Leveraging Causal Policy-Reward Entropy for Enhanced Exploration

Tianying Ji[1*] **Yan Zeng**[2*] **Yu Luo**[1] **Xianyuan Zhan**[1†]
[1] Tsinghua University [2] Beijing Technology and Businesses University
[*]Equal contribution. [†]Corresponding author. zhanxianyuan@air.tsinghua.edu.cn

Abstract

The impacts of taking different actions in reinforcement learning (RL) tasks often dynamically vary during the policy learning process. We exploit the causal relationship between actions and potential reward gains, proposing a causal policy-reward entropy term. This term could effectively identify and prioritize actions with high potential impacts, thus enhancing exploration efficiency. Moreover, it could be seamlessly incorporated into any Max-Entropy RL framework. Our instantiation, termed Causal Actor-Critic (CAC), showcases superior performance across a range of continuous control tasks and provides insightful explanations for the actions.

1 Introduction

Effective exploration lies in the core of reinforcement learning (RL) for optimal decision-making (Lopes et al., 2012; Sutton & Barto, 2018; Ladosz et al., 2022). A prominent framework is Maximum Entropy RL, which promotes exploration by maximizing specific entropy terms (Zhang et al., 2021b; Agarwal et al., 2021a), often guided by the Optimism in the Face of Uncertainty (OFU) principle (Cassel et al., 2022), prioritizing less frequented actions and states, thereby broadening the visitation space. Intriguingly, we find in this paper that the conventional policy entropy (Mnih et al., 2016; Haarnoja et al., 2018; Ji et al., 2023) simply aggregates uncertainty across all action dimensions, failing to account for the varying significance of each action dimension in the policy optimization process over the course of training, hence might lead to inefficient exploration.

A concrete example is provided in Figure 1, where a robotic arm is trained to hammer a screw, revealing the varying importance of each action dimension, such as torque and positions, at different policy learning stages. Emphasizing the exploration of actions with greater potential for reward gain at the current stage could lead to enhanced learning efficiency. For example, at the stage indicated by ● in Figure 1, the arm struggles with grasping the hammer. By concentrating exploration on the gripper finger's torque, the agent develops a grasping policy more efficiently, requiring fewer samples than if it were to explore the end-effector's x and y positions. This example underscores the limitation of traditional policy entropy: relying solely on uncertainty may result in unproductive exploration, as interactions would be wasted in actions that have little impact on current performance.

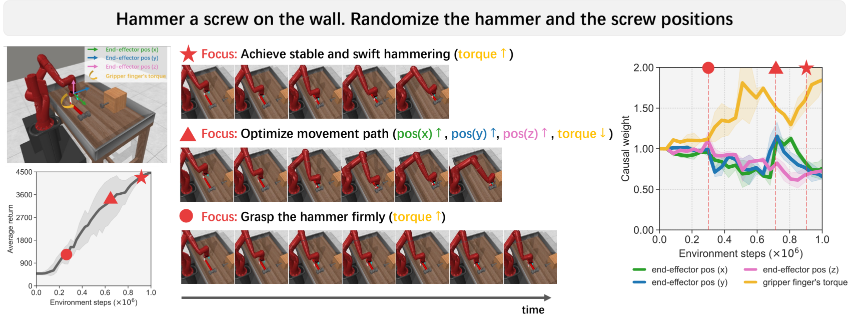


Figure 1: **Motivating example.** The task involves a robotic arm hammering a screw into a wall. ● Initially, the robotic arm struggles with hammer grasping, making torque exploration a priority for a stable grip. ▲ As the training advances, the arm is able to perform the task, but not at an optimal level. The focus shifts to optimizing movement, prioritizing end-effector position over torque. ★ Finally, potential improvements lie in the stable and swift hammering, shifting focus back to torque. The evolving causal weights, depicted on the left, reflect these changing priorities and align with human cognition.

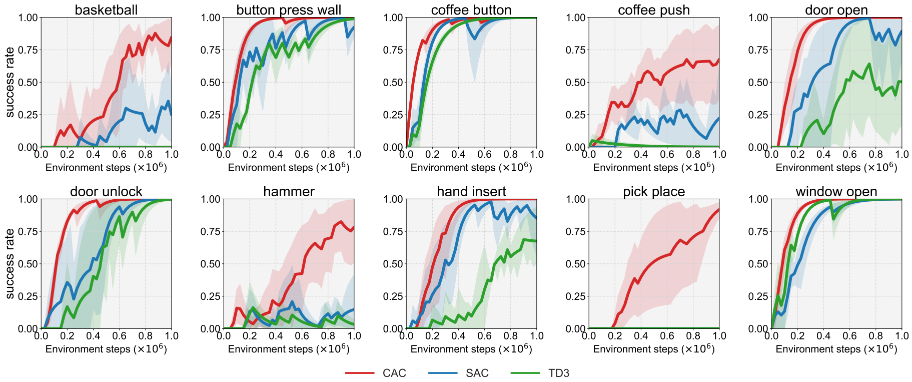


Figure 2: Success rate of CAC SAC, TD3 on 10 manipulation tasks from Meta-World (sorted alphabetically). Solid curves depict the mean of 7 runs, and shaded regions correspond to the standard deviation.

Leveraging the interpretability inherent in causality, we introduce a versatile causal policy-reward entropy. This entropy effectively identifies and prioritizes actions with a high potential for reward, enhancing exploration efficiency. Furthermore, it facilitates understanding of the agent's actions, reflecting the adaptive learning mechanisms akin to human cognition. Our implementation, CAC, achieves strong performance in various benchmark tasks.

