ACE : Off-Policy Actor-Critic with Causality-Aware Entropy Regularization

Tianying Ji [*1] **Yongyuan Liang** [*2] **Yan Zeng** [3] **Yu Luo** [1] **Guowei Xu** [1] **Jiawei Guo** [1] **Ruijie Zheng** [2] **Furong Huang** [2] **Fuchun Sun** [1] **Huazhe Xu** [1 4]

Abstract

The varying significance of distinct primitive behaviors during the policy learning process has been overlooked by prior model-free RL algorithms. Leveraging this insight, we explore the causal relationship between different action dimensions and rewards to evaluate the significance of various primitive behaviors during training. We introduce a causality-aware entropy term that effectively identifies and prioritizes actions with high potential impacts for efficient exploration. Furthermore, to prevent excessive focus on specific primitive behaviors, we analyze the gradient dormancy phenomenon and introduce a dormancy-guided reset mechanism to further enhance the efficacy of our method. Our proposed algorithm, **ACE**: Off-policy **A**ctor-critic with **C**ausality-aware **E**ntropy regularization, demonstrates a substantial performance advantage across 29 diverse continuous control tasks spanning 7 domains compared to model-free RL baselines, which underscores the effectiveness, versatility, and efficient sample efficiency of our approach. Benchmark results and videos are available at https://ace-rl.github.io/.

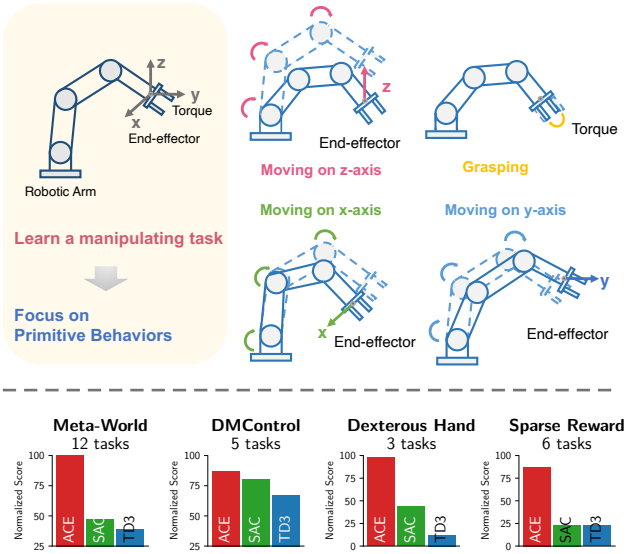


Figure 1: *(Top)*: **Learning process of a manipulator.** A robotic arm learns to manipulate objects in a manner akin to human learning. This arm would be programmed with four primitive behaviors for its end-effector: vertical movements along the z-axis (up and down), horizontal movements along the x-axis (left and right), depth movements along the y-axis (forward and backward), and grasping (apply torque). *(Bottom)*: **Comparison of normalized score.** Our **ACE** demonstrates a significant superiority over the widely used model-free RL baselines SAC and TD3 with a single set of hyperparameters.

1. Introduction

Reinforcement Learning (RL) has made remarkable strides in addressing complex decision-making problems, ranging from video games (Mnih et al., 2013; Silver et al., 2016) to robot control (Schulman et al., 2015; 2017b; Lee et al., 2020). Despite this, a persistent challenge in RL is high sample complexity, which poses a formidable impediment to the practical application of deep RL in real-world scenarios. Effective exploration lies in the core of reinforcement learning (RL) for optimal decision-making as well as sample efficiency (Lopes et al., 2012; Sutton & Barto, 2018; Ladosz et al., 2022). Ineffective exploration would lead to unsatisfactory sample efficiency, as the agent may spend excessive interactions in low-value or irrelevant areas that do not contribute to performance improvement. In order to enhance sample efficiency, previous works have proposed various exploration strategies, such as upper-confidence bounds (UCB) based exploration (Chen et al., 2017), curiosity-driven exploration for sparse reward tasks (Pathak et al., 2017), random network distillation (Burda et al., 2018), maximum entropy RL (Haarnoja et al., 2018), etc.

Intriguingly, existing exploration methods typically simply aggregate uncertainty across all action dimensions, failing to account for the varying significance of each primitive behavior in the policy optimization process over the course

[*]Equal contribution [1]Tsinghua University [2]University of Maryland [3]Beijing Technology and Businesses University [4]Shanghai Qi Zhi Institute. Correspondence to: Tianying Ji <jity20@mails.tsinghua.edu.cn>, Yongyuan Liang <cheryll-Liang@outlook.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

of training. These methods may neglect a fundamental aspect: mastering a motor task often involves multiple stages, each requiring proficiency in different primitive behaviors, similar to how humans learn (Roy et al., 2022). Consider a simple example, as shown in Figure 1: a manipulator should initially learn to lower its arm and grasp the object, then shift attention to learning the movement direction of the arm towards the final goal. Therefore, it is crucial to emphasize the exploration of the most significant primitive behaviors at distinct stages of policy learning. The deliberate focus on various primitive behaviors during exploration can accelerate the agent's learning of essential primitive behaviors at each stage, thus improving the efficiency of mastering the complete motor task.

How can we identify the most crucial primitive behaviors at each stage of policy learning? To answer this question, we evaluate the significance of individual primitive behaviors by quantifying their impact on rewards through the analysis of causal relationships. Within each task, the action spaces of RL agents comprise dimensions representing individual primitive behaviors—such as the 3D positioning of the end-effector and manipulation of gripper fingers. When a particular action dimension demonstrates a larger influence on rewards, it indicates its higher importance in the ongoing learning stage, and vice versa. We introduce a causal policy-reward structural model to compute the causal weights on action spaces and provide theoretical analyses to ensure the identifiability of the causal structure. In Figure 2, we illustrate the changes in causal weights on four action dimensions in a manipulator task, contrasting them with the agent's behaviors at different time points. This effectively elucidates how the agent's focus on different primitive behaviors changes during the learning process, as reflected in the causal weights.

Therefore, causal weights naturally guide agents to conduct more efficient exploration, encouraging exploration for action dimensions with larger causal weights, indicating greater significance on rewards, and reducing exploration for those with smaller causal weights. Our methodology builds on the maximum entropy framework (Haarnoja et al., 2018; Ziebart et al., 2008; Zhao et al., 2019), incorporating a regularization term for policy entropy in the objective function. The general maximum entropy objective lacks awareness of the importance of distinctions between primitive behaviors at different learning stages, potentially resulting in inefficient exploration. To address this limitation, we introduce a policy entropy weighted by causal weights as a causality-aware entropy maximization objective, effectively strengthening the exploration of significant primitive behaviors and leading to more efficient exploration.

To mitigate the risk of overfitting due to excessive focus on certain behaviors, we introduce a gradient-dormancy-based reset mechanism based on our analysis of the gradient dormancy phenomenon. This mechanism intermittently perturbs the agent's neural networks with a factor determined by the gradient dormancy degree. The integration of causality-aware exploration with this novel reset mechanism aims to facilitate more efficient and effective exploration, ultimately enhancing the agent's overall performance.

We conduct a comprehensive evaluation of our method across 28 diverse continuous control tasks, spanning 7 domains, including tabletop manipulation (Yu et al., 2019a; Gallouédec et al., 2021b), locomotion control (Todorov et al., 2012a; Tassa et al., 2018a; Ahn et al., 2020c), and dexterous hand manipulation tasks (Rajeswaran et al., 2018; Plappert et al., 2018), covering a broad spectrum of task difficulties including sparse reward tasks. In comparison with popular model-free RL algorithms - TD3 (Fujimoto et al., 2018) and SAC (Haarnoja et al., 2018) and exploration method RND (Burda et al., 2018), our results consistently demonstrate that **ACE** outperforms all the baselines across various task types. Specifically, our method achieves performance improvements of *2.1x* on highly challenging manipulator tasks, *1.1x* on locomotion tasks, *2.2x* on dexterous hand tasks, and *3.7x* on tasks with sparse rewards as shown in Figure 1. Our work introduces a novel perspective to RL by deconstructing the learning process and analyzing the varying significance of primitive behaviors over the course of policy optimization. Our method leverages these insights to enhance sample efficiency via efficient exploration, and we empirically validate its effectiveness through an extensive set of experiments.

Our **contributions** can be summarized as follows:

- We propose a causal policy-reward structural model to calculate the impact on rewards from different primitive behaviors. This provides insight that agents purposefully focus on various primitive behaviors during RL training.
- We introduce a technology that employs causal weights on policy entropy as a causality-aware entropy objective to enhance exploration efficiency.
- Additionally, we analyze the gradient dormancy phenomenon and propose a gradient-dormancy-guided reset mechanism to prevent agents from excessively overfitting to certain behaviors.
- Extensive experiments across a wide range of tasks and domains consistently achieve superior performance against model-free RL baselines with a single set of hyperparameters, showing the effectiveness of our method. Our work offers valuable insights into enhancing sample efficiency more systematically by focusing on the efficient exploration of significant primitive behaviors.

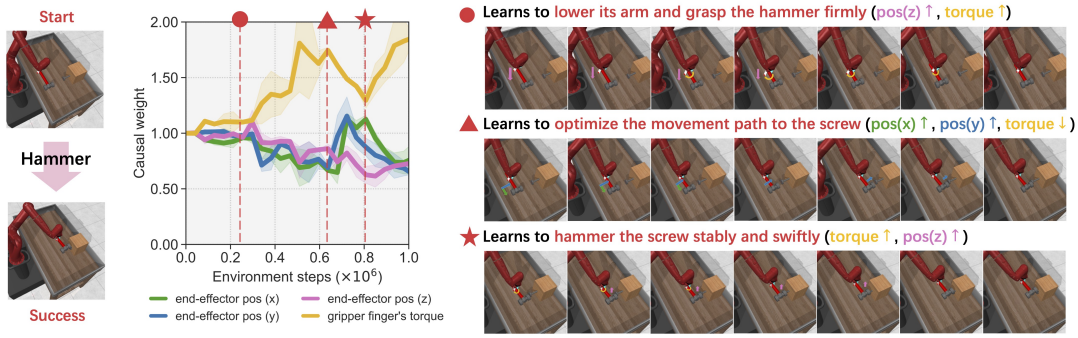


Figure 2: **Motivating example.** This task involves a robotic arm hammering a screw into a wall. ● Initially, the robotic arm approaches the desk moving on the z-axis and struggles with torque grasping, making z-axis positioning ↑ and torque exploration ↑ a priority. ▲ As the training advances, the agent's focus shifts to optimizing movement, prioritizing end-effector position (x-axis ↑ and y-axis ↑). ★ Finally, potential improvements lie in the stable and swift hammering, shifting focus back to torque ↑ and placing down the object ↑. The evolving causal weights, depicted on the left, reflect these changing priorities. See more examples in Appendix D.

2. Preliminary

Reinforcement Learning (RL). Within the standard framework of the Markov decision process (MDP), RL can be formulated as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. Here, $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $r : \mathcal{S} \times \mathcal{A} \in [-R_{max}, R_{max}]$ the reward function, and $\gamma \in (0, 1)$ the discount factor, and $P(\cdot \mid s, a)$ stands for transition dynamics. The objective of an RL agent is to learn an optimal policy π that maximizes the expected discounted sum of rewards, formulated as $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$. For any $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, the value of action a under state s is given by the action-value function $Q^\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)]$. In deep RL, the policy π and the value function Q are represented by neural network function approximations.

Soft Actor-Critic (SAC). Soft actor-critic (SAC) is a popular off-policy maximum entropy deep reinforcement learning algorithm based on soft policy iteration that maximizes the entropy-augmented policy objective function. Its 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, stated as below,

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

3. Off-policy Actor-critic with Causality-aware Entropy regularization

Overview. Our approach builds upon maximum entropy RL (Haarnoja et al., 2018), with the primary aim of enhancing the sample efficiency of off-policy RL algorithms. Initially, we construct a causal policy-reward structural model, providing a theoretical guarantee for its identifiability. Subsequently, based on this causal model, we introduce causality-aware entropy to present a causality-aware variant of SAC named CausalSAC. To address potential overfitting and further enhance exploration efficiency, we analyze the gradient dormancy phenomenon and propose a gradient-dormancy-guided reset mechanism. The integration of this reset mechanism with CausalSAC constitutes our proposed **ACE**: Off-policy **A**ctor-critic with **C**ausality-aware **E**ntropy regularization algorithm.

3.1. Causal Discovery on Policy-Reward Relationship

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 and provide theoretical analyses to ensure the identifiability of the causal structure.

Causal policy-reward structural modeling. Suppose we have sequences of observations $\{\mathbf{s}_t, \mathbf{a}_t, r_t\}_{t=1}^T$, where $\mathbf{s}_t = (s_{1,t}, ..., s_{\dim\mathcal{S},t})^T \subseteq \mathcal{S}$ denote the perceived $\dim\mathcal{S}$-dimensional states at time t, $\mathbf{a}_t = (a_{1,t}, ..., a_{\dim\mathcal{A},t})^T \subseteq \mathcal{A}$ are the executed $\dim\mathcal{A}$-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., 2022b). To integrate such relationships in MDP, we explicitly encode the causal structures over variables into the reward function

$$r_t = r_{\mathcal{M}} \left(\mathbf{B}_{\mathbf{s} \to r|\mathbf{a}} \odot \mathbf{s}_t, \mathbf{B}_{\mathbf{a} \to r|\mathbf{s}} \odot \mathbf{a}_t, \epsilon_t \right), \quad (1)$$

where $\mathbf{B}_{\mathbf{s} \to r|\mathbf{a}} \in \mathbb{R}^{\dim\mathcal{S} \times 1}$ and $\mathbf{B}_{\mathbf{a} \to r|\mathbf{s}} \in \mathbb{R}^{\dim\mathcal{A} \times 1}$ 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. Here $\odot$ denotes the element-wise product while ϵ_t are i.i.d. noise

[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.

terms.

Specifically, under the causal Markov condition and faithfulness assumption (Pearl, 2009), we establish conditions for the causal relationship existence in Proposition 3.3, then the true causal graph $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$ could be identified from observational data alone, as guaranteed in Theorem 3.4.

Assumption 3.1 (Global Markov Condition (Spirtes et al., 2000; Pearl, 2009))**.** The distribution p over a set of variables $\mathbf{V} = (s_{1,t}, ..., s_{\dim\mathcal{S},t}, a_{1,t}, ..., a_{\dim\mathcal{A},t}, r_t)^T$ 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 3.2 (Faithfulness Assumption (Spirtes et al., 2000; Pearl, 2009))**.** For a set of variables $\mathbf{V} = (s_{1,t}, ..., s_{\dim\mathcal{S},t}, a_{1,t}, ..., a_{\dim\mathcal{A},t}, r_t)^T$, 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 3.3. *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 $a_{i,t} \not\perp\!\!\!\perp r_t|\mathbf{s}_t, \mathbf{a}_{-i,t}$, where $\mathbf{a}_{-i,t}$ are states of $\mathbf{a}_t$ except $a_{i,t}$.*

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

Theorem 3.4. *Suppose s_t, a_t, and r_t follow the MDP model with no unobserved confounders, as in Eq.(1). Under the Markov condition and faithfulness assumption, the structural vectors $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$ are identifiable.*

Note that such a theorem guarantees the identifiability of the correct graph in an asymptotic manner. Additionally, by imposing further assumptions on the data generation mechanism, we could uniquely identify the causal effects. See the proof in Appendix A.1 for details.

3.2. Causality-aware Bellman Operator

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

$$\mathcal{H}_c(\pi(\cdot|\mathbf{s})) = -\mathbb{E}_{\mathbf{a}\in\mathcal{A}}\left[\sum_{i=1}^{\dim\mathcal{A}} \mathbf{B}_{a_i\to r|\mathbf{s}}\pi(a_i|\mathbf{s})\log\pi(a_i|\mathbf{s})\right], \quad \mathbf{a} = (a_1, \ldots, a_{\dim\mathcal{A}}). \tag{2}$$

Based on the causality-aware 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,

$$\mathcal{T}_c^\pi Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\mathbb{E}_{\mathbf{s}_{t+1}\sim P}[\mathbb{E}_{\mathbf{a}_t\sim\pi}[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \alpha\mathcal{H}_c(\pi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}))]]. \tag{3}$$

For a better understanding of our operator, we conduct a theoretical analysis of its dynamic programming properties in the tabular MDP setting, covering policy evaluation, policy improvement, and policy iteration. All proofs are included in Appendix A.2.

Proposition 3.5 (Policy evaluation)**.** *Consider an initial $Q_0 : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ with $|\mathcal{A}| < \infty$, and Q-vlaue iterates by $Q_{k+1} = \mathcal{T}_c^\pi Q_k$. Then the sequence $\{Q_k\}$ converges to a fixed point Q^π as $k\to\infty$.*

Proposition 3.6 (Policy improvement)**.** *Let π_k be the policy at iteration k, and π_{k+1} be the updated policy (maximize of the Q-value). Then for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $|\mathcal{A}| < \infty$, we have $Q^{\pi_{k+1}}(\mathbf{s},\mathbf{a}) \geq Q^{\pi_k}(\mathbf{s},\mathbf{a})$.*

Proposition 3.7 (Policy iteration)**.** *Assume $|\mathcal{A}| < \infty$, by repeating iterations of the policy evaluation and policy improvement, any initial policy converge to the optimal policy π^*, s.t. $Q^{\pi^*}(\mathbf{s}_t,\mathbf{a}_t) \geq Q^\pi(\mathbf{s}_t,\mathbf{a}_t), \forall\pi\in\Pi, \forall(\mathbf{s}_t,\mathbf{a}_t)$.*

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

3.3. Gradient-dormancy-guided Reset

Guided by causality-aware entropy, the agent efficiently explores and masters primitive behaviors for different learning stages. However, causality-aware exploration introduces the risk of getting stuck in local optima and overfitting to specific primitive behaviors. To address this challenge, we analyze the gradient dormancy phenomenon during RL training and introduce a soft reset mechanism. This mechanism, guided by gradient dormancy, regularly perturbs the agents' neural networks to maintain network expressivity, thereby improving the agent's performance.

The dormancy phenomenon of neural networks in RL, signifying a loss of expressive capacity, has been previously discussed in existing works (Sokar et al., 2023; Xu et al., 2023). However, the dormant phenomenon defined in these works is not evident in state-based RL, and it cannot effectively enhance our algorithm in exploration scheduling. We are the first to investigate dormancy from the perspective of gradients. Here, we introduce definitions for gradient-

dormant neurons and the gradient dormancy degree of a neural network.

Definition 3.8 (Gradient-dormant Neurons)**.** For a fully connected layer in a neural network, where N^l represents the number of neurons in layer l, the L2 norm of gradients of the weights for neuron i is denoted as n_i^l. Neuron i is classified as a gradient-dormant neuron if it satisfies

$$\frac{n_i^l(x)}{\frac{1}{N^l}\sum_{k\in l} n_k^l} \leq \tau, \tag{4}$$

where τ is a constant serving as a threshold to determine the gradient dormancy of neurons in each layer.

Definition 3.9 (τ-Dormancy Degree α_τ)**.** Denote the number of all neurons in the neural network identified as gradient-dormant neurons as N_τ^l. The α_τ for the neural network is defined as:

$$\alpha_\tau = \frac{\sum_{l\in\phi} N_\tau^l}{\sum_{l\in\phi} N^l}. \tag{5}$$

The τ-dormancy degree α_τ indicates the percentage of gradient-dormant neurons with the τ threshold in the fully-connected neural network.

In Figure 3, we depict the occurrence of the gradient dormancy phenomenon at the initial learning stages of SAC. However, in algorithms with superior sample efficiency like CausalSAC and **ACE** , there is a notable reduction in the dormancy degree during training. In tasks that CausalSAC can effectively solve, the gradient dormancy degree can decrease to near-zero levels, as shown in Figure 3. Particularly, when addressing challenging tasks such as pick place walls, which were previously unsolved by existing baselines, dormancy degrees persist at high levels without reset interventions in SAC and CausalSAC. Hence, we speculate that this may represent a potential local optimum for causal-aware exploration. We consider that dormancy degrees may impact sample efficiency and employ a soft reset method (Xu et al., 2023; Ash & Adams, 2020) to further decrease dormancy degree by periodically perturbing the policy network and critic network with a reset factor η, representing the magnitude of weight resetting:

$$\theta_t = (1-\eta)\theta_{t-1} + \eta\phi_i,\ \phi_i \sim \text{initializer}. \tag{6}$$

Intuitively, a higher dormancy degree should correspond to a more substantial degree of weight refresh. The value of η is determined by the gradient-dormancy degree α_τ and regulated as $\eta = \text{clip}(\alpha_\tau, 0, \eta_{\text{max}})$, with $\eta_{\text{max}} \leq 1$ as a constant. Compared with prior works (Sokar et al., 2023; Xu et al., 2023), our gradient-dormancy-guided reset method employs minimal hyperparameters and is highly adaptable in different RL settings, as analyzed in Appendix J.

3.4. Algorithm instantiation

Combining the causality-aware entropy regularization and gradient-dormancy-guided reset mechanism, we propose our algorithm **ACE** : off-policy **A**ctor-critic with **C**ausality-aware **E**ntropy regularization. The pseudocode and further implementation details are provided in Appendix E.

Instantiating **ACE** involves specifying three main components: 1) effectively recognizing the causal weights of $\mathbf{a} \to \mathbf{r}|\mathbf{s}$; 2) incorporating causal weights and the corresponding causality-aware entropy term into policy optimization. 3) periodically resetting the network based on our gradient dormancy degree.

To effectively compute $\mathbb{B}_{a_i \to r|s}$, we adopt the well-regarded DirectLiNGAM (Shimizu et al., 2011) method. The main implementation idea of training 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. 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.

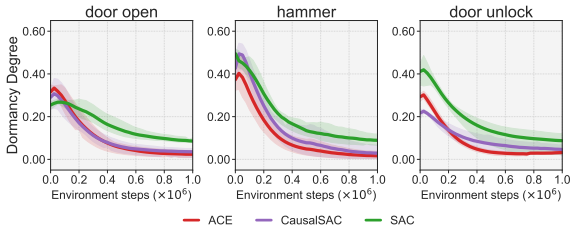
Given the causal weight matrix $\mathbf{B}_{\mathbf{a}\to\mathbf{r}|\mathbf{s}}$, we could obtain the causality-aware entropy $\mathcal{H}_c(\pi(\cdot|\mathbf{s}))$ through Eq.(2). Based on the causality-aware entropy, then the Q-value for a fixed policy π could be computed iteratively by applying $\mathcal{T}_c^\pi$. 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 CausalSAC and **ACE**.

For each reset interval, we calculate the gradient dormancy degree, initialize a random network with weights ϕ_i, and soft reset the policy network π_θ and the Q network Q_ϕ.

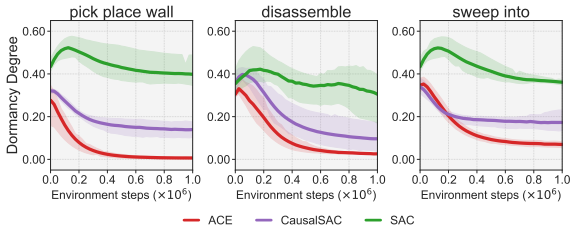
In the next section, we empirically substantiate the effectiveness and efficiency of our proposed causality-aware actor-critic, coupled with the gradient-dormancy-guided reset mechanism, through an extensive array of experiments.

4. Experiments

Our experiments aim to investigate the following questions: 1) How effective is the proposed **ACE** in diverse continuous control tasks, spanning locomotion and manipulation skills and covering both sparse and dense reward settings, compared to model-free RL baselines? 2) What role does each component of our method play in achieving the final performance? 3) How do hyperparameters and the use of



(a) In simpler tasks, CausalSAC performs adequately, with observed gradient dormancy degrees decreasing to near-zero. This indicates that in less complex environments, the issue of gradient dormancy becomes negligible.



(b) In challenging tasks, both SAC and CausalSAC exhibit high dormancy degrees without achieving expected performance. The gradient dormancy-guided reset mechanism in ACE, however, effectively lowers the gradient dormancy degree.

Figure 3: Dormancy degree curves for SAC, CausalSAC, and **ACE** in MetaWorld tasks, which indicates that the gradient-dormancy-guided reset mechanism effectively reduces gradient dormancy degrees, contributing to the best performance of **ACE** .

different causal inference methods impact our results?

4.1. Evaluation on various benchmark suites

We evaluate **ACE** across 29 diverse continuous control tasks spanning 7 task domains using a single set of hyperparameters: **MuJoCo** (Todorov et al., 2012a), **Meta-World** (Yu et al., 2019a), **Deepmind Control Suite** (Tassa et al., 2018a), **Adroit** (Rajeswaran et al., 2018), **Shadow Dexterous Hand** (Plappert et al., 2018), **Panda-gym** (Gallouédec et al., 2021b), and **ROBEL** (Ahn et al., 2020b).Our experimental tasks include high-dimensional state and action spaces, sparse rewards, multi-object manipulation, and locomotion for diverse embodiments, and cover a broad spectrum of task difficulties. **ACE** can be applied to a variety continuous control problems covering high-dimensional state and action spaces, sparse rewards, multi-object manipulation, and locomotion for diverse embodiments, without any need for hyperparameter-tuning.

Baselines. We compare our method with three popular data-efficient model-free RL baselines on all the tasks and an efficient exploration method on tasks with sparse rewards: 1) Soft Actor-Critic (SAC) (Haarnoja et al., 2018), a model-free off-policy actor-critic algorithm with maximum entropy regularization. 2) Twin Delayed DDPG (TD3) (Fujimoto et al., 2018): An advanced version of DDPG, which trains a deterministic policy in an off-policy way, incorporating Double Q-learning tricks. 3) Random Network Distillation (RND) (Burda et al., 2018), an efficient exploration bonus based on the error of predicting features of observations, which is particularly effective for sparse reward tasks.

We also provide a comparison with BAC (Ji et al., 2023a), which leverages the value of past successes to enhance Q-value estimation and policy learning, in Appendix F.1. Additionally, we integrate our causality-aware entropy and reset mechanism into the BAC algorithm and find that the ad-hoc ACE-BAC also outperforms the original BAC, as shown in Appendix F.4, which further showcases the generalizability of our method. Detailed performance curves on different benchmark suites are provided in Appendix F.

Tabletop Manipulation. We conducted experiments on tabletop manipulation tasks in MetaWorld, tackling 14 tasks with dense rewards, spanning 4 very hard, 7 hard, and 3 medium tasks, including all types of tasks and all levels of task difficulties, as shown in Figure 4. Notably, **ACE** exhibited a substantial lead of over 70% in very hard tasks, coupled with noteworthy performance improvements exceeding 30% in hard and medium tasks. Traditional model-free RL baselines often struggle to accomplish very hard tasks like pick-place-wall, stick push and disassemble. In contrast, our approach not only demonstrates superior learning efficiency but also achieves a flawless 100% success rate in these challenging tasks. These results prominently highlight the high sample efficiency of **ACE** in tabletop manipulation tasks, underscoring the crucial role of causality-aware entropy in improving sample efficiency, along with the positive impact of our reset mechanism on effective exploration, particularly in challenging exploration tasks.

Locomotion. Another important task category involves locomotion. We conducted experiments on four MuJoCo tasks and five DeepMind Control Suite tasks, encompassing diverse embodiments as presented in Figure 5. Our algorithm achieves state-of-the-art performance and demonstrates improvements across all locomotion tasks, with particularly notable advancements in the Walker2d and HalfCheetah tasks. It is noteworthy to observe that, due to the relatively smaller total number of training steps required for locomotion tasks, the gradient dormancy phenomenon is less pronounced compared to manipulation tasks. Consequently, our reset mechanism may not exhibit its optimal effect. To provide further evaluation, we include a detailed comparison between our **ACE** CausalSAC and baselines in Figure 26. The results underscore the significant performance enhancement that causality-aware entropy brings to off-policy RL in locomotion tasks.

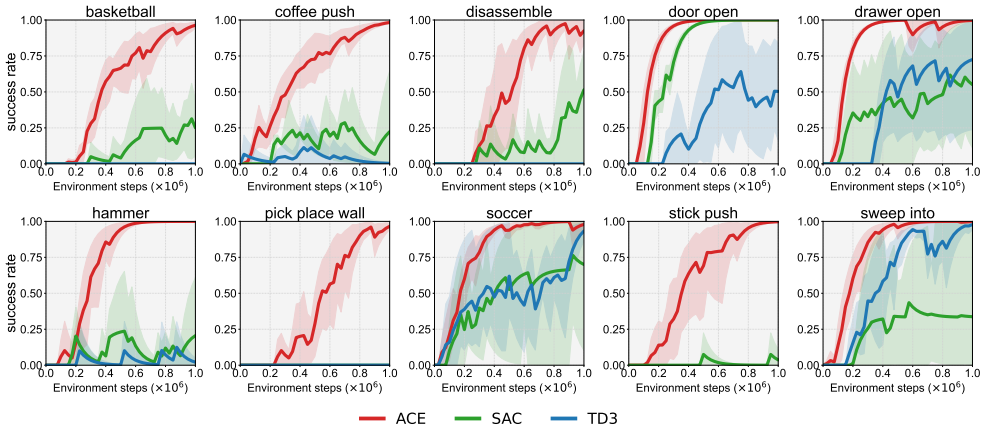


Figure 4: **Manipulation tasks.** Success rate of **ACE**, SAC, TD3 on manipulation tasks from the MetaWorld benchmark suite. Solid curves depict the mean of six trials, and shaded regions correspond to the one standard deviation. More results are in Appendix Figure 22.

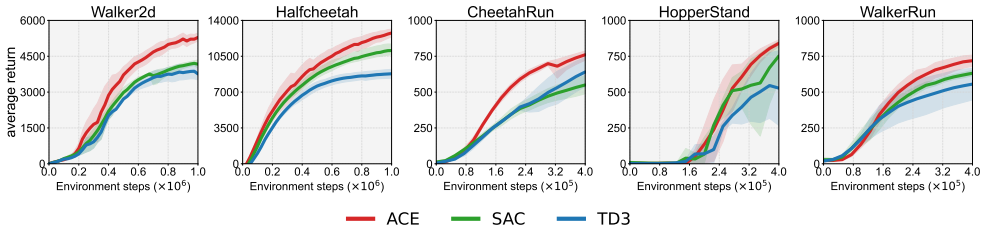


Figure 5: **Locomotion tasks.** Average return of **ACE** SAC, TD3 on locomotion tasks provided by MuJoCo and DMControl benchmark suites. Solid curves depict the mean of six trials, and shaded regions correspond to the one standard deviation. See Figure 23 and 24 in the Appendix for an overall comparison of locomotion tasks.

Dexterous hand manipulation. To evaluate our method on high-dimensional tasks, we compare **ACE** with baselines on three dexterous hand manipulation tasks, including Adroit (Rajeswaran et al., 2018), which involves controlling a robotic hand with up to 30 actuated degrees of freedom ($\mathcal{A} \in \mathbb{R}^{28}$), and Shadow Dexterous Hand (Plappert et al., 2018), a robotic hand with 24 degrees of freedom ($\mathcal{A} \in \mathbb{R}^{20}$). Notably, tasks with Shadow Dexterous Hand include multi-goal manipulation, and we train all the algorithms without goal information on these tasks. As shown in Figure 6, **ACE** consistently outperforms baselines by a significant margin on all the dexterous hand manipulation tasks. This outstanding performance underscores the effectiveness of our causal model in computing meaningful causal weights in high-dimensional action spaces, leading to a substantial improvement in exploration efficiency for these tasks.

Hard exploration tasks with sparse rewards. To better illustrate the effectiveness of our proposed method in improving exploration efficiency, we evaluate our approach against baselines and the efficient exploration method RND (Burda et al., 2018) on tasks with sparse rewards. These tasks pose significant challenges for online RL exploration, covering both complex robot locomotion (Panda-gym (Gallouédec et al., 2021b) and ROBEL (Ahn et al., 2020b)) and manipulation (MetaWorld (Yu et al., 2019a)), as illustrated in Figure 7. Across 6 sparse reward tasks, **ACE** outperforms the efficient exploration method RND and surpasses all other baselines by a significant margin, particularly excelling in tasks where the baselines completely fail to learn. These results showcase the superior sample efficiency of our method and further support the versatility of **ACE** in various challenging exploration tasks.

4.2. Ablation Studies

To substantiate the efficacy of the design decisions in our algorithm, we conduct extensive ablation experiments.

Effects of each mechanism. We conduct ablation studies on the MetaWorld tasks with dense and sparse rewards to evaluate the contribution of each component to our method, including causality-aware entropy and gradient-dormancy-guided reset. Our ablation results are shown in Figure 9.

We observe that all components of our proposed methodology significantly contribute to the final performance of **ACE** . Even without the reset mechanism, CausalSAC (vanilla ACE) demonstrates superior performance. Further, results on sparse reward tasks show that CausalSAC outperforms other SOTA exploration techniques. This highlights its potential in challenging environments.

Simultaneously, our reset mechanism also plays a crucial role in further improving performance, especially in sparse reward tasks. Interestingly, applying the gradient-dormancy-

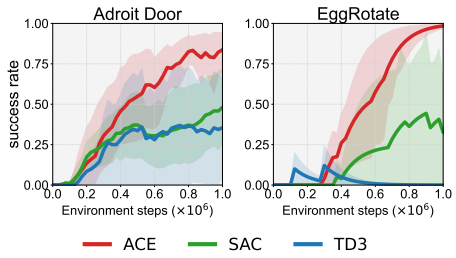


Figure 6: **Shadow hand manipulation tasks.** Success rate of **ACE** , SAC, TD3 on challenging shadow hand manipulation tasks from Adroit and Shadow Dexterous Hand suites. 6 seeds. All the results are in Figure 25.

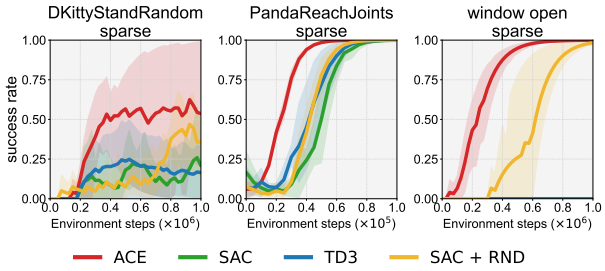


Figure 7: **Sparse reward tasks.** Success rate of **ACE**, SAC, TD3, SAC+RND on sparse reward tasks from ROBEL, Panda-gym and Metaworld benchmark suites. 6 seeds. Full performance on sparse reward tasks is in Figure 27.

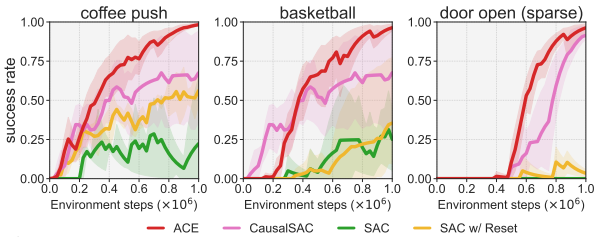


Figure 8: **Ablation experiments.** We ablate each component of **ACE** and show that each mechanism effectively combines to contribute to the overall effectiveness of **ACE** .

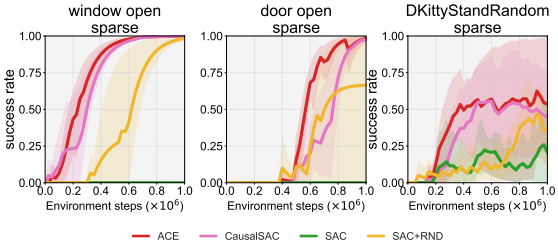


Figure 9: **Ablation studies on SOTA exploration techniques.** Learning curves of ACE, CausalSAC, SAC, and SOTA exploration technique RND on three sparse reward tasks.

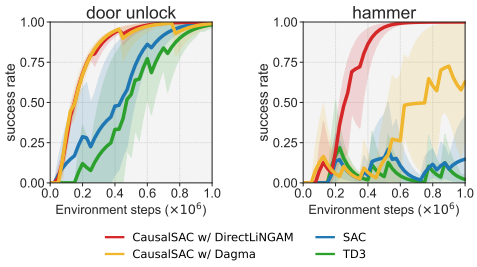


Figure 10: **Different causal inference methods.** Performance curves of **CausalSAC** employing DirectLiNGAM or Dagma. Runs over 6 random seeds.

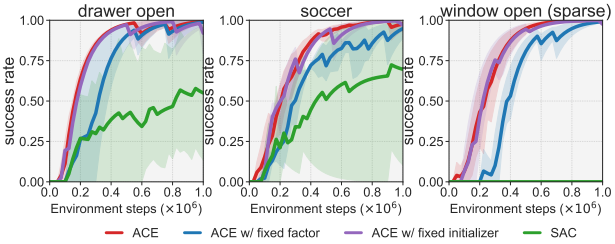


Figure 11: **Ablation on reset mechanism factors.** Learning curves of **ACE** w/wo fixed reset factor η and fixed initializer Φ.

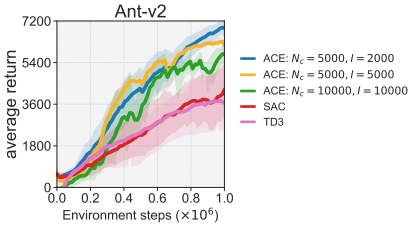


Figure 12: **Hyperparameter study.** Performance curves of **ACE** with different hyperparameters for calculating causal weights.

guided reset solely to SAC demonstrates improved performance and sample efficiency, suggesting that the gradient dormant phenomenon might indeed be one of the factors contributing to SAC's suboptimal performance in hard tasks.

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 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 10 indicate that the integration of Dagma into our method produces outcomes that are on par with those obtained using DirectLiNGAM. This suggests that our framework is versatile and can effectively work with various causal inference methods. Notably, to eliminate the possible influence from other factors except the underlying causal inference method, we deactivate the reset mechanism in **ACE**, thus excluding its impact. We employ CausalSAC for ablations related to the causal inference method.

Ablation studies on reset mechanism factors. We conduct ablation studies on the fixed reset factor η and the fixed network initializer ϕ across three tasks. As shown in Figure 11, fixing the reset factor (i.e., no longer guiding it by the dormancy degree) results in a significant performance

drop. This highlights the importance of the proposed dormancy degree in our method. On the other hand, fixing the network initializer ϕ does not affect the performance of the dormancy-guided reset mechanism in ACE.

Hyperparameter study. The extra hyperparameters introduced by **ACE** 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 12. We see that reducing the causality computation interval may increase the performance yet cause more computation cost. And the performance of **ACE** is not highly sensitive to the hyperparameters.

5. 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).

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. 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.

Exploration in RL. Efficient exploration in online RL, especially in high-dimensional environments with sparse rewards, remains a significant challenge. Exploration strategies can be broadly categorized into two major groups based on key ideas and principles. One category is uncertainty-oriented exploration (Jin et al., 2020; Ménard et al., 2021; Kaufmann et al., 2021), utilizing techniques like the upper confidence bound (UCB) (Chen et al., 2017) to incorporate value estimate uncertainty for guiding exploration. The other category is intrinsic motivation-oriented exploration, incentivizing agents to explore by maximizing intrinsic rewards (Pathak et al., 2017; Burda et al., 2018; Sekar et al., 2020; Badia et al., 2020), count-based state novelty (Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017), or maximizing state entropy as an intrinsic reward (Lee et al., 2019; Hazan et al., 2019b; Mutti et al., 2022; Yang & Spaan, 2023). In contrast to exploration methods that focus on sampling data, the concept of resetting the neural network has gained attention in RL as a unique approach to mitigate the loss of network expressivity. Nikishin et al. (2022) address the primacy bias by periodically reinitializing the parameters, and Nikishin et al. (2023) propose to temporarily freeze the current network and leverage newly initialized weights. The dormant neuron phenomenon, as discussed by Sokar et al. (2023), suggests resetting dormant neurons to preserve network expressivity during training. For visual RL, Xu et al. (2023) propose a perturbation technique and exploration strategy based on the dormant neuron ratio. In contrast to these prior exploration methods, our approach combines causality-aware max-entropy RL with a reset mechanism, adjusting exploration through causality between rewards and policy, and employs a novel gradient dormancy degree for resetting. We also discuss extensive related works about maximum-entropy RL in Appendix B.

6. Conclusions and Discussion

This paper introduces insights into the consideration of the significance of various primitive behaviors throughout the policy learning process. Building upon this understanding, we propose a causal policy-reward structural model to quantify the impact of each action dimension on rewards. We introduce a causality-aware off-policy actor-critic algorithm with a novel gradient-dormancy-guided reset mechanism, achieving efficient and effective exploration and establishing a substantial lead across various domains and tasks.

Looking ahead, there are several directions for further exploration stemming from this work. First, despite our analysis of computing causal weights in the latent space and their application to DrM (Xu et al., 2023) in Appendix J, there is untapped potential for exploring additional applications of causality-aware exploration to enhance the sample efficiency of visual RL. Second, it is worth discussing the prospect of leveraging dynamic models in model-based RL to facilitate more efficient computation of causality between rewards and policy. Concerning time efficiency, a detailed analysis is provided in Appendix L, confirming the high efficiency and cost-effectiveness of our algorithm. We believe that the implications of causality-aware exploration extend beyond our current work, offering more effective solutions to enhance the sample efficiency of online RL.

Impact Statement

This research advances both cognition and application of Reinforcement Learning, particularly in the domain of off-policy actor-critic. Our exploration uncovers a fascinating insight: by harnessing the causal influences of each action dimension on potential rewards, we can dramatically boost the efficiency of exploration. Moreover, the resetting technique we devised is not just innovative; it echoes the renewal mechanism found in the human brain, empowering RL agents to escape the confines of local optima with remarkable agility. This breakthrough not only deepens our understanding of the agent's decision-making process but also mirrors the dynamic and adaptive learning capabilities akin to human cognition.

The introduction of causality entropy and the resetting mechanism as flexible enhancements to existing models introduce an exhilarating possibility for the evolution of RL. However, it is worth noting that venturing into real-world environments with these RL agents brings to light a significant challenge: navigating the need for stringent safety measures to curb any risky behaviors during exploration.

Acknowledgements

This work was funded by the National Science and Technology Major Project (2021ZD0113804), partly by Tsinghua University-Toyota Joint Research Center for AI Technology of Automated Vehicle (TTAD 2024-03), the National Natural Science Foundation of China under Grant (U22A2057) and Sino-German Collaborative Research Project "Crossmodal Learning" (NSFC 62061136001/ DFG TRR169). The authors would also like to thank the anonymous reviewers for their careful reading and their many insightful comments.

References

Ahn, M., Zhu, H., Hartikainen, K., Ponte, H., Gupta, A., Levine, S., and Kumar, V. Robel: Robotics benchmarks for learning with low-cost robots. In *Conference on Robot Learning*, 2020a.

Ahn, M., Zhu, H., Hartikainen, K., Ponte, H., Gupta, A., Levine, S., and Kumar, V. Robel: Robotics benchmarks for learning with low-cost robots. In *Conference on robot learning*, 2020b.

Ahn, M., Zhu, H., Hartikainen, K., Ponte, H., Gupta, A., Levine, S., and Kumar, V. Robel: Robotics benchmarks for learning with low-cost robots. In *Conference on robot learning*, pp. 1300–1313. PMLR, 2020c.

Ash, J. and Adams, R. P. On warm-starting neural network training. *Advances in neural information processing systems*, 33:3884–3894, 2020.

Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., and Blundell, C. Never give up: Learning directed exploration strategies. In *ICLR*. OpenReview.net, 2020.

Bannon, J., Windsor, B., Song, W., and Li, T. Causality and batch reinforcement learning: Complementary approaches to planning in unknown domains. *arXiv preprint arXiv:2006.02579*, 2020.

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *NIPS*, pp. 1471–1479, 2016.

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

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Chen, R. Y., Sidor, S., Abbeel, P., and Schulman, J. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Chen, X., Wang, C., Zhou, Z., and Ross, K. W. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2020.

Deng, Z., Jiang, J., Long, G., and Zhang, C. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

Feng, F., Huang, B., Zhang, K., and Magliacane, S. Factored adaptation for non-stationary reinforcement learning. *arXiv preprint arXiv:2203.16582*, 2022.

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

Gallouédec, Q., Cazin, N., Dellandréa, E., and Chen, L. panda-gym: Open-source goal-conditioned environments for robotic learning. *arXiv preprint arXiv:2106.13687*, 2021a.

Gallouédec, Q., Cazin, N., Dellandréa, E., and Chen, L. panda-gym: Open-source goal-conditioned environments for robotic learning. *arXiv preprint arXiv:2106.13687*, 2021b.

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

Grau-Moya, J., Leibfried, F., and Vrancx, P. Soft q-learning with mutual-information regularization. In *International conference on learning representations*, 2018.

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

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

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*, 2018.

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

Han, S. and Sung, Y. A max-min entropy framework for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:25732–25745, 2021b.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691. PMLR, 2019a.

Hazan, E., Kakade, S. M., Singh, K., and Soest, A. V. Provably efficient maximum entropy exploration. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2681–2691. PMLR, 2019b.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

Huang, B., Feng, F., Lu, C., Magliacane, S., and Zhang, K. Adarl: What, where, and how to adapt in transfer reinforcement learning. *International Conference on Learning Representations*, 2022a.

Huang, B., Lu, C., Leqi, L., Hernández-Lobato, J. M., Glymour, C., Schölkopf, B., and Zhang, K. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, 2022b.

Ji, T., Luo, Y., Sun, F., Zhan, X., Zhang, J., and Xu, H. Seizing serendipity: Exploiting the value of past success in off-policy actor-critic. *arXiv preprint arXiv:2306.02865*, 2023a.

Ji, T., Luo, Y., Sun, F., Zhan, X., Zhang, J., and Xu, H. Seizing serendipity: Exploiting the value of past success in off-policy actor-critic. *arXiv preprint arXiv:2306.02865*, 2023b.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 2020.

Jinnai, Y., Park, J. W., Machado, M. C., and Konidaris, G. Exploration in reinforcement learning with deep covering options. In *International Conference on Learning Representations*, 2019.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *ALT*, volume 132 of *Proceedings of Machine Learning Research*, pp. 865–891. PMLR, 2021.

Ladosz, P., Weng, L., Kim, M., and Oh, H. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., and Hutter, M. Learning quadrupedal locomotion over challenging terrain. *Sci. Robotics*, 5(47):5986, 2020.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Levine, S. and Koltun, V. Guided policy search. In *International conference on machine learning*, pp. 1–9. PMLR, 2013.

Li, Q., Peng, Z., Wu, H., Feng, L., and Zhou, B. Human-ai shared control via policy dissection. *Advances in Neural Information Processing Systems*, 35:8853–8867, 2022.

Liu, X., Wang, G., Liu, Z., Liu, Y., Liu, Z., and Huang, P. Hierarchical reinforcement learning integrating with human knowledge for practical robot skill learning in complex multi-stage manipulation. *IEEE Transactions on Automation Science and Engineering*, 2023.

Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in neural information processing systems*, 2012.

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

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

Ménard, P., Domingues, O. D., Shang, X., and Valko, M. UCB momentum q-learning: Correcting the bias without forgetting. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7609–7618. PMLR, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

Mutti, M., Santi, R. D., and Restelli, M. The importance of non-markovianity in maximum state entropy exploration. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16223–16239. PMLR, 2022.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. In *Advances in neural information processing systems*, 2018.

Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P., and Courville, A. C. The primacy bias in deep reinforcement learning. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16828–16847. PMLR, 2022.

Nikishin, E., Oh, J., Ostrovski, G., Lyle, C., Pascanu, R., Dabney, W., and Barreto, A. Deep reinforcement learning with plasticity injection. *arXiv preprint arXiv:2305.15555*, 2023.

O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V. Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*, 2016.

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. Count-based exploration with neural density models. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2721–2730. PMLR, 2017.

Pateria, S., Subagdja, B., Tan, A.-h., and Quek, C. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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

Plappert, M., Andrychowicz, M., Ray, A., McGrew, B., Baker, B., Powell, G., Schneider, J., Tobin, J., Chociej, M., Welinder, P., Kumar, V., and Zaremba, W. Multi-goal reinforcement learning: Challenging robotics environments and request for research, 2018.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems*, 2018.

Roy, D. S., Park, Y.-G., Kim, M. E., Zhang, Y., Ogawa, S. K., DiNapoli, N., Gu, X., Cho, J. H., Choi, H., Kamentsky, L., et al. Brain-wide mapping reveals that engrams for a single memory are distributed across multiple brain regions. *Nature communications*, 13(1):1799, 2022.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Chen, X., and Abbeel, P. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017a.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592. PMLR, 2020.

Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P., and Lee, K. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, pp. 9443–9454. PMLR, 2021.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(4):1225–1248, 2011.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sokar, G., Agarwal, R., Castro, P. S., and Evci, U. The dormant neuron phenomenon in deep reinforcement learning.

In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32145–32168. PMLR, 2023.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2000.

Sun, H., Han, L., Yang, R., Ma, X., Guo, J., and Zhou, B. Optimistic curiosity exploration and conservative exploitation with linear reward shaping. In *Advances in Neural Information Processing Systems*, 2022.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

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

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

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. #exploration: A study of count-based exploration for deep reinforcement learning. In *NIPS*, pp. 2753–2762, 2017.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018a.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018b.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012a.

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

Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *the AAAI conference on artificial intelligence*, 2016.

Xu, G., Zheng, R., Liang, Y., Wang, X., Yuan, Z., Ji, T., Luo, Y., Liu, X., Yuan, J., Hua, P., et al. Drm: Mastering visual reinforcement learning through dormant ratio minimization. *arXiv preprint arXiv:2310.19668*, 2023.

Yang, Q. and Spaan, M. T. J. CEM: constrained entropy maximization for task-agnostic safe exploration. In *AAAI*, pp. 10798–10806. AAAI Press, 2023.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100. PMLR, 2019a.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2019b.

Zeng, Y., Cai, R., Sun, F., Huang, L., and Hao, Z. A survey on causal reinforcement learning. *arXiv preprint arXiv:2302.05209*, 2023.

Zhao, R., Sun, X., and Tresp, V. Maximum entropy-regularized multi-goal reinforcement learning. In *International Conference on Machine Learning*, pp. 7553–7562. PMLR, 2019.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A. Theoretical Analyses

A.1. Causal identifiability

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

Assumption A.1 (Global Markov Condition (Spirtes et al., 2000; Pearl, 2009))**.** The distribution p over a set of variables $\mathbf{V} = (s_{1,t}, ..., s_{\dim\mathcal{S},t}, a_{1,t}, ..., a_{\dim\mathcal{A},t}, r_t)^T$ 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 A.2 (Faithfulness Assumption (Spirtes et al., 2000; Pearl, 2009))**.** For a set of variables $\mathbf{V} = (s_{1,t}, ..., s_{\dim\mathcal{S},t}, a_{1,t}, ..., a_{\dim\mathcal{A},t}, r_t)^T$, 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 A.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 $a_{i,t} \not\perp\!\!\!\perp r_t|\mathbf{s}_t, \mathbf{a}_{-i,t}$, 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 $a_{i,t} \not\perp\!\!\!\perp r_t|\mathbf{s}_t, \mathbf{a}_{-i,t}$. 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 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 A.4. *Suppose $\mathbf{s}_t$, $\mathbf{a}_t$, and r_t follow the MDP model with Eq.(1). 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. □

A.2. Approximate dynamic programming properties

Proposition A.5 (Policy evaluation)**.** *Consider an initial $Q_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathcal{A}| < \infty$, and define $Q_{k+1} = \mathcal{T}_c^\pi Q_k$. Then the sequence $\{Q_k\}$ converges to a fixed point Q^π as $k \to \infty$.*

Proof. First, let us show that our causal policy-reward Bellman operator $\mathcal{B}$ is a γ-contraction operator in the $\mathcal{L}_\infty$ norm.

Let Q_1 and Q_2 be two arbitrary Q functions, for the Bellman operator $\mathcal{T}_c$, we have,

$$\begin{aligned}\|\mathcal{T}_c^{\pi}Q_1 - \mathcal{T}_c^{\pi}Q_2\|_\infty =& \max_{\mathbf{s},\mathbf{a}} |\gamma \mathbb{E}_{\mathbf{s}'}\left[\mathbb{E}_{\mathbf{a}'\sim\pi}Q_1(\mathbf{s}',\mathbf{a}') - \gamma\mathbb{E}_{\mathbf{a}'\sim\pi}Q_2(\mathbf{s}',\mathbf{a}')\right]| \\ \leq& \gamma \max_{\mathbf{s},\mathbf{a}} \mathbb{E}_{\mathbf{s}'}|\mathbb{E}_{\mathbf{a}'\sim\pi}Q_1(\mathbf{s}',\mathbf{a}') - \mathbb{E}_{\mathbf{a}'\sim\pi}Q_2(\mathbf{s}',\mathbf{a}')| \\ \leq& \gamma \max_{\mathbf{s},\mathbf{a}} \mathbb{E}_{\mathbf{s}'}\mathbb{E}_{\mathbf{a}'\sim\pi}|Q_1(\mathbf{s}',\mathbf{a}') - Q_2(\mathbf{s}',\mathbf{a}')| \\ \leq& \gamma \max_{\mathbf{s},\mathbf{a}} \|Q_1 - Q_2\|_\infty = \gamma\|Q_1 - Q_2\|_\infty\end{aligned}$$

we conclude that the Bellman operator $\mathcal{T}_c$ satisfies γ -contraction property, which naturally leads to the conclusion that any initial Q function will converge to a unique fixed point by repeatedly applying $\mathcal{T}_c^{\pi}$. □

Proposition A.6 (Policy improvement)**.** *Let π_k be the policy at iteration k, and π_{k+1} be the updated policies, where π_{k+1} is the greedy policy of the Q-value. Then for all $(\mathbf{s},\mathbf{a}) \in \mathcal{S}\times\mathcal{A}$, $|\mathcal{A}| < \infty$, we have $Q^{\pi_{k+1}}(\mathbf{s},\mathbf{a}) \geq Q^{\pi_k}(\mathbf{s},\mathbf{a})$.*

Proof. At iteration k, π_k denotes the policy, and the corresponding value function is $Q^{\{\mu,\pi\}}$. We update the policy from π_k to π_{k+1}, where π_{k+1} is the greedy policy w.r.t $J_{\pi_k}(\pi)$, *i.e.*, $\pi_{k+1} = \arg\max_\pi \mathbb{E}_{\mathbf{a}\sim\pi}[Q^{\pi_k}(\mathbf{s},\mathbf{a}) + \alpha\mathcal{H}_c(\pi(\mathbf{a}|\mathbf{s}))]$.

Since $\pi_{k+1} = \arg\max_\pi J_{\pi_k}(\pi)$, we have that $J_{\pi_k}(\pi_{k+1}) \geq J_{\pi_k}(\pi_k)$. Expressing $J_{\pi_k}(\pi_{k+1})$ and $J_{\pi_k}(\pi_k)$ by their definition, we have $\mathbb{E}_{\mathbf{a}\sim\pi_{k+1}}[Q^{\pi_k}(\mathbf{s},\mathbf{a}) + \alpha\mathcal{H}_c(\pi_{k+1}(\mathbf{a}|\mathbf{s}))] \geq \mathbb{E}_{\mathbf{a}\sim\pi_k}[Q^{\pi_k}(\mathbf{s},\mathbf{a}) + \alpha\mathcal{H}_c(\pi_k(\mathbf{a}|\mathbf{s}))]$.

In a similar way to the proof of the soft policy improvement (Haarnoja et al., 2017), we come to the following inequality:

$$\begin{aligned}Q^{\pi_k}(\mathbf{s}_t,\mathbf{a}_t) =& r(\mathbf{s}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{\mathbf{s}_{t+1}}\left\{\mathbb{E}_{a_{t+1}\sim\pi_k}[Q^{\pi_k}(\mathbf{s}_{t+1},\mathbf{a}_{t+1}) + \alpha\mathcal{H}_c(\pi_k(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})]\right\} \\ \leq& r(\mathbf{s}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{\mathbf{s}_{t+1}}\{\mathbb{E}_{a_{t+1}\sim\pi_{k+1}}[Q^{\pi_k}(\mathbf{s}_{t+1},\mathbf{a}_{t+1}) + \alpha\mathcal{H}_c(\pi_{k+1}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})]\} \\ &\vdots \\ \leq& Q^{\pi_{k+1}}(\mathbf{s}_t,\mathbf{a}_t)\end{aligned}$$

Here, the inequality is obtained by repeatedly expanding Q^{π_k} on the RHS through $Q^{\pi_k}(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \gamma\mathbb{E}_{\mathbf{s}'}[\mathbb{E}_{a'\sim\pi_k}[Q^{\pi_k}(\mathbf{s}',\mathbf{a}') + \alpha\mathcal{H}_c(\pi_k(\mathbf{a}'|\mathbf{s}')]]$ and applying the inequality $\mathbb{E}_{\mathbf{a}\sim\pi_{k+1}}[Q^{\pi_k}(\mathbf{s},\mathbf{a}) - \omega(\mathbf{s},\mathbf{a}|\pi_{k+1})] \geq \mathbb{E}_{\mathbf{a}\sim\pi_k}[Q^{\pi_k}(\mathbf{s},\mathbf{a}) + \mathcal{H}_c(\pi_k(\mathbf{a}|\mathbf{s})]$. Finally, we arrive at convergence to $Q^{\pi_{k+1}}(\mathbf{s}_t,\mathbf{a}_t)$ and finish the proof. □

Proposition A.7 (**Policy iteration**)**.** *Assume $|\mathcal{A}| < \infty$, repeated application of the policy evaluation and policy improvement to any initial policy converges to a policy π^*, s.t. $Q^{\pi^*}(\mathbf{s}_t,\mathbf{a}_t) \geq Q^{\pi'}(\mathbf{s}_t,\mathbf{a}_t), \forall\pi' \in \Pi, \forall(\mathbf{s}_t,\mathbf{a}_t) \in \mathcal{S}\times\mathcal{A}$.*

Proof. Let Π be the space of policy distributions and let π_i be the policies at iteration i. By the policy improvement property in Proposition 3.6, the sequence Q^{π_i} is monotonically increasing. Also, for any state-action pair $(\mathbf{s}_t,\mathbf{a}_t) \in \mathcal{S}\times\mathcal{A}$, each Q^{π_i} is bounded due to the discount factor γ. Thus, the sequence of π_i converges to some π^* that are local optimum. We will still need to show that π^* are indeed optimal; we assume finite MDP, as typically assumed for convergence proof in usual policy iteration (Sutton, 1988). At convergence, we get $J_{\pi^*}(\pi^*)[\mathbf{s}] \geq J_{\pi^*}(\pi')[\mathbf{s}], \forall\pi' \in \Pi$. Using the same iterative augument as in the proof of Proposition 3.6, we get $Q^{\pi^*}(\mathbf{s},\mathbf{a}) \geq Q^{\pi'}(\mathbf{s},\mathbf{a})$ for all $(\mathbf{s},\mathbf{a}) \in \mathcal{S}\times\mathcal{A}$. Hence, π^* are optimal in Π. □

B. Extensive Related Works

Maximum-Entropy Reinforcement Learning. Maximizing entropy in Reinforcement Learning (RL) aims to optimize policies for both maximizing the expected return and the expected entropy of the policy. This approach has found application in various RL contexts, ranging from inverse RL (Ziebart et al., 2008) to multi-goal RL (Zhao et al., 2019). Guided policy search (Levine & Koltun, 2013) utilizes maximum entropy to guide policy learning towards high-reward regions. The incorporation of entropy regularization establishes a connection between value-based and policy-based RL (O'Donoghue et al., 2016; Schulman et al., 2017a).

In the domain of off-policy RL, soft Q-learning (Haarnoja et al., 2017) and its variants (Schulman et al., 2017a; Grau-Moya et al., 2018) learn a softened value function by replacing the hard maximum operator in the Q-learning update with a softmax operator. Soft actor-critic (SAC) (Haarnoja et al., 2017) introduces a maximum entropy actor-critic algorithm,

offering a balance between sample efficiency and stability. Moreover, Han & Sung (2021b) propose a max-min entropy framework to encourage visiting states with low entropy and maximize the entropy of these low-entropy states to enhance exploration (Han & Sung, 2021b). Another direction is maximum state entropy exploration. (Hazan et al., 2019a) present an efficient algorithm for optimizing intrinsically state entropy objectives, and Seo et al. (2021) use state entropy as an intrinsic reward to improve exploration. Our proposed approach, falling within the spectrum of SAC variants, introduces causality-aware weighted entropy to selectively enhance exploration for different primitive behaviors, showcasing superior efficiency and effectiveness.

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., 2023). 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. Further, an interesting work is Human-AI shared control via Policy Dissection (Li et al., 2022). While both ACE and Policy Dissection draw inspiration from neuroscience, particularly in the realm of motion primitives, they operate in different paradigms. Policy Dissection fosters human-AI shared control, enabling human collaboration with RL agents in intricate environments. It establishes an interpretable interface on the agent, allowing humans to directly influence agent behavior and facilitate shared control.

In contrast to explicit multi-stage approaches or human-guided task stage division, our proposed method **ACE** 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.

C. Environment Setup

We evaluate **ACE** across diverse continuous control tasks, spanning MuJoCo (Todorov et al., 2012b), ROBEL (Ahn et al., 2020a), DMControl (Tassa et al., 2018b), Meta-World (Yu et al., 2019b), Adroit (Rajeswaran et al., 2018), panda-gym (Gallouédec et al., 2021b), Shadow Dexterous Hand (Plappert et al., 2018). It excels in both locomotion and manipulation tasks, both sparse reward and dense reward settings. Visualizations of these tasks are provided in Figure 13, Figure 14, Figure 15, Figure 16, and Figure 17.

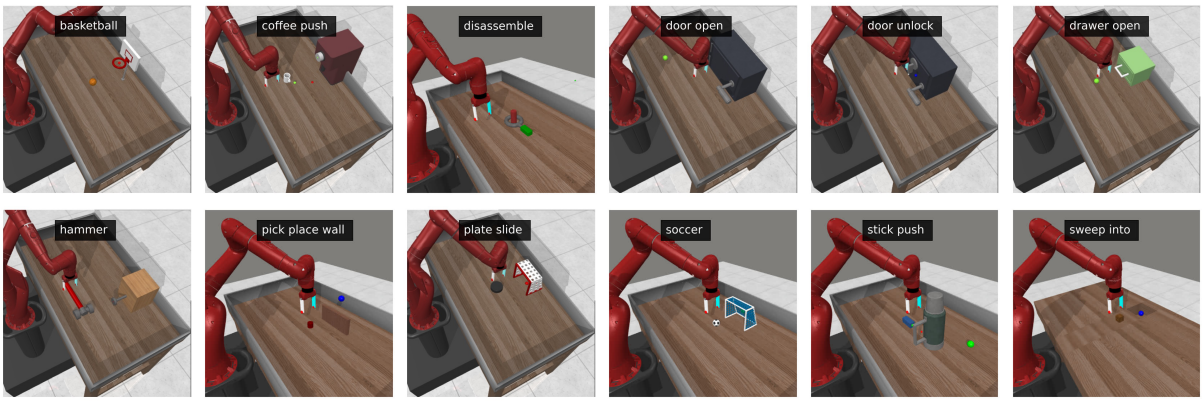


Figure 13: Meta-World benchmark tasks.

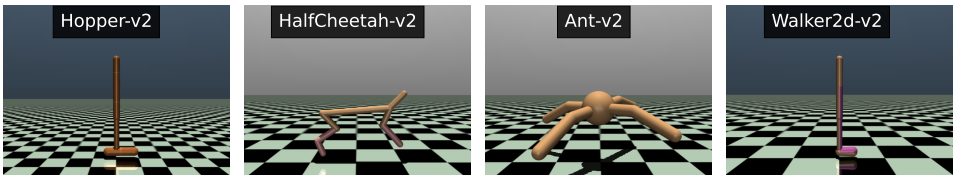


Figure 14: MuJoCo benchmark tasks.

Figure 15: DMControl benchmark tasks.

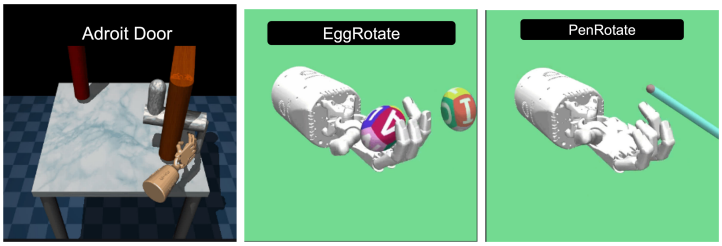
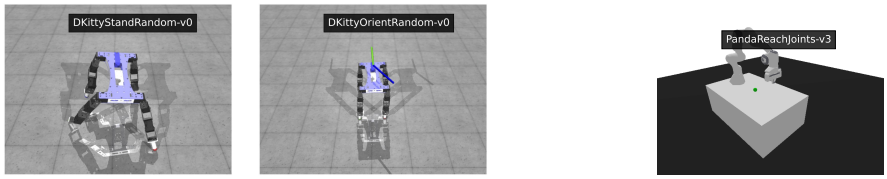


Figure 16: Adroit and Shadow Dexterous Hand benchmark tasks.

(d) MetaWorld benchmark tasks (sparse reward)



(d) ROBEL benchmark tasks (sparse reward) (e) panda-gym benchmark tasks (sparse reward)

Figure 17: ROBEL and pand-gym benchmark tasks (sparse reward).

D. More Examples of Causal Weights

Causal weight facilitates understanding of the agent's actions, reflecting the adaptive learning mechanisms akin to human cognition. Here, we provide more examples for illustration.

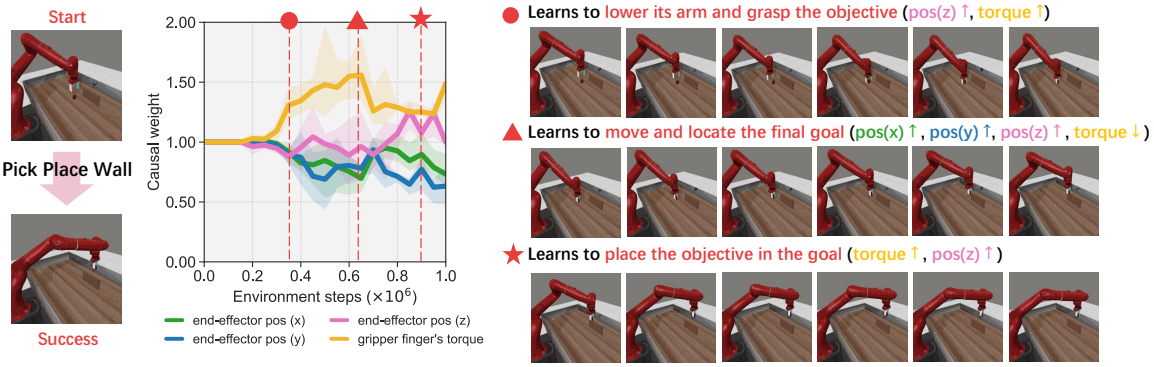


Figure 18: **Visualization of causal weights.** We visualize the varying causal weights on the pick-place-wall task during the training stages.

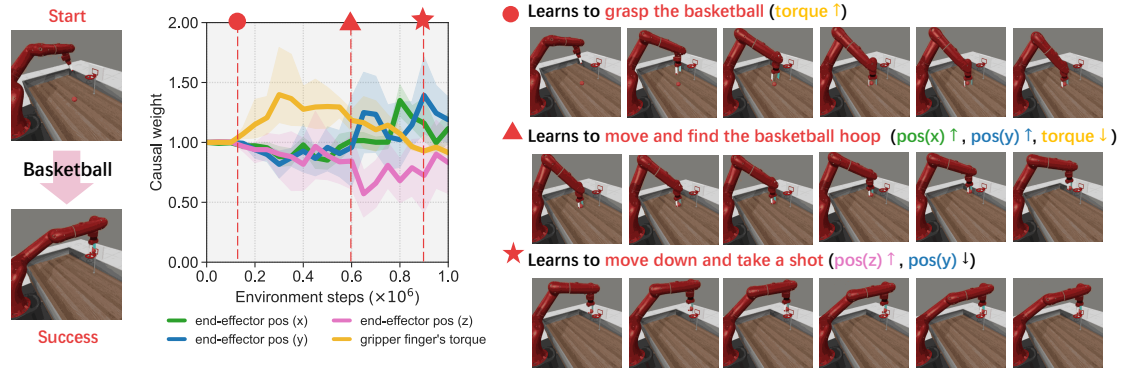


Figure 19: **Visualization of causal weights.** We visualize the varying causal weights on the basketball task during the training stages.

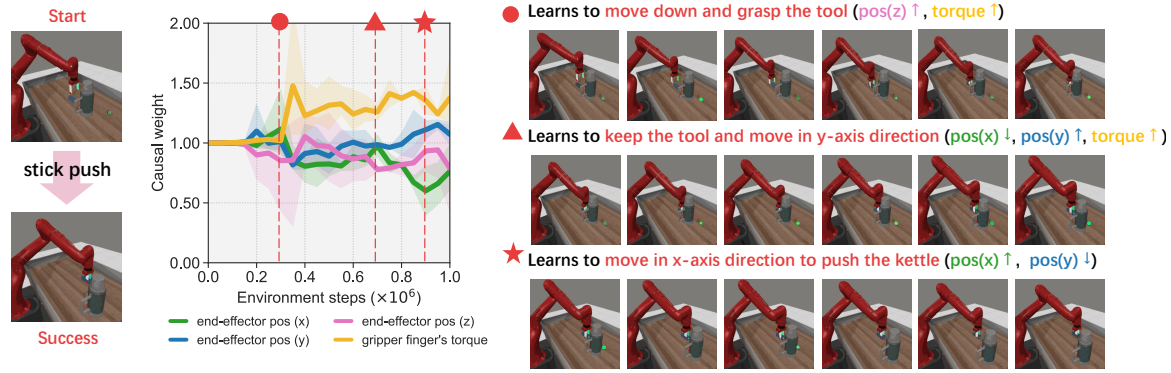


Figure 20: **Visualization of causal weights.** We visualize the varying causal weights on the stick-push task during the training stages.

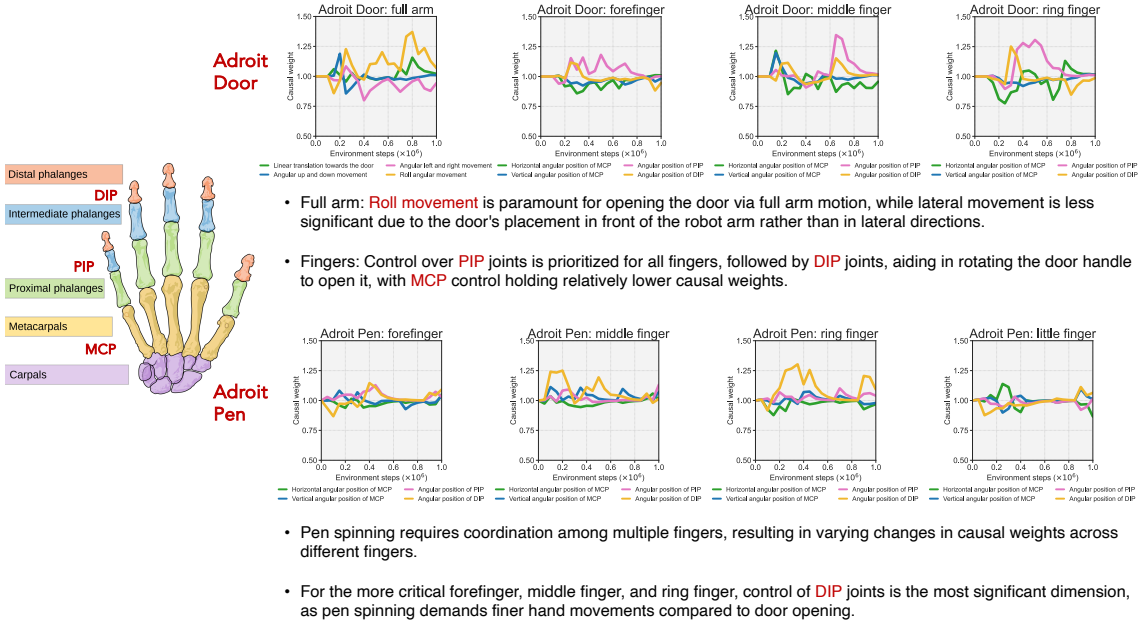


Figure 21: **Visualizations of causal weights in Adroit Door and Pen tasks (29 action dimensions).** These curves demonstrate how causal weights control individual finger joints alongside the entire arm, with detailed explanations provided. Notably, the crucial finger joints differ across tasks. This clearly highlights the effectiveness of our causal graph calculation in high-dimensional settings involving multi-object control.

E. Implementation Details

E.1. Algorithm instantiation

Instantiating **ACE** involves specifying three main components: 1) effectively recognizing the causal weights of $\mathbf{a} \rightarrow \mathbf{r}|\mathbf{s}$; 2) incorporating causal weights and the corresponding causality-aware entropy term into policy optimization. 3) periodically resetting the network based on our gradient dormancy degree.

- The pseudocode of our proposed **ACE** is provided in Algorithm 1.
- The variant named **CausalSAC** simplifies ACE by removing the reset mechanism.
- By incorporating the causality entropy term and the reset mechanism into the BAC backbone algorithm, we created the variant **ACE-BAC**.

Causal discovery on $\mathbf{a} \rightarrow \mathbf{r}|\mathbf{s}$. To effectively compute $\mathbf{B}_{\mathbf{a}\rightarrow\mathbf{r}|\mathbf{s}}$, 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 $\mathbf{B}_{\mathbf{a}\rightarrow\mathbf{r}|\mathbf{s}}$, we could obtain the causality-aware entropy $\mathcal{H}_c(\pi(\cdot|\mathbf{s}))$ through Eq.(2). 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 causality-aware 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,

$$\mathcal{T}_c^\pi Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi}[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \alpha\mathcal{H}_c(\pi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}))]\right]. \tag{7}$$

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, 2021a; Sun et al., 2022; Ji et al., 2023b) 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 CausalSAC and **ACE**.

