uncertainty set encountered in the real world is illustrated as the blue region in Figure 3, which could have a complicated structure. Since the uncertainty set in RMDPs is homogeneous (illustrated by the green circles), one often faces the dilemma of being either too conservative (when σ is large) or too reckless (when σ is small). In contrast, the proposed RSC-MDPs –

Table 1. Testing reward on shifted environments. Bold font means the best reward.

Method	Brightness	Behavior	Crossing	CarType	Lift	Stack	Wipe	Door
SAC	0.56±0.13	0.13±0.03	0.81±0.13	0.63±0.14	0.58±0.13	0.26±0.12	0.16±0.20	0.08±0.07
RMDP-G	0.55±0.15	0.16±0.04	0.47±0.13	0.53±0.16	0.31±0.08	0.33±0.15	0.06±0.17	0.07±0.03
RMDP-U	0.54±0.19	0.13±0.05	0.60±0.15	0.39±0.13	0.51±0.17	0.23±0.11	0.06±0.17	0.10±0.13
MoCoDA	0.50±0.14	0.16±0.05	0.22±0.14	0.23±0.12	0.46±0.14	0.29±0.11	0.01±0.24	0.09±0.14
ATLA	0.48±0.11	0.14±0.03	0.61±0.14	0.52±0.14	0.61±0.18	0.21±0.12	0.29±0.18	0.28±0.19
DBC	0.52±0.18	0.16±0.03	0.68±0.12	0.45±0.10	0.12±0.02	0.03±0.02	0.19±0.35	0.01±0.01
RSC-SAC	**0.99**±0.11	**1.02**±0.09	**1.04**±0.02	**1.03**±0.02	**0.98**±0.04	**0.77**±0.20	**0.85**±0.12	**0.61**±0.17

shown in Figure 3(b) – take advantage of the semantic uncertainty set (illustrated by the orange region) enabled by the underlying SCM, which can potentially lead to much better estimation of the true uncertainty set. Specifically, the varying unobserved confounder induces diverse perturbation to different portions of the state through the structural causal function, enabling heterogeneous and structural uncertainty sets over the state space.

To theoretically understand the advantages of the proposed robust formulation RSC-MDPs with comparison to RMDPs, Theorem 1 verifies that RSC-MDPs enable additional robustness to fierce semantic attack besides small model perturbation or noise considered in RMDPs, which is specified with proof in Appendix D.2.

4. Empirically Solve RSC-MDPs: RSC-SAC

Solving RSC-MDPs could be challenging as the semantic uncertainty set is induced by the causal effect of perturbing the confounder. The precise characterization of this semantic uncertainty set is difficult since neither the unobserved confounder nor the true causal graph of the observable variables is accessible, both of which are necessary for intervention or counterfactual reasoning. Therefore, we choose to approximate the causal effect of perturbing the confounder by learning from the data collected during training.

In this section, we propose an intuitive yet effective empirical approach named RSC-SAC for solving RSC-MDPs, which is outlined in Algorithm 1. The detailed algorithm can be found in Appendix C. We first estimate the effect of perturbing the distribution P^c of the confounder to generate new states (Section C.1). Then, we learn the structural causal model $\mathcal{P}_t^i$ to predict rewards and the next states given the perturbed states (Section C.2). By combining these two components, we construct a data generator capable of simulating novel transitions (s_t, a_t, r_t, s_{t+1}) from the semantic uncertainty set. To learn the optimal policy, we construct the data buffer with a mixture of the original data and the generated data and then use the Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018) to optimize the policy.

5. Experiments and Evaluation

5.1. Environments with spurious correlation

To the best of our knowledge, no existing benchmark addresses the issues of spurious correlation in the state space of RL. To bridge the gap, we design a benchmark consisting of eight novel tasks in self-driving and manipulation domains using the Carla (Dosovitskiy et al., 2017) and Robosuite (Zhu et al., 2020) platforms. Tasks are designed to include spurious correlations in terms of human common sense, which is ubiquitous in decision-making applications and could cause safety issues. We leave the full descriptions of the tasks in Appendix F.3 and the description of baselines in Appendix F.1.

5.2. Results Analysis

To comprehensively evaluate the performance of the proposed method RSC-SAC, we conduct experiments to answer the following question: **Q1.** Can RSC-SAC eliminate the harmful effect of spurious correlation in learned policy? **R1.** RSC-SAC **is robust against spurious correlation.** The testing results of our proposed method with comparisons to the baselines are presented in Table 1, where the rewards are normalized by the episode reward of SAC in the nominal environment. The results reveal that RSC-SAC significantly outperforms other baselines in shifted test environments. An interesting and even surprising finding, as shown in Table 1, is that although RMDP-G, RMDP-U, and ATLA are trained desired to be robust against small perturbations, their performance drops more than non-robust SAC in some tasks. This indicates that using the samples generated from the traditional robust algorithms could harm the policy performance when the test environment is outside of the prescribed uncertainty set considered in the robust algorithms.

6. Conclusion

This work focuses on robust reinforcement learning against spurious correlation in state space, which broadly exists in (sequential) decision-making tasks. We propose robust SC-MDPs as a general framework to break spurious correlations by perturbing the value of unobserved confounders. We not only theoretically show the advantages of the framework compared to existing robust works in RL, but also design an empirical algorithm to solve robust SC-MDPs by approximating the causal effect of the confounder perturbation. The experimental results demonstrate that our algorithm is robust to spurious correlation – outperforms the baselines when the value of the confounder in the test environment derivates from the training one.

References

Abid, A., Yuksekgonul, M., and Zou, J. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, pp. 66–88. PMLR, 2022.

Agarwal, S. and Chinchali, S. P. Synthesizing adversarial visual scenarios for model-based robotic control. In *Conference on Robot Learning*, pp. 800–811. PMLR, 2023.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Badrinath, K. P. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

Bahadori, M. T. and Heckerman, D. E. Debiasing concept-based explanations with causal analysis. *arXiv preprint arXiv:2007.11500*, 2020.

Bai, C., Wang, L., Han, L., Garg, A., Hao, J., Liu, P., and Wang, Z. Dynamic bottleneck for robust self-supervised exploration. *Advances in Neural Information Processing Systems*, 34:17007–17020, 2021.

Bontempelli, A., Teso, S., Giunchiglia, F., and Passerini, A. Concept-level debugging of part-prototype networks. *arXiv preprint arXiv:2205.15769*, 2022.

Chebotar, Y., Handa, A., Makoviychuk, V., Macklin, M., Issac, J., Ratliff, N., and Fox, D. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8973–8979. IEEE, 2019.

Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble-based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

Clavier, P., Pennec, E. L., and Geist, M. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

De Haan, P., Jayaraman, D., and Levine, S. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Deng, Z., Fu, Z., Wang, L., Yang, Z., Bai, C., Wang, Z., and Jiang, J. Score: Spurious correlation reduction for offline reinforcement learning. *arXiv preprint arXiv:2110.12468*, 2021.

Derman, E. and Mannor, S. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

Ding, W., Lin, H., Li, B., and Zhao, D. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. *arXiv preprint arXiv:2207.09081*, 2022.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Goyal, V. and Grand-Clement, J. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Hagos, M. T., Curran, K. M., and Mac Namee, B. Identifying spurious correlations and correcting them with an explanation-based learning. *arXiv preprint arXiv:2211.08285*, 2022.

Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Han, S., Su, S., He, S., Han, S., Yang, H., and Miao, F. What is the solution for state adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*, 2022.

Hansen, N. and Wang, X. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13611–13617. IEEE, 2021.

Hansen, N., Su, H., and Wang, X. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in neural information processing systems*, 34:3680–3693, 2021.

Ho, C. P., Petrik, M., and Wiesemann, W. Fast bellman updates for robust mdps. In *International Conference on Machine Learning*, pp. 1979–1988. PMLR, 2018.

Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

James, S., Wohlhart, P., Kalakrishnan, M., Kalashnikov, D., Irpan, A., Ibarz, J., Levine, S., Hadsell, R., and Bousmalis, K. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12627–12637, 2019.

Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., and Fan, L. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

Kaufman, D. L. and Schaefer, A. J. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

Kaushik, D., Hovy, E., and Lipton, Z. C. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Kostrikov, I., Yarats, D., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Larsen, K. G. and Skou, A. Bisimulation through probabilistic testing (preliminary report). In *Proceedings of the 16th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pp. 344–352, 1989.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020.

Le, H., Jiang, N., Agarwal, A., Dudík, M., Yue, Y., and Daumé III, H. Hierarchical imitation and reinforcement learning. In *International conference on machine learning*, pp. 2917–2926. PMLR, 2018.

Lu, C., Huang, B., Wang, K., Hernández-Lobato, J. M., Zhang, K., and Schölkopf, B. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*, 2020.

Lu, Y., Shen, Y., Zhou, S., Courville, A., Tenenbaum, J. B., and Gan, C. Learning task decomposition with ordered memory policy network. *arXiv preprint arXiv:2103.10972*, 2021.

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., and Bergstra, J. Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pp. 561–591. PMLR, 2018.

Mehta, B., Diaz, M., Golemo, F., Pal, C. J., and Paull, L. Active domain randomization. In *Conference on Robot Learning*, pp. 1162–1176. PMLR, 2020.

Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., and Peters, J. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022.

Nauta, M., Walsh, R., Dubowski, A., and Seifert, C. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40, 2021.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Pitis, S., Creager, E., and Garg, A. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

Pitis, S., Creager, E., Mandlekar, A., and Garg, A. Mocoda: Model-based counterfactual data augmentation. *arXiv preprint arXiv:2210.11287*, 2022.

Plumb, G., Ribeiro, M. T., and Talwalkar, A. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021.

Qiaoben, Y., Zhou, X., Ying, C., and Zhu, J. Strategically-timed state-observation attacks on deep reinforcement learning agents. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., and Fergus, R. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:5402–5415, 2021.

Ruiz, N., Schulter, S., and Chandraker, M. Learning to simulate. *arXiv preprint arXiv:1810.02513*, 2018.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Seo, S., Lee, J.-Y., and Han, B. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16742–16751, 2022.

Shi, L. and Chi, Y. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Smirnova, E., Dohmatob, E., and Mary, J. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Sun, K., Liu, Y., Zhao, Y., Yao, H., Jui, S., and Kong, L. Exploring the training robustness of distributional reinforcement learning against noisy state observations. *arXiv preprint arXiv:2109.08776*, 2021.

Tamar, A., Mannor, S., and Xu, H. Scaling up robust mdps using function approximation. In *International conference on machine learning*, pp. 181–189. PMLR, 2014.

Tan, K. L., Esfandiari, Y., Lee, X. Y., Sarkar, S., et al. Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*, pp. 3959–3964. IEEE, 2020.

Tennenholtz, G., Hallak, A., Dalal, G., Mannor, S., Chechik, G., and Shalit, U. On covariate shift of latent confounders in imitation and reinforcement learning. *arXiv preprint arXiv:2110.06539*, 2021.

Tessler, C., Efroni, Y., and Mannor, S. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224. PMLR, 2019.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Wang, K., Kang, B., Shao, J., and Feng, J. Improving generalization in reinforcement learning with mixture regularization. *Advances in Neural Information Processing Systems*, 33:7968–7978, 2020.

Wang, L., Yang, Z., and Wang, Z. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34:21164–21175, 2021.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. A finite sample complexity bound for distributionally robust q-learning. *arXiv preprint arXiv:2302.13203*, 2023.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34, 2021.

Weng, J., Chen, H., Yan, D., You, K., Duburcq, A., Zhang, M., Su, Y., Su, H., and Zhu, J. Tianshou: A highly modularized deep reinforcement learning library. *Journal of Machine Learning Research*, 23(267):1–6, 2022. URL http://jmlr.org/papers/v23/21-1127.html.

Wiesemann, W., Kuhn, D., and Rustem, B. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Wolff, E. M., Topcu, U., and Murray, R. M. Robust control of uncertain markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3372–3379. IEEE, 2012.

Wu, S., Yuksekgonul, M., Zhang, L., and Zou, J. Discover and cure: Concept-aware mitigation of spurious correlation. *arXiv preprint arXiv:2305.00650*, 2023.

Xie, A., Sodhani, S., Finn, C., Pineau, J., and Zhang, A. Robust policy learning over multiple uncertainty sets. In *International Conference on Machine Learning*, pp. 24414–24429. PMLR, 2022.

Xiong, Z., Eappen, J., Zhu, H., and Jagannathan, S. Defending observation attacks in deep reinforcement learning via detection and denoising. *arXiv preprint arXiv:2206.07188*, 2022.

Xu, H. and Mannor, S. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Xu, M., Huang, P., Niu, Y., Kumar, V., Qiu, J., Fang, C., Lee, K.-H., Qi, X., Lam, H., Li, B., et al. Group distributionally robust reinforcement learning with hierarchical latent variables. In *International Conference on Artificial Intelligence and Statistics*, pp. 2677–2703. PMLR, 2023a.

Xu, Z., Panaganti, K., and Kalathil, D. Improved sample complexity bounds for distributionally robust reinforcement learning. *arXiv preprint arXiv:2303.02783*, 2023b.

Yang, W., Zhang, L., and Zhang, Z. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

Yoo, M., Cho, S., and Woo, H. Skills regularized task decomposition for multi-task offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 37432–37444, 2022.

Zakharov, S., Kehl, W., and Ilic, S. Deceptionnet: Network-driven domain randomization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 532–541, 2019.

Zhang, A., McAllister, R., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020a.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33: 21024–21037, 2020b.

Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021a.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Zhang, Y., Gong, M., Liu, T., Niu, G., Tian, X., Han, B., Schölkopf, B., and Zhang, K. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021b.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

Zhu, Y., Wong, J., Mandlekar, A., Martín-Martín, R., Joshi, A., Nasiriany, S., and Zhu, Y. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

Appendix

A. Additional Preliminaries

Value function and Q-function of standard MDPs. To represent the long-term cumulative reward, the value function $V_t^{\pi,P}: \mathcal{S} \to \mathbb{R}$ and Q-value function $Q_t^{\pi,P}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ associated with policy π at step t are defined as $V_t^{\pi,P}(s) = \mathbb{E}_{\pi,P}\big[\sum_{k=t}^{T} r_k(s_k,a_k) \mid s_k = s\big]$ and $Q_t^{\pi,P}(s,a) = \mathbb{E}_{\pi,P}\big[\sum_{k=t}^{T} r_k(s_k,a_k) \mid s_t = s, a_t = a\big]$, where the expectation is taken over the sample trajectory $\{(s_t,a_t)\}_{1\le t\le T}$ generated following $a_t \sim \pi_t(\cdot \mid s_t)$ and $s_{t+1} \sim P_t(\cdot \mid s_t, a_t)$.