2 Policy Learning with Causal Policy-Reward Entropy

Causal discovery on $\mathbf{a} \to \mathbf{r}|\mathbf{s}$. To explore the causal relationships between each action dimension $\mathbf{a}_i$ and its potential impact on reward gains r, we first establish a causal policy-reward structural model, $r_t = r_{\mathcal{M}}\left(\mathbf{B}_{\mathbf{s}\to r|\mathbf{a}} \odot \mathbf{s}_t, \mathbf{B}_{\mathbf{a}\to r|\mathbf{s}} \odot \mathbf{a}_t, \epsilon_t\right)$, and provide theoretical analyses to ensure the identifiability of the causal structure $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$ in Appendix B. Specifically, under the causal Markov condition and faithfulness assumption (Pearl, 2009), we establish conditions for the causal relationship existence in Proposition B.3, then the true causal graph $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$ could be identified from observational data alone, as guaranteed in Theorem B.4. Given the policy π impacts the distribution of $(\mathbf{s}, \mathbf{a})$, the corresponding causal weights also undergo dynamic evolution as it is computed on the fresh data.

Causal policy-reward entropy $\mathcal{H}_c$. By infusing the explainable causal weights $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$ into policy entropy, we propose the causal policy-reward entropy $\mathcal{H}_c$ for enhanced exploration. $\mathcal{H}_c$, defined as,

$$\mathcal{H}_c(\pi(\cdot|\mathbf{s})) = -\mathbb{E}_{\mathbf{a}\in\mathcal{A}}\left[\sum_{i=1}^{\dim\mathcal{A}} \mathbf{B}_{a_i\to r|\mathbf{s}}\pi(a_i|\mathbf{s})\log\pi(a_i|\mathbf{s})\right], \mathbf{a} = (a_1, \ldots, a_{\dim\mathcal{A}}). \tag{2.1}$$

Policy optimization with $\mathcal{H}_c$. Our causal policy-reward entropy provides a flexible solution that can be seamlessly incorporated into any Max-Entropy RL framework. For example, as a plug-and-play component, CAC can be implemented within SAC (Haarnoja et al., 2018) by integrating our $\mathcal{H}_c$ into the policy optimization objective, $J(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho(\pi)}\left[\gamma^t(r(\mathbf{s}_t, \mathbf{a}_t) + \alpha\mathcal{H}_c(\pi(\cdot|\mathbf{s}_t))\right]$.

▷ Preliminaries are detailed in Appendix A. For theoretical analyses, please refer to Appendix B. And practical implementation details are provided in Appendix E.

3 Experiments

We compare CAC to two popular model-free baselines, SAC (Haarnoja et al., 2018) and TD3 (Fujimoto et al., 2018) on a set of Meta-World (Yu et al., 2019) continuous control tasks. As shown in Figure 2, CAC surpasses SAC and TD3 by a large margin in terms of success rates. Extensive results on DMControl (Tassa et al., 2018), MuJoCo (Todorov et al., 2012), ROBEL (Ahn et al., 2020), and panda-gym (Gallouédec et al., 2021) benchmark suites are provided in Appendix G.3.

4 Conclusion

In this paper, we unveil that leveraging the causal effects of each action dimension toward potential reward gain could remarkably enhance exploration efficiency. In our experiments, we show that the proposed entropy term emerges as a versatile tool that boosts policy learning performance.

URM Statement

The authors acknowledge that all the key authors of this work meet the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021a.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021b.

Michael Ahn, Henry Zhu, Kristian Hartikainen, Hugo Ponte, Abhishek Gupta, Sergey Levine, and Vikash Kumar. Robel: Robotics benchmarks for learning with low-cost robots. In *Conference on Robot Learning*, 2020.

James Bannon, Brad Windsor, Wenbo Song, and Tao Li. Causality and batch reinforcement learning: Complementary approaches to planning in unknown domains. *arXiv preprint arXiv:2006.02579*, 2020.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 34, 2021.

Asaf B Cassel, Alon Cohen, and Tomer Koren. Efficient online linear control with stochastic convex costs and unknown dynamics. In *Conference on Learning Theory*, 2022.

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

Fan Feng, Biwei Huang, Kun Zhang, and Sara Magliacane. Factored adaptation for non-stationary reinforcement learning. *arXiv preprint arXiv:2203.16582*, 2022.

S Fujimoto, H van Hoof, and D Meger. Addressing function approximation error in actor-critic methods. *Proceedings of Machine Learning Research*, 80:1587–1596, 2018.

Quentin Gallouédec, Nicolas Cazin, Emmanuel Dellandréa, and Liming Chen. panda-gym: Open-source goal-conditioned environments for robotic learning. *arXiv preprint arXiv:2106.13687*, 2021.

Samuel J Gershman. Reinforcement learning and causal models. *The Oxford handbook of causal reasoning*, 1:295, 2017.

Lin Guan, Sarath Sreedharan, and Subbarao Kambhampati. Leveraging approximate symbolic models for reinforcement learning via skill diversity. In *International Conference on Machine Learning*, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

Seungyul Han and Youngchul Sung. Diversity actor-critic: Sample-aware entropy regularization for sample-efficient exploration. In *International Conference on Machine Learning*, 2021.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *International Conference on Learning Representations*, 2022a.

Biwei Huang, Chaochao Lu, Liu Leqi, José Miguel Hernández-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*. International Machine Learning Society, 2022b.

Biwei Huang, Chaochao Lu, Liu Leqi, José Miguel Hernández-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, 2022c.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pp. 3040–3049. PMLR, 2019.