Gradient-dormancy-guided reset mechanism. For each reset interval, we calculate the gradient dormancy degree, initialize a random network with weights ϕ_i and soft reset the policy network π_θ and the Q network Q_ϕ.

E.2. Hyperparameters

The hyperparameters used for training **ACE** are outlined in Table 2. We conduct all experiments with this single set of hyperparameters.

The hyperparameters used for training **ACE** are outlined in Table 2. We conduct all experiments with this single set of hyperparameters.

Algorithm 1 off-policy Actor-critic with Causality-aware Entropy (**ACE**)

initialize: Q network Q_ϕ, policy network π_θ, replay buffer $\mathcal{D}$; local buffer $\mathcal{D}_c$ with size N_c, causal weight matrix $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$, perturb factor f;
for each environment step t **do**
  Collect data with π_θ from real environment
  Add to replay buffer $\mathcal{D}$ and local buffer $\mathcal{D}_c$
end for

// Causal discovery
if every I environment step **then**
  Sample all N_c transitions from local buffer $\mathcal{D}_c$
  Update causal weight matrix $\mathbf{B}_{\mathbf{a}\to r|\mathbf{s}}$
end if

// Policy optimization
for each gradient step **do**
  Sample N transitions $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$ from $\mathcal{D}$
  Compute causality-aware entropy $\mathcal{H}_c(\pi(\cdot|\mathbf{s}))$
  Calculate the target Q value
  Update Q_ϕ by $\min_\phi \left(\mathcal{T}_c Q_\phi - Q_\phi\right)^2$
  Update π_θ by $\max_\theta Q_\phi(s, a)$
end for

// Reset mechanism
if every reset interval **then**
  Calculate the gradient dormant degree β_γ
  Initialize a random network with weights ϕ_i
  Soft reset π_θ by $\theta_t = (1-\eta)\theta_{t-1} + \eta\phi_i$
  Soft reset Q_ϕ by $\phi_t = (1-\eta)\phi_{t-1} + \eta\phi_i$
  Reset the state of the policy optimizer and Q optimizer
end if

Table 1: Hyperparameter settings for **ACE**.

Hyper-parameter	**Value**
Q-value network	MLP with hidden size 512
V-value network	MLP with hidden size 512
policy network	Gaussian MLP with the 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	10000
causality computation interval I	10000
max reset factor α_{max}	0.8
reset interval	200000
dormancy threshold τ	0.025

Table 2: Hyperparameter settings for **ACE**.

Hyper-parameter	**Value**
Q-value network	MLP with hidden size 512
V-value network	MLP with hidden size 512
policy network	Gaussian MLP with the 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	10000
causality computation interval I	10000
max reset factor α_{max}	0.8
reset interval	200000
dormancy threshold τ	0.025

F. More Benchmark Results

We conduct experiments on the more complex locomotion and manipulation tasks from **MuJoCo** (Todorov et al., 2012a), **DMControl** (Tassa et al., 2018a), **Meta-World** (Yu et al., 2019b), **Adroit** (Rajeswaran et al., 2018), **Shadow Dexterous Hand** (Plappert et al., 2018) for further evaluation of **ACE**and the baselines. Currently, several tasks in these benchmarks pose a formidable challenge that stumps most model-free methods. Notably, **ACE** has demonstrated its effectiveness by successfully solving many of these challenging tasks.

F.1. Evaluation on MetaWorld benchmark tasks

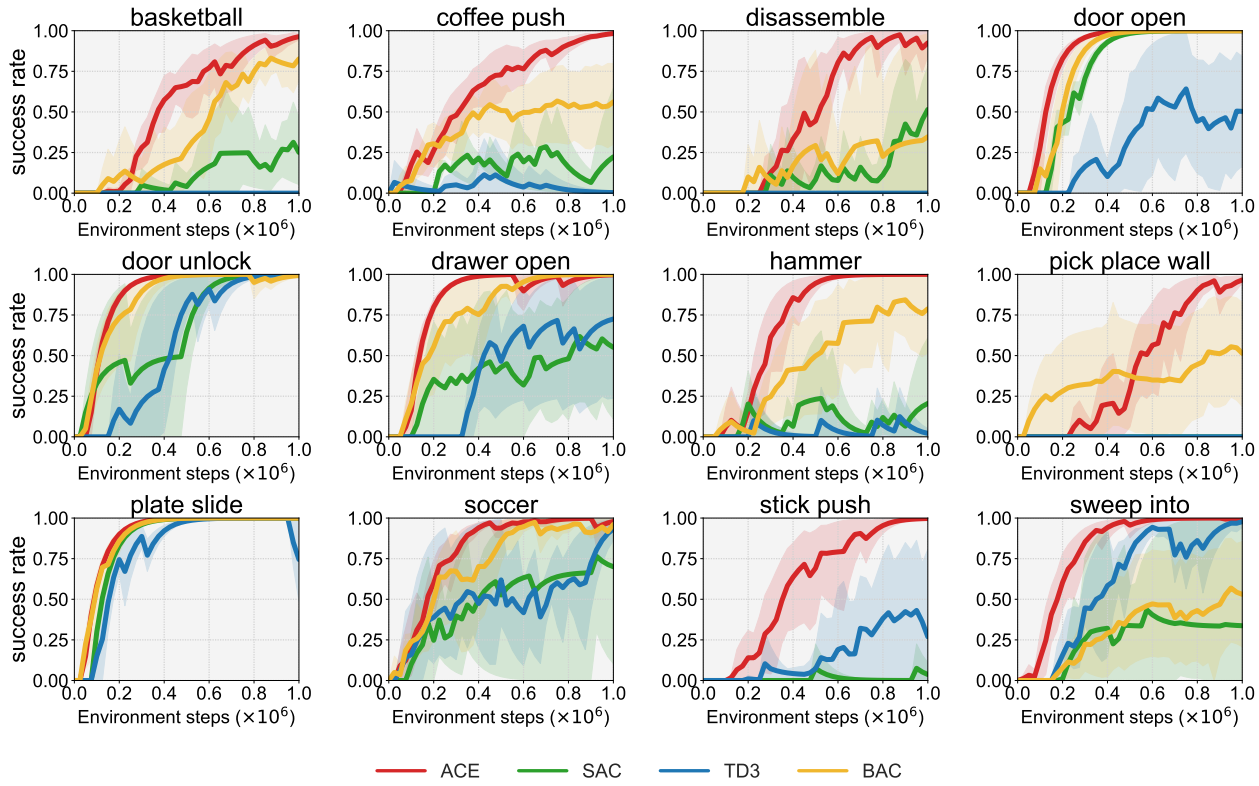


Figure 22: **MetaWorld tasks.** Success rate of **ACE**, BAC, SAC, TD3 in MetaWorld tasks. Solid curves depict the mean of 6 trials, and shaded regions correspond to the one standard deviation.

F.2. Evaluation on MuJoCo benchmark tasks

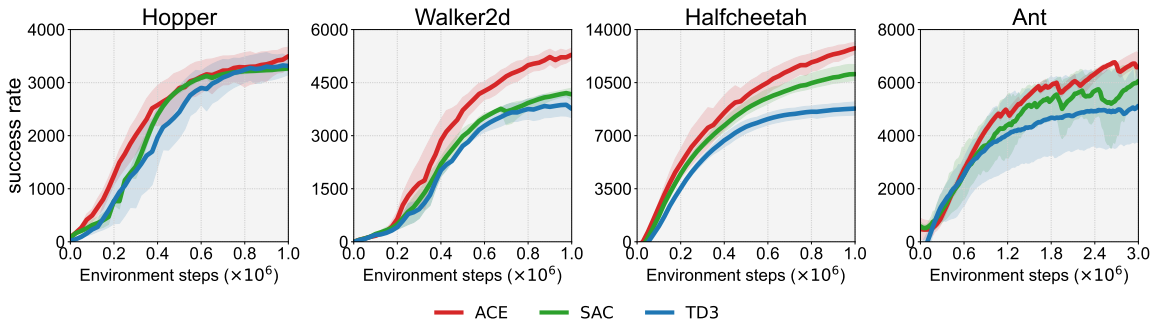


Figure 23: **MuJoCo benchmark tasks.** Training curves of **ACE**, SAC, TD3 in MuJoCo benchmark tasks. Solid curves depict the mean of 6 trials and shaded regions correspond to the one standard deviation.

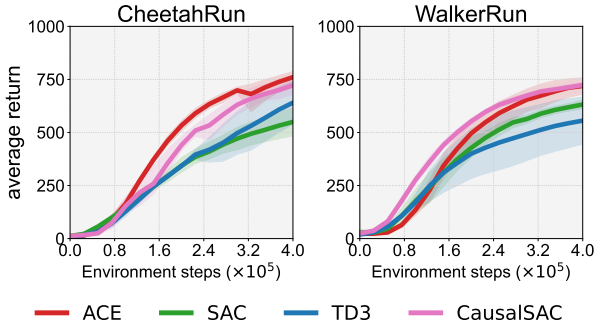


Figure 26: **Ablation of reset mechanism on locomotion tasks.** Average return of **ACE**, CausalSAC, SAC, TD3 in two locomotion tasks. CausalSAC is a reduced form of ACE by eliminating the reset mechanism of ACE.

F.3. Evaluation on DMControl benchmark tasks

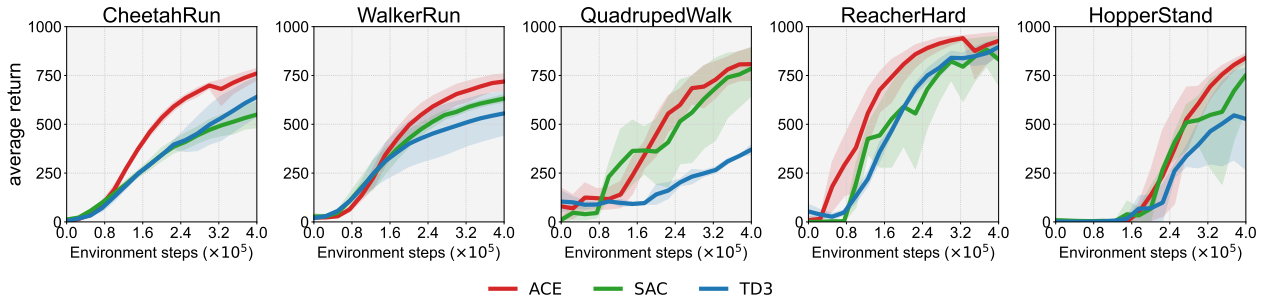


Figure 24: **DMControl benchmark tasks.** Training curves of **ACE** , SAC, TD3 in DMControl benchmark tasks. Solid curves depict the mean of 6 trials and shaded regions correspond to the one standard deviation.

F.4. Evaluation on Adroit and Shadow Dexterous Hand benchmark tasks

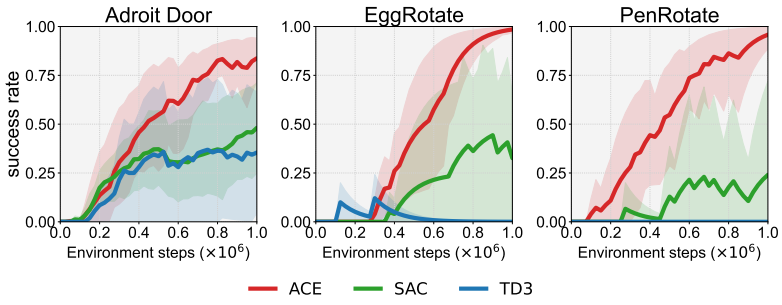


Figure 25: **Dexterous hand tasks.** Training curves of **ACE** , SAC, TD3 in dexterous hand tasks. Solid curves depict the mean of 6 trials, and shaded regions correspond to the one standard deviation.

G. Extensive Ablation Studies on Locomotion Tasks

We present the results of CausalSAC on a subset of locomotion tasks, demonstrating comparable or even superior performance compared to **ACE** . Due to the fact that the total number of steps in these tasks is only 40% of that in manipulation tasks, the policy converges faster. We attribute this to the key reason for the less pronounced impact of our reset mechanism in these tasks. Nevertheless, it is worth noting that the causal weighted entropy we propose still exhibits significant significance in improving sample efficiency in locomotion tasks.

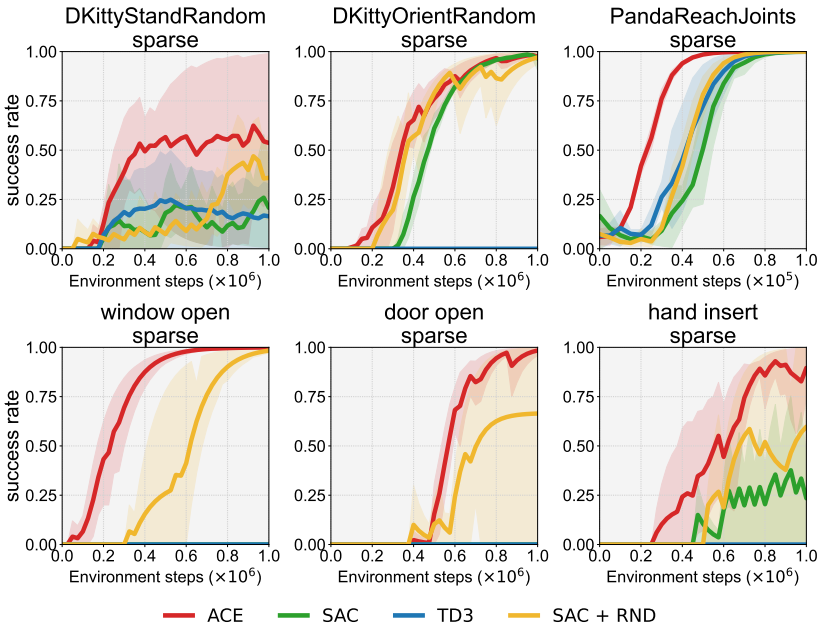


Figure 27: **Sparse reward tasks.** Training curves of **ACE**, SAC, TD3 and SAC+RND in sparse reward tasks. Solid curves depict the mean of 6 trials, and shaded regions correspond to the one standard deviation.

H. Effectiveness in Sparse Reward Settings

We conduct experiments in sparse reward tasks to showcase the efficiency of **ACE**. We evaluate both robot locomotion and manipulation tasks based on the sparse reward version of benchmark tasks from MetaWorld (Yu et al., 2019a), ROBEL (Ahn et al., 2020a) and panda-gym (Gallouédec et al., 2021a). MetaWorld manipulation tasks are based on a Shawyer robot arm with end-effector control. 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 27, our **ACE** surpasses the baselines by a large margin.

I. Generalizability of ACE

The proposed causality-aware entropy and gradient-dormancy-guided reset mechanism are versatile and effective plug-ins, thus they could be integrated into various RL backbone algorithms and techniques.

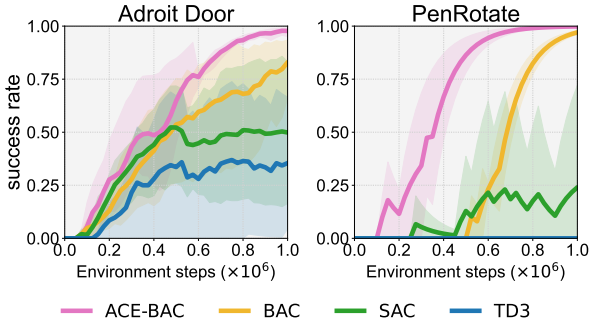


Figure 28: **Performance of ad-hoc ACE-BAC.** Training curves of ACE-BAC, BAC, SAC, and TD3 in challenging dexterous hand manipulation reward tasks. Solid curves depict the mean of 6 trials, and shaded regions correspond to the one standard deviation.

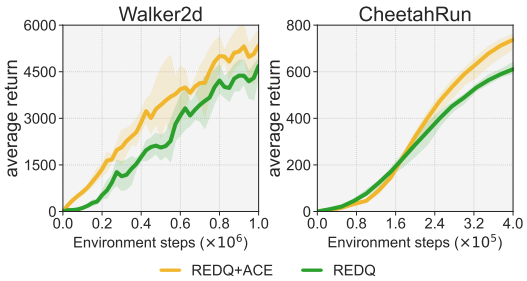


Figure 29: **Performance of ACE with higher-UTD methods.** We applied ACE to REDQ by integrating causality-aware entropy and a dormancy-guided reset mechanism. The results show that our method significantly improves REDQ's sample efficiency and performance in multiple tasks. Solid curves depict the mean of 6 trials, and shaded regions correspond to the one standard deviation.

	halfcheetah-medium-v2	hopper-medium-v2	walker2d-medium-v2	maze2d-medium-v1
CQL	47.04 ± 0.22	59.08 ± 3.77	80.75 ± 3.28	86.11 ± 9.68
CQL+ACE	50.23 ± 0.43	60.62 ± 3.97	85.27 ± 3.68	89.61 ± 18.72

Table 3: Performance comparison of CQL and CQL+ACE across various tasks.

Compatibility with Max-Entropy algorithms. We integrate **ACE** into a recent BAC algorithm, which is effective in challenging failure-prone tasks. Results on the challenging dexterous hand manipulation tasks show that the ad-hoc **ACE-BAC** algorithm could outperform the BAC algorithm, refer to Figure 28.

Compatibility with high-UTD techniques. Higher-UTD (Update-to-Data) methods are popular and effective in improving the sample efficiency and performance of reinforcement learning algorithms. We integrate **ACE** into the popular REDQ (Chen et al., 2020) algorithm, which is known for its efficiency in continuous control tasks. Results on the Walker2d and CheetahRun benchmark tasks, shown in Figure 29 demonstrate that **ACE** significantly enhances REDQ's sample efficiency and overall performance across a range of tasks.

Compatibility with offline RL setting. Excitingly, we observed that CQL with an ACE backbone outperforms CQL. This not only underscores the effectiveness and versatility of ACE but also highlights a promising direction for future exploration.

These observations demonstrate the effectiveness and generalizability of our proposed mechanisms and also shed light on the further applications of our mechanisms in enhancing existing RL algorithms.

J. ACE for Visual RL Application

We extend our method to facilitate visual RL learning with image inputs. The challenges we address in visual RL include: 1) determining how to compute causality between rewards and policy using image inputs, as our causal model cannot directly handle high-dimensional observations. 2) applying our methodology to existing visual RL baselines that do not utilize entropy regularization.

Our proposed solutions are as follows: 1) Utilizing the features extracted from the encoder output for causality computation. Since these features have only 50 dimensions, they can be effectively processed by our causal model. Our experiments demonstrate that causal weights calculated in this manner also provide effective guidance for exploration. 2) Using causal weights to guide action sampling, applying more significant noise to action dimensions with larger causal weights to encourage exploration, and vice versa. Simultaneously, we incorporate our proposed gradient dormancy degree into the exploration schedule. Leveraging these, we make the following improvements to the state-of-the-art visual RL algorithm DrM (Xu et al., 2023): 1. Replacing the dormant ratio in DrM, which guides perturbation and exploration schedule, with our proposed gradient dormancy degree. 2. Utilizing causal weights to guide action sampling in DrM.

Figure 30 depicts the dormancy curves using two distinct dormancy definitions, applied to state-based and visual-based

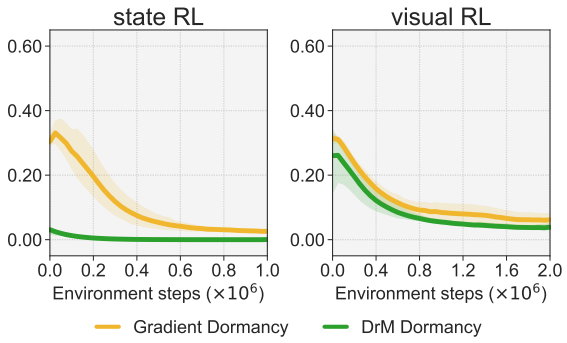


Figure 30: **Comparison between curves of different dormancy degrees in state-based and visual-based RL.** In state-based RL, the dormant ratio proposed by (Xu et al., 2023) consistently approaches zero, while our gradient dormancy degree undergoes notable variations.

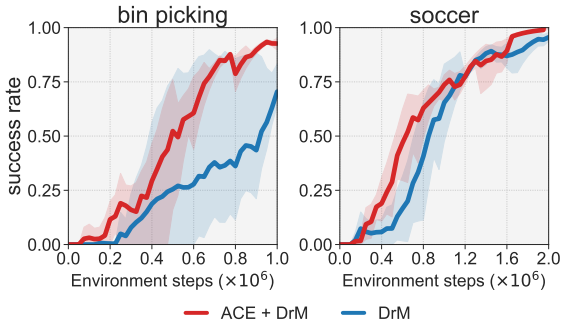


Figure 31: **Performance comparison of ACE + DrM and DrM.**

soccer task in MetaWorld, respectively, during training with **ACE** and DrM algorithms. It is apparent that in state-based RL, our proposed gradient dormancy degree effectively reflects the network's capacity. In contrast, DrM's dormant ratio, as defined by Sokar et al. (2023), consistently remains close to zero in state-input tasks and fails to provide effective guidance for state-based RL exploration. In visual RL, the two dormancy curves are closely aligned, effectively guiding exploration and improving the sample efficiency of visual RL. This underscores the adaptability of our proposed gradient dormancy, suitable for any RL scenario.

As shown in Figure 31, our refinements to DrM, leveraging causal weights and the gradient dormancy degree, effectively enhance its performance and learning efficiency. Moving forward, we will continue exploring the potential applications of causality-aware exploration in image-based RL.

K. The effectiveness of ACE in solving hard exploration problems.

The effectiveness of ACE is particularly evident in scenarios where the causal relationship between action and reward, conditioned on the state, can be misleading and hurt performance.

Addressing Dummy Actions with the Reset Mechanism. For scenarios involving dummy actions, the resetting mechanism helps to solve it. To be specific, such dummy actions would trap the exploration in local optima and hence the gradient dormant would be high (i.e., low policy gradients, high gradient dormancy degree). Then our reset mechanism intervenes by perturbing the network, encouraging unbiased exploration.

A concrete example is the pick-and-place-wall task, where the wall initially biases the agents' exploration towards vertical and grasping actions. In contrast to the high dormancy degree and poor performance observed with all other baselines, ACE overcomes this challenge with its dormancy-guided reset mechanism, as evidenced by the dormancy degree curves in Figure 3 (y-axis=dormancy degree) and the performance data in Figure 4.

Handling Misleading or Imperfect Rewards. For scenarios with misleading or imperfect rewards, we could simply turn it into a sparse reward setting. Our experiments, as shown in Figures 7 and 27, reveal ACE's effectiveness in sparse reward environments. This advantage primarily arises from ACE's causal-aware entropy, which discerns the implicit action-reward relationship based on the state.

Furthermore, ACE's versatility also allows it to be combined with alternative intrinsic reward approaches, enhancing its applicability across various scenarios. This adaptability ensures that specific challenges, such as those involving misleading rewards.

L. 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 32 presents the computational time comparison between our algorithm **ACE** and SAC on 12 MetaWorld benchmark tasks. Compared to SAC, the total training time of **ACE** only increased by an average of 0.69 hours, hence, the additional costs are acceptable. Further, for practical use, **ACE** requires fewer interactions for similar performance, which may lower the needed computation time in practice.

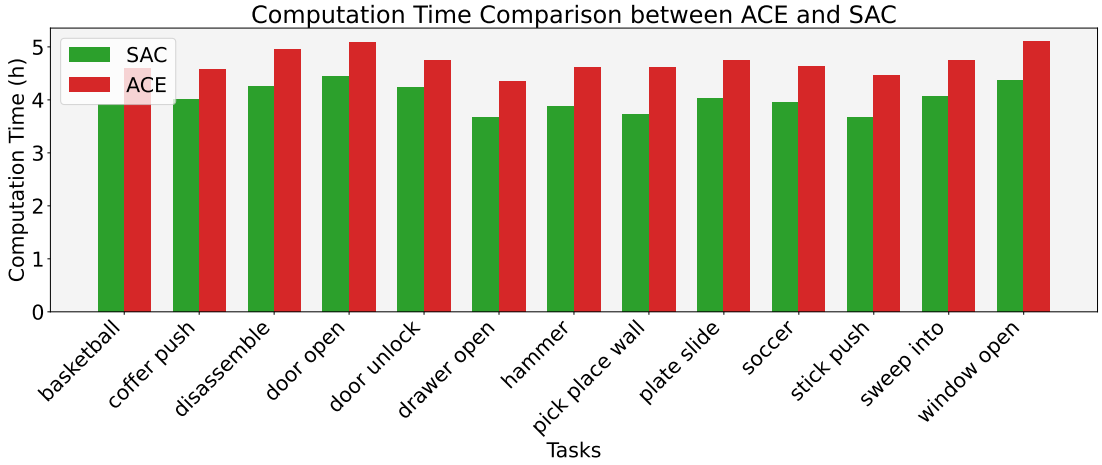


Figure 32: **Computation time comparison.** Computation time comparison between **ACE** and SAC in ten MetaWorld tasks, each averaged on 6 trials.