Robust Markov decision processes (RMDPs). As a robust variant of standard MDPs motivated by distributionally robust optimization, RMDP is a natural formulation to promote robustness to the uncertainty of the transition probability kernel (Iyengar, 2005; Shi & Chi, 2022), represented as $\mathcal{M}_{\mathsf{rob}} = \big\{\mathcal{S}, \mathcal{A}, T, r, \mathcal{U}^\sigma(P^0)\big\}$. Here, we reuse the definitions of $\mathcal{S}, \mathcal{A}, T, r$ in standard MDPs, and denote $\mathcal{U}^\sigma(P^0)$ as an uncertainty set of probability transition kernels centered around a nominal transition kernel $P^0 = \{P_t^0\}_{1\le t\le T}$ measured by some 'distance' function ρ with radius σ. In particular, the uncertainty set obeying the (s,a)-rectangularity (Wiesemann et al., 2013) can be defined over all (s,a) state-action pairs at each time step t as

$$\mathcal{U}^\sigma(P^0) \coloneqq \otimes\, \mathcal{U}^\sigma(P_{t,s,a}^0), \qquad \mathcal{U}^\sigma(P_{t,s,a}^0) \coloneqq \big\{P_{t,s,a} \in \Delta(\mathcal{S}) : \rho\left(P_{t,s,a}, P_{t,s,a}^0\right) \le \sigma\big\}, \tag{4}$$

where $\otimes$ denotes the Cartesian product. Here, $P_{t,s,a} \coloneqq P_t(\cdot \mid s,a) \in \Delta(\mathcal{S})$ and $P_{t,s,a}^0 \coloneqq P_t^0(\cdot \mid s,a) \in \Delta(\mathcal{S})$ denote the transition kernel P_t or P_t^0 at each state-action pair (s,a) respectively. Consequently, the next state s_{t+1} follows $s_{t+1} \sim P_t(\cdot \mid s_t, a_t)$ for any $P_t \in \mathcal{U}^\sigma(P_{t,s_t,a_t}^0)$, namely, s_{t+1} can be generated from any transition kernel belonging to the uncertainty set $\mathcal{U}^\sigma(P_{t,s_t,a_t}^0)$ rather than a fixed one in standard MDPs. As a result, for any policy π, the corresponding *robust value function* and *robust Q function* are defined as

$$V_t^{\pi,\sigma}(s) \coloneqq \inf_{P \in \mathcal{U}^\sigma(P^0)} V_t^{\pi,P}(s), \qquad Q_t^{\pi,\sigma}(s,a) \coloneqq \inf_{P \in \mathcal{U}^\sigma(P^0)} Q_t^{\pi,P}(s,a), \tag{5}$$

which characterize the cumulative reward in the worst case when the transition kernel is within the uncertainty set $\mathcal{U}^\sigma(P^0)$. Using samples generated from the nominal transition kernel P^0, the goal of RMDPs is to find an optimal robust policy that maximizes $V_1^{\pi,\sigma}$ when $t = 1$, i.e., perform optimally in the worst case when the transition kernel of the test environment lies in a prescribed uncertainty set $\mathcal{U}^\sigma(P^0)$.

B. Additional Related Works

In this section, besides the most related formulation, robust RL introduced in Sec 3.3, we also introduce some other related RL problem formulations partially shown in Figure 3. Then, we limit our discussion to mainly two lines of work that are related to ours: (1) promoting robustness in RL; (2) concerning the spurious correlation issues in RL.

B.1. Related RL formulations

Robustness to noisy state: POMDPs and SA-MDPs. State-noisy MDPs refer to the RL problem that the agent can only access and choose the action based on a noisy observation rather than the true state at each step, including two existing types of problems: Partially observable MDPs (POMDPs) and state-adversarial MDPs (SA-MDPs), shown in Figure 3(b). In particular, at each step t, in POMDPs, the observation o_t is generated by a fixed probability transition $\mathcal{O}(\cdot \mid s_t)$ (we refer to the case that o_t only depends on the state s_t but not action); for state-adversarial MDPs, the observation is an adversary $\nu(s_t)$ against and thus determined by the conducted policy, leading to the worst performance by perturbing the state in a small set around itself. To against the state perturbation, both POMDPs, and SA-MDPs are indeed robust to the noisy observation, or called agent-observed state, but not the real state that transitions to the environment and next steps. In contrast, our RSC-MDPs propose the robustness to the real state shift that will directly transition to the next state in the environment, involving additional challenges induced by the appearance of out-of-distribution states.

Robustness to unobserved confounder: MDPUC and confounded MDPs. To address the misleading spurious correlations hidden in components of RL, people formulate RL problems as MDPs with some additional components – unobserved confounders. In particular, the Markov decision process with unobserved confounders (MDPUC) (Wu et al., 2023) serves as a general framework to concern all types of possible spurious correlations in RL problems – at each step, the state, action, and reward are all possibly influenced by some unobserved confounder, shown in Figure 2(d); confounded MDPs (Wang et al., 2021) mainly concerns the misleading correlation between the current action and the next state, illustrated in Figure 3(e). The proposed state-confounded MDPs (SC-MDPs) can be seen as a specified type of MDPUC that focus on breaking the spurious correlation between different parts of the state space itself (different from confounded MDPs which consider the correlation between action and next state), motivated by various real-world applications in self-driving and control tasks. In addition, the proposed formulation is more flexible and can work in both online and offline RL settings.

Contexual MDPs (CDMPs). A contextual MDP (CMDP) (Hallak et al., 2015) is basically a set of standard MDPs sharing the same state and action space but specified by different contexts within a context space. In particular, the transition kernel, reward, and action of a CMDP are all determined by a (possibly unknown) fixed context. The proposed robust state-confounded MDPs (RSC-MDPs) are similar to CMDPs if we cast the unobserved confounder as the context in CMDPs, while different in two aspects: (1) In a CMDP, the context is fixed throughout an episode, while the unobserved confounder in RSC-MDPs can vary as $\{c_t\}_{1\le t\le T}$; (2) In the online setting, the goal of CMDP is to beat the optimal policy depending on the context, while RSC-MDPs seek to learn the optimal policy that does not depend on the confounder $\{c_t\}_{1\le t\le T}$.

B.2. Related literature of robustness in RL

Robust RL (robust MDPs). Concerning the robust issues in RL, a large portion of works focus on robust RL with explicit uncertainty of the transition kernel, which is well-posed and a natural way to consider the uncertainty of the environment (Iyengar, 2005; Xu & Mannor, 2010; Wolff et al., 2012; Kaufman & Schaefer, 2013; Ho et al., 2018; Smirnova et al., 2019; Ho et al., 2021; Goyal & Grand-Clement, 2022; Derman & Mannor, 2020; Tamar et al., 2014; Badrinath & Kalathil, 2021). However, to define the uncertainty set for the environment, most existing works use task structure-agnostic and heuristic 'distance' such as KL divergence and total variation (Yang et al., 2022; Panaganti & Kalathil, 2022; Zhou et al., 2021; Shi & Chi, 2022; Xu et al., 2023b; Wang et al., 2023; Clavier et al., 2023; Wang & Zou, 2021) to measure the shift between the training and test transition kernel, leading to a homogeneous (almost structure-free) uncertainty set around the state space. In contrast, we consider a more general uncertainty set that enables the robustness to a task-dependent heterogeneous uncertainty set shaped by unobserved confounder and causal structure, in order to break the spurious correlation hidden in different parts of the state space.

Robustness in RL Despite the remarkable success that standard RL has achieved, current RL algorithms are still limited since the agent is vulnerable if the deployed environment is subject to uncertainty and even structural changes. To address these challenges, a recent line of RL works begins to concern robustness to the uncertainty or changes over different components of MDPs – state, action, reward, and transition kernel, where a review (Moos et al., 2022) can be referred to. Besides robust RL framework concerning the shift of the transition kernel and reward, to promote robustness in RL, there exist various works (Tessler et al., 2019; Tan et al., 2020) that consider the robustness to action uncertainty, i.e., the deployed action in the environment is distorted by an adversarial agent smoothly or circumstantially; some works (Zhang et al., 2020b; 2021a; Han et al., 2022; Qiaoben et al., 2021; Sun et al., 2021; Xiong et al., 2022) investigate the robustness to the state uncertainty including but not limited to the introduced POMDPs and SA-MDPs in Appendix B.1, where the agent chooses the action based on observation – the perturbed state determined by some restricted noise or adversarial attack. The

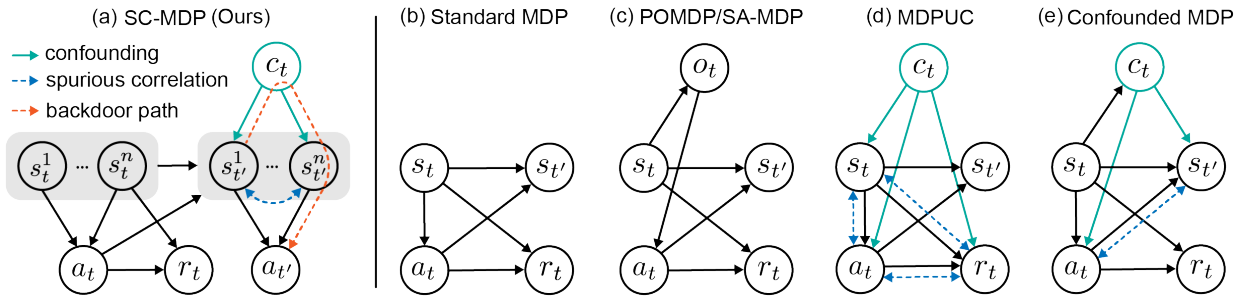

Figure 4. The probabilistic graphs of our formulation (SC-MDP) and other related formulations.

proposed RSC-MDPs can be regarded as addressing the state uncertainty since the shift of the unobserved confounder leads to state perturbation. In contrast, RSC-MDPs consider the out-of-distribution of the real state that will directly influence the subsequent transition in the environment, but not the observation in POMDPs and SA-MDPs that will not directly influence the environment.

B.3. Related literature of spurious correlation in RL

Confounder in RL. These works mainly focus on the confounder between action (treatment) and state (effect), which is a long-standing problem that exists in the causal inference area. However, we find that the confounder may cause problems from another perspective, where the confounder is built upon different dimensions of the state variable. Some people focus on the confounder between action and state, which is common in offline settings since the dataset is fixed and intervention is not allowed. But in the online setting, actions are controlled by an agent and intervention is available to eliminate spurious correlation. (Deng et al., 2021) reduces the spurious correlation between action and state in the offline setting. (Bai et al., 2021) deal with environment-irrelevant white noise; possible shift + causal (Tennenholtz et al., 2021). The confounder problem is usually easy to solve since agents can interact with the environment to do interventions. However, different from most existing settings, we find that even with the capability of intervention, the confounding between dimensions in states cannot be fully eliminated. Then the learned policy is heavily influenced if these confounder change during testing.

Invariant Feature learning. The problem of spurious correlation has attracted attention in the supervised learning area for a long time and many solutions are proposed to learn invariant features to eliminate spurious correlations. A general framework to remedy the ignorance of spurious correlation in empirical risk minimization (ERM) is invariant risk minimization (IRM) (Arjovsky et al., 2019). Other works tackle this problem with group distributional robustness (Sagawa et al., 2019), adversarial robustness (Zhang et al., 2021b), and contrastive learning (Zhang et al., 2022). These methods are also adapted to sequential settings. The idea of increasing the robustness of RL agents by training agents on multiple environments has been shown in previous works (Xie et al., 2022; Zhang et al., 2020a;a). However, a shared assumption among these methods is that multiple environments with different values of confounder are accessible, which is not always true in the real world.

Counterfactual Data Augmentation in RL. One way to simulate multiple environments is data augmentation. However, most data augmentation works (Laskin et al., 2020; Wang et al., 2020; Yarats et al., 2021; Kostrikov et al., 2020; Hansen et al., 2021; Raileanu et al., 2021; Hansen & Wang, 2021) apply image transformation to raw inputs, which requires strong domain knowledge for image manipulation and cannot be applied to other types of inputs. In RL, the dynamic model and reward model follow certain causal structures, which allow counterfactual generation of new transitions based on the collected samples. This line of work, named counterfactual data augmentation, is very close to this work. Deep generative models (Lu et al., 2020) and adversarial examples (Agarwal & Chinchali, 2023) are considered for the generation to improve sample efficiency in model-based RL. CoDA (Pitis et al., 2020) and MocoDA (Pitis et al., 2022) leverage the concept of locally factored dynamics to randomly stitch components from different trajectories. However, the assumption of local causality may be limited.

Domain Randomization. If we are allowed to control the data generation process, e.g., the underlying mechanism of the simulator, we can apply the golden rule in causality – Randomized Controlled Trial (RCT). The well-known technic, domain randomization (Tobin et al., 2017), exactly follows the idea of RCT, which randomly perturb the internal state of the experiment in simulators. Later literature follows this direction and develops variants including randomization

Algorithm 1 RSC-SAC Training

Input: policy π, data buffer $\mathcal{D}$, transition model P_θ, ratio of modified data β
for $t \in [1, T]$ **do**
  Sample action $a_t \sim \pi(\cdot|s_t)$
  $(s_{t+1}, r_t) \leftarrow \text{Env}(s_t, a_t)$
  Add buffer $\mathcal{D} = \mathcal{D} \cup \{s_t, a_t, s_{t+1}, r_t\}$
  for sample batch $\mathcal{B} \in \mathcal{D}$ **do**
    Randomly select $\beta\%$ data in $\mathcal{B}$
    Modify s_t in selected data with (6)
    $(\hat{s}_{t+1}, \hat{r}_t) \sim P_\theta(s_t, a_t, \mathcal{G}_\phi)$
    Replace data with $(s_t, a_t, \hat{s}_{t+1}, \hat{r}_t)$
    $\mathcal{L} = \|s_{t+1} - \hat{s}_{t+1}\|_2^2 + \|r_t - \hat{r}_t\|_2^2$
    Update θ and ϕ with $\mathcal{L} + \lambda\|\boldsymbol{G}\|_p$
    Update π with SAC algorithm
  end for
end for

guided by downstream tasks in the target domain (Ruiz et al., 2018; Mehta et al., 2020), randomization to match real-world distributions (James et al., 2019; Chebotar et al., 2019), and randomization to minimize data divergence (Zakharov et al., 2019). However, it is usually impossible to randomly manipulate internal states in most situations in the real world. In addition, determining which variables to randomize is even harder given so many factors in complex systems.

Discovering Spurious Correlations Detecting spurious correlations helps models remove features that are harmful to generalization. Usually, domain knowledge is required to find such correlations (Clark et al., 2019; Kaushik et al., 2019; Nauta et al., 2021). However, when prior knowledge is accessible, technics such as clustering can also be used to reveal spurious attributes (Wu et al., 2023; Sohoni et al., 2020; Seo et al., 2022). When human inspection is available, recent works (Plumb et al., 2021; Hagos et al., 2022; Abid et al., 2022) also use explainability techniques to find spurious correlations. Another area for discovery is concept-level and interactive debugging (Bontempelli et al., 2022; Bahadori & Heckerman, 2020), which leverage concepts or human feedback to perform debugging.

C. Details about RSC-SAC Algorithm

C.1. Distribution of confounder

As we have no prior knowledge about the confounder, we choose to approximate the effect of perturbing them without explicitly estimating the distribution P^c. We first randomly select a dimension i from the state s_t to apply perturbation and then assign the dimension i of s_t with a heuristic rule. We select the value from another sample s_k that has the most different value from s_t in dimension i and the most similar value to s_t in the remaining dimensions. Formally, this process solves the following optimization problem to select sample k from a batch of K samples:

$$s_t^i \leftarrow s_k^i,\ k = \arg\max \frac{\|s_t^i - s_k^i\|_2^2}{\sum_{\neg i} \|s_t^{\neg i} - s_k^{\neg i}\|_2^2},\ \ k \in \{1, ..., K\} \tag{6}$$

where s_t^i and $s_t^{\neg i}$ means dimension i of s_t and other dimensions of s_t except for i, respectively. Intuitively, permuting the dimension of two samples breaks the spurious correlation and remains the most semantic meaning of the state space. However, this permutation sometimes also breaks the true cause and effect between dimensions, leading to a performance drop. The trade-off between robustness and performance (Xu et al., 2023a) is a long-standing dilemma in the robust optimization framework, which we will leave to future work.

C.2. Learning of structural causal model

With the perturbed state s_t, we then learn an SCM to predict the next state and reward considering the effect of the action on the previous state. This model contains a causal graph to achieve better generalization to unseen state-action pairs. Specifically, we simultaneously learn the model parameter and discover the underlying causal graph in a fully differentiable way with $(\hat{s}_{t+1}, \hat{r}_t) \sim P_\theta(s_t, a_t, \mathcal{G}_\phi)$, where θ is the parameter of the neural network of the dynamic model

and $\phi \in \mathbb{R}^{(n+d_{\mathcal{A}})\times(n+1)}$ is the parameter to represent causal graph $\mathcal{G}$ between $\{s_t, a_t\}$ and $\{s_{t+1}, r_t\}$. This graph is represented by a binary adjacency matrix $\boldsymbol{G}$, where $1/0$ means the existence/absence of an edge. P_θ has an encoder-decoder structure with matrix $\boldsymbol{G}$ as an intermediate linear transformation. The encoder takes state and action in and outputs features $f_e \in \mathbb{R}^{(n+d_{\mathcal{A}})\times d_f}$ for each dimension, where d_f is the dimension of the feature. Then, the causal graph is multiplied to generate the feature for the decoder $f_d = f_e^T \boldsymbol{G} \in \mathbb{R}^{d_f \times (n+1)}$. The decoder takes in f_d and outputs the next state and reward. The detailed architecture of this causal transition model can be found in Appendix F.2.

The objective for training this model consists of two parts, one is the supervision signal from collected data $\|s_{t+1} - \hat{s}_{t+1}\|_2^2 + \|r_t - \hat{r}_t\|_2^2$, and the other is a penalty term $\lambda\|\boldsymbol{G}\|_p$ with weight λ to encourage the sparsity of the matrix $\boldsymbol{G}$. The penalty is important to break the spurious correlation between dimensions of state since it forces the model to eliminate unnecessary inputs for prediction.

D. Theoretical Analyses

D.1. Theoretical guarantees of RSC-MDPs: advantages of semantic uncertainty

To theoretically understand the advantages of the proposed robust formulation RSC-MDPs with comparison to prior works, especially RMDPs, the following theorem verifies that RSC-MDPs enable additional robustness to fierce semantic attack besides small model perturbation or noise considered in RMDPs. The proof is postponed to Appendix D.2.

Theorem 1. *Consider some standard MDPs $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P^0, T, r\}$, equivalently represented as an SC-MDP $\mathcal{M}_{\mathsf{sc}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, P^c\}$ with $\mathcal{C} \coloneqq [0,1]$, and the widely-used total deviation as the 'distance' function ρ to measure the uncertainty set (namely, the admissible uncertainty level obeys $\sigma \in [0,1]$). For the corresponding RMDP $\mathcal{M}_{\mathsf{rob}}$ with the uncertainty set $\mathcal{U}^{\sigma_1}(P^0)$, and the proposed RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, \mathcal{U}^{\sigma_2}(P^c)\}$, the optimal robust policy $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}$ associated with $\mathcal{M}_{\mathsf{rob}}$ and $\pi_{\mathsf{RSC}}^{\star,\sigma_2}$ associated with $\mathcal{M}_{\mathsf{sc\text{-}rob}}$ obey: given $\sigma_2 \in \left(\frac{3}{4}, 1\right]$, there exist RSC-MDPs with some initial distribution ϕ such that*

$$\widetilde{V}_1^{\pi_{\mathsf{RSC}}^{\star,\sigma_2},\sigma_2}(\phi) - \widetilde{V}_1^{\pi_{\mathsf{RMDP}}^{\star,\sigma_1},\sigma_2}(\phi) \geq \frac{T}{4}, \quad \forall \sigma_1 \in [0,1]. \tag{7}$$

In words, Theorem 1 reveals a fact about the proposed RSC-MDPs: *RSC-MDPs could succeed in intense semantic attacks while RMDPs fail*. As shown by (7), when fierce semantic shifts appear between the training and test scenarios – perturbing the unobserved confounder in a large uncertainty set $\mathcal{U}^{\sigma_2}(P^c)$, solving RSC-MDPs with $\pi_{\mathsf{RSC}}^{\star,\sigma_2}$ succeed in testing while $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}$ trained by solving RMDPs can fail catastrophically. The proof is achieved by constructing hard constants of RSC-MDPs that RMDPs could not cope with due to inherent limitations. Moreover, this advantage of RSC-MDPs is consistent with and verified by the empirical performance evaluation in Section 5.2 **R1**.

D.2. Proof of Theorem 1

Constructing a hard instance of the standard MDP. In this section, we consider the following standard MDP instance $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P^0, T, r\}$, where $\mathcal{S} = \{[0,0], [0,1], [1,0], [1,1]\}$ is the state space consisting of four elements in dimension $n = 2$, and $\mathcal{A} = \{0,1\}$ is the action space with only two options. The transition kernel $P^0 = \{P_t^0\}_{1\leq t\leq T}$ at different time steps $1 \leq t \leq T$ is defined as

$$P_1^0(s' \mid s, a) = \begin{cases} \mathbb{1}(s' = [0,0])\mathbb{1}(a=0) + \mathbb{1}(s' = [0,1])\mathbb{1}(a=1) & \text{if} (s,a) = ([0,0], a) \\ \mathbb{1}(s' = s) & \text{otherwise} \end{cases}, \tag{8}$$

and

$$P_t^0(s' \mid s, a) = \mathbb{1}(s' = s), \quad \forall (t, s, a) \in \{2, 3, \cdots, T\} \times \mathcal{S} \times \mathcal{A}. \tag{9}$$

Note that this transition kernel P^0 ensures the next state transitioned from the state $[0,0]$ is either $[0,0]$ or $[0,1]$. The reward function is specified as follows: for all time steps $1 \leq t \leq T$,

$$r_t(s,a) = \begin{cases} 1 & \text{if } s = [0,0] \text{ or } s = [1,1] \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

The equivalence to one SC-MDP. Then, we shall show that the constructed standard MDP $\mathcal{M}$ can be equivalently represented by one SC-MDP $\mathcal{M}_{\mathsf{sc}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, P^c\}$ with $\mathcal{C} \coloneqq [0,1]$, which yields the sequential observations

$\{s_t, a_t, r_t\}_{1\le t\le T}$ induced by any policy and any initial state distribution in two processes are identical. To specify, $\mathcal{S}, \mathcal{A}, T, r$ are kept the same as $\mathcal{M}$. Here, $\{\mathcal{P}_t^i\}$ shall be specified in a while, which determines the transition to each dimension of the next state conditioned on the current state, action, and confounder for all time steps, i.e., $s_{t+1}^i \sim \mathbb{E}_{c_t\sim P_t^c}\left[\mathcal{P}_t^i(\cdot \mid s_t, a_t, c_t)\right]$ for any i-th dimension of the state ($i \in \{1, 2\}$ and all timestep $1 \le t \le T$. For convenience, we denote $\mathcal{P}_t := [\mathcal{P}_t^1, \mathcal{P}_t^2] \in \Delta(\mathcal{S})$ as the transition kernel towards the next state, namely, $s_{t+1} \sim \mathbb{E}_{c_t\sim P_t^c}\left[\mathcal{P}_t(\cdot \mid s_t, a_t, c_t)\right]$.

To ensure the marginalized transition probability from any state-action pair (s_t, a_t) to the next state s_{t+1} in $\mathcal{M}_{\mathsf{sc}}$ aligns with the one in the MDP $\mathcal{M}$, we set

$$P_t^c(c) = \mathbb{1}(c = 0), \quad \forall 1 \le t \le T. \tag{11}$$

In addition, before introducing the transition kernel $\{\mathcal{P}_t^i\}$ of the SC-MDP $\mathcal{M}_{\mathsf{sc}}$, we introduce an auxiliary transition kernel $P^{\mathrm{sc}} = \{P_t^{\mathrm{sc}}\}$ as follows:

$$P_1^{\mathrm{sc}}(s' \mid s, a) = \begin{cases} \mathbb{1}(s' = [1, 0])\mathbb{1}(a = 0) + \mathbb{1}(s' = [1, 1])\mathbb{1}(a = 1) & \text{if}(s, a) = ([0, 0], 0) \\ \mathbb{1}(s' = s) & \text{otherwise} \end{cases}, \tag{12}$$

and

$$P_t^{\mathrm{sc}}(s' \mid s, a) = \mathbb{1}(s' = s), \quad \forall (t, s, a) \in \{2, 3, \cdots, T\} \times \mathcal{S} \times \mathcal{A}. \tag{13}$$

It can be observed that P^{sc} is similar to P^0 except for the transition in the state $[0, 0]$.

Armed with this transition kernel P^{sc}, the $\{\mathcal{P}_t^i\}$ of the SC-MDP $\mathcal{M}_{\mathsf{sc}}$ is set to obey

$$\mathcal{P}_1(s' \mid s, a, c) = \begin{cases} (1 - c)P_1^0(s' \mid s, a) + cP_1^{\mathrm{sc}}(s' \mid s, a) & \text{if}(s, a) = ([0, 0], a) \\ \mathbb{1}(s' = s) & \text{otherwise} \end{cases}, \tag{14}$$

and

$$\mathcal{P}_t(s' \mid s, a, c) = \mathbb{1}(s' = s), \quad \forall (t, s, a, c) \in \{2, 3, \cdots, T\} \times \mathcal{S} \times \mathcal{A} \times \mathcal{C}. \tag{15}$$

With the above preparation, we are ready to verify that the marginalized transition from the current state and action to the next state in the SC-MDP $\mathcal{M}_{\mathsf{sc}}$ is identical to the one in MDP $\mathcal{M}$: for all $(t, s_t, a_t, s_{t+1}) \in \{1, 2, \cdots, T\} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$\mathbb{P}(s_{t+1} \mid s_t, a_t) = \mathbb{E}_{c_t\sim P_t^c}\left[\mathcal{P}_t(s_{t+1} \mid s_t, a_t, c_t)\right] = \mathcal{P}_t(s_{t+1} \mid s_t, a_t, 0) = P^0(s_{t+1} \mid s_t, a_t) \tag{16}$$

where the second equality holds by the definition of P^c in equation 11, and the last equality holds by the definitions of P^0 (see equation 8 and equation 9) and $\mathcal{P}$ (see equation 14 and equation 15).

In summary, we verified that the standard MDP $\mathcal{M} = \left\{\mathcal{S}, \mathcal{A}, P^0, T, r\right\}$ is equal to the above specified SC-MDP $\mathcal{M}_{\mathsf{sc}}$.

Defining the corresponding RMDP and RSC-MDP. Equipped with the equivalent MDP $\mathcal{M}$ and SC-MDP $\mathcal{M}_{\mathsf{sc}}$, people could consider the robust variants of them respectively — a RMDP $\mathcal{M}_{\mathsf{rob}} = \left\{\mathcal{S}, \mathcal{A}, \mathcal{U}^{\sigma_1}(P^0), T, r\right\}$ with the uncertainty level σ_1, and the proposed RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}} = \left\{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, \mathcal{U}^{\sigma_2}(P^c)\right\}$ with the uncertainty level σ_2.

In this section, without loss of generality, we consider total deviation as the 'distance' function ρ for the uncertainty sets of both RMDP $\mathcal{M}_{\mathsf{rob}}$ and RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}}$, i.e., for any probability vectors $P', P \in \Delta(\mathcal{C})$ (or $P', P \in \Delta(\mathcal{S})$), $\rho(P', P) := \frac{1}{2}\left\|P' - P\right\|_1$. Consequently, for any uncertainty set $\sigma \in [0, 1]$, the uncertainty set $\mathcal{U}^{\sigma_1}(P^0)$ of the RMDP (see equation 4) and $\mathcal{U}^{\sigma_2}(P^c)$ of the RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}}$ (see equation 2) are defined as follows:

$$\begin{aligned} \mathcal{U}^{\sigma}(P^0) &:= \otimes\, \mathcal{U}^{\sigma}(P_{t,s,a}^0), \qquad \mathcal{U}^{\sigma}(P_{t,s,a}^0) := \left\{P_{t,s,a} \in \Delta(\mathcal{S}) : \frac{1}{2}\left\|P_{t,s,a} - P_{t,s,a}^0\right\|_1 \le \sigma\right\}, \\ \mathcal{U}^{\sigma}(P^c) &:= \otimes\, \mathcal{U}^{\sigma}(P_t^c), \qquad \mathcal{U}^{\sigma}(P_t^c) := \left\{P \in \Delta(\mathcal{C}) : \frac{1}{2}\left\|P - P_t^c\right\|_1 \le \sigma\right\}. \end{aligned} \tag{17}$$

To continue, the proof is established by specifying the robust optimal policy $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}$ associated with $\mathcal{M}_{\mathsf{rob}}$ and $\pi_{\mathsf{RSC}}^{\star,\sigma_2}$ associated with $\mathcal{M}_{\mathsf{sc\text{-}rob}}$ and then compare their performance on RSC-MDP with some initial state distribution.

The performance comparisons between $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}$ of RMDP $\mathcal{M}_{\mathsf{rob}}$ and $\pi_{\mathsf{RSC}}^{\star,\sigma_2}$ of RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}}$.

To begin, we introduce the following lemma which specifies the robust optimal policy $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}$ associated with the RMDP $\mathcal{M}_{\mathsf{rob}}$.

Lemma 1. *For any $\sigma_1 \in (0,1]$, the robust optimal policy and its corresponding robust SC-value functions satisfy*

$$\pi^{\star,\sigma_1}_{\mathsf{RMDP}}(0 \,|\, s) = 1, \qquad \text{for } s \in \mathcal{S}. \tag{18a}$$

In addition, we characterize the robust SC-value functions of the RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}}$ associated with any policy, combined with the robust optimal policy $\pi^{\star,\sigma_2}_{\mathsf{RSC}}$ of $\mathcal{M}_{\mathsf{sc\text{-}rob}}$ — the optimal robust SC-value functions, shown in the following lemma.

Lemma 2. *Consider any $\sigma_2 \in (\frac{3}{4}, 1]$ and the RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}} = \big\{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}^i_t\}, \mathcal{U}^{\sigma_2}(P^c)\big\}$. For any policy π, the corresponding robust SC-value functions satisfy*

$$\widetilde{V}^{\pi,\sigma_2}_1([0,0]) = 1 + (T-1) \inf_{P \in \mathcal{U}^{\sigma}(P^c_1)} \mathbb{E}_{c_1 \sim P}\Big[\pi_1(0 \,|\, [0,0])(1-c_1) + \pi_1(1 \,|\, [0,0])c_1\Big]. \tag{19a}$$

In addition, the optimal robust SC-value function and the robust optimal policy $\pi^{\star,\sigma_2}_{\mathsf{RSC}}$ of the RMDP $\mathcal{M}_{\mathsf{sc\text{-}rob}}$ obeys:

$$\widetilde{V}^{\pi^{\star,\sigma_2}_{\mathsf{RSC}},\sigma_2}_1([0,0]) = \widetilde{V}^{\star,\sigma_2}_1([0,0]) = 1 + \frac{T-1}{2}. \tag{20}$$

Applying Lemma 2 with policy $\pi = \pi^{\star,\sigma_1}_{\mathsf{RMDP}}$ in Lemma 1, one has

$$\widetilde{V}^{\pi^{\star,\sigma_1}_{\mathsf{RMDP}},\sigma_2}_1([0,0]) = 1 + (T-1) \inf_{P \in \mathcal{U}^{\sigma}_2(P^c_1)} \mathbb{E}_{c_1 \sim P}\Big[1 - c_1\Big] \le 1 + \frac{T-1}{4}, \tag{21}$$

where the last inequality holds by the probability distribution P obeying $P_1(0) = \frac{1}{4}$ and $P_1(1) = \frac{3}{4}$ is inside the uncertainty set $\mathcal{U}^{\sigma}_2(P^c_1)$.

Finally, putting equation 21 and equation 20 together, we complete the proof by showing that with the initial state distribution ϕ define as $\rho(s_1 = [0,0]) = 1$, we arrive at

$$\widetilde{V}^{\pi^{\star,\sigma_2}_{\mathsf{RSC}},\sigma_2}_1(\phi) - \widetilde{V}^{\pi^{\star,\sigma_1}_{\mathsf{RMDP}},\sigma_2}_1(\phi) = \widetilde{V}^{\star,\sigma_2}_1(\phi) - \widetilde{V}^{\pi^{\star,\sigma_1}_{\mathsf{RMDP}},\sigma_2}_1(\phi) \ge \frac{T-1}{4} \approx \frac{T}{4}. \tag{22}$$

D.2.1. Proof of Lemma 1

Specifying the minimum of the robust value functions in different states. For any uncertainty set $\sigma_1 \in (0,1]$, we first characterize the robust value function of any policy π over different states. To start, we denote the minimum of the robust value function over states at each time step t as below:

$$V^{\pi,\sigma_1}_{\min,t} \coloneqq \min_{s \in \mathcal{S}} V^{\pi,\sigma_1}_t(s) \ge 0, \tag{23}$$

where the last inequality holds by that the reward function defined in equation 10 is always non-negative. Obviously, there exists at least one state $s^{\pi}_{\min,t}$ that satisfies $V^{\pi,\sigma_1}_t(s^{\pi}_{\min,t}) = V^{\pi,\sigma_1}_{\min,t}$.

With this in mind, we shall verify that for any policy π,

$$\forall 1 \le t \le T: \quad V^{\pi,\sigma_1}_t([0,1]) = V^{\pi,\sigma_1}_t([1,0]) = 0. \tag{24}$$

To achieve this, we will use a recursive argument. First, the base case can be verified since when $t+1 = T+1$, the value functions are all zeros at $T+1$ step, i.e., $V^{\pi,\sigma_1}_{t+1}(s) = V^{\pi,\sigma_1}_{T+1}(s) = 0$ for all $s \in \mathcal{S}$. Then, the goal is to verify the following fact

$$V^{\pi,\sigma_1}_t([0,1]) = V^{\pi,\sigma_1}_t([1,0]) = 0 \tag{25}$$

with the assumption that $V^{\pi,\sigma_1}_{t+1}(s) = 0$ for any state $s = \{[0,1],[1,0]\}$. It is easily observed that for any policy π, the robust value function when state $s = \{[0,1],[1,0]\}$ at any time step t obeys

$$0 \le V^{\pi,\sigma_1}_t(s) = \mathbb{E}_{a \sim \pi_t(\cdot \,|\, s)}\left[r_t(s,a) + \inf_{P \in \mathcal{U}^{\sigma_1}(P^0_{t,s,a})} P V^{\pi,\sigma_1}_{t+1}\right]$$

$$\begin{aligned}&\overset{\text{(i)}}{=} 0+(1-\sigma_1)V_{t+1}^{\pi,\sigma_1}(s)+\sigma_1 V_{\min,t+1}^{\pi,\sigma_1} \overset{\text{(ii)}}{=} 0+\sigma_1 V_{\min,t+1}^{\pi,\sigma_1}\\ &\le 0+\sigma_1 V_{t+1}^{\pi,\sigma_1}(s)=0\end{aligned} \tag{26}$$

where (i) holds by $r_t(s,a)=0$ for all $s=\{[0,1],[1,0]\}$, the fact $P_t^0(s\,|\,s,a)=1$ (see equation 8 and equation 9), and the definition of the uncertainty set $\mathcal{U}^{\sigma_1}(P^0)$ in equation 17, (ii) follows from the recursive assumption $V_{t+1}^{\pi,\sigma_1}(s)=0$ for any state $s=\{[0,1],[1,0]\}$, and the last equality holds by $V_{\min,t+1}^{\pi,\sigma_1}\le V_{t+1}^{\pi,\sigma_1}(s)$ (see equation 23). Until now, we complete the proof for equation 25 and then verify equation 24.

Note that equation 24 direcly leads to

$$\forall 1\le t\le T:\quad V_{\min,t}^{\pi,\sigma_1}=0. \tag{27}$$

Considering the robust value function at state $[0,0]$**.** Armed with above facts, we are now ready to derive the robust value function for the state $[0,0]$.

When $2\le t\le T$, one has

$$\begin{aligned}V_t^{\pi,\sigma_1}([0,0])&=\mathbb{E}_{a\sim\pi_t(\cdot\,|\,[0,0])}\left[r_t([0,0],a)+\inf_{P\in\mathcal{U}^{\sigma_1}(P_{t,[0,0],a})}PV_{t+1}^{\pi,\sigma_1}\right]\\ &\overset{\text{(i)}}{=}1+\left[(1-\sigma_1)V_{t+1}^{\pi,\sigma_1}([0,0])+\sigma_1 V_{\min,t+1}^{\pi,\sigma_1}\right]\\ &=1+(1-\sigma_1)V_{t+1}^{\pi,\sigma_1}([0,0])\end{aligned} \tag{28}$$

where (i) holds by $r_t([0,0],a)=1$ for all $a\in\{0,1\}$ and the definition of P^0 (see equation 8 and equation 9), and the last equality arises from equation 27 .

Applying equation 28 recursively for $t,t+1,\cdots,T$ yields that

$$V_t^{\pi,\sigma_1}([0,0])=\sum_{k=t}^{T}(1-\sigma_1)^{k-t}\ge 1. \tag{29}$$

At the first step, the robust value function obeys:

$$\begin{aligned}V_1^{\pi,\sigma_1}([0,0])&=\mathbb{E}_{a\sim\pi_1(\cdot\,|\,[0,0])}\left[r_t([0,0],a)+\inf_{P\in\mathcal{U}^{\sigma_1}(P_{1,[0,0],a})}PV_2^{\pi,\sigma_1}\right]\\ &\overset{\text{(i)}}{=}1+\pi_1(0\,|\,[0,0])\inf_{P\in\mathcal{U}^{\sigma_1}(P_{1,[0,0],0})}PV_2^{\pi,\sigma_1}+\pi_1(1\,|\,[0,0])\inf_{P\in\mathcal{U}^{\sigma_1}(P_{1,[0,0],1})}PV_2^{\pi,\sigma_1}\\ &\overset{\text{(ii)}}{=}1+\pi_1(0\,|\,[0,0])\left[(1-\sigma_1)V_2^{\pi,\sigma_1}([0,0])+\sigma_1 V_{\min,2}^{\pi,\sigma_1}\right]\\ &\quad+\pi_1(1\,|\,[0,0])\left[(1-\sigma_1)V_2^{\pi,\sigma_1}([0,1])+\sigma_1 V_{\min,2}^{\pi,\sigma_1}\right]\\ &=1+\pi_1(0\,|\,[0,0])(1-\sigma_1)V_2^{\pi,\sigma_1}([0,0])\end{aligned} \tag{30}$$

where (i) holds by $r_t([0,0],a)=1$ for all $a\in\{0,1\}$, (ii) follows from the definition of P^0 (see equation 8 and equation 9), and the last equality arises from equation 24 and equation 27.

The optimal policy $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}$**.** Observing that the positive value of $V_2^{\pi,\sigma_1}([0,0])$ verified in equation 29, as $V_1^{\pi,\sigma_1}([0,0])$ is increasing monotically as $\pi_1(0\,|\,[0,0])$ is larger, we directly have that $\pi_{\mathsf{RMDP}}^{\star,\sigma_1}(0\,|\,[0,0])=1$.

Considering that the action does not influence the state transition for all other states $s\neq[0,0]$, without loss of generality, we choose the robust optimal policy to obey

$$\forall s\in\mathcal{S}:\quad \pi_{\mathsf{RMDP}}^{\star,\sigma_1}(0\,|\,s)=1. \tag{31}$$

D.2.2. Proof of Lemma 2

To begin with, for any uncertainty level $\sigma_2\in(\frac{1}{2},1]$ and any policy $\pi=\{\pi_t\}$, we consider the robust SC-value function $\widetilde{V}_1^{\pi,\sigma_2}$ of the RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}}$.

Deriving $\widetilde{V}_t^{\pi,\sigma_2}$ for $2 \le t \le T$. Towards this, for any $2 \le t \le T$ and $s \in \mathcal{S}$, one has

$$\begin{aligned}
\widetilde{V}_t^{\pi,\sigma_2}(s) &\overset{\text{(i)}}{=} \inf_{P \in \mathcal{U}^\sigma(P^c)} \widetilde{V}_t^{\pi,P}(s) = \inf_{P \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}_{a\sim\pi_t(s)}\left[\widetilde{Q}_t^{\pi,P}(s,a)\right] \\
&\overset{\text{(ii)}}{=} \inf_{P \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}_{a\sim\pi_t(s)}\left[r_t(s,a) + \mathbb{E}_{c_t\sim P}\left[\mathcal{P}_{t,s,a,c_t}\widetilde{V}_{t+1}^{\pi,\sigma}\right]\right] \\
&\overset{\text{(iii)}}{=} r_t(s,a) + \inf_{P \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}_{c_t\sim P}\left[\mathcal{P}_{t,s,a,c_t}\widetilde{V}_{t+1}^{\pi,\sigma}\right] \\
&= r_t(s,a) + \widetilde{V}_{t+1}^{\pi,\sigma}(s),
\end{aligned} \tag{32}$$

where (i) holds by the definition in equation 3, (ii) follows from the *state-confounded* Bellman consistency equation in equation 47, (iii) follows from that the reward function r and $\mathcal{P}_t$ are all independent from the action (see equation 10, equation 11 and equation 15), and the last inequality holds by $\mathcal{P}_t(s' \,|\, s,a,c) = \mathbb{1}(s'=s)$ is independent from c_t (see equation 15).

Applying the above fact recursively for $t, t+1, \cdots, T$ leads to that for any $s \in \mathcal{S}$,

$$\begin{aligned}
\widetilde{V}_t^{\pi,\sigma_2}(s) &= r_t(s,a_t) + \widetilde{V}_{t+1}^{\pi,\sigma}(s) = r_t(s,a) + r_{t+1}(s,a_{t+1}) + \widetilde{V}_{t+2}^{\pi,\sigma}(s) \\
&= \cdots = r_t(s,a_t) + \sum_{k=t+1}^{T} r_k(s_k,a_k),
\end{aligned} \tag{33}$$

which directly yields

$$\widetilde{V}_2^{\pi,\sigma_2}([0,0]) = \widetilde{V}_2^{\pi,\sigma_2}([1,1]) = T-1 \quad \text{and} \quad \widetilde{V}_2^{\pi,\sigma_2}([0,1]) = \widetilde{V}_2^{\pi,\sigma_2}([1,0]) = 0. \tag{34}$$

Characterizing $\widetilde{V}_1^{\pi,\sigma_2}([0,0])$ for any policy π. In this section, we are especially interested in the value of $\widetilde{V}_1^{\pi,\sigma_2}$ on the state $[0,0]$. To proceed, one has

$$\begin{aligned}
\widetilde{V}_1^{\pi,\sigma_2}([0,0]) &\overset{\text{(i)}}{=} \inf_{P \in \mathcal{U}^\sigma(P^c)} \widetilde{V}_1^{\pi,P}([0,0]) = \inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{a\sim\pi_1([0,0])}\left[\widetilde{Q}_1^{\pi,P}([0,0],a)\right] \\
&\overset{\text{(ii)}}{=} \inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{a\sim\pi_t([0,0])}\left[r_1([0,0],a) + \mathbb{E}_{c_t\sim P}\left[\mathcal{P}_{1,[0,0],a,c_t}\widetilde{V}_2^{\pi,\sigma}\right]\right] \\
&\overset{\text{(iii)}}{=} 1 + \inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{c_1\sim P}\left[\left(\pi_1(0\,|\,[0,0])\mathcal{P}_{1,[0,0],0,c_1} + \pi_t(1\,|\,[0,0])\mathcal{P}_{1,[0,0],1,c_1}\right)\widetilde{V}_2^{\pi,\sigma}\right] \\
&\overset{\text{(iv)}}{=} 1 + \inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{c_1\sim P}\Bigg[\pi_1(0\,|\,[0,0])\left((1-c_1)P^0_{1,[0,0],0} + c_1 P^{\text{sc}}_{1,[0,0],0}\right)\widetilde{V}_2^{\pi,\sigma} \\
&\quad + \pi_1(1\,|\,[0,0])\left((1-c_1)P^0_{1,[0,0],1} + c_1 P^{\text{sc}}_{1,[0,0],1}\right)\widetilde{V}_2^{\pi,\sigma}\Bigg] \\
&\overset{\text{(v)}}{=} 1 + \inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{c_1\sim P}\Bigg[\pi_1(0\,|\,[0,0])\left((1-c_1)\widetilde{V}_2^{\pi,\sigma}([0,0]) + c_1\widetilde{V}_2^{\pi,\sigma}([1,0])\right) \\
&\quad + \pi_1(1\,|\,[0,0])\left((1-c_1)\widetilde{V}_2^{\pi,\sigma}([0,1]) + c_1\widetilde{V}_2^{\pi,\sigma}([1,1])\right)\Bigg] \\
&= 1 + (T-1)\inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{c_1\sim P}\Big[\pi_1(0\,|\,[0,0])(1-c_1) + \pi_1(1\,|\,[0,0])c_1\Big] \\
&= 1 + (T-1)\pi_1(0\,|\,[0,0]) + (T-1)\inf_{P \in \mathcal{U}^\sigma(P_1^c)} \mathbb{E}_{c_1\sim P}\Big[c_1\big(1-2\pi_1(0\,|\,[0,0])\big)\Big],
\end{aligned} \tag{35}$$

where (i) holds by the definition in equation 3, (ii) follows from the *state-confounded* Bellman consistency equation in equation 47, (iii) follows from $r_1([0,0],a) = 1$ for all $a \in \{0,1\}$ which is independent from c_t. (iv) arises from the

definition of $\mathcal{P}$ in equation 14, (v) can be verified by plugging in the definitions from equation 8 and equation 12, and the penultimate equality holds by equation 34.

Characterizing the optimal robust SC-value functions.

To further consider equation 35, we recall the fact that $\mathcal{U}^{\sigma}(P_1^c) = \left\{P \in \Delta(\mathcal{C}) : \frac{1}{2}\left\|P - P_1^c\right\|_1 \leq \sigma_2\right\}$.

Observing from equation 35 that for any fixed $\pi_1(0 \,|\, [0,0])$, $c_1\big(1 - 2\pi_1(0 \,|\, [0,0])\big)$ is monotonously increasing with c_1 when $1 - 2\pi_1(0 \,|\, [0,0]) >= 0$ and decreasing with c_1 otherwise, it is easily verified that the solution of

$$f(\pi_1(0 \,|\, [0,0])) \coloneqq (T-1) \inf_{P \in \mathcal{U}^{\sigma}(P_1^c)} \mathbb{E}_{c_1 \sim P}\left[c_1\big(1 - 2\pi_1(0 \,|\, [0,0])\big)\right] \tag{36}$$

satisfies

$$f(\pi_1(0 \,|\, [0,0])) = \begin{cases} 0 & \text{if } \pi_1(0 \,|\, [0,0]) \geq \frac{1}{2} \\ (T-1)\sigma_2\big(1 - 2\pi_1(0 \,|\, [0,0])\big) & \text{otherwise} \end{cases}. \tag{37}$$

And note that the value of $\widetilde{V}_1^{\pi,\sigma_2}([0,0])$ only depends on $\pi_1(\cdot \,|\, [0,0])$ which can be represent by $\pi_1(0 \,|\, [0,0])$. Plugging in equation 37 into equation 35, we have that when $\pi_1(0 \,|\, [0,0]) \geq \frac{1}{2}$,

$$\begin{aligned} \max_{\pi} \widetilde{V}_1^{\pi,\sigma_2}([0,0]) &= \max_{\pi_1(0 \,|\, [0,0]) \geq \frac{1}{2}} 1 + (T-1)\pi_1(0 \,|\, [0,0]) + (T-1)\sigma_2\big(1 - 2\pi_1(0 \,|\, [0,0])\big) \\ &= 1 + (T-1)\sigma_2 + (T-1) \max_{\pi_1(0 \,|\, [0,0]) \geq \frac{1}{2}} (1 - 2\sigma_2)\pi_1(0 \,|\, [0,0]) \\ &= 1 + \frac{T-1}{2}, \end{aligned} \tag{38}$$

where the last equality holds by $\sigma_2 > \frac{1}{2}$ and letting $\pi_1(0 \,|\, [0,0]) = \frac{1}{2}$. Similarly, when $\pi_1(0 \,|\, [0,0]) < \frac{1}{2}$,

$$\max_{\pi} \widetilde{V}_1^{\pi,\sigma_2}([0,0]) = \max_{\pi_1(0 \,|\, [0,0]) < \frac{1}{2}} 1 + (T-1)\pi_1(0 \,|\, [0,0]) < 1 + \frac{T-1}{2}. \tag{39}$$

Consequently, we complete the proof by concluding that

$$\widetilde{V}_1^{\pi_{\mathsf{RSC}}^{\star,\sigma_2},\sigma_2}([0,0]) = \widetilde{V}_1^{\star,\sigma_2}([0,0]) = \max_{\pi} \widetilde{V}_1^{\pi,\sigma_2}([0,0]) = 1 + \frac{T-1}{2}. \tag{40}$$

D.3. Theorem 2

A natural question is: does there exist an optimal policy that maximizes the robust SC-value function $\widetilde{V}_t^{\pi,\sigma}$ for any RSC-MDP so that we can target to learn? To answer this, we introduce the following theorem that ensures the existence of the optimal policy for all RSC-MDPs.

Theorem 2. *Let Π be the set of all non-stationary and stochastic policies. Consider any RSC-MDP, there exists at least one optimal policy $\pi^{\mathsf{sc},\star} = \{\pi_t^{\mathsf{sc},\star}\}_{1 \leq t \leq T}$ such that for all $s \in \mathcal{S}$ and $1 \leq t \leq T$, one has*

$$\widetilde{V}_t^{\pi^{\mathsf{sc},\star},\sigma}(s) = \widetilde{V}_t^{\star,\sigma}(s) \coloneqq \sup_{\pi \in \Pi} \widetilde{V}_t^{\pi,\sigma}(s) \quad \textit{and} \quad \widetilde{Q}_t^{\pi^{\mathsf{sc},\star},\sigma}(s,a) = \widetilde{Q}_t^{\star,\sigma}(s,a) \coloneqq \sup_{\pi \in \Pi} \widetilde{Q}_t^{\pi,\sigma}(s,a).$$

Proof. The proof follows the pipeline of proving the existence of the optimal policy for standard MDPs but tailored for RSC-MDPs since the additional components confounder C_{s} and the infimum operator. To begin with, recall that the goal is to find a policy $\widetilde{\pi} = \{\widetilde{\pi}_t\}_{1 \leq t \leq T}$ such that:

$$\widetilde{V}_t^{\widetilde{\pi},\sigma}(s) = \widetilde{V}_t^{\star,\sigma}(s) \coloneqq \sup_{\pi \in \Pi} \widetilde{V}_t^{\pi,\sigma}(s) \quad \text{and} \quad \widetilde{Q}_t^{\widetilde{\pi},\sigma}(s,a) = \widetilde{Q}_t^{\star,\sigma}(s,a) \coloneqq \sup_{\pi \in \Pi} \widetilde{Q}_t^{\pi,\sigma}(s,a). \tag{41}$$

Towards this, we start from the first claim in equation 41. Before proceeding, we let $\{S_t, A_t, R_t, C_t\}$ denote the random variables at time step t for all $1 \leq t \leq T$. Then due to the Markov properties, we know that conditioned on current state

s_t, the future state, action, and reward are all independent from the previous $s_1, a_1, r_1, c_1, \cdots, s_{t-1}, a_{t-1}, r_{t-1}, c_{t-1}$. For convenience, we introduce the following notation:

$$\forall 1 \le t \le T: \quad P_{+t} \coloneqq \{P_k\}_{t \le k \le T} \quad \text{and} \quad \mathcal{U}^\sigma(P^c_{+t}) \coloneqq \{\mathcal{U}^\sigma(P^c_k)\}_{t \le k \le T} \tag{42}$$

to represent the collection of variables from time step t to the end of the episode, and choose $\widetilde{\pi}$ to obey

$$\forall 1 \le t \le T: \quad \pi_t(s) \coloneqq \operatorname{argmax}_{a \in \mathcal{A}} \mathbb{E}\left[r_t(s,a) + \inf_{P_t \in \mathcal{U}^\sigma(P^c_{t,s,a})} \mathbb{E}_{c_t \sim P_t}\left[\widetilde{V}^{\star,\sigma}_{t+1}(s_{t+1}) \right] \right] \tag{43}$$

With the above preparation in mind, for any $(t,s) \in \{1,2,\cdots,T\} \times \mathcal{S}$, one has

$$\begin{aligned}
&\widetilde{V}^{\star,\sigma}_t(s) \\
&\overset{\text{(i)}}{=} \sup_{\pi \in \Pi} \inf_{P_{+t} \in \mathcal{U}^\sigma\left(P^c_{+t}\right)} \widetilde{V}^{\pi,P}_t(s) \overset{\text{(ii)}}{=} \sup_{\pi \in \Pi} \inf_{P_{+t} \in \mathcal{U}^\sigma\left(P^c_{+t}\right)} \mathbb{E}_{\pi, P_{+t}}\left[\sum_{k=t}^{T} r_k(s_k, a_k) \right] \\
&\overset{\text{(iii)}}{=} \sup_{\pi \in \Pi} \inf_{P_{+t} \in \mathcal{U}^\sigma\left(P^c_{+t}\right)} \mathbb{E}_{\pi_t}\Bigg[r_t(s,a_t) \\
&\quad + \mathbb{E}_{c_t \sim P_t} \mathbb{E}\left[\sum_{k=t+1}^{T} r_k(s_k,a_k) \,|\, \pi, P_{+(t+1)}, (S_t, A_t, R_t, C_t, S_{t+1}) = (s, a_t, r_t, c_t, s_{t+1}) \right] \Bigg] \\
&= \sup_{\pi \in \Pi} \mathbb{E}_{\pi_t}\Bigg[\inf_{P_{+t} \in \mathcal{U}^\sigma\left(P^c_{+t}\right)} r_t(s,a_t) + \inf_{P_{+t} \in \mathcal{U}^\sigma(P^c_{+t})} \mathbb{E}_{c_t \sim P_t} \\
&\quad \mathbb{E}\left[\sum_{k=t+1}^{T} r_k(s_k,a_k) \,|\, \pi, P_{+(t+1)}, (S_t, A_t, R_t, C_t, S_{t+1}) = (s, a_t, r_t, c_t, s_{t+1}) \right] \Bigg]
\end{aligned}$$

where (i) and (ii) holds by the definitions in equation 3 and equation 1 respectively, and (iii) follows from expressing the term of interest by moving one step ahead and $\mathbb{E}_{\pi_t}$ is taken with respect to $a_t \sim \pi_t(\cdot \,|\, S_1 = s_1, A_1 = a_1, \cdots, S_t = s)$, and the last equality arises from we can exchange the operators $\mathbb{E}_{\pi_t}$ and $\inf_{P \in \mathcal{U}^\sigma(P^c)}$ since they are independent.

To continue, we observe that the above equation can be rewritten and controlled as follows:

$$\begin{aligned}
&\widetilde{V}^{\star,\sigma}_t(s) \\
&= \sup_{\pi \in \Pi} \mathbb{E}_{\pi_t}\Bigg[\inf_{P_{+t} \in \mathcal{U}^\sigma(P^c_{+t})} r_t(s,a_t) + \inf_{P_t \in \mathcal{U}^\sigma(P^c_t)} \mathbb{E}_{c_t \sim P_t} \inf_{P_{+(t+1)} \in \mathcal{U}^\sigma(P^c_{+(t+1)})} \\
&\quad \mathbb{E}\left[\sum_{k=t+1}^{T} r_k(s_k,a_k) \,|\, \pi', P_{+(t+1)}, (S_t, A_t, R_t, C_t, S_{t+1}) = (s, a_t, r_t, c_t, s_{t+1}) \right] \Bigg] \\
&\le \sup_{\pi \in \Pi} \mathbb{E}_{\pi_t}\Bigg[\inf_{P_{+t} \in \mathcal{U}^\sigma(P^c_{+t})} r_t(s,a_t) + \inf_{P_t \in \mathcal{U}^\sigma(P^c_t)} \mathbb{E}_{c_t \sim P_t} \sup_{\pi' \in \Pi} \inf_{P_{+(t+1)} \in \mathcal{U}^\sigma(P^c_{+(t+1)})} \\
&\quad \mathbb{E}\left[\sum_{k=t+1}^{T} r_k(s_k,a_k) \,|\, \pi', P_{+(t+1)}, (S_t, A_t, R_t, C_t, S_{t+1}) = (s, a_t, r_t, c_t, s_{t+1}) \right] \Bigg] \\
&\overset{\text{(i)}}{=} \sup_{\pi \in \Pi} \mathbb{E}_{\pi_t}\left[r_t(s,a_t) + \inf_{P_t \in \mathcal{U}^\sigma(P^c_t)} \mathbb{E}_{c_t \sim P_t}\left[\sup_{\pi' \in \Pi} \inf_{P_{+(t+1)} \in \mathcal{U}^\sigma(P^c_{+(t+1)})} \mathbb{E}_{\pi', P_{+(t+1)}}\left[\sum_{k=t+1}^{T} r_k(s_k,a_k) \right] \right] \right] \\
&= \sup_{\pi \in \Pi} \mathbb{E}_{\pi_t}\left[r_t(s,a_t) + \inf_{P_t \in \mathcal{U}^\sigma(P^c_t)} \mathbb{E}_{c_t \sim P_t}\left[\widetilde{V}^{\star,\sigma}_{t+1}(s_{t+1}) \right] \right] \\
&= \sup_{a_t \in \mathcal{A}} \mathbb{E}_{a_t}\left[r_t(s,a_t) + \inf_{P_t \in \mathcal{U}^\sigma(P^c_t)} \mathbb{E}_{c_t \sim P_t}\left[\widetilde{V}^{\star,\sigma}_{t+1}(s_{t+1}) \right] \right]
\end{aligned}$$

$$= \inf_{P_t \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}\left[r_t(s,a_t) + \mathbb{E}_{c_t \sim P_t}\left[\widetilde{V}_{t+1}^{\star,\sigma}(s_{t+1})\right] \mid a_t = \widetilde{\pi}_t(s)\right], \tag{44}$$

where (i) holds by the Markov decision such that the rewards $\{r_k(s_k,a_k)\}_{t+1\le k\le T}$ conditioned on determined $(S_t, A_t, R_t, C_t, S_{t+1})$ or S_{t+1} are the same, and the last equality follows from the definition of $\widetilde{\pi}$ in equation 43 and the exchangeability of $\inf_{P_t \in \mathcal{U}^\sigma(P_t^c)}$ and $\mathbb{E}_{a_t}[\cdot]$.

Applying equation 44 recursively for $t+1, \cdots T$, we arrive at

$$\begin{aligned}
\widetilde{V}_t^{\star,\sigma}(s) &\le \inf_{P_t \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}\left[r_t(s,a_t) + \mathbb{E}_{c_t \sim P_t}\left[\widetilde{V}_{t+1}^{\star,\sigma}(s_{t+1})\right] \mid a_t = \widetilde{\pi}_t(s)\right] \\
&\le \inf_{P_t \in \mathcal{U}^\sigma(P_t^c)} \inf_{P_{t+1} \in \mathcal{U}^\sigma(P_{t+1}^c)} \mathbb{E}\Bigg[r_t(s,a_t) + \\
&\quad \mathbb{E}_{c_t \sim P_t}\left[r_{t+1}(s_{t+1},a_{t+1}) + \mathbb{E}_{c_{t+1} \sim P_{t+1}}\left[\widetilde{V}_{t+2}^{\star,\sigma}(s_{t+1})\right]\right] \mid (a_t, a_{t+1}) = (\widetilde{\pi}_t(s), \widetilde{\pi}_{t+1}(s_{t+1}))\Bigg] \\
&\le \cdots \le \inf_{P_{+t} \in \mathcal{U}^\sigma(P_{+t}^c)\}} \mathbb{E}_{\pi,P}\left[\sum_{k=t}^{T} r_k(s_k,a_k)\right] = \widetilde{V}_t^{\widetilde{\pi},\sigma}(s).
\end{aligned} \tag{45}$$

where (i) holds by the Markov properties of the rewards.

Observing from equation 45 that

$$\forall s \in \mathcal{S}: \quad \widetilde{V}_t^{\star,\sigma}(s) \le \widetilde{V}_t^{\widetilde{\pi},\sigma}(s) \le \sup_{\pi \in \Pi} \widetilde{V}_t^{\pi,\sigma}(s) = \widetilde{V}_t^{\star,\sigma}(s), \tag{46}$$

which directly verifies the first assertion in equation 41 $\widetilde{V}_t^{\widetilde{\pi},\sigma}(s) = \widetilde{V}_t^{\star,\sigma}(s)$ for all $s \in \mathcal{S}$. The second assertion in equation 41 can be achieved analogously. Until now, we verify that there exists at least a policy $\widetilde{\pi}$ that obeys equation 41, which we refer it as an optimal policy since its value is equal to or larger than any other non-stationary and stochastic policies over all states $s \in \mathcal{S}$.

■

D.4. Auxiliary results of SC-MDPs and RSC-MDPs

Facts about SC-MDPs. For any state-confounded MDPs (SC-MDPs) $\mathcal{M}_{\mathsf{SC}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, P^c\}$, denoting the optimal policy as $\pi^\star$ and the corresponding optimal SC-value function as $\widetilde{V}$, any policy π satisfies the corresponding *state-confounded* Bellman consistency equation as below:

$$\widetilde{Q}_t^{\pi,P^c}(s,a) = r_t(s,a) + \mathbb{E}_{c_t \sim P_t^c}\left[\mathcal{P}_{t,s,a,c_t} \widetilde{V}_{t+1}^{\pi,\sigma}\right], \tag{47}$$

where $\mathcal{P}_{t,s,a,c_t} \in \mathbb{R}^{1\times S}$ such that $\mathcal{P}_{t,s,a,c_t}(s') \coloneqq \mathcal{P}_t(s' \mid s,a,c_t)$ for $s' \in \mathcal{S}$.

Facts about RSC-MDPs. It is easily verified that for any RSC-MDP $\mathcal{M}_{\mathsf{sc\text{-}rob}} = \{\mathcal{S}, \mathcal{A}, T, r, \mathcal{C}, \{\mathcal{P}_t^i\}, \mathcal{U}^{\sigma_2}(P^c)\}$, any policy π and the optimal policy $\pi^\star$ satisfy the corresponding *robust state-confounded* Bellman consistency equation and Bellman optimality equation shown below, respectively:

$$\begin{aligned}
\widetilde{Q}_t^{\pi,\sigma}(s,a) &= r_t(s,a) + \inf_{P \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}_{c_t \sim P}\left[\mathcal{P}_{t,s,a,c_t} \widetilde{V}_{t+1}^{\pi,\sigma}\right], \\
\widetilde{Q}_t^{\star,\sigma}(s,a) &= r_t(s,a) + \inf_{P \in \mathcal{U}^\sigma(P_t^c)} \mathbb{E}_{c_t \sim P}\left[\mathcal{P}_{t,s,a,c_t} \widetilde{V}_{t+1}^{\star,\sigma}\right],
\end{aligned} \tag{48}$$

where $\mathcal{P}_{t,s,a,c_t} \in \mathbb{R}^{1\times S}$ such that $\mathcal{P}_{t,s,a,c_t}(s') \coloneqq \mathcal{P}_t(s' \mid s,a,c_t)$ for $s' \in \mathcal{S}$, and $\widetilde{V}_{t+1}^{\star,\sigma}(s) = \max_a \widetilde{Q}_{t+1}^{\star,\sigma}(s,a)$.

E. Additional Experiment Results

The training curves of four environments are displayed in Figure 5, showing that RSC-SAC achieves similar rewards compared to non-robust SAC although converges slower than it.

Table 2. Testing reward on nominal environments. Underline means the reward is over 0.9.

Method	Brightness	Behavior	Crossing	CarType	Lift	Stack	Wipe	Door
SAC	1.00±0.09	1.00±0.08	1.00±0.02	1.00±0.03	1.00±0.03	1.00±0.09	1.00±0.12	1.00±0.03
RMDP-G	1.04±0.09	1.00±0.11	0.78±0.05	0.79±0.05	0.92±0.07	0.86±0.14	0.99±0.13	0.99±0.06
RMDP-U	1.02±0.09	1.04±0.07	0.90±0.03	0.88±0.03	0.97±0.05	0.92±0.12	0.97±0.14	0.88±0.31
MoCoDA	0.65±0.17	0.78±0.15	0.57±0.07	0.55±0.13	0.79±0.11	0.72±0.08	0.69±0.13	0.41±0.22
ATLA	0.99±0.11	0.98±0.11	0.89±0.05	0.88±0.04	0.94±0.08	0.88±0.10	0.96±0.12	0.97±0.05
DBC	0.75±0.12	0.78±0.10	0.85±0.08	0.86±0.06	0.27±0.04	0.12±0.08	0.31±0.21	0.01±0.01
RSC-SAC	0.92±0.31	1.06±0.07	0.96±0.03	0.96±0.03	0.96±0.05	1.04±0.08	0.92±0.14	0.98±0.05

Table 3. Influence of modules

Method	Lift	Behavior	Crossing
w/o $\boldsymbol{G}_\phi$	0.79±0.15	0.51±0.24	0.87±0.10
w/o P_θ	0.75±0.13	0.41±0.28	0.89±0.08
w/o P^c	0.90±0.09	0.66±0.21	0.96±0.04
Full model	0.98±0.04	1.02±0.09	1.04±0.02

RSC-SAC **maintains great performance in the training environments.** Previous literature (Xu et al., 2023a) finds out that there usually exists a trade-off between the performance in the nominal environment and the robustness against uncertainty. To evaluate the performance of RSC-SAC in the nominal environment, we conduct experiments and summarize results in Table 2, which shows that RSC-SAC still performs well in the training environment.

Both the distribution of confounder and the structural causal model are critical. To assess the impact of each module in our algorithm, we conduct three additional ablation studies (in Table 3), where we remove the causal graph $\boldsymbol{G}_\phi$, the transition model P_θ, and the distribution of the confounder P^c respectively. The results demonstrate that the learnable causal graph $\boldsymbol{G}_\phi$ is critical for the performance that enhances the prediction of the next state and reward, thereby facilitating the generation of high-quality next states with current perturbed states. The transition model without $\boldsymbol{G}_\phi$ may still retain numerous spurious correlations, resulting in a performance drop similar to the one without P_θ, which does not alter the next state and reward. In the third row of Table 3, the performance drop indicates that the confounder P^c also plays a crucial role in preserving semantic meaning and avoiding policy training distractions.

RSC-SAC **is also robust to random perturbation.** The final investigation aims to assess the generalizability of our method to cope with random perturbation that is widely considered in robust RL (RMDPs). Towards this, we evaluate the proposed algorithm in the test environments added with random noise under the Gaussian distribution with two varying scales in the *Lift* environment. In Table 4, *Lift-0* indicates the nominal training environment, while *Lift-0.01* and *Lift-0.1* represent the environments perturbed by the Gaussian noise with standard derivation 0.01 and 0.1, respectively. The results indicate that our RSC-SAC achieves comparable robustness compared to RMDP-0.01 in both large and small perturbation settings and outperforms RMDP methods in the nominal training environment.

F. Experiment Details

F.1. Baselines

We use a non-robust RL and four representative algorithms of robust RL as baselines, all of which are implemented on top of the SAC (Haarnoja et al., 2018) algorithm. **Non-robust RL (SAC):** This serves as a basic baseline without considering any robustness during training; **Solving robust MDP:** We generate the samples to cover the uncertainty set over the state

Table 4. Random perturbation

Method	Lift-0	Lift-0.01	Lift-0.1
SAC	1.00±0.03	0.77±0.13	0.46±0.23
RMDP-0.01	0.97±0.05	0.96±0.06	0.51±0.21
RMDP-0.1	0.85±0.12	0.82±0.09	0.39±0.15
RSC-SAC	0.96±0.05	0.94±0.06	0.44±0.18

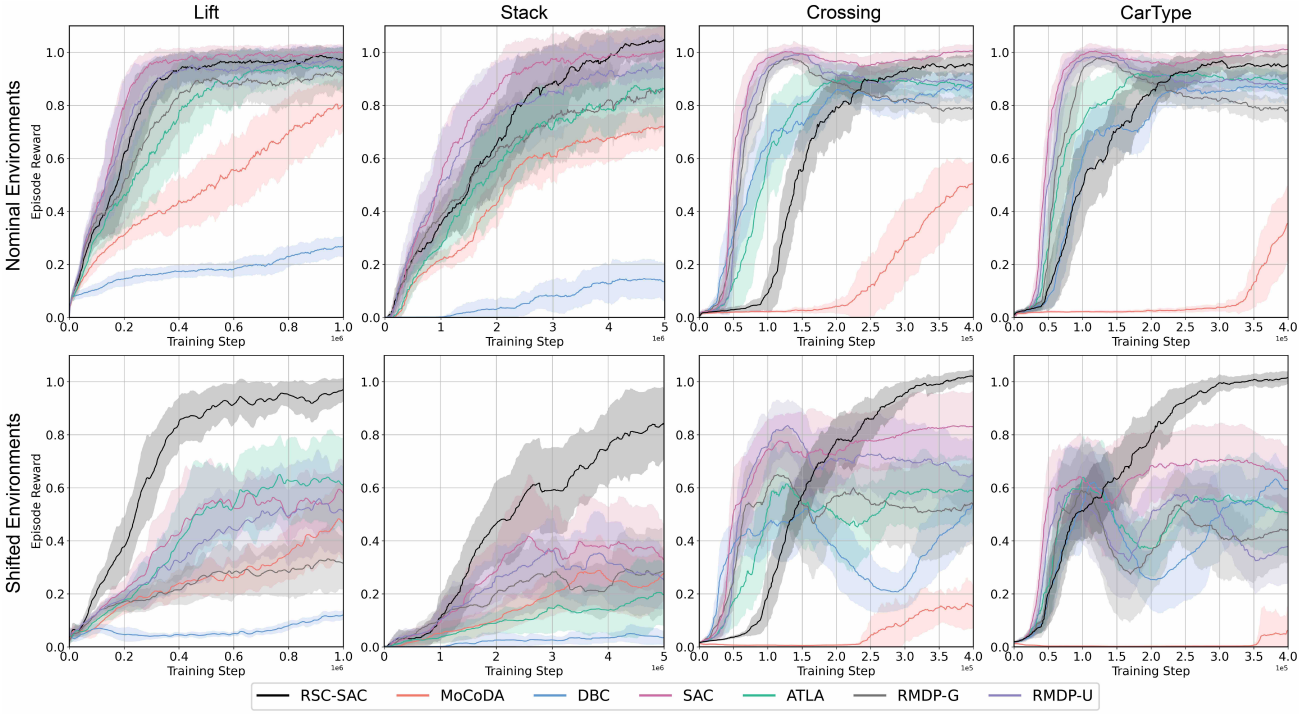

Figure 5. The first row shows the testing reward on the nominal environments, while the second row shows the testing reward on the shifted environments.

space by adding perturbation around the nominal states that follows two distribution, i.e., uniform distribution (RMDP-U) and Gaussian distribution (RMDP-G). **Solving SA-MDP:** We select ATLA (Zhang et al., 2021a), a strong algorithm that generates adversarial states using an optimal adversary within the uncertainty set. **Invariant feature learning:** We choose DBC (Zhang et al., 2020a), which learns invariant features using the bi-simulation metric (Larsen & Skou, 1989). **Counterfactual data augmentation:** We select MoCoDA (Pitis et al., 2022), which identifies local causality to switch components and generate counterfactual samples to cover the targeted uncertainty set. We adapt this algorithm using an approximate causal graph rather than the true causal graph.

F.2. Architecture of the structural causal model

We plot the architecture of the structural causal model we used in our method in Figure 6. In normal neural networks, the input is treated as a whole to pass through linear layers or convolution layers. However, this structure blends all information in the input, making the causal graph useless to separate cause and effect. Thus, in our model, we design an encoder that is shared across all dimensions of the input. Since different dimensions could have exactly the same values, we add a learnable position embedding to the input of the encoder. In summary, the input dimension of the encoder is $1 + d_{pos}$, where d_{pos} is the dimension of the position embedding.

After the encoder, we obtain a set of independent features for each dimension of the input. We now multiply the features with a learnable binary causal graph $\boldsymbol{G}$. The element (i, j) of the graph is sampled from a Gumbel-Softmax distribution with parameter $\phi_{i,j}$ to ensure the loss function is differentiable w.r.t ϕ.

The multiplication of the causal graph and the input feature creates a linear combination of the input feature with respect to the causal graph. The obtained features are then passed through a decoder to predict the next state and reward. Again, the decoder is shared across all dimensions to avoid information leaking between dimensions. Position embedding is included in the input to the decoder and the output dimension of the decoder is 1.

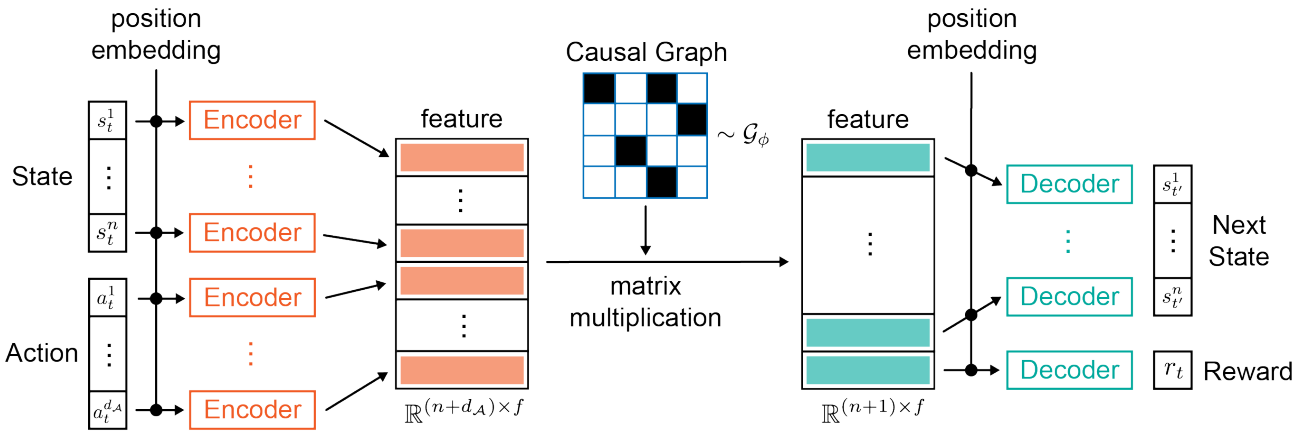

Figure 6. Model architecture of the structural causal model. Encoder, Decoder, position embedding, and Causal Graph are learnable during the training stage.

F.3. Environments

We design four self-driving tasks in the Carla simulator (Dosovitskiy et al., 2017) and four manipulation tasks in the Robosuite platform (Zhu et al., 2020). All of these realistic tasks contain strong spurious correlations that are explicit to humans. We categorize the tasks into *distraction correlation* and *composition correlation* according to the type of spurious correlation. We specify these two types of correlation below.

- **Distraction correlation** is between task-relevant and task-irrelevant portions of the state. The task-irrelevant part could distract the policy model from learning important features and lead to a performance drop. A typical method to avoid distraction is background augmentation (Laskin et al., 2020; Yarats et al., 2021). We design four tasks with this category of correlation, i.e., *Lift*, *Wipe*, *Brightness*, and *CarType*.
- **Composition correlation** is between two task-relevant portions of the state. This correlation usually exists in compositional generalization, where states are re-composed to form novel tasks during testing. Typical examples are multi-task RL (Jiang et al., 2022; Lu et al., 2021) and hierarchical RL (Yoo et al., 2022; Le et al., 2018). We design four tasks with this category of correlation, i.e., *Stack*, *Door*, *Behavior*, and *Crossing*.

Brightness. The nominal environments are shown in the 1^{th} column of Figure 7, where the brightness and the traffic density are correlated. When the ego vehicle drives in the daytime, there are many surrounding vehicles (first row). When the ego vehicle drives in the evening, there is no surrounding vehicle (second row). The shifted environment swaps the brightness and traffic density in the nominal environment, i.e., many surrounding vehicles in the evening and no surrounding vehicles in the daytime.

Behavior. The nominal environments are shown in the 2^{nd} column of Figure 7, where the other vehicle has aggressive driving behavior. When the ego vehicle is in front of the other vehicle, the other vehicle always accelerates and overtakes the ego vehicle in the left lane. When the ego vehicle is behind the other vehicle, the other vehicle will always accelerate. In the shifted environment, the behavior of the other vehicle is conservative, i.e., the other vehicle always decelerates to block the ego vehicle.

Crossing. The nominal environments are shown in the 3^{rd} column of Figure 7, where the pedestrian follows the traffic rule and only cross the road when the traffic light is green. In the shifted environment, the pedestrian disobeys the traffic rule and crosses the road when the traffic light is red.

CarType. The nominal environments are shown in the 4^{th} column of Figure 7, where the type of vehicle and the speed of the vehicle are correlated. When the vehicle is a truck, the speed is low and when the vehicle is a motorcycle, the speed is high. In the shifted environment, the truck drives very fast and the motorcycle drives very slow.

Lift. The nominal environments are shown in the 1^{th} column of Figure 8, where the position of the cube and the color of the cube are correlated. When the cube is in the left part of the table, the color of the cube is green, when the cube is in the

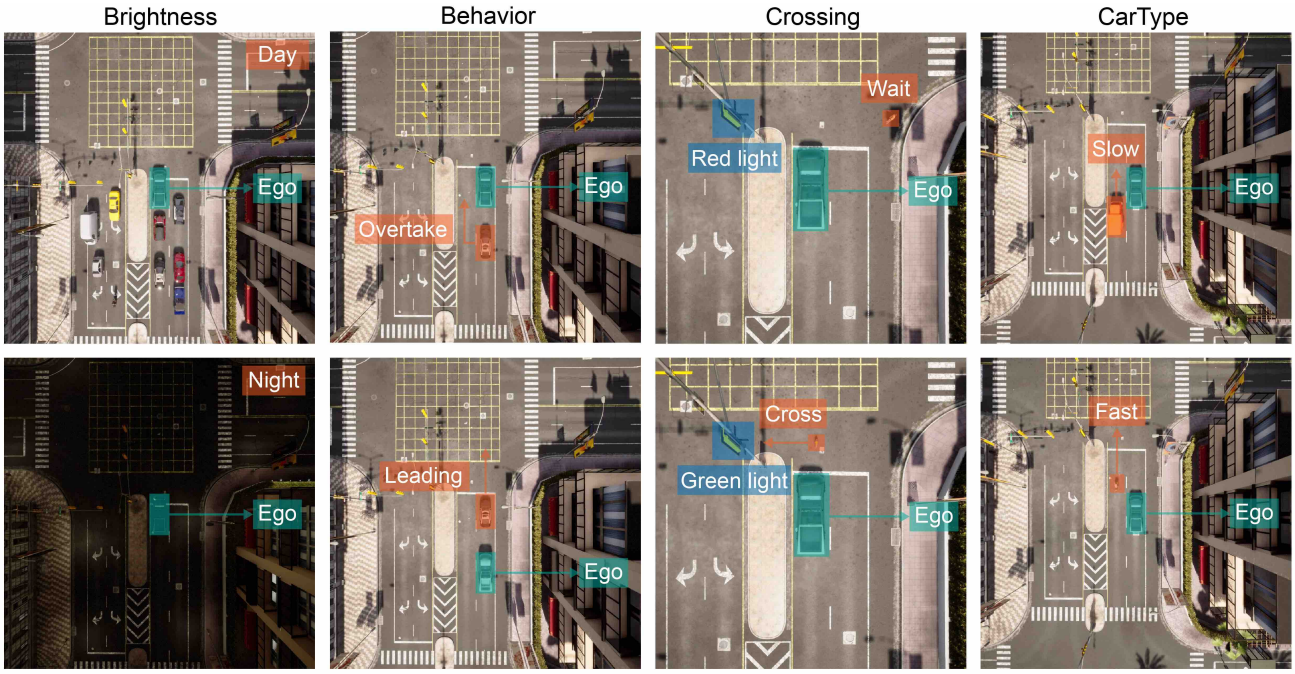

Figure 7. Illustration of tasks in the Carla simulator.

right part of the table, the color of the cube is red. The shifted environment swaps the color and position of the cube in the nominal environment, i.e., the cube is green when it is in the right part and the cube is red when it is in the left part.

Stack. The nominal environments are shown in the 2$^{\text{nd}}$ column of Figure 8, where the position of the red cube and green plate are correlated. When the cube is in the left part of the table, the plate is also in the left part; when the cube is in the right part of the table, the plate is also in the right part. In the shifted environment, the relative position of the cube and the plate changes, i.e., When the cube is in the left part of the table, the plate is in the right part; when the cube is in the right part of the table, the plate is in the left part.

Wipe. The nominal environments are shown in the 3$^{\text{rd}}$ column of Figure 8, where the shape of the dirty region is correlated to the position of the cube. When the dirty region is diagonal, the cube is on the right-hand side of the robot arm. When the dirty region is anti-diagonal, the cube is on the left-hand side of the robot arm. In the shifted environment, the correlation changes, i.e., when the dirty region is diagonal, the cube is on the left-hand side of the robot arm; when the dirty region is anti-diagonal, the cube is on the right-hand side of the robot arm.

Door. The nominal environments are shown in the 4$^{\text{th}}$ column of Figure 8, where the height of the handle and the position of the door is correlated. When the door is closed to the robot arm, the handle is in a low position. When the door is far from the robot arm, the handle is in a high position. In the shifted environment, the correlation changes, i.e., when the door is closed to the robot arm, the handle is in a high position; when the door is far from the robot arm, the handle is in a low position.

F.4. Computation resources

Our algorithm is implemented on top of the Tianshou (Weng et al., 2022) package. All of our experiments are conducted on a machine with an Intel i9-9900K CPU@3.60GHz (16 core) CPU, an NVIDIA GeForce GTX 1080Ti GPU, and 64GB memory.

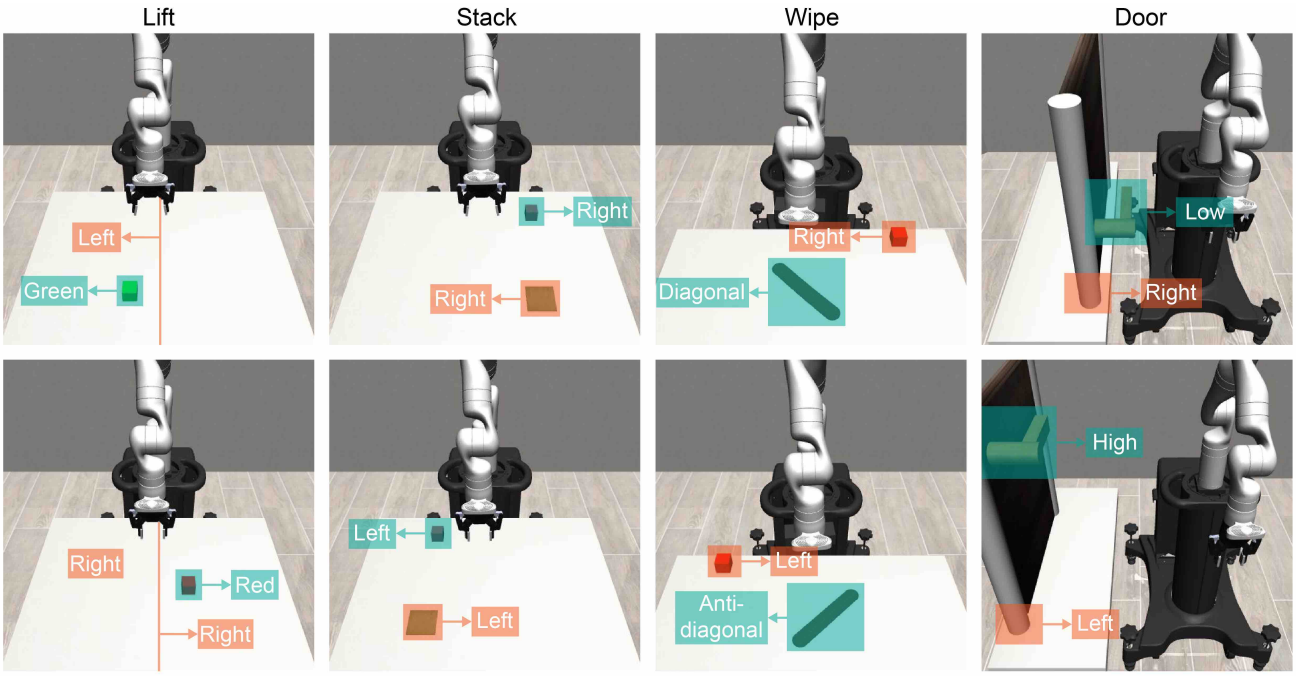

Figure 8. Illustration of tasks in the Robosuite simulator.

F.5. Hyperparameters

We summarize all hyper-parameters used in the Carla experiments (Table 5) and Robosuite experiments (Table 6). The source code of experiments will be released after the double-blind review.

F.6. Discovered Causal Graph in SCM

To show the performance of our learned SCM, we plot the estimated causal graphs of all experiments in Figure 9, Figure 10, Figure 11, Figure 12, and Figure 13.

Table 5. Hyper-parameters in Carla experiments

Parameters	Notation	Environment			
		Brightness	Behavior	Crossing	CarType
Horizon steps	T	100	100	100	100
State dimension	n	24	12	12	12
Action dimension	$d_{\mathcal{A}}$	2	2	2	2
Max training steps		1×10^5	1×10^5	5×10^5	5×10^5
Weight of $\|\boldsymbol{G}\|_p$	λ	0.1	-	-	-
norm of $\|\boldsymbol{G}\|_p$	p	0.1	-	-	-
Actor learning rate		3×10^{-4}	-	-	-
Critic learning rate		1×10^{-3}	-	-	-
Batch size		256	-	-	-
Discount factor	γ in SAC	0.99	-	-	-
Soft update weight	τ in SAC	0.005	-	-	-
Weight of entropy	α in SAC	0.1	-	-	-
Hidden layers		[256, 256, 256]	-	-	-
Returns estimation step		4	-	-	-
Buffer size		1×10^5	-	-	-
Steps per update		10	-	-	-

Table 6. Hyper-parameters in Robosuite experiments

Parameters	Notation	Environment			
		Lift	Stack	Door	Wipe
Horizon steps	T	300	300	300	500
Control frequency (Hz)		20	20	20	20
State dimension	n	50	110	22	30
Action dimension	$d_{\mathcal{A}}$	4	4	8	7
Controller type		OSC position	OSC position	Joint velocity	Joint velocity
Max training steps		1×10^6	5×10^6	1×10^6	1×10^6
Weight of $\|\boldsymbol{G}\|_p$	λ	0.01	-	-	-
norm of $\|\boldsymbol{G}\|_p$	p	0.1	-	-	-
Actor learning rate		3×10^{-4}	-	-	-
Critic learning rate		1×10^{-3}	-	-	-
Batch size		128	-	-	-
Discount factor	γ in SAC	0.99	-	-	-
Soft update weight	τ in SAC	0.005	-	-	-
alpha learning rate	lr_α in SAC	3×10^{-4}	-	-	-
Hidden layers		[256, 256, 256]	-	-	-
Returns estimation step		4	-	-	-
Buffer size		1×10^6	-	-	-
Steps per update		10	-	-	-

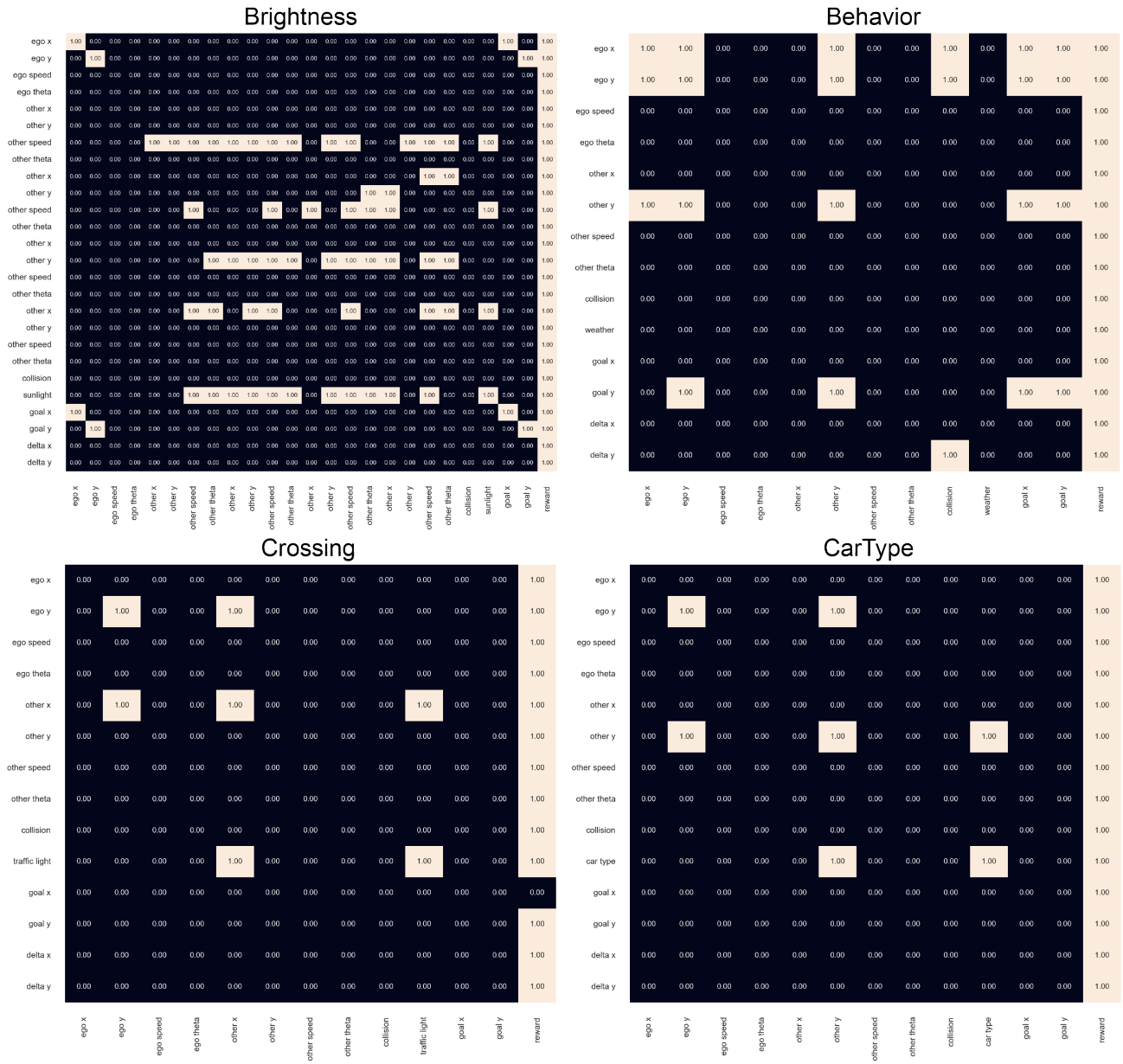

Figure 9. Estimated Causal Graphs of four tasks in Carla.

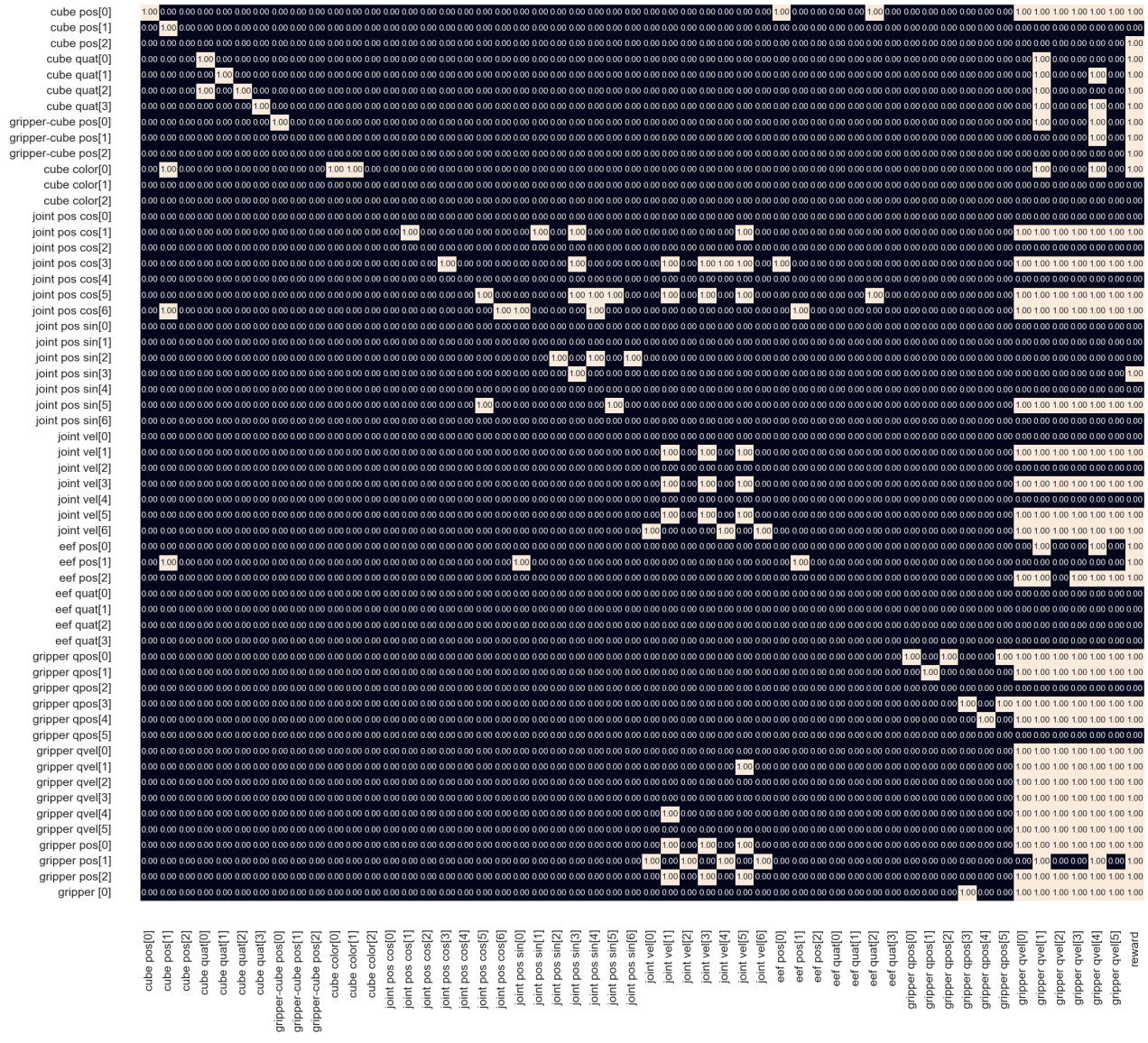

Figure 10. Estimated Causal Graphs of the Lift task in Robosuite.

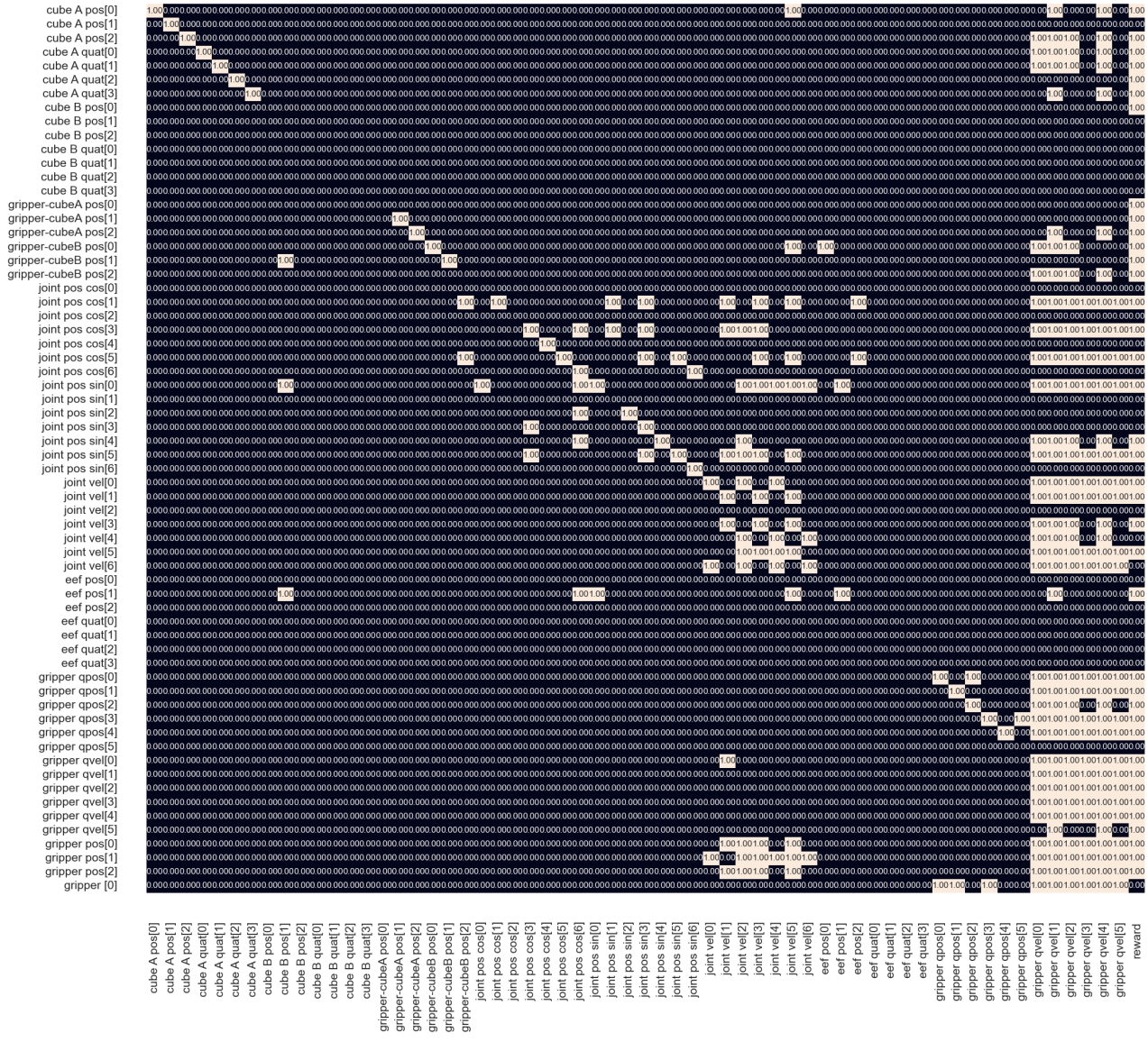

Figure 11. Estimated Causal Graphs of the Stack task in Robosuite.

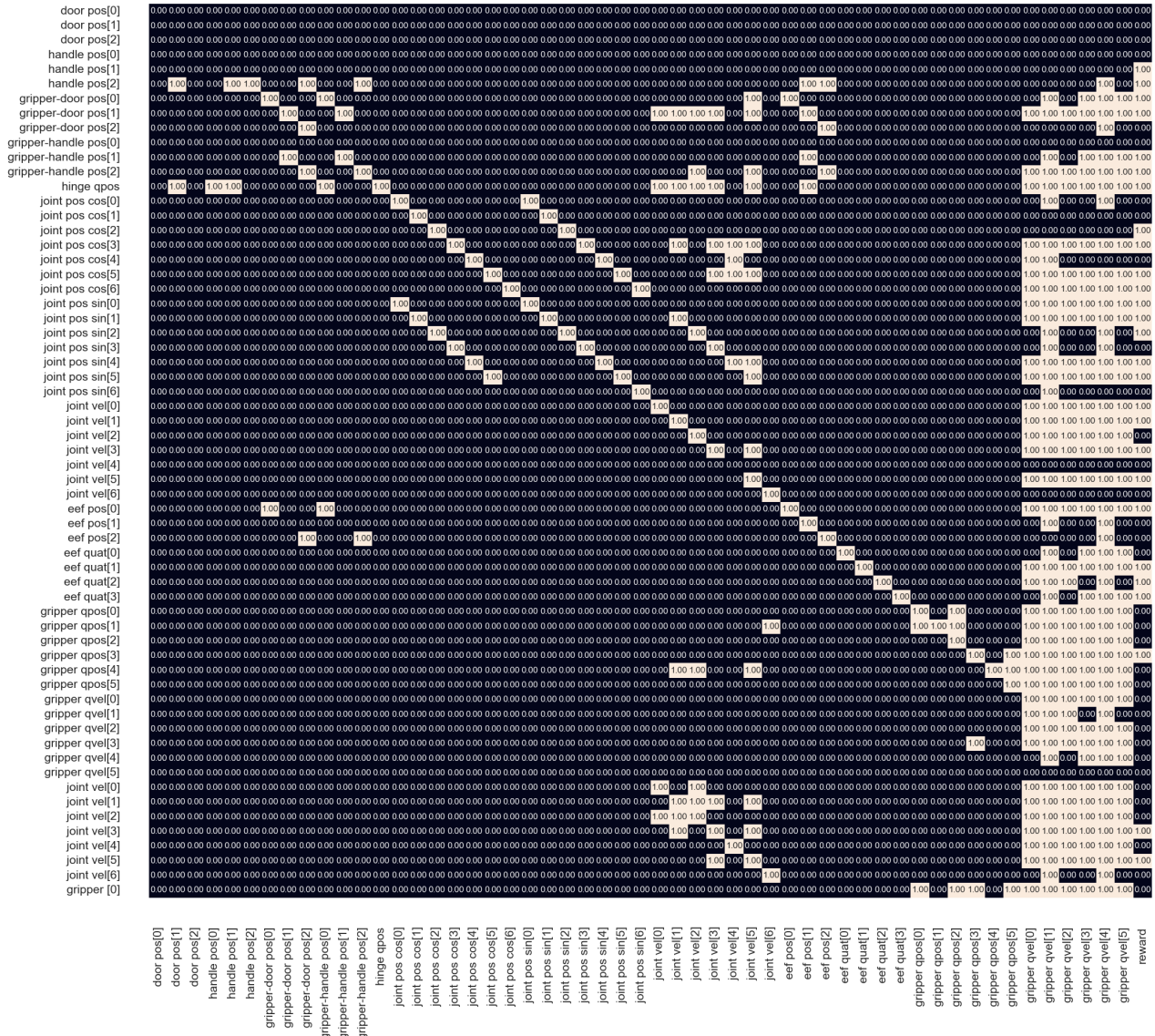

Figure 12. Estimated Causal Graphs of the Door task in Robosuite.

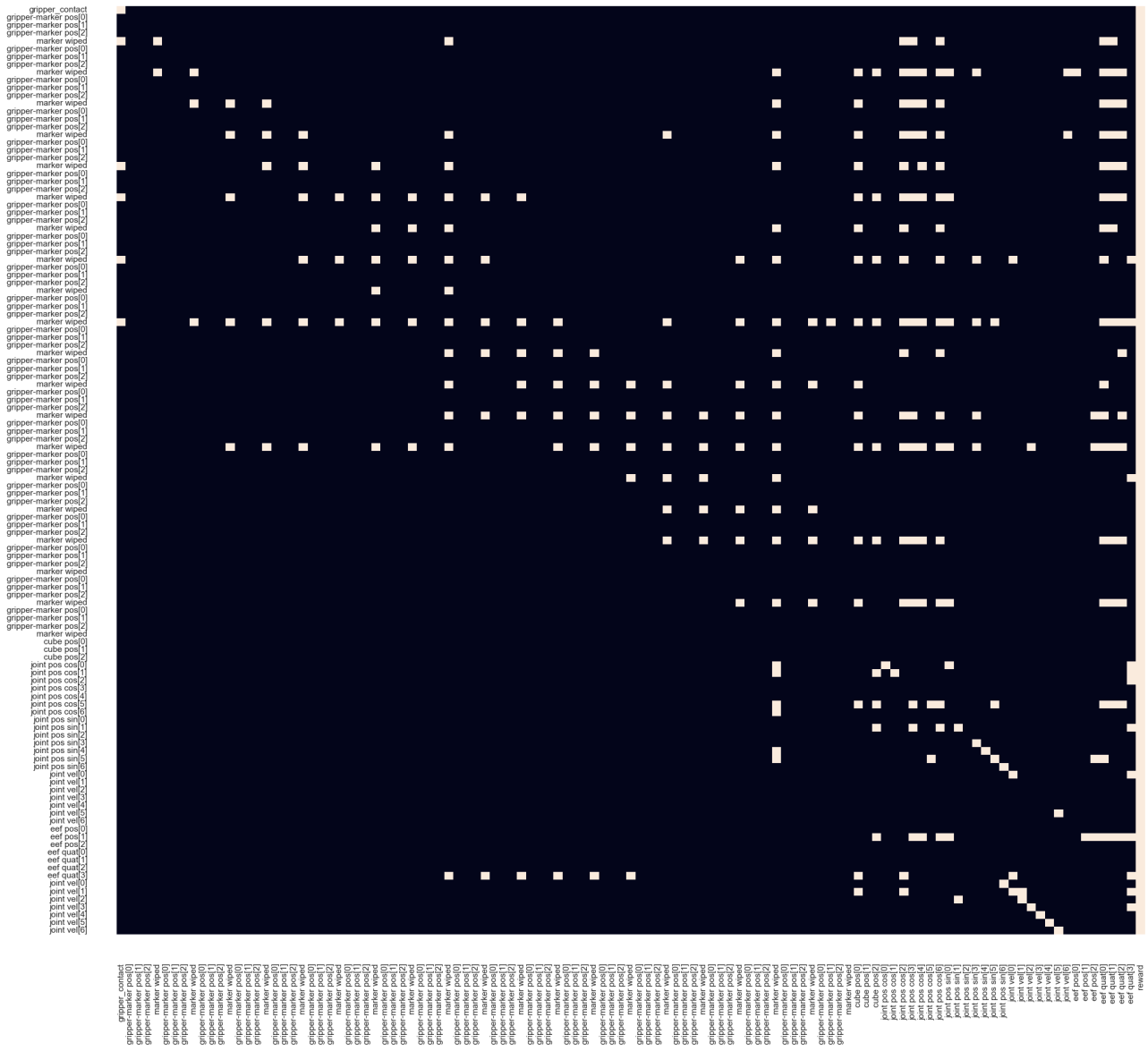

Figure 13. Estimated Causal Graphs of the Wipe task in Robosuite.