Tianying Ji, Yu Luo, Fuchun Sun, Xianyuan Zhan, Jianwei Zhang, and Huazhe Xu. Seizing serendipity: Exploiting the value of past success in off-policy actor-critic. *arXiv preprint arXiv:2306.02865*, 2023.

Yuu Jinnai, Jee Won Park, Marlos C Machado, and George Konidaris. Exploration in reinforcement learning with deep covering options. In *International Conference on Learning Representations*, 2019.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

Xing Liu, Gaozhao Wang, Zihao Liu, Yu Liu, Zhengxiong Liu, and Panfeng Huang. Hierarchical reinforcement learning integrating with human knowledge for practical robot skill learning in complex multi-stage manipulation. *IEEE Transactions on Automation Science and Engineering*, 2023a.

Yu-Ren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. *Advances in Neural Information Processing Systems*, 2023b.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in neural information processing systems*, 2012.

Vincent Luczkow. *Structural Causal Models for Reinforcement Learning*. McGill University (Canada), 2021.

Marlos C Machado, Andre Barreto, Doina Precup, and Michael Bowling. Temporal abstraction in reinforcement learning with the successor representation. *Journal of Machine Learning Research*, 24(80):1–69, 2023.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2493–2500, 2020.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Aditya Mohan, Amy Zhang, and Marius Lindauer. Structure in reinforcement learning: A survey and open problems. *arXiv preprint arXiv:2306.16021*, 2023.

Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *Advances in neural information processing systems*, 2018.

Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

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

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

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

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(4):1225–1248, 2011.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

Hao Sun, Lei Han, Rui Yang, Xiaoteng Ma, Jian Guo, and Bolei Zhou. Optimistic curiosity exploration and conservative exploitation with linear reward shaping. In *Advances in Neural Information Processing Systems*, 2022.

Yuewen Sun, Kun Zhang, and Changyin Sun. Model-based transfer reinforcement learning based on graphical model representations. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 2012.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *the AAAI conference on artificial intelligence*, 2016.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2019.

Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *arXiv preprint arXiv:2302.05209*, 2023.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pp. 11214–11224. PMLR, 2020.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021a.

Tianjun Zhang, Paria Rashidinejad, Jiantao Jiao, Yuandong Tian, Joseph E Gonzalez, and Stuart Russell. Made: Exploration via maximizing deviation from explored regions. In *Advances in Neural Information Processing Systems*, 2021b.

Zheng-Mao Zhu, Xiong-Hui Chen, Hong-Long Tian, Kun Zhang, and Yang Yu. Offline reinforcement learning with causal structured world models. *arXiv preprint arXiv:2206.01474*, 2022a.

Zheng-Mao Zhu, Shengyi Jiang, Yu-Ren Liu, Yang Yu, and Kun Zhang. Invariant action effect model for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9260–9268, 2022b.

A Preliminaries

Markov Decision Process. We denote a Markov decision process (MDP) as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $r : \mathcal{S} \times \mathcal{A} \to [-R_{max}, R_{max}]$ the reward function, and $\gamma \in (0, 1)$ the discount factor, $P(\cdot|\mathbf{s}, \mathbf{a})$ the transition dynamics. Let $\rho(\pi)$ denote the state-action marginals of the trajectory distribution induced by $\pi(\mathbf{a}_t|\mathbf{s}_t)$. The Q-value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, corresponds to the discounted returns obtained by starting from the state $\mathbf{s}$ and action $\mathbf{a}$, and then following the policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$.

Maximum Entropy Reinforcement Learning. Maximum Entropy Reinforcement Learning maximizes policy entropy or relative entropy in addition to the standard RL objective. A general maximum entropy objective includes a policy entropy regularization term in the objective function with the aim of performing more diverse actions for each given state and visiting states with higher entropy for better exploration. The entropy-regularized policy objective function based on any policy entropy term $\mathcal{H}(\pi(\cdot|\mathbf{s}_t))$ is given as below, here α is a temperature parameter, which determines the relative importance of the entropy term against the reward.

$$J(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho(\pi)} \left[\gamma^t (r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\mathbf{s}_t))) \right] \tag{A.1}$$

Causal Policy-Reward Structural Modeling. Suppose we have sequences of observations $\{\mathbf{s}_t, \mathbf{a}_t, r_t\}_{t=1}^T$, where $\mathbf{s}_t = (s_{1,t}, ..., s_{\mathrm{dim}\mathcal{S},t})^T \subseteq \mathcal{S}$ denote the perceived $\mathrm{dim}\mathcal{S}$-dimensional states at time t, $\mathbf{a}_t = (a_{1,t}, ..., a_{\mathrm{dim}\mathcal{A},t})^T \subseteq \mathcal{A}$ are the executed $\mathrm{dim}\mathcal{A}$-dimensional actions and r_t is the reward. Note that the reward variable r_t may not be influenced by every dimension of $\mathbf{s}_t$ or $\mathbf{a}_t$, and there are causal structural relationships between $\mathbf{s}_t$, $\mathbf{a}_t$ and r_t (Huang et al., 2022c). To integrate such relationships in MDP, we explicitly encode the causal structures over variables into the reward function

$$r_t = r_{\mathcal{M}} \left(\mathbf{B}_{\mathbf{s} \to r|\mathbf{a}} \odot \mathbf{s}_t, \mathbf{B}_{\mathbf{a} \to r|\mathbf{s}} \odot \mathbf{a}_t, \epsilon_t \right), \tag{A.2}$$

where $\mathbf{B}_{\mathbf{s} \to r|\mathbf{a}} \in \mathbb{R}^{\mathrm{dim}\mathcal{S} \times 1}$ and $\mathbf{B}_{\mathbf{a} \to r|\mathbf{s}} \in \mathbb{R}^{\mathrm{dim}\mathcal{A} \times 1}$ are vectors that represent the graph structure [1] from $\mathbf{s}_t$ to r_t given $\mathbf{a}_t$ and from $\mathbf{a}_t$ to r_t given $\mathbf{s}_t$, respectively. $\odot$ denotes the element-wise product while ϵ_t are i.i.d. noise terms.

B Theoretical Analyses

We first give definitions of the Markov condition and faithfulness assumption, which will be used in our theoretical analyses.

Assumption B.1 (Global Markov Condition (Spirtes et al., 2000; Pearl, 2009)). *The distribution p over a set of variables $\mathbf{V} = (s_{1,t}, ..., s_{\mathrm{dim}\mathcal{S},t}, a_{1,t}, ..., a_{\mathrm{dim}\mathcal{A},t}, r_t)^T$ satisfies the global Markov condition on the graph if for any partition $(\mathbf{S}, \mathbf{A}, \mathbf{R})$ in $\mathbf{V}$ such that if $\mathbf{A}$ d-separates $\mathbf{S}$ from $\mathbf{R}$, then $p(\mathbf{S}, \mathbf{R}|\mathbf{A}) = p(\mathbf{S}|\mathbf{A})p(\mathbf{R}|\mathbf{A})$.*

Assumption B.2 (Faithfulness Assumption (Spirtes et al., 2000; Pearl, 2009)). *For a set of variables $\mathbf{V} = (s_{1,t}, ..., s_{\mathrm{dim}\mathcal{S},t}, a_{1,t}, ..., a_{\mathrm{dim}\mathcal{A},t}, r_t)^T$, there are no independencies between variables that are not entailed by the Markovian Condition.*

With these two assumptions, we provide the following proposition to characterize the condition of the causal relationship existence so that we are able to uncover those key actions from conditional independence relationships.

Proposition B.3. *Under the assumptions that the causal graph is Markov and faithful to the observations, there exists an edge from $a_{i,t}$ to r_t if and only if $a_{i,t} \not\perp r_t|\mathbf{s}_t, \mathbf{a}_{-i,t}$, where $\mathbf{a}_{-i,t}$ are states of $\mathbf{a}_t$ except $a_{i,t}$.*

Proof. (i) We first prove that if there exists an edge from $a_{i,t}$ to r_t, then $a_{i,t} \not\perp r_t|\mathbf{s}_t, \mathbf{a}_{-i,t}$. We prove it by contradiction. Suppose that $a_{i,t}$ is independent of r_t given $\mathbf{s}_t, \mathbf{a}_{-i,t}$. According to the

[1] Please note that $\mathbf{B}_{\cdot \to \cdot}$ encodes information of both causal directions and causal effects. For example, $B^i_{\mathbf{a} \to r} = 0$ means there is no edge between $a_{i,t}$ and r_t; and $B^i_{\mathbf{a} \to r} = c$ implies that $a_{i,t}$ causally influences r_t with effects c. Causal effects are called causal weights as well in this paper.

faithfulness assumption, we get that from the graph, $a_{i,t}$ does not have a directed path to r_t, i.e., there is no edge between $a_{i,t}$ and r_t. It contradicts our statement about the existence of the edge.

(ii) We next prove that if $a_{i,t} \not\perp\!\!\!\perp r_t | \mathbf{s}_t, \mathbf{a}_{-i,t}$, then there exists an edge from $a_{i,t}$ to r_t. Similarly, by contradiction, we suppose that $a_{i,t}$ does not have a directed path to r_t. From the definition of our MDP, we see in the graph that the path from $a_{i,t}$ to r_t could be blocked by $\mathbf{s}_t$ and $\mathbf{a}_{-i,t}$. According to the global Markov condition, $a_{i,t}$ is independent of r_t given $\mathbf{s}_t$ and $\mathbf{a}_{-i,t}$, which contradicts the assumption about the dependence between $a_{i,t}$ and r_t. □

We next provide the theorem to guarantee the identifiability of the proposed causal structure.

Theorem B.4. *Suppose $\mathbf{s}_t$, $\mathbf{a}_t$, and r_t follow the MDP model with Eq.(A.2). Under the Markov condition, and faithfulness assumption, the structural vectors $\mathbf{B}_{\mathbf{s}\to r|\mathbf{a}}$ and $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$ are identifiable.*

Proof. We prove it motivated by (Huang et al., 2022a). Denote all variable dimensions in the MDP by $\mathbf{V}$, with $\mathbf{V} = \{s_{1,t}, ...s_{\dim \mathcal{S},t}, a_{1,t}, ..., a_{\dim \mathcal{A},t}, r_t\}$, and these variables form a dynamic Bayesian network (Murphy, 2002). Note that our theorem only involves possible edges from state dimensions $s_{i,t} \in \mathbf{s}_t$ to the reward r_t or from action dimensions $a_{j,t} \in \mathbf{a}_t$ to the reward r_t. (Huang et al., 2020) showed that under the Markov condition and faithfulness assumption, even with non-stationary data, for every $V_i, V_j \in \mathbf{V}$, V_i and V_j are not adjacent in the graph if and only if they are independent conditional on some subset of other variables in $\mathbf{V}$, i.e., $\{V_l | l \neq i, l \neq j\}$. Based on this, we can asymptotically identify the correct graph skeleton over $\mathbf{V}$. Besides, due to the property of dynamic Bayesian networks that future variables can not affect past ones, we can determine the directions as $a_{i,t} \to r_t$ if $a_{i,t}$ and r_t are adjacent. So does $s_{j,t}$ and r_t. Thus, the structural vectors $\mathbf{B}_{\mathbf{s}\to r|\mathbf{a}}$ and $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$, which are parts of the graph in $\mathbf{V}$, are identifiable. Note that, following results of (Shimizu et al., 2011), if we further assume the linearity of observations as well as the non-Gaussianity of the noise terms, we can uniquely identify $\mathbf{B}_{\mathbf{s}\to r|\mathbf{a}}$ and $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$, including both causal directions and causal effects. □

C Related Works

Causal Reinforcement Learning. In the past decades, causality and reinforcement learning have independently undergone significant theoretical and technical advancements, yet the potential for a synergistic integration between the two has been underexplored (Zeng et al., 2023). Recently, recognizing the substantial capabilities of causality in addressing data inefficiency and interpretability challenges within RL, there has been a surge of research in the domain of causal reinforcement learning (Gershman, 2017; Bannon et al., 2020; Zeng et al., 2023; Deng et al., 2023; Mohan et al., 2023).

While existing methods in this area can be categorized based on whether causal information is explicitly given or not, our work falls into the more challenging, practical, and realistic category where the causal structure and effects are not explicitly provided. There exist as well other flourishing causal reinforcement learning approaches to identify and exploit causal information. These approaches are usually built upon causal influence detection (Seitzer et al., 2021; Jaques et al., 2019; Madumal et al., 2020), invariance representation learning (Zhu et al., 2022b; Zhang et al., 2021a; 2020; Bica et al., 2021; Huang et al., 2022a), factorization learning (Liu et al., 2023b; Pitis et al., 2020; Feng et al., 2022), and causal structure learning (Huang et al., 2022b; Sun et al., 2021; Zhu et al., 2022a), etc. Most approaches here assume a binary adjacency vector to characterize the existence of causal directions, however, we model with causal weights in a more refined way for $\mathbf{B}_{\mathbf{a}\to\mathbf{r}|\mathbf{s}}$. Additionally, our CAC method is most similar to the one in (Seitzer et al., 2021), which derived a measure of causal action influence and integrated it into RL algorithms to improve exploration. The key difference lies in the fact that they exploited the independent relationships between action and the next states, while we aim at inferring the causal effects between actions and reward given states, and design the causal policy-reward entropy term to enhance exploration. The proposed term works as a plug-and-play component with causal information, which is simpler and more adaptive. In causal reinforcement learning, one of the challenges lies in the characterization of shifting structures and effects from data, which might affect the performances of policy learning. To this end, alternatives may involve incorporating the changing structures between states (Luczkow, 2021) into policy learning or modeling the changes using some dynamic factors (Huang et al., 2022a; Feng et al., 2022).

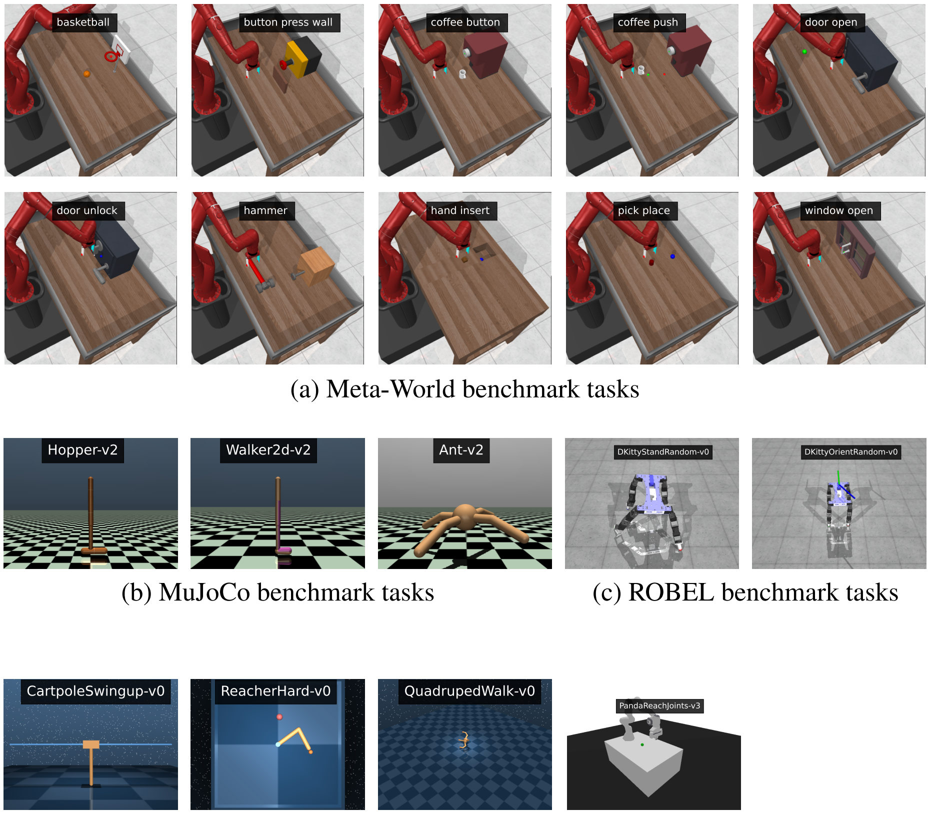
Our approach, however, focuses on capturing changing causal effects, a nuanced facet that improves policy exploration.

RL in multi-stage learning. Inspired by human cognition, dividing the original task into multiple stages for policy learning has been explored in RL through heuristic methods tailored to specific tasks (Jinnai et al., 2019; Liu et al., 2023a). For instance, hierarchical RL (Sutton et al., 1999; Pateria et al., 2021) has implemented this idea by introducing additional subgoal spaces or leveraging the semi-Markov assumption. This allows the agent to segment tasks into different subtasks (Nachum et al., 2018), options (Machado et al., 2023) or skills (Guan et al., 2022) as multiple stages, enabling policies to possess distinct exploration capabilities across various stages, ultimately enhancing the success of agents in complex continuous control tasks.

In contrast to explicit multi-stage approaches, our proposed method CAC does not necessitate a clear task stage division. Instead, it can identify and prioritize actions with a high potential for reward through causal discovery dynamically, emphasizing the various importance of each action dimension in different stages for enhanced exploration and performance.

D Environment Setup

We evaluate CAC across 19 diverse continuous control tasks, spanning MuJoCo (Todorov et al., 2012), Robel (Ahn et al., 2020), DMControl (Tassa et al., 2018), and Meta-World (Yu et al., 2019). It excels in both locomotion and manipulation tasks. Visualizations of these tasks are provided in Figure 3.



(a) Meta-World benchmark tasks

(b) MuJoCo benchmark tasks

(c) ROBEL benchmark tasks

(d) DMControl benchmark tasks

(e) panda-gym benchmark tasks

Figure 3: Visualization of 19 benchmark tasks.

E Practical Implementation

Instantiating CAC amounts to specifying two main components: 1) how to effectively recognize the causal weights of $\mathbf{a} \to \mathbf{r}|\mathbf{s}$; 2) how to incorporate causal weights and the corresponding causal policy-reward entropy term into policy optimization. The pseudocode of our proposed CAC is provided in Algorithm 1.

Causal discovery on $\mathbf{a} \to \mathbf{r}|\mathbf{s}$. To effectively compute $\mathbf{B}_{\mathbf{a}\to\mathbf{r}|\mathbf{s}}$, we adopt the well-regarded DirectLiNGAM method (Shimizu et al., 2011). While alternative score-based methods that simultaneously learn causal effects can also be employed, we opt for DirectLiNGAM for two main reasons: 1) Empirical validation confirms its remarkably exceptional performance, prioritizing actions with higher reward potential and aligning with human cognition in executing complex tasks. 2) Under the linearity assumption, one can straightforwardly and practically learn coefficients as causal effects. Moreover, the non-Gaussianity assumption facilitates the unique identification of the causal structure. The main implementation idea of DirectLiNGAM is as follows. In the first phase, it estimates a causal ordering for all variables of interest (i.e., state, action, and reward variables), based on the independence and non-Gaussianity characteristics of the root variable. The causal ordering is a sequence that implies the latter variable cannot cause the former one. In the second phase, DirectLiNGAM estimates the causal effects between variables, using some conventional covariance-based methods such as least squares and maximum likelihood approaches. Its convergence is guaranteed theoretically under some assumptions. Besides, we formulate a training regime wherein we iteratively adjust the causal weights for the policy at regular intervals I on a local buffer $\mathcal{D}_c$ with fresh transitions to reduce computation cost.

Policy optimization. Given the causal weight matrix $\mathbf{B}_{\mathbf{a}\to\mathbf{r}|\mathbf{s}}$, we could obtain the causal policy-reward entropy $\mathcal{H}_c(\pi(\cdot|\mathbf{s}))$ through Eq.(2.1). Note that to ease the computation burden of updating the causal weight matrix, we opt to conduct causal discovery with a fixed interval.

Based on the causal policy-reward entropy, then the Q-value for a fixed policy π could be computed iteratively by applying a modified Bellman operator $\mathcal{T}_c^\pi$ with $\mathcal{H}_c(\pi(\cdot|\mathbf{s}))$ term as stated below,

$$\mathcal{T}_c^\pi Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi}[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \alpha \mathcal{H}_c(\pi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}))]\right]. \tag{E.1}$$

In particular, we parameterize two Q-networks and train them independently, and then adopt the commonly used double-Q-techniques (Van Hasselt et al., 2016; Fujimoto et al., 2018; Haarnoja et al., 2018; Han & Sung, 2021; Sun et al., 2022; Ji et al., 2023) to obtain the minimum of the Q-functions for policy optimization. Based on the policy evaluation, we can adopt many off-the-shelf policy optimization oracles; we chose SAC as the backbone technique primarily for its simplicity in our primary implementation of CAC.

F Hyper-parameters

The hyperparameters used for training CAC are outlined in Table 1.

Algorithm 1: Primary Implementation of Causal Actor-Critic (CAC)

```
initialize : Q network $Q_\phi$, policy network $\pi_\theta$, replay buffer $\mathcal{D}$;
             local buffer $\mathcal{D}_c$ with size $N_c$, causal weight matrix $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$;
for each environment step $t$ do
    Collect data with $\pi_\theta$ from real environment
    Add to replay buffer $\mathcal{D}$ and local buffer $\mathcal{D}_c$
/* Causal discovery                                                  */
if every $I$ environment step then
    Sample all $N_c$ transitions from local buffer $\mathcal{D}_c$
    Update causal weight matrix $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$
for each gradient step do
    Sample $N$ transitions $(s, a, r, s')$ from $\mathcal{D}$
    /* Policy evaluation                                              */
    Compute causal policy-reward entropy $\mathcal{H}_c(\pi(\cdot|\mathbf{s}))$ by Eq.(2.1)
    Calculate the target $Q$ value by Eq.(E.1)
    Update $Q_\phi$ by $\min_\phi (\mathcal{B}Q_\phi - Q_\phi)^2$
    /* Policy optimization                                            */
    Update $\pi_\theta$ by $\max_\theta Q_\phi(s, a)$
```

Table 1: Hyperparameter settings for CAC.

Hyper-parameter	Value
Q-value network	MLP with hidden size 512
V-value network	MLP with hidden size 512
policy network	Gaussian MLP with hidden size 512
discounted factor γ	0.99
soft update factor τ	0.005
learning rate α	0.0003
batch size N	512
policy updates per step	1
value target updates interval	2
sample size for causality N_c	5000
causality computation interval I	5000

G Additional Experimental Results

G.1 Additional Metrics

We report additional (aggregate) performance metrics of CAC and baselines on the set of 10 Meta-World tasks using the rliable toolkit (Agarwal et al., 2021b). As shown in Figure 37, CAC outperforms SAC and TD3 in terms of Median, interquantile mean (IQM), Mean, and Optimality Gap.

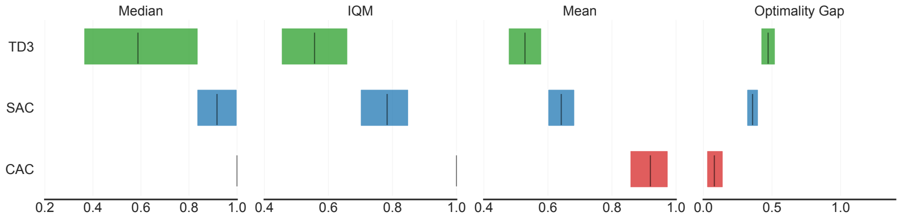


Figure 4: **Rliable metrics.** Median, IQM, Mean (higher values are better), and Optimality Gap (lower values are better) of CAC and baselines on the 10 Meta-World manipulation tasks. 7 random seeds.

G.2 Ablation Studies

Hyperparameter study. The extra hyperparameters introduced by CAC are sample size for causality N_c and causality computation interval I. The primary choices for both hyperparameters are guided by the objective of achieving a balanced trade-off between computational efficiency and algorithmic performance. And they are sufficient to achieve strong performance throughout all our experiments.

We conduct experiments on these two hyperparameters; refer to Figure 5. We see that reducing the causality computation interval may increase the performance yet cause more computation cost. And the performance of CAC is not highly sensitive to the hyperparameters.

Different causal inference methods. We initially opted for DirectLiNGAM due to its simplicity and efficacy in learning causal effects. However, to explore the adaptability of our CAC framework with other score-based causal inference methods, we conducted additional experiments using Dagma (Bello et al., 2022). These experiments were aimed at assessing whether different causal inference techniques could yield comparable results within our framework. The results in Figure 6 indicate that the integration of Dagma into the CAC method produces outcomes that are on par with those obtained using DirectLiNGAM. This suggests that our CAC framework is versatile and can effectively work with various causal inference methods.

G.3 Generalizability and Effectiveness of CAC

Figure 2 in the main paper illustrates the superiority of CAC in dealing with manipulation tasks with end-effector control. Here, we conduct more experiments on the various locomotion and manipulation tasks to further demonstrate the generalizability of CAC. We evaluate CAC and baselines in robot locomotion tasks based on MuJoCo (Todorov et al., 2012) and DMControl (Tassa et al., 2018) benchmark tasks, as shown in Figure 7a and 7b, our CAC outperforms in terms of both the eventual performance and the sample efficiency.

Moreover, we conduct experiments in sparse reward tasks to showcase the efficiency of CAC. We evaluate in both robot locomotion and manipulation tasks, based on the sparse reward version of benchmark tasks from ROBEL (Ahn et al., 2020) and panda-gym (Gallouédec et al., 2021). Panda-gym manipulation tasks are based on a Franka Emika Panda robot with joint angle control. ROBEL quadruped locomotion tasks are based on a D'Kitty robot with 12 joint positions control. As shown in Figure 7c, our CAC surpasses the baselines by a large margin.

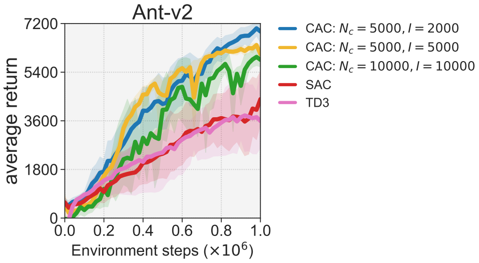


Figure 5: **Hyperparameter study.** Performance curves of CAC with different hyperparameters over 7 random seeds.

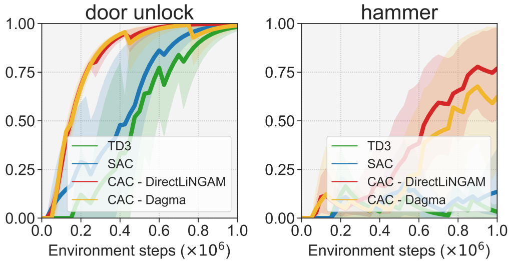
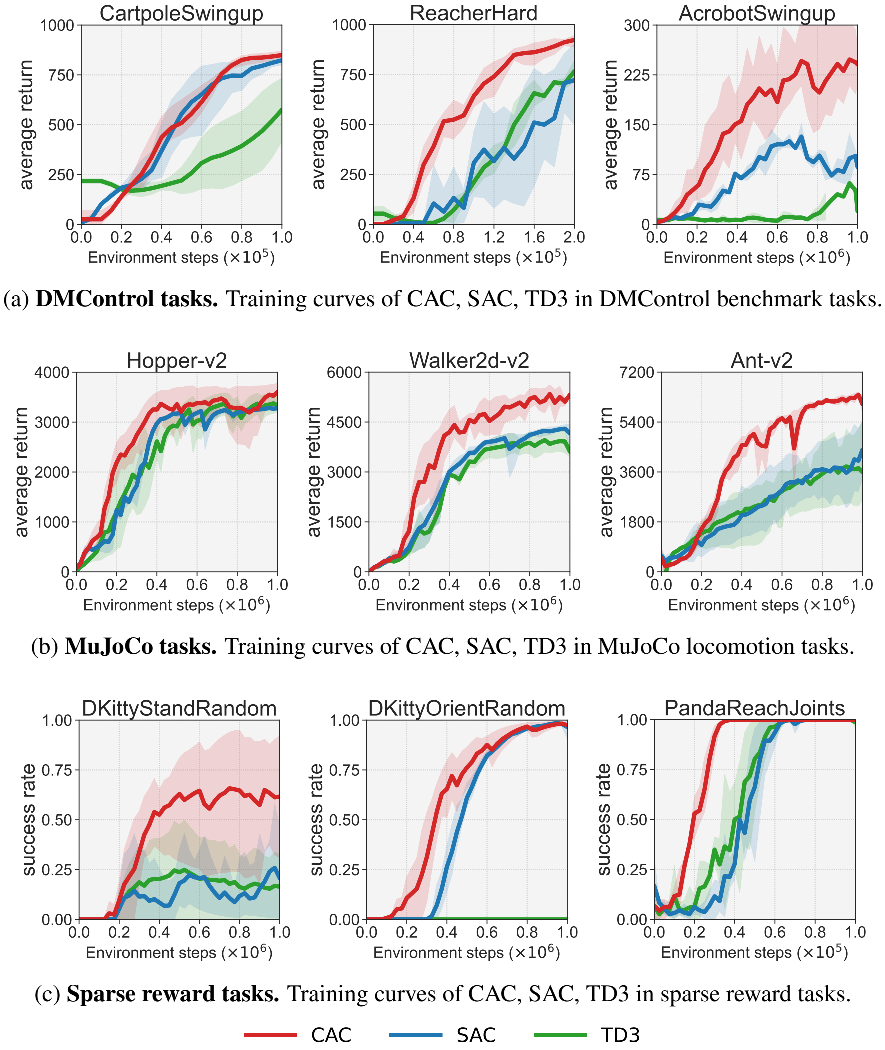


Figure 6: **Different causal inference methods.** Performance curves of CAC employing DirectLiNGAM or Dagma. Runs over 7 random seeds.



(a) **DMControl tasks.** Training curves of CAC, SAC, TD3 in DMControl benchmark tasks.

(b) **MuJoCo tasks.** Training curves of CAC, SAC, TD3 in MuJoCo locomotion tasks.

(c) **Sparse reward tasks.** Training curves of CAC, SAC, TD3 in sparse reward tasks.

Figure 7: Performance comparison of CAC and baselines on various continuous control tasks on the **DMControl**, **MuJoCo**, **ROBEL**, and **panda-gym** robotics platforms. These tasks include manipulation and locomotion tasks with dense and sparse reward types. Solid curves depict the mean of 4 trials and shaded regions correspond to the standard deviations.

H Computing Infrastructure and Computational Time

Our experiments were conducted on a server equipped with an `AMD EPYC 7763 64-Core Processor (256 threads)` and four `NVIDIA GeForce RTX 3090 GPUs`.

Figure 8 presents the computational time comparison between our algorithm CAC and SAC on 10 Meta-World benchmark tasks. Compared to SAC, the total training time of CAC only increased by an average of 0.67 hours, hence, the additional costs are acceptable. Further, for practical use, CAC requires fewer interactions for similar performance, which may lower the needed computation time.

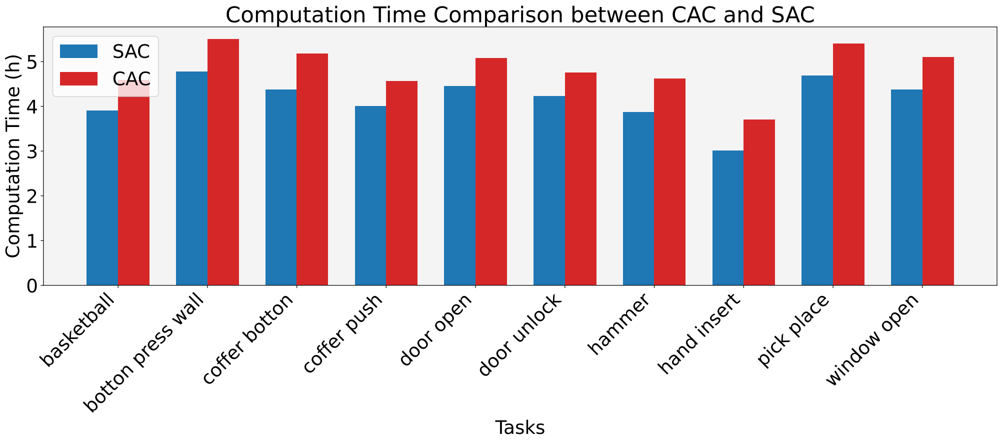


Figure 8: **Computation time comparison.** Computation time comparison between CAC and SAC in ten Meta-World tasks, each averaged on 4 trials.