FROM CAUSAL GRAPHS TO MARKOVIAN STATES: MULTI-ORDER STATE EXPOSURE FOR DEEP RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Online reinforcement learning (RL) relies on the Markov property for guaranteed performance, but real-world applications often lack well-defined states given raw observed variables. While causal RL has attracted growing interest, existing work typically assumes Markovian states are provided and uses causality only to accelerate learning. This leaves a fundamental gap: *given a longitudinal causal graph over observed variables, how does one construct MDP states that provably satisfy the Markov property?* We provide a procedure that constructs a provably minimal such state. In deep RL, the minimal representation alone empirically fails to improve performance, indicating that neural networks cannot automatically exploit Markovian minimality. We therefore propose **MOSE** (Multi-Order State Exposure), which exposes a single shared Q-function to history-based state constructions of orders 1 through W simultaneously. An attention-based selection layer scores the dimensions of this combined input by their estimated relevance to value-function estimation and selects the high-scoring ones. Causal MOSE further constrains the selection layer to retain every variable of the minimal state, guaranteeing sufficiency while leaving the model free to admit useful redundancy. MOSE consistently outperforms both the minimal state construction and single-window policies on common benchmarks and synthetic datasets. Including the minimal representation alongside MOSE can further improve performance. Our results establish a core principle for causal deep RL: minimal sufficiency is not enough. *Controlled redundancy* gives the model flexibility to learn its state representation end-to-end. Causal state information provides structural guidance that can make this learning more efficient.

1 INTRODUCTION

Reinforcement learning (RL) algorithms typically assume that the agent's state satisfies the Markov property: the future is conditionally independent of the past given the present (Sutton & Barto, 1998). In practice, however, raw observations (e.g., sensor streams) do not inherently form a Markovian state, because the system's dynamics may depend on an unknown amount of history. For instance, sepsis treatment depends on vital-sign trajectories, not just current readings (Komorowski et al., 2018); similarly, robotic manipulation may require tracking occluded objects across frames (Chung et al., 2026). Using non-Markovian observations as states can, in theory, lead to suboptimal policies and unsafe decisions.

A large body of prior work addresses this challenge under the partially observable Markov decision process (POMDP) framework, where a latent Markovian state exists but is not directly observed (Monahan, 1982). Methods in this line of work learn a belief state using sequential architectures such as LSTMs (Hausknecht & Stone, 2015; Kapturowski et al., 2018; Badia et al., 2020) or Transformers (Zambaldi et al., 2019; Chen et al., 2021b; Zheng et al., 2022). Though in principle these approaches can recover the optimal policy given sufficient history, they offer no guidance on which past observations are actually needed. Instead, this is learned from data, often requiring large samples. Further, it is unclear if the learned representation satisfies Markovian Property which makes state selection on the learned representation necessary. Meanwhile, *frame stacking* arises as a pragmatic heuristic that has become widespread in deep RL practice. Concretely, frame stacking concatenates a fixed number k of recent observations into a single input, reducing partial observability and rendering the process approximately Markovian over k time steps (Shi et al., 2020). Frame

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

stacking can be effective when dependencies span a short, known horizon. For example, stacking four frames in Atari is assumed to sufficiently capture object motion and works well in practice (Mnih et al., 2015; Hessel et al., 2018; Kaiser et al., 2020). However, as the true window length k is often unknown, k must be tuned or learned. When k is too small, we lose critical information, while arbitrarily large k requires excess computation and memory and leads to poor sample efficiency (Tasse et al., 2025).

To move beyond these heuristics when constructing the state space, a principled approach would need to determine exactly which past variables are causally relevant for the future. While causal *directed acyclic graphs* (DAGs) have been used in RL to improve learning, these DAGs are typically static (non-longitudinal) and assume a pre-specified, Markovian state representation. These methods do not ask the more basic question: *how does one construct Markovian states from raw observations?* In this work, we bridge this gap by establishing a direct mapping from a known longitudinal causal graph to a *minimal Markovian state representation*: a representation that satisfies the Markov property and is minimal in the sense that removing any variable would, for some consistent environment, either break the Markov property or reduce the optimal value. This establishes the long-missing connection from longitudinal causal graphs to Markov decision process (MDP) states.

While prior causal RL methods assume access to a valid MDP (Li et al., 2025a;b), we instead ask whether the state we construct can itself improve learning. The answer, however, is not straightforward in deep RL. Deep networks are known to overfit to spurious correlations and rely on redundant features which affect the expressivity of neural networks (Zhang et al., 2018; Nikishin et al., 2022; Sokar et al., 2023; Cobbe et al., 2019). By stripping away all statistical slack, a minimal representation can lead to sharp minima and poor generalization (Zhang et al., 2018; Nikishin et al., 2022; Sokar et al., 2023; Cobbe et al., 2019). Indeed, in our experiments, we discover that feeding only the minimal representation often fails to improve performance and can even degrade it.

Motivated by this, we reconsider the role of redundancy. We propose MOSE (Multi-Order State Exposure): we add a selective layer that will operate on the belief history either learned by sequential architecture or the raw observation that is markovian (W^{th} order) so the model can select the redundant information which helps learning value function. To calculate importance score, we calculate the attention between a learnable query and each element of the history and then compare with a learnable threshold to decide whether to include that element as input or not. This design is supported by a range of findings: (i) attention can calculate the relationship between elements (Zambaldi et al., 2019; Vaswani et al., 2017); (ii) exploration signals improve deep RL performance (Fortunato et al., 2018; Laskin et al., 2020b); and (iii) multi-view learning helps neural networks build richer representations (Li et al., 2019; Chen et al., 2021a; Jangir et al., 2022; Hwang et al., 2023). Empirically, MOSE substantially outperforms both the minimal state and standard frame stacking in multiple environments. Moreover, we introduce Causal-MOSE, which always inputs minimal markovian state while the model can decide whether to retain other information. Causal-MOSE performed as well or better than MOSE in experiments, suggesting that controlled redundancy can unlock the benefit of causal state information.

Contributions This work presents three main contributions.

1. *Minimal Markovian state construction from time series causal DAGs.* We prove a graphical criterion for constructing a valid state space given a time series causal DAG. We provide a procedure that returns a provably minimal state space satisfying validity (Algorithm G.1).
2. *Multi-Order State Exposure (MOSE).* We provide a heuristic method for state space construction that does not require knowledge of the causal DAG (Algorithm G.2). We show that in deep RL, MOSE improves upon both the naive frame stacking heuristic and the minimal markovian state provided by Algorithm G.1.
3. *Causal-MOSE.* Finally, we show that combining the minimal Markovian state with MOSE performs as well or better on the experimental settings explored in this work.

2 Related Works

This work lies at the intersection of RL, causality, and deep learning. See Appendix B for extended related works.

Markovian Tests Non-Markovianity is a known issue in RL, yet testing it requires new test statistics as classical conditional independence tests (CITs) assume i.i.d. data. Recent efforts have addressed this by developing asymptotically valid test procedures (e.g., conditional characteristic-function tests (Chen & Hong, 2012) and forward-backward sequential tests (Shi et al., 2020)), embedding such tests in deep generative frameworks for high-dimensional settings (Zhou et al., 2023). These Markovian tests generalize CITs to time series, enabling us to directly test whether a future variable is conditionally independent of the full past given a candidate parent set. We further discuss tests of Markovianity in our framework in Section 3.

Causal Reinforcement Learning A growing literature explores the intersection of causality and RL. Assuming access to a ground truth DAG, causal graphs have been explored in *reward design* for mediator-based surrogate rewards (Zou et al., 2025), robust reward modeling (Srivastava et al., 2025), and causal Bellman equations under confounded offline data (Li et al., 2025a; Juliani et al., 2026). Causal perspectives have been leveraged in *policy space reduction* to eliminate reward-irrelevant actions and shrink the policy space Zhang & Bareinboim (2020), and in *state space representation* Wang et al. (2022); Zhang et al. (2020); Wang et al. (2024); Gao et al. (2025) to aggregate similar states (a.k.a., state abstraction (Andre et al., 2002; Abel et al., 2018; Ni et al., 2024)) and learn latent representations Lesort et al. (2018); Ota et al. (2020). All these methods assume a valid Markovian state representation is already given. Our work is complementary: we provide a procedure to construct a minimal Markovian state from raw observations using a longitudinal causal graph, and we propose MOSE to make it usable in deep RL. The resulting representation can serve as a foundation upon which existing causal RL techniques could be applied to potentially further improve performance.

Redundancy and Exploration in Deep RL A separate line of work has shown that controlled redundancy via structured noise, data augmentation, or multi-view inputs can improve learning and generalization in deep RL. Coarse representations may encourage the learner to ignore impertinent information (Sutton & Barto, 1998). Noise-based exploration methods (e.g., NoisyNet (Fortunato et al., 2018) and curiosity-driven exploration (Pathak et al., 2017)) inject stochasticity into the agent's parameters or reward signal, smoothing the optimization landscape and preventing premature convergence. Deep RL agents have been observed to learn more efficiently when increasing the input dimensionality through high-dimensional representation learning (Ota et al., 2020). Data augmentation (e.g., random shifts, cropping) acts as an effective regularizer in visual RL, improving generalization (Laskin et al., 2020b; Yarats et al., 2021; Stooke et al., 2021). Multi-view RL (Li et al., 2019; Chen et al., 2021a; Jangir et al., 2022; Hwang et al., 2023) and contrastive representation learning (Sermanet et al., 2018; Laskin et al., 2020a; Stooke et al., 2021; Laskin et al., 2022; Eysenbach et al., 2022) further demonstrate that providing multiple, partially redundant views of the same observation yields richer, more robust representations. Collectively, these findings support the intuition that deep networks benefit not from information-theoretic minimalism, but from carefully designed redundancy. MOSE operationalizes this basic principle.

3 PRELIMINARIES

Structural Causal Models This work uses the formalisms of structural causal models (SCMs) (Pearl, 2000) to bridge causal DAGs and MDPs. For explication, we begin by defining stationary SCMs for i.i.d. data and subsequently extend them to the time series setting.

Definition 3.1 (Structural causal model (SCM), Bareinboim et al. 2022)**.** An SCM is a tuple $\mathcal{M} := \langle \mathbf{U}, p(\mathbf{u}), \mathbf{X}, \mathcal{F} \rangle$ where $\mathbf{U} = \{U^i\}_{i=1}^n$ is the set of exogenous variables explained by mechanisms external to $\mathcal{M}$, $p(\mathbf{u})$ is the distribution over $\mathbf{U}$, $\mathbf{X} = \{X^i\}_{i=1}^n$ is the set of endogenous variables explained by variables in $\mathbf{U} \cup \mathbf{X}$, and $\mathcal{F} = \{f^i\}_{i=1}^n$ is the set of structural functions such that $x^i = f^i(\mathbf{pa}_{X^i}, u^i)$ for endogenous parent set $\mathbf{pa}_{X^i}$ and exogenous context u^i.

Given an SCM, $\mathcal{M}$ (Def. 3.1), its causal structure is encoded by a directed graph $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ over variables $\mathbf{X}$, and $\mathbf{E}$ is the edge set. For a $\mathcal{G}$ that is linked with $\mathcal{M}$, an edge $X^i \to X^j$ iff $X^i \in \mathbf{pa}_{X^j}$. We restrict our attention to *directed acyclic graphs* (DAGs). We assume a positive SCM: $p(x^i) > 0$ for every realization $X^i \in \mathbf{X} = x^i$ (Pearl, 1999). We additionally assume causal sufficiency, i.e., the absence of latent confounding in $\mathcal{G}$. We assume a nonparametric SCM, but do not require jointly independent noise variables and are therefore not restricted to the *non-parametric structural equation model with independent errors* (NPSEM-IE) regime (Wang et al., 2025).

Extension to Time Series This work considers the setting where SCM $\mathcal{M}$ represents a discrete-time multivariate stochastic process $\mathbf{X}, \mathbf{R}, \mathbf{A}$ over a finite horizon, so observations are temporally dependent and not i.i.d. The endogenous variables are

$$\mathbf{X} \coloneqq \{\mathbf{X}_t\}_{t=0}^{T} = \{\mathbf{X}_0, \ldots, \mathbf{X}_T\}, \mathbf{R} \coloneqq \{R_t\}_{t=0}^{T} = \{R_0, \ldots, R_T\}, \mathbf{A} \coloneqq \{A_t\}_{t=0}^{T} = \{A_0, \ldots, A_T\}, \tag{1}$$

where $\mathbf{X}_t \coloneqq \{X_t^i\}_{i=0}^{m-1}$, with m variables at each time step, and $\mathbf{X}^i \coloneqq \{X_t^i\}_{t=0}^{T}$ for discrete time indices $t \in \mathbb{N}$. As in Definition 3.1, X_t^i is a function of its causal parents and an exogenous noise variable U_t^i. $\mathcal{G}$ is now a time series causal DAG over endogenous variables $\{\mathbf{X}_t\}_{t=0}^{T} \cup \{R_t\}_{t=0}^{T} \cup \{A_t\}_{t=0}^{T}$. The graph containing every variable in the time series SCM is the *full time graph*.

Definition 3.2 (Full time graph, Runge 2018; Assaad et al. 2022; Hasan et al. 2023)**.** Let $\mathcal{G} = (\mathbf{X} \cup \mathbf{R} \cup \mathbf{A}, \mathbf{E})$ be a full time graph of multivariate process $\mathbf{X}, \mathbf{R}, \mathbf{A}$. Nodes are $X_t^i \in \mathbf{X}$, $R_t \in \mathbf{R}$, $A_t \in \mathbf{A}$ for all time $t \in [T]$. Edges $\mathbf{E}$ are lag-specific directed links $X_{t-k}^i \to X_t^j$ with a time lