Fast Counterfactual Inference for History-Based Reinforcement Learning

Haichuan Gao[1], **Tianren Zhang**[1], **Zhile Yang**[2],
Yuqing Guo[1], **Jinsheng Ren**[1], **Shangqi Guo**[1,3*], **Feng Chen**[1,4*]
[1]Department of Automation, Tsinghua University, Beijing, China
[2]School of Computing, University of Leeds, Leeds, UK
[3]Department of Precision Instrument, Tsinghua University, Beijing, China
[4]LSBDPA Beijing Key Laboratory, Beijing, China
{ghc18, zhang-tr19}@mails.tsinghua.edu.cn, sczy@leeds.ac.uk, {gyq18, rjs17}@mails.tsinghua.edu.cn,
shangqi_guo@foxmail.com, chenfeng@mail.tsinghua.edu.cn

Abstract

Incorporating sequence-to-sequence models into history-based Reinforcement Learning (RL) provides a general way to extend RL to partially-observable tasks. This method compresses history spaces according to the correlations between historical observations and the rewards. However, they do not adjust for the confounding correlations caused by data sampling and assign high beliefs to uninformative historical observations, leading to limited compression of history spaces. Counterfactual Inference (CI), which estimates causal effects by single-variable intervention, is a promising way to adjust for confounding. However, it is computationally infeasible to directly apply the single-variable intervention to a huge number of historical observations. This paper proposes to perform CI on observation sub-spaces instead of single observations and develop a coarse-to-fine CI algorithm, called Tree-based History Counterfactual Inference (T-HCI), to reduce the number of interventions exponentially. We show that T-HCI is computationally feasible in practice and brings significant sample efficiency gains in various challenging partially-observable tasks, including Maze, BabyAI, and robot manipulation tasks.

Introduction

Integrating historical observations into states provides a general way to scale up Reinforcement Learning (RL) to partially-observable tasks (Majeed and Hutter 2018). Nevertheless, a core problem in this setting is that the large scale of history spaces (Joelle, Geoffrey, and Thrun 2003; Smith and Simmons 2005) leads to sample-inefficient policy learning. Prior works encode historical observations into hidden states by sequence-to-sequence (seq2seq) models (Graves et al. 2016; Hausknecht and Stone 2015; Parisotto et al. 2020). Seq2seq models assign high beliefs on the historical observations of high correlations with learning signals and attenuate those with low correlation (Zhu et al. 2018b; Ramos et al. 2021; Zhang et al. 2022b). However, they do not adjust for confounding correlations, i.e., the high correlation of uninformative historical observations induced by the biases in data sampling, which limits their compression capability.

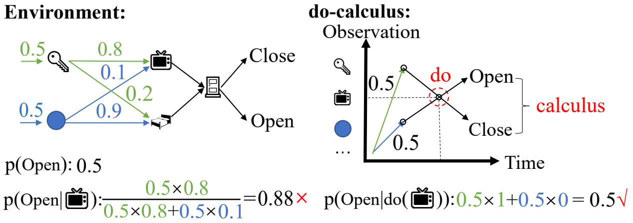

Figure 1: Key-to-door example. The high correlation on TV caused by sampling can be eliminated by do-calculus which separates confounders (key and ball).

It is promising to mine the historical observations with causality using Counterfactual Inference (CI) (Pearl 2013) to eliminate confounding correlations caused by data sampling. Consider the key-to-door example shown in Figure 1, whether the agent can open the door only depends on whether the key is collected. However, if the current sampling policy is that when the agent starts from the key, the agent tends to pass through the TV, then the TV will have a high correlation with opening the door, i.e., $p(\text{Open} \mid \text{TV}) = 0.88$. This is because the effect estimation is influenced by the confounding correlation between Key and TV. This confounding can be eliminated by do-calculus (Pearl 2013) which separates the variables (key and ball) that may cause confounding (i.e., the backdoor variables following backdoor adjustment formula (Pearl 2009)) and estimates the causal effect of TV by integrating the probabilities of Open, resulting in $p(\text{Open} \mid \text{do}(\text{TV})) = 0.5$. Since historical observations with causality are often relatively sparse, mining them by CI can greatly compress history spaces.

Prior researchers perform CI on Markovian RL tasks for feature selection (Zhang et al. 2021; Tomar et al. 2021) or credit assignment (Mesnard et al. 2021), where the states are Markovian. However, for history-based RL, we should intervene for each non-Markovian observation-and-time combination o_t and estimate its causal effect. CI estimates causal effects by single variable intervention, leading to high computational complexity due to the large scale of historical observations.

This paper proposes to perform CI on observation sub-

[*]Corresponding authors: Shangqi Guo and Feng Chen

spaces by simultaneously intervening in all historical observations belonging to an observation sub-space instead of on a single time-specific historical observation. However, the challenge of this CI regime lies in that multiple simultaneously intervened historical observations have no common backdoor variable. To overcome this challenge, we propose a novel *step-backdoor adjustment* tailored to history-based RL to estimate the causal effect. We develop a coarse-to-fine CI algorithm, called Tree-based History Counterfactual Inference (T-HCI), based on the CI on observation sub-spaces. We prove that T-HCI exponentially reduces the number of interventions and is computationally feasible to be combined with RL.

We empirically verify the effectiveness of T-HCI on various popular history-based tasks and show that 1) T-HCI substantially improves computational efficiency compared with vanilla CI, making it computationally feasible to perform CI in history-based RL; 2) T-HCI eliminates confounding correlations and mines semantically causal observations, bringing significant sample efficiency gains compared with representative seq2seq approaches.

Preliminaries

HDP. We consider a History-based Decision Process (HDP) (Majeed and Hutter 2021) tuple $\langle \mathcal{O}, \mathcal{H}, \mathcal{A}, T, P, R, \gamma \rangle$, where $\mathcal{O}$ is the observation space, $\mathcal{H}$ is the history space, $\mathcal{A}$ is the action space, T is the horizon, $P(o_{t+1}|h_t, a_t)$ is the history-action transition probability, $R: \mathcal{H} \to [0,1]$ is the reward function, and γ is the discount factor. Following commonly used benchmarks (Parisotto et al. 2020; Loynd et al. 2020; Chen et al. 2021b), this paper assumes that the causal effects of historical actions are reflected in historical observations and denotes $h_t \in \mathcal{H}$ by a sequence of observations $\{o_1, \cdots, o_t\}$. We denote a policy by $\pi: \mathcal{H} \times \mathcal{A} \to [0,1]$. Let $Q_\pi: \mathcal{H} \times \mathcal{A} \to \mathbb{R}$ denote the Q function and $V_\pi: \mathcal{H} \to \mathbb{R}$ denote the value function, i.e., $V_\pi(h) := \mathbb{E}_\pi \left[\sum_{i=t}^{T} \gamma^{i-t} R(h_i) \big| h_t = h \right]$. We use $\mathcal{H}_t$ to denote the history subspace containing all histories h_t at time step t. We use $|\cdot|$ to denote the cardinality of a set and use $\Omega(\cdot)$ as the complexity notation: for two non-negative sequences $\{a_n\}, \{b_n\}$, $a_n = \Omega(b_n)$ means that there exists a positive constant C such that $a_n \leq C b_n$.

Do-calculus and Backdoor Formula. In causal inference, the random variables constituting the nodes of a causal diagram $\mathcal{G}$ can be divided into a set of covariates $\boldsymbol{X} := \{X_1, \cdots, X_n\}$ and a response variable Y of interest. To estimate the causal effect, the intervention operation $\mathrm{do}(X_j = x)$ is adopted by imposing a certain value x on one of the covariates X_j. Then, the causal effect of $\mathrm{do}(X_j = x)$ can be estimated with the well-known backdoor formula, which separates the backdoor variables to adjust for confounders (Pearl 2013):

$$p\big(Y|\mathrm{do}(X_j = x)\big) = \int p\big(Y|X_j = x, \boldsymbol{X}^{\mathrm{ba}}\big) \mathrm{d}p(\boldsymbol{X}^{\mathrm{ba}}), \tag{1}$$

where $\boldsymbol{X}^{\mathrm{ba}} \subseteq \boldsymbol{X}$ are the backdoor variables relative to (X_j, Y), if 1) $\boldsymbol{X}^{\mathrm{ba}}$ contains no descendant of X_j, and 2) $\boldsymbol{X}^{\mathrm{ba}}$ blocks each path between X_j and Y.

Fine-Grained History Counterfactual Inference

Prior research performs CI in the setting of Markovian RL (Seitzer, Schölkopf, and Martius 2021; Mesnard et al. 2021). This section extends CI to history-based RL by investigating dynamics changes. We treat the past observations, the current observation, and the action $\{h_{t-1}, o_t, a_t\}$ as covariates, where h_{t-1} denotes the historical observations from o_1 to o_{t-1}, and the immediate reward and next observation $\{r_{t+1}, o_{t+1}\}$ as response variables. We use timestamped symbols (e.g., o_t) for variables and non-timestamped symbols (e.g., o) for values. Then, for any historical observation $o_j (j < t)$, the history h_{j-1} meets the backdoor formula:

Proposition 1 (Fine-grained CI)**.** *Given o_t and a_t, h_{j-1} satisfies the backdoor formula relative to $(o_j, \{o_{t+1}, r_{t+1}\})$ for any historical observation $o_j (j < t)$, and the causal effect of $\mathrm{do}(o_j = o)$ can be estimated with*

$$\begin{aligned} & p\big(o_{t+1}, r_{t+1}|\mathrm{do}(o_j = o), o_t, a_t\big) \\ &= \int_{h_{j-1} \in \mathcal{H}_{j-1}} p\big(o_{t+1}, r_{t+1}|o_j = o, h_{j-1}, o_t, a_t\big) \mathrm{d}p(h_{j-1}). \end{aligned} \tag{2}$$

All the proofs of the propositions and theorems are provided in Appendix C. We approximate $p(h_{j-1})$ with history distribution in the replay buffer.

A common approach of causal inference developed based on the do-calculus is CI (Zhang et al. 2017), which performs counterfactual intervention by assigning treatment to a certain value to assess its effect. Following the method of (Chen et al. 2021a), we perform CI by assigning zero vectors $\mathbf{0}$ to some values of a historical observation o_j of interest. Then, the resulting counterfactual probability is

$$\begin{aligned} & p\big(o_{t+1}, r_{t+1}|\mathrm{do}(o_j = \mathbf{0}), o_t, a_t\big) \\ &= \int_{h_{j-1} \in \mathcal{H}_{j-1}} p(o_{t+1}, r_{t+1}|o_j = \mathbf{0}, h_{j-1}, o_t, a_t) \mathrm{d}p(h_{j-1}). \end{aligned} \tag{3}$$

The effect of the counterfactual intervention is estimated by comparing the counterfactual distribution $p(\cdot|\mathrm{do}(o_j = \mathbf{0}), o_t, a_t)$ with the factual distribution $p(\cdot|o_j, o_t, a_t)$. For example, in the task shown in Figure 1, the factual distribution is that $p(\mathrm{open}|\mathrm{key}, \mathrm{door}) = 1$ and $p(\mathrm{open}|\mathrm{ball}, \mathrm{door}) = 0$. After setting the observation values key and ball as zero vectors, we can get a counterfactual probability $p(\mathrm{open}|\mathbf{0}, \mathrm{door}) = 0.5 \times 1 + 0.5 \times 0 = 0.5$. The difference between the counterfactual and factual distributions is referred to as the Average Treatment Effect (ATE) (Liu, Ma, and Wang 2018; Pearl 2009). For simplicity, we begin by considering that any two values o and o' of o_j are set as zero vectors and use $\overline{o_j^{o,o'}}$ to denote that the observation values o and o' of an observation variable o_j are set as zero vectors. Then, the ATE is

$$\begin{aligned} \mathrm{ATE}(\overline{o_j^{o,o'}}) := \sum_{t=j+1}^{T} \mathbb{E}_{\substack{o_j \in \{o, o'\}, \\ o_t \in \mathcal{O}, a_t \in \mathcal{A}}} \\ \big|\big|p(\cdot|o_j, o_t, a_t) - p\big(\cdot|\mathrm{do}(o_j = \mathbf{0}), o_t, a_t\big)\big|\big|_1. \end{aligned} \tag{4}$$

Causal observations can be mined by repeating the above CI, e.g., with the wiped-out o and o', continually setting a new value o'' from o_j as zero vectors (i.e., $\overline{o, o', o''}$) and estimating its corresponding $\mathrm{ATE}(o_j^{\overline{o,o',o''}})$ until all observation values in $\mathcal{O}$ are traversed. Values that cause ATE to be zero have no causal effect and should be eliminated from memory.

However, performing CI on historical observations leads to high computational complexity in complex tasks with long time horizons and numerous observations. The time span is from 0 to T, and at each time step the observation scale is $\Omega(|\mathcal{O}|)$. The number of interventions has a complexity of $\Omega(T \cdot |\mathcal{O}|)$, making CI computationally infeasible.

Coarse-to-Fine History Counterfactual Inference

In this section, we introduce a coarse-to-fine CI method to reduce the computational complexity of fine-grained CI. Our method is based on two mild assumptions: the observations with causality are sparse in the observation space and causal effects are independent of time steps. Note that the two assumptions hold in many popular history-based RL benchmarks (Oh et al. 2016; Chevalier-Boisvert et al. 2019; Botea, Müller, and Schaeffer 2002). However, the causal historical observations are commonly not sparse in the time dimension, e.g., a key state may be observed at many time steps, which hinders us from performing coarse-grained CI. Thus, we aim to eliminate this redundancy in the time dimension for causal effect estimation and, further, develop a fast CI method through the following two steps. First, we estimate the causal effects of observations without timestamps, which reduces the number of interventions from $\Omega(T \cdot |\mathcal{O}|)$ to $\Omega(|\mathcal{O}|)$. Second, we perform CI on observation sub-spaces, which reduces the number of interventions from $\Omega(|\mathcal{O}|)$ to $\Omega(\log|\mathcal{O}|)$.

Counterfactual Inference on Observations

This section proposes a method to estimate the causal effects of observations without timestamps. We use $\mathrm{Do}(o)$ to represent the interventions on all the historical observations with the value o. Let $o_{I_m} := \{o_{t_1}, o_{t_2}, \cdots, o_{t_m}\}$ denote m historical observations with the same value, where $I_m := t_1, t_2 \cdots t_m$ denotes m timestamps belonging to $[1, t-1]$. Since m observations at different timestamps I_m may have different causal effects, we first need to estimate the average causal effect of the interventions at different sets of m timestamps. Given the number of timestamps m, we use Ξ_m to denote the complete collection of m time stamps. Meanwhile, m takes values in the range $[1, t-1]$. Then, the causal effect of $\mathrm{Do}(o)$ is the weighted average of each causal effect of simultaneous intervention $\mathrm{do}(o_{I_m} = o)$:

$$\begin{aligned} &p\big(o_{t+1}, r_{t+1}|\mathrm{Do}(o), o_t, a_t\big) \\ &= \sum_{m=1}^{t-1} \sum_{I_m \in \Xi_m} w(o_{I_m}) p\big(o_{t+1}, r_{t+1}|\mathrm{do}(o_{I_m} = o), o_t, a_t\big), \end{aligned} \tag{5}$$

where each weight $w(o_{I_m})$ is a weighting factor defined on the histories that contain historical observations with value o, i.e., $w(o_{I_m}) = p(I_m)/\sum_{m'=1}^{t-1} \sum_{I_{m'} \in \Xi_{m'}} p(I_{m'})$ where $p(I_m)$ denotes the probability of historical observations on timestamps I_m taking the value o. If each component of the overall causal effect $p(o_{t+1}, r_{t+1}|\mathrm{do}(o_{I_m} = o), o_t, a_t)$ is estimated, we can obtain the overall causal effect.

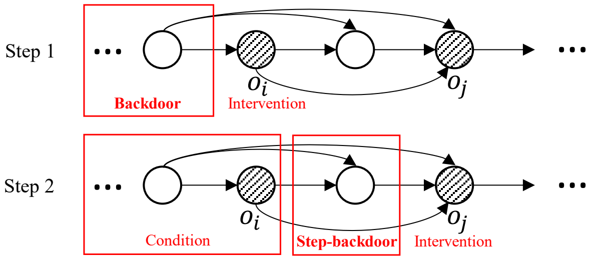

Figure 2: Step-backdoor in the HDP causal diagram of intervened o_i and o_j.

However, the estimation of these components is difficult because multiple intervened historical observations have no common backdoor variables: as shown in Figure 2, for any two intervened historical observations o_i and o_j $(i < j)$, a part of the backdoor variables of o_j are the children of o_i, which conflicts with the backdoor formula relative to o_i. While many general adjustment formulas for single-variable intervention exist (Pearl 2013; Chernozhukov, Fernández-Val, and Melly 2013; Schochet 2010), for multi-variable intervention, there are currently only task-specific adjustment formulas requiring additional assumptions, such as the independence between feature dimensions (Lu 2016b; Witte and Didelez 2019; Lu 2016a), which do not apply to our setting. For multiple variable interventions in the history-based RL domain, this paper proposes a step-backdoor adjustment formula to estimate the causal effect. For two variables X_i and X_j $(i < j)$ and a response variable Y, we consider backdoor variables $\boldsymbol{X}_i^{\mathrm{ba}}$ and $\boldsymbol{X}_j^{\mathrm{ba}}$ relative to (X_i, Y) and (X_j, Y), respectively. We first define the backdoor adjustment formula for two-variable intervention as follows:

Definition 1 (Step-backdoor adjustment formula). *$\boldsymbol{X}_j^{\mathrm{s-ba}} = \boldsymbol{X}_j^{\mathrm{ba}} \setminus (\boldsymbol{X}_i^{\mathrm{ba}} \cup \{X_i\})$ is step-backdoor relative to (X_j, Y) if 1) $\boldsymbol{X}_j^{\mathrm{s-ba}}$ has no descendant of X_j, 2) $\boldsymbol{X}_i^{\mathrm{ba}}$, X_i, and $\boldsymbol{X}_j^{\mathrm{s-ba}}$ block each path between X_j and Y, and 3) conditioned on X_i and $\boldsymbol{X}_i^{\mathrm{ba}}$, the distribution of $\boldsymbol{X}_j^{\mathrm{s-ba}}$ is identifiable.*

With the step-backdoor adjustment formula, we further estimate the causal effects of more than two intervened variables: considering a new variable X_k intervened together with X_i and X_j $(i < j < k)$, then the overall causal effect can be estimated by additionally estimating the causal effect of $\mathrm{do}(X_k = x'')$ conditioned on $(X_i, \boldsymbol{X}_i^{\mathrm{ba}}, X_j, \boldsymbol{X}_j^{\mathrm{s-ba}})$ with new step-backdoor $\boldsymbol{X}_k^{\mathrm{s-ba}}$, as the following theorem:

Theorem 1. *Given a set of intervened variables with different timestamps, if every two temporally adjacent variables meet the step-backdoor adjustment formula, then the overall* *causal effect can be estimated with*

$$\begin{aligned} & p\big(Y|\mathrm{do}(X_i = x, X_j = x', X_k = x'', \cdots)\big) \\ & = \int \cdots \int p\big(Y|\boldsymbol{X}_i^{\mathrm{ba}}, \boldsymbol{X}_j^{\mathrm{s-ba}}, \boldsymbol{X}_k^{\mathrm{s-ba}}, \cdots, \\ & \qquad X_i = x, X_j = x', X_k = x'', \cdots\big) \qquad (6) \\ & \qquad \mathrm{d}p(\boldsymbol{X}_i^{\mathrm{ba}}) \\ & \qquad \mathrm{d}p(\boldsymbol{X}_j^{s-\mathrm{ba}}|X_i = x, \boldsymbol{X}_i^{\mathrm{ba}}) \\ & \qquad \mathrm{d}p(\boldsymbol{X}_k^{s-\mathrm{ba}}|X_i = x, \boldsymbol{X}_i^{\mathrm{ba}}, X_j = x', \boldsymbol{X}_j^{s-\mathrm{ba}}) \cdots \end{aligned}$$

Here we analyze why the causal effect of $\mathrm{do}(o_{I_m} = o)$ can be estimated with the step-backdoor adjustment formula. We notice the forward nature of the causal effects in the HDP causal diagram: an earlier historical observation has a causal effect on the latter one, but the latter has no causal effect on the earlier so conditions 1) and 2) in Definition 1 are satisfied. Meanwhile, the distributions of later historical observations conditioned on the earlier are identifiable. As shown in Figure 2, for two simultaneously intervened historical observations o_i and $o_j (i < j)$, the distributions of the step-backdoor variables $(h_{j-1} \setminus h_i)$ conditioned on $\mathrm{do}(o_i = o)$ and h_{i-1} are identifiable. For intervention $\mathrm{do}(o_{I_m} = o)$, we can use the historical observations between every two adjacent intervention time steps as the step backdoor. Based on this, the causal effect of $\mathrm{Do}(o)$ can be calculated with the following theorem.

Theorem 2 (CI on observations)**.** *Given o_t and a_t, the causal effect of $\mathrm{Do}(o)$ can be estimated by*

$$\begin{aligned} & p\big(o_{t+1}, r_{t+1}|\mathrm{Do}(o), o_t, a_t\big) \qquad (7) \\ & = \int_{h_{t-1} \in \mathcal{H}_{t-1}^{o}} p\big(o_{t+1}, r_{t+1}|h_{t-1}, o_t, a_t\big) \mathrm{d}p(h_{t-1}|\mathcal{H}_{t-1}^{o}), \end{aligned}$$

where $\mathcal{H}_{t-1}^{o}$ denotes the history sub-space where each history $h_{t-1} \in \mathcal{H}_{t-1}^{o}$ contains at least one observation with value o.

Theorem 2 enables estimating the causal effect of observations in HDPs, which can change the number of interventions from $\Omega(T \cdot |\mathcal{O}|)$ to $\Omega(|\mathcal{O}|)$. However, this is still unsatisfying because $|\mathcal{O}|$ is often large in complex tasks. In the next section, we will introduce the CI on observation sub-spaces and develop a coarse-to-fine CI method to exponentially reduce the number of interventions.

Counterfactual Inference on Sub-Spaces

Based on the method proposed in the previous section, we further exploit the sparsity of causal observations to perform CI at different scales of observation sub-spaces. We begin by extending $\mathrm{Do}(o)$ to $\mathrm{Do}(\mathcal{O}^i)$, which is the simultaneous intervention of historical observations belonging to an observation sub-space $\mathcal{O}^i \subseteq \mathcal{O}$. We show that the causal effect can be estimated in a similar way to Theorem 2, as follows:

Proposition 2 (CI on observation sub-spaces)**.** *Given o_t and a_t, the causal effect of $\mathrm{Do}(\mathcal{O}^i)$ can be estimated by*

$$\begin{aligned} & p\big(o_{t+1}, r_{t+1}|\mathrm{Do}(\mathcal{O}^i), o_t, a_t\big) \qquad (8) \\ & = \int_{h_{t-1} \in \mathcal{H}_{t-1}^{\mathcal{O}^i}} p\big(o_{t+1}, r_{t+1}|h_{t-1}, o_t, a_t\big) \mathrm{d}p(h_{t-1}|\mathcal{H}_{t-1}^{\mathcal{O}^i}), \end{aligned}$$

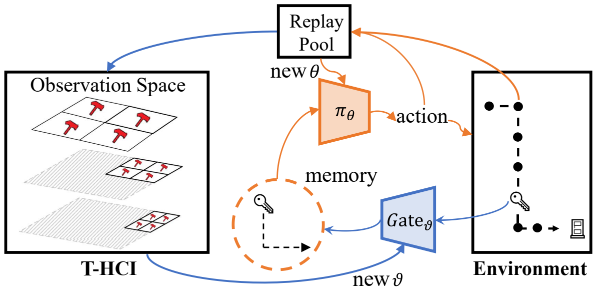

Figure 3: T-HCI Algorithm framework. The blue and orange lines respectively mark the CI loop and RL loop.

where $\mathcal{H}_{t-1}^{\mathcal{O}^i}$ represents the history subspace where each history $h_{t-1} \in \mathcal{H}_{t-1}^{\mathcal{O}^i}$ contains at least one observation belonging to the observation subspace $\mathcal{O}^i$.

Coarse-to-fine CI is developed by continually performing CI on observation sub-spaces. Let $\mathcal{O}^{\mathrm{Ca}}$ denote the observation sub-space with historical causality. Suppose that an observation space is divided into $Z \geq |\mathcal{O}^{\mathrm{Ca}}|$ sub-spaces and there are at least $Z - |\mathcal{O}^{\mathrm{Ca}}|$ parts containing no causal observation. We eliminate them and continue the process: dividing the rest observation space and carrying on CI. The process terminates until the observation space contains less than Z observations. This regime exponentially reduces the number of interventions, as shown by the following proposition:

Proposition 3 (Coarse-to-fine CI)**.** *If $Z \geq |\mathcal{O}^{\mathrm{Ca}}|$, the number of interventions for coarse-to-fine CI is $\Omega(\log|\mathcal{O}|)$.*

Considering that Proposition 3 holds for any division method, we choose a simple one, i.e., evenly dividing the observation space. Proposition 3 implies a tractable computational complexity, which will also be verified by some numerical experiments in Appendix J.

Combining RL and Coarse-to-Fine Counterfactual Inference

The coarse-to-fine CI forms a tree with depth $\Omega(\log|\mathcal{O}|)$ and width $\Omega(Z\log|\mathcal{O}|)$; we thus name it *Tree-based History Counterfactual Inference (T-HCI)*. This section combines T-HCI with RL to develop a concrete algorithm, of which the structure is illustrated in Figure 3 and its details are shown in Appendix A. Our algorithm includes two loops, the T-HCI loop and the RL loop, which are executed alternately. In the RL loop, an RL algorithm is performed for a certain number of episodes, and the roll-outs are stored in a replay pool; in the T-HCI loop, T-HCI is performed on the replay pool.

T-HCI is built on discrete observations but can also handle continuous observation spaces by using observation discretization techniques such as hashing (Zhu et al. 2018a; Hong et al. 2021; Burda et al. 2019b). The detailed observation discretization techniques are provided in Appendix A.2. We use $\widehat{\mathcal{O}}$ to denote the constructed discrete observation space, $\mathcal{O}^i$ to denote the current intervened observation subspace from $\widehat{\mathcal{O}}$, and $\mathcal{O}^{\mathrm{w}}$ to denote the observations that have been eliminated by previous inference. We set observations

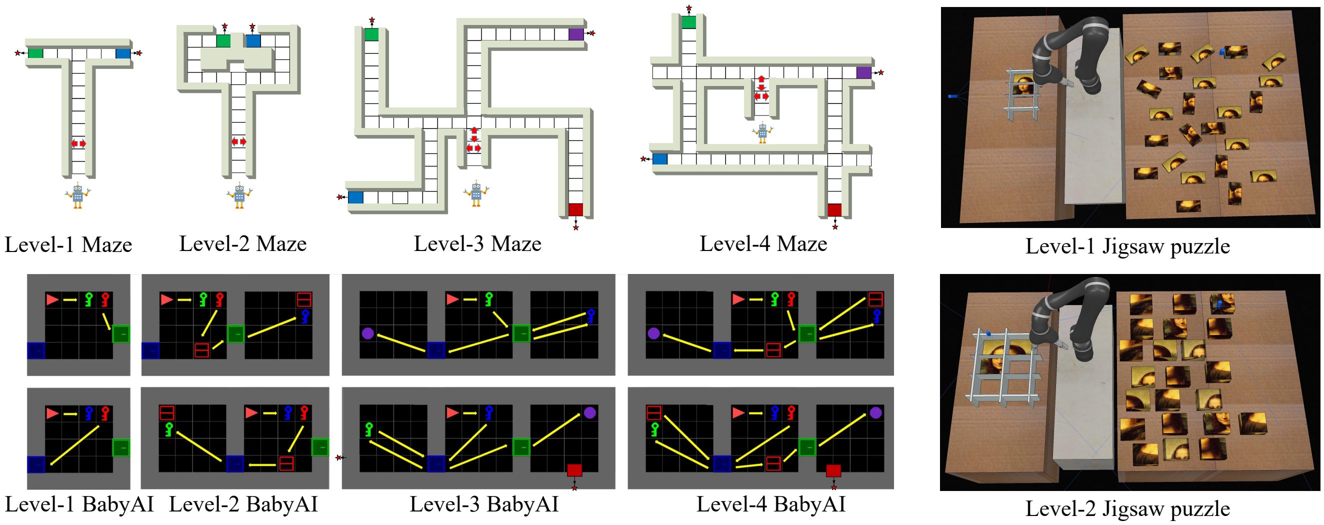

Figure 4: Environments of Maze, BabyAI, and Jigsaw Puzzle tasks.

from $\mathcal{O}^i \cup \mathcal{O}^{\text{w}}$ in h_{t-1} as zero vectors and get $h_{t-1}^{\overline{\mathcal{O}^i,\mathcal{O}^{\text{w}}}}$, and use a dynamics predictor p_ζ parameterized with ζ to estimate the counterfactual distribution. Let $\hat{p}(\cdot|h_t, a_t)$ denote the factual distributions of (o_{t+1}, r_{t+1}) on the samples in the replay buffer. Then, we can get the ATE by optimizing ζ with

$$\text{ATE}(\overline{\mathcal{O}^i,\mathcal{O}^{\text{w}}}) \tag{9}$$

$$= \min_\zeta \sum_{t=1}^{T} \mathbb{E}_{\substack{h_t \in \mathcal{H}_t \\ a_t \in \mathcal{A}}} \left|\left|p_\zeta(\cdot|h_{t-1}^{\overline{\mathcal{O}^i,\mathcal{O}^{\text{w}}}}, o_t, a_t) - \hat{p}(\cdot|h_t, a_t)\right|\right|_1.$$

As samples increase, the empirically constructed observation space and the estimated ATE will approach the true space and the true value, respectively. In practice, ATE can be estimated by inverse dynamics to avoid the reconstruction of high-dimensional observations $p(a_t|h_t, o_{t+1})$ (Pathak et al. 2017; Jordan and Rumelhart 1992; Sun et al. 2019; Burda et al. 2019a).Appendix A.2 shows the detailed derivation of ATE, as well as the techniques for further speeding up CI.

If $\mathcal{O}^i$ includes observations with causality, ATE cannot be optimized to zero; otherwise, the ATE will approach zero. A threshold ξ is set to determine whether an ATE is significant. If ATE$\leq \xi$, $\mathcal{O}^i$ are assigned to the set of observations $\mathcal{O}^{\text{w}}$ without causality, i.e., $\mathcal{O}^{\text{w}} \leftarrow \mathcal{O}^{\text{w}} \cup \mathcal{O}^i$. Based on $\mathcal{O}^{\text{w}}$, we can construct a memory of causal historical observations. The memory for policy learning is realized recurrently with a gate function Gate_ϑ parameterized by ϑ, with $\text{Gate}_\vartheta(o_j)$ taking value 1 if o_j belongs to $\mathcal{O}^{\text{w}}$ and 0 otherwise. The gate function is trained with the following loss function:

$$\mathcal{L}_{\text{Gate}}(\vartheta) = \frac{1}{|\widehat{\mathcal{O}}|} \sum_{o \in \widehat{\mathcal{O}}} |\text{Gate}_\vartheta(o) - I(o \in \mathcal{O}^{\text{w}})|, \tag{10}$$

where $I(\cdot)$ is the indicator function. Policy learning is built on the memories of causal historical observations, which are constructed by a function mapping $\phi : \{o_0, \cdots, o_{t-1}, o_t\} \mapsto \{\text{Gate}(o_0) \ominus o_0, \cdots, \text{Gate}(o_{t-1}) \ominus o_{t-1}, o_t\}$, w.r.t. $\phi : h \mapsto \psi$. The operator $\ominus$ denotes whether o_j is removed: o_j is removed when $\text{Gate}(o_j) = 1$ and retained vice versa. We name ψ as causal memory. The Gate_ϑ leads to a parameterized causal mapping ϕ_ϑ, upon which a policy $\pi_\theta(a_t|\phi_\vartheta(h_t))$ parameterized by θ is built. Since causal memories are compact, we use a classic LSTM model to encode them in practice. The policy is trained with three representative algorithms, Deep Monte Carlo (DMC) (Sutton and Barto 2018), A2C (Mohammad et al. 2017), and PPO (Schulman et al. 2017). The loss functions are provided in Appendix A.1.

At the cost of computational complexity, the main advantage of CI is filtering information without causality to greatly compress histories for sample efficiency gaining. This can be further verified by some theoretical analysis of sample complexity, which is provided in Appendix C.6. In the next section, we experimentally verify the effectiveness of T-HCI in terms of sample efficiency and computational feasibility.

Experiments

In this section, we evaluate T-HCI on various RL tasks with partial observability. Our experiments are designed to answer the following three questions: 1) Can T-HCI improve the sample efficiency of RL methods? 2) Is the computational overhead of T-HCI acceptable in practice? 3) Can T-HCI mine observations with causal effects?

Environmental Settings

Three popular types of tasks are used to evaluate T-HCI's effectiveness to adjust for confounding: Maze, BabyAI, and Jigsaw puzzle. Maze and BabyAI tasks are commonly used as grid-like partially-observable tests (Oh et al. 2016; Loynd et al. 2020; Chevalier-Boisvert et al. 2019). To validate T-HCI under different degrees of causal sparsity, we employ four levels of Maze and BabyAI sub-tasks as shown in Figure 4. In the Maze tasks, the agent needs to navigate to the exit. The exit location and the signposts near the entrance are randomly generated at the beginning of each episode.

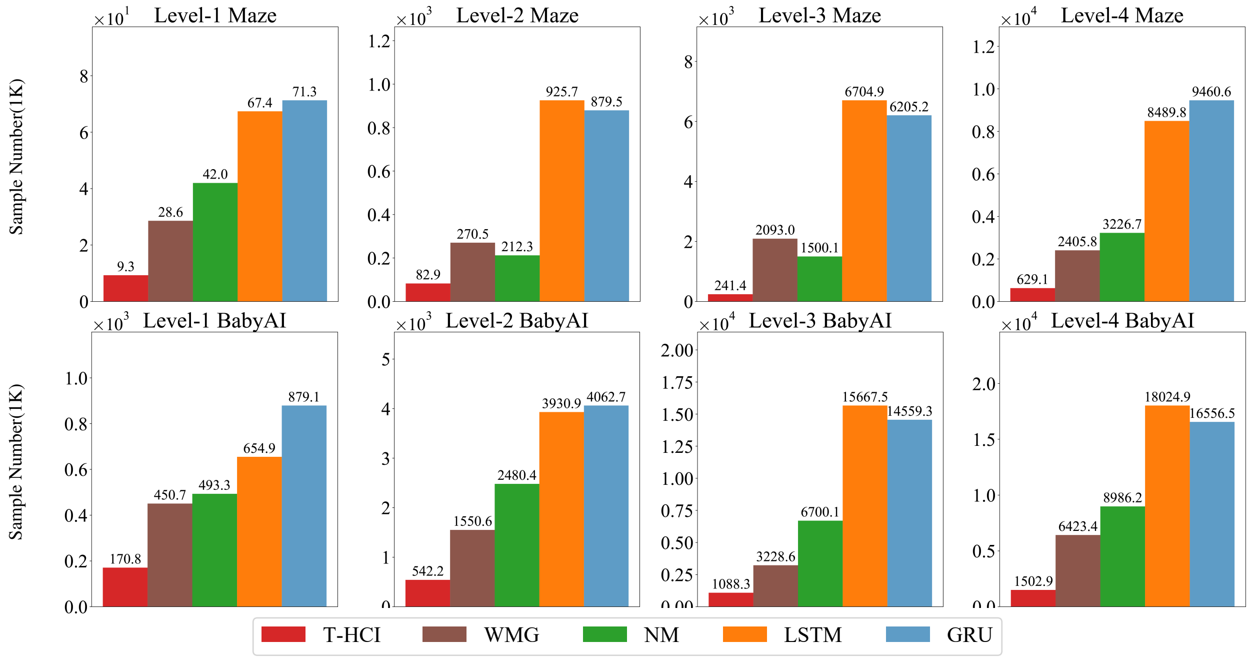

Figure 5: Comparison of average numbers of samples (thousand) in 10 trials of the Maze and BabyAI tasks.

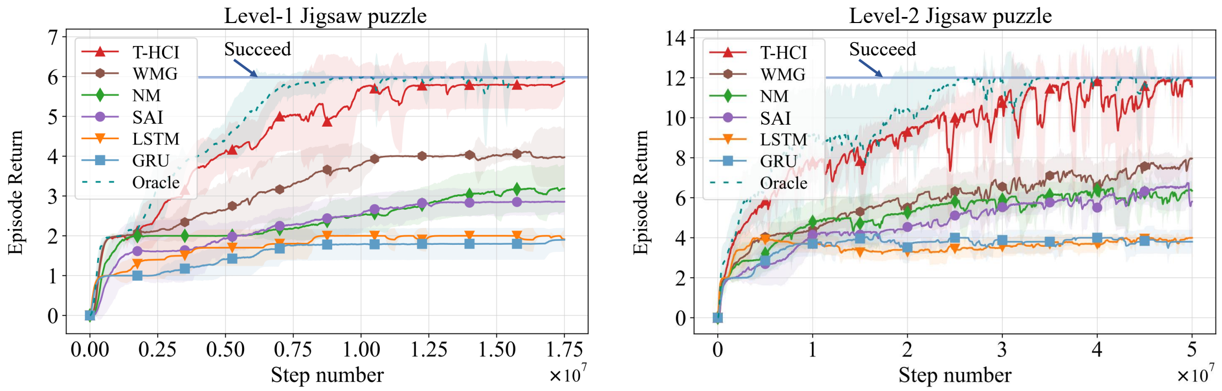

Figure 6: Learning curves of the Jigsaw puzzle tasks.

The *narrow* terrain in the maze and the *sparsity* of causal observations increase the confounding correlations of historical observations. As for BabyAI tasks, we focus on UnlockToUnlock (Chevalier-Boisvert et al. 2019), including key-to-door, key-to-box, and go-to-ball navigation tasks. In these tasks, the agent needs to depart and return from the middle room, which increases the confounding correlations for *bottleneck states* near the doors and makes memories abundant and hinders sample-efficient RL. Tasks of the puzzle domain (Kristensen, Valdivia, and Burelli 2020; Jain, Szot, and Lim 2020; Kulharia et al. 2016; Botea, Müller, and Schaeffer 2002; Kapturowski et al. 2019) are commonly used for high-level challenges for partially-observable tests due to rich environmental factors. As shown in Figure 4, we focus on 3D Jigsaw puzzle with continuous observation spaces built on Coppeliasim (Rohmer, Singh, and Freese 2013; Bogaerts et al. 2020; Gao et al. 2022). Its challenges are two-fold, i.e., the uncertainty of robot grasping and inserting and the huge scale of history spaces caused by vast observations and long-term horizons.

Comparative Experiments

We compare T-HCI with three different styles of baselines to evaluate the performance. The three styles of baselines include LSTM and GRU (Hausknecht and Stone 2015) of Gated RNN style, Neural Map (NM) (Parisotto and Salakhutdinov 2018) of Neural Turing Machine (NTM) style, and Working Memory Graph (WMG) (Loynd et al. 2020) of recurrent Transformer style. Herein, NM and WMG are respectively the state-of-the-art algorithms for Maze tasks and BabyAI tasks. More details of the baselines and parameter settings are in Appendix E.

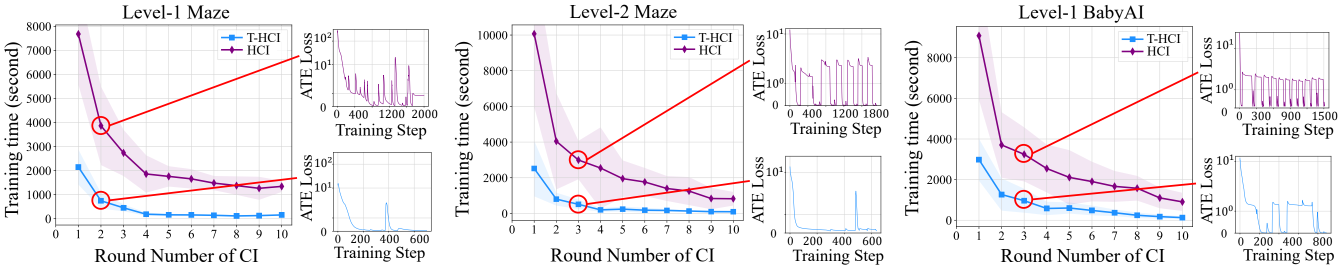

Figure 7: Training times at first 10 CI rounds and training losses in some representative rounds.

Task	Maze				BabyAI			
	Level-1	Level-2	Level-3	Level-4	Level-1	Level-2	Level-3	Level-4
LSTM	277	2361	10950	13950	1560	6846	18909	22506
GRU	213	2180	9442	11080	1743	6496	16683	20442
NM	405	1043	4339	6951	1684	4725	10846	14730
WMG	363	1797	5163	5152	1723	3982	7683	10155
HCI	23146	33714	>70000	>70000	36775	64330	>70000	>70000
T-HCI	4627	6169	10663	12570	7166	10846	14822	19804

Table 1: Average training wall-clock time (second) to complete the Maze tasks and BabyAI tasks.

TAIE	Level-3 Maze	Level-3 BabyAI
HAI	0.38	0.32
SAI:head-1	0.17	0.16
SAI:head-2	0.15	0.17
SAI:head-3	0.11	0.18
SAI:head-4	0.19	0.16
T-HCI	0.00	0.00

Table 2: Quantitative evaluation of inference results of various algorithms based on TAIE.

To evaluate sample complexity, we average the number of samples over ten independent trials for each task. Figure 5 shows that T-HCI achieves the best sample efficiency in every sub-task. The sample efficiency of T-HCI is almost five times higher than that of advanced NM and WMG in the complex Level 3-4 sub-tasks. The learning curves are provided in Appendix G. Because T-HCI improves sample efficiency at the cost of computational overhead, we propose to apply T-HCI in robotic Jigsaw puzzle tasks where sample overhead is more expensive. We add an oracle baseline to compare the effect of utilizing complete observation. Figure 6 shows the learning curves for the Level 1-2 Jigsaw puzzle tasks. Only T-HCI and Oracle can complete these complex tasks, and T-HCI yields similar performance to Oracle, which indicates the adaptability of T-HCI to more realistic and complex problems. Videos of the T-HCI agent playing Jigsaw puzzle are provided in the supplementary material.

To evaluate computational complexity, we deploy T-HCI and vanilla HCI (CI on observations) in the Maze and BabyAI tasks. We average the training wall-clock times over ten independent trials for each algorithm. Figure 7 and Table 1 show that the high computational complexity of HCI prevents its direct use for complex tasks, while T-HCI can significantly reduce computational complexity. Although T-HCI exhibit higher computational complexity in Level 1-2 tasks, its complexity grows slower when the task complexity increases than the compared methods; for Level 3-4 tasks T-HCI is more efficient than some of the baselines, showing that the computational cost of T-HCI is acceptable.

Ablation Study and Visualization

We also compare T-HCI with the following variants to qualitatively and quantitatively evaluate the inference results. (1) Oracle: a variant that leverages causal historical observations provided by experts and needs no memory. (2) SAI: a variant that implements reasoning with a soft-attention mechanism (Vaswani et al. 2017) on the replay buffer instead of CI. SAI is with four attention heads and one transformer layer for visualization of the attention weights of these heads. (3) HAI: a variant that mines informative observations with a hard-attention mechanism (Xu et al. 2015) on the replay buffer.

We visualize the heatmaps of inference results in Figure 8, which indicate that T-HCI does mine the observations with semantic causality. In comparison, the attentions of SAI and HAI are distracted by the uninformative historical observations, e.g., those around the doorway. As shown in Table 2, we propose Timely Averaged Inference Error (TAIE) to quantitatively evaluate the difference between the inference results and ground truth. Let P_{π^*} denote the distribution of the trajectories followed by π^*, and T_τ denote the length of a trajectory. We then define TAIE $= \mathbb{E}_{\tau\sim P_{\pi^*}}\left[\frac{\sum_{t=0}^{T_\tau}|\nu(o_t)-\nu^*(o_t)|}{T_\tau}\right]$, where ν is the attention weight, and ν^* is the ground truth (the weights of observations with causal effects are 1 while others are 0, labeled by humans). We show the average training times and numbers of samples in Level-3 BabyAI in Figure 9. Although HAI and SAI use the replay buffer, the lack of adjustment for confounders causes high computational

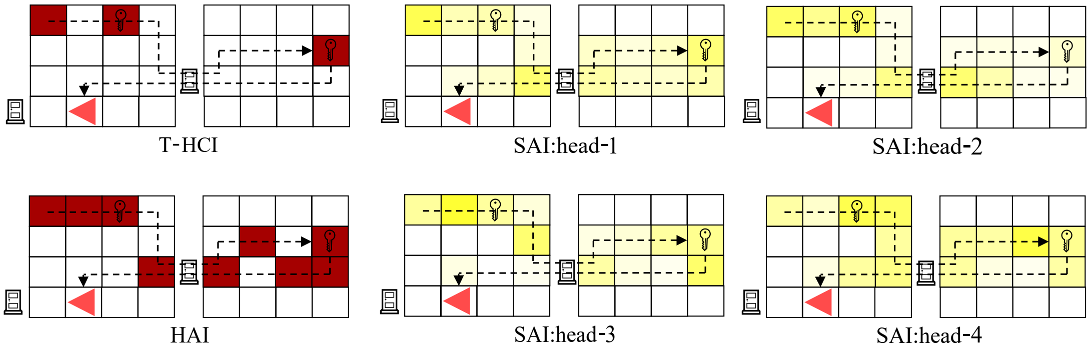

Figure 8: Heatmaps of positions on the illustrated trajectory in the Level-3 BabyAI task, where warmer yellow colors represent larger attention weights and the brown colors represent hard attention.

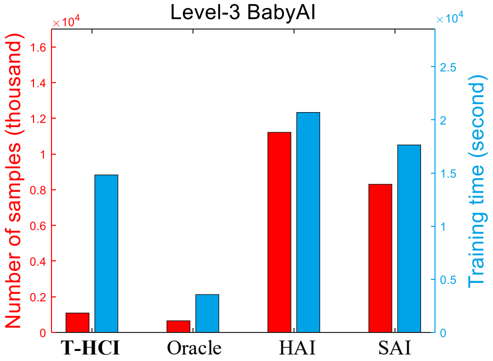

Figure 9: Ablation study of sample complexity and computational complexity.

and sample complexity. The sample efficiency of T-HCI approaches that of Oracle, suggesting that T-HCI infers causal observations with a small number of samples.

Related Work

Unlike traditional sliding-window-like history-based RL (Oh and Kaneko 2018; Mnih et al. 2015; Jiang et al. 2017), advanced seq2seq-based methods encode complete histories into hidden states (Majeed and Hutter 2018; Hutter 2008). The mainstream seq2seq models in the RL community can be divided into three categories: Gated RNN (Hausknecht and Stone 2015; Peng et al. 2018; Gao et al. 2020), NTM (Graves et al. 2016; Yang and Rush 2017), and Attention (Etchart, Ladosz, and Mulvaney 2019; Oriol Vinyals et al. 2019; Zhong, Rocktäschel, and Grefenstette 2019; Mishra et al. 2018). In earlier research, NTM-style methods outperform Gated RNN-style methods in grid-like tasks such as Maze (Oh et al. 2016; Parisotto and Salakhutdinov 2018). Recently, transformer-style methods achieve good results in many partially-observable tasks (Goyal et al. 2021; Parisotto et al. 2020; Chen et al. 2021b,c). However, learning a compact historical representation with seq2seq models remains a challenge (Mishra et al. 2018; Oh and Kaneko 2018).

In POMDPs, the hidden states (James, Singh, and Littman 2004; Monahan 1982) in general symbolize sufficient statistics for optimal decision-making, which are commonly not observable but estimable. These hidden states correspond to more compact subsets of histories than the causal historical observations mined by T-HCI. To further learn a historical representation of these hidden states, T-HCI resorts to performing LSTM methods over the mined causal historical observations. Since T-HCI filters out those historical observations without causality in a gating way, T-HCI can learn more compact representation than encoding complete histories with seq2seq models.

Our work draws inspiration from state abstraction and CI over time. Specifically, state abstraction methods that use CI (Zhang et al. 2021, 2020; Suau et al. 2020) mainly analyze the Structural Causal Model (SCM (Pearl 2013)) and construct loss functions so that the factored states with causality can be directly obtained through gradient descent, following the idea of Invariant Causal Prediction (ICP (Peters, Buhlmann, and Meinshausen 2016)). The main challenge of CI for history abstraction compared to CI for state abstraction is that historical observations have a posteriori effect on the distributions of multi-step dynamics transitions. Instead, our approach counts the causal effects on future transition probabilities, as shown in our ATE losses that are built upon the dynamics model-invariant metrics (Tomar et al. 2021). The methods of CI over time (Bica et al. 2020; Zhang et al. 2022a) have been developed, and recently a few studies prove that causal inference in partially observable environments is theoretically feasible in imitation learning domains (Kumor, Zhang, and Bareinboim 2021; Zhang, Kumor, and Bareinboim 2020; Etesami and Geiger 2020; de Haan, Jayaraman, and Levine 2019). However, these studies rely on sufficient expert data, which is not available in RL domains. Besides, these methods are built on single-variable intervention and cannot control the computational complexity of CI when being applied to complex RL tasks with long histories.

Discussions

One of the main challenges of history-based RL is that some historical information is cause for confounding but is hard to be adjusted. T-HCI successfully combines CI with history-based RL to effectively adjust for confounders. This technique is helpful in developing more efficient RL methods. We show that T-HCI exponentially reduces the number of interventions and is computationally feasible to be combined with RL in practice. We believe this work would potentially benefit both RL and causality communities (Bica et al. 2020; Zhang et al. 2022a).

T-HCI is built upon two mild assumptions, i.e., the sparsity and the time independence of the causal historical observations. To our knowledge, they hold in a wide range of applications and benchmarks including Maze, BabyAI, and Puzzle domains (Oh et al. 2016; Chevalier-Boisvert et al. 2019; Botea, Müller, and Schaeffer 2002). Additionally, T-HCI has two extra limitations that can be relaxed in future work. First, although we use historical observations to estimate the hidden state instead of historical observation-action pairs, this paper assumes that the effects of historical actions are reflected in the historical observations, and thus the historical observations are sufficient for decision-making. We would also like to note that T-HCI can be applied to tasks with observation-action histories by redefining the historical observation as the historical observation-action pair. We will test T-HCI in these tasks in future work. Second, the technique of setting certain values to zero vectors in T-HCI is an approximate CI method followed by this paper and other research (Chen et al. 2021a). Empirically, we found it effective to assess the treatment effect in our tasks.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0200300, in part by the National Natural Science Foundation of China under Grant 61836004 and 62176133, in part by the Tsinghua-Guoqiang research program under Grant 2019GQG0006, and in part by the National Natural Science Foundation of China under Grant 62206151, the China National Postdoctoral Program for Innovative Talents under grant BX20220167, and the Shuimu Tsinghua Scholar Program.

References

Bica, I.; Alaa, A. M.; Jordon, J.; and van der Schaar, M. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *ICLR*.

Bogaerts, B.; Sels, S.; Vanlanduit, S.; and Penne, R. 2020. Connecting the CoppeliaSim robotics simulator to virtual reality. *SoftwareX*.

Botea, A.; Müller, M.; and Schaeffer, J. 2002. Using Abstraction for Planning in Sokoban. In *Computers and Games*.

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A. J.; Darrell, T.; and Efros, A. A. 2019a. Large-Scale Study of Curiosity-Driven Learning. In *ICLR*.

Burda, Y.; Edwards, H.; Storkey, A. J.; and Klimov, O. 2019b. Exploration by random network distillation. In *ICLR*.

Chen, J.; Wu, X.; Hu, Y.; and Luo, J. 2021a. Spatial-temporal Causal Inference for Partial Image-to-video Adaptation. In *AAAI*.

Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I. 2021b. Decision Transformer: Reinforcement Learning via Sequence Modeling. *CoRR*.

Chen, S.; Guhur, P.; Schmid, C.; and Laptev, I. 2021c. History Aware Multimodal Transformer for Vision-and-Language Navigation. *CoRR*.

Chernozhukov, V.; Fernández-Val, I.; and Melly, B. 2013. Inference on counterfactual distributions. *Econometrica*.

Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T. H.; and Bengio, Y. 2019. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. In *ICLR*.

de Haan, P.; Jayaraman, D.; and Levine, S. 2019. Causal Confusion in Imitation Learning. In *NeurIPS*.

Etchart, M.; Ladosz, P.; and Mulvaney, D. 2019. Spatio-Temporal Attention Deep Recurrent Q-Network for POMDPs. In *EPIA*.

Etesami, J.; and Geiger, P. 2020. Causal Transfer for Imitation Learning and Decision Making under Sensor-Shift. In *AAAI*.

Gao, H.; Yang, Z.; Su, X.; Tan, T.; and Chen, F. 2020. Adaptability Preserving Domain Decomposition for Stabilizing Sim2Real Reinforcement Learning. In *IROS*.

Gao, H.; Yang, Z.; Tan, T.; Zhang, T.; Ren, J.; Sun, P.; Guo, S.; and Chen, F. 2022. Partial Consistency for Stabilizing Undiscounted Reinforcement Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Goyal, A.; Lamb, A.; Hoffmann, J.; Sodhani, S.; Levine, S.; Bengio, Y.; and Schölkopf, B. 2021. Recurrent Independent Mechanisms. In *ICLR*.

Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwinska, A.; Colmenarejo, S. G.; Grefenstette, E.; Ramalho, T.; Agapiou, J. P.; Badia, A. P.; Hermann, K. M.; Zwols, Y.; Ostrovski, G.; Cain, A.; King, H.; Summerfield, C.; Blunsom, P.; Kavukcuoglu, K.; and Hassabis, D. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*.

Hausknecht, M. J.; and Stone, P. 2015. Deep Recurrent Q-Learning for Partially Observable MDPs. In *AAAI*.

Hong, Z.-W.; Chen, T.; Lin, Y.-C.; Pajarinen, J.; and Agrawal, P. 2021. Topological Experience Replay for Fast Q-Learning. *Workshop in ICML*.

Hutter, M. 2008. Feature Markov Decision Processes. *CoRR*.

Jain, A.; Szot, A.; and Lim, J. J. 2020. Generalization to New Actions in Reinforcement Learning. In *ICML*.

James, M. R.; Singh, S.; and Littman, M. L. 2004. Planning with predictive state representations. In *ICMLA*.

Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2017. Contextual Decision Processes with low Bellman rank are PAC-Learnable. In *ICML*.

Joelle, P.; Geoffrey, J. G.; and Thrun, S. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI*.

Jordan, M. I.; and Rumelhart, D. E. 1992. Forward Models: Supervised Learning with a Distal Teacher. *Cogn. Sci.*

Kapturowski, S.; Ostrovski, G.; Quan, J.; Munos, R.; and Dabney, W. 2019. Recurrent Experience Replay in Distributed Reinforcement Learning. In *ICLR*.

Kristensen, J. T.; Valdivia, A.; and Burelli, P. 2020. Estimating Player Completion Rate in Mobile Puzzle Games Using Reinforcement Learning. In *Conference on Games*.

Kulharia, V.; Ghosh, A.; Patil, N.; and Rai, P. 2016. Neural Perspective to Jigsaw Puzzle Solving. *Department of Computer Science*.

Kumor, D.; Zhang, J.; and Bareinboim, E. 2021. Sequential causal imitation learning with unobserved confounders. In *NeurIPS*.

Liu, J.; Ma, Y.; and Wang, L. 2018. An alternative robust estimator of average treatment effect in causal inference. *Biometrics*.

Loynd, R.; Fernandez, R.; Celikyilmaz, A.; Swaminathan, A.; and Hausknecht, M. J. 2020. Working Memory Graphs. In *ICML*.

Lu, J. 2016a. Covariate adjustment in randomization-based causal inference for 2K factorial designs. *Statistics & Probability Letters*.

Lu, J. 2016b. On randomization-based and regression-based inferences for 2k factorial designs. *Statistics & Probability Letters*.

Majeed, S. J.; and Hutter, M. 2018. On Q-learning Convergence for Non-Markov Decision Processes. In *IJCAI*.

Majeed, S. J.; and Hutter, M. 2021. Exact Reduction of Huge Action Spaces in General Reinforcement Learning. In *AAAI*.

Mesnard, T.; Weber, T.; Viola, F.; Thakoor, S.; Saade, A.; Harutyunyan, A.; Dabney, W.; Stepleton, T. S.; Heess, N.; Guez, A.; Moulines, E.; Hutter, M.; Buesing, L.; and Munos, R. 2021. Counterfactual Credit Assignment in Model-Free Reinforcement Learning. In *ICML*.

Mishra, N.; Rohaninejad, M.; Chen, X.; and Abbeel, P. 2018. A Simple Neural Attentive Meta-Learner. In *ICLR*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; and et al. 2015. Human-level control through deep reinforcement learning. *Nature*.

Mohammad, B.; Iuri, F.; Stephen, T.; Jason, C.; and Jan, K. 2017. Reinforcement Learning through Asynchronous Advantage Actor-Critic on a GPU. In *ICLR*.

Monahan, G. E. 1982. A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms. In *Management Science*.

Oh, H.; and Kaneko, T. 2018. Deep Recurrent Q-Network with Truncated History. In *TAAI*.

Oh, J.; Chockalingam, V.; Singh, S. P.; and Lee, H. 2016. Control of Memory, Active Perception, and Action in Minecraft. In *ICML*.

Oriol Vinyals, I. B.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J. P.; Jaderberg, M.; Vezhnevets, A. S.; Leblond, R.; Pohlen, T.; Dalibard, V.; Budden, D.; Sulsky, Y.; Molloy, J.; Paine, T. L.; Gülçehre, Ç.; Wang, Z.; Pfaff, T.; Wu, Y.; Ring, R.; Yogatama, D.; Wünsch, D.; McKinney, K.; Smith, O.; Schaul, T.; Lillicrap, T. P.; Kavukcuoglu, K.; Hassabis, D.; Apps, C.; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*.

Parisotto, E.; and Salakhutdinov, R. 2018. Neural Map: Structured Memory for Deep Reinforcement Learning. In *ICLR*.

Parisotto, E.; Song, H. F.; Rae, J. W.; Pascanu, R.; Gülçehre, Ç.; Jayakumar, S. M.; Jaderberg, M.; Kaufman, R. L.; Clark, A.; Noury, S.; Botvinick, M.; Heess, N.; and Hadsell, R. 2020. Stabilizing Transformers for Reinforcement Learning. In *ICML*.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *ICML*.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics surveys*.

Pearl, J. 2013. Structural counterfactuals: A brief introduction. *Cognitive science*.

Peng, X. B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *ICRA*.

Peters, J.; Buhlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society*.

Ramos, A. G. C. P.; Mehrotra, A.; Lane, N. D.; and Bhattacharya, S. 2021. Conditioning Sequence-to-sequence Networks with Learned Activations. In *ICLR*.

Rohmer, E.; Singh, S. P. N.; and Freese, M. 2013. V-REP: A versatile and scalable robot simulation framework. In *IROS*.

Schochet, P. Z. 2010. Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*.

Seitzer, M.; Schölkopf, B.; and Martius, G. 2021. Causal Influence Detection for Improving Efficiency in Reinforcement Learning. In *NeurIPS*.

Smith, T.; and Simmons, R. G. 2005. Point-Based POMDP Algorithms: Improved Analysis and Implementation. In *UAI*.

Suau, M.; Congeduti, E.; He, J.; Starre, R. A. N.; Czechowski, A.; and Oliehoek, F. A. 2020. Influence-aware Memory for Deep RL in POMDPs. *NeurIPS Workshop*.

Sun, H.; Li, Z.; Liu, X.; Zhou, B.; and Lin, D. 2019. Policy Continuation with Hindsight Inverse Dynamics. In *NeurIPS*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*.

Tomar, M.; Zhang, A.; Calandra, R.; Taylor, M. E.; and Pineau, J. 2021. Model-invariant state abstractions for model-based reinforcement learning. *arXiv preprint arXiv:2102.09850*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*.

Witte, J.; and Didelez, V. 2019. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yang, G.; and Rush, A. M. 2017. Lie-Access Neural Turing Machines. In *ICLR*.

Zhang, A.; Lyle, C.; Sodhani, S.; Filos, A.; Kwiatkowska, M.; Pineau, J.; Gal, Y.; and Precup, D. 2020. Invariant Causal Prediction for Block MDPs. In *ICML*.

Zhang, A.; McAllister, R. T.; Calandra, R.; Gal, Y.; and Levine, S. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *ICLR*.

Zhang, J.; Han, Y.; Tang, J.; Hu, Q.; and Jiang, J. 2017. Semi-Supervised Image-to-Video Adaptation for Video Action Recognition. *IEEE Trans. Cybern.*

Zhang, J.; Kumor, D.; and Bareinboim, E. 2020. Causal Imitation Learning With Unobserved Confounders. In *NeurIPS*.

Zhang, Y.-F.; Zhang, H.; Lipton, Z. C.; Erran, L.; Eric, L.; and Xing, P. 2022a. Exploring Transformer Backbones for Heterogeneous Treatment Effect Estimation. In *Arxiv*.

Zhang, Z.; Ding, L.; Cheng, D.; Liu, X.; Zhang, M.; and Tao, D. 2022b. BLISS: Robust Sequence-to-Sequence Learning via Self-Supervised Input Representation. *CoRR*.

Zhong, V.; Rocktäschel, T.; and Grefenstette, E. 2019. RTFM: Generalising to Novel Environment Dynamics via Reading. *CoRR*.

Zhu, Y.; Wang, Z.; Merel, J.; Rusu, A. A.; Erez, T.; Cabi, S.; Tunyasuvunakool, S.; Kramár, J.; Hadsell, R.; de Freitas, N.; and Heess, N. 2018a. Reinforcement and Imitation Learning for Diverse Visuomotor Skills. In *RSS*.

Zhu, Z.; Wu, W.; Zou, W.; and Yan, J. 2018b. End-to-End Flow Correlation Tracking With Spatial-Temporal Attention. In *CVPR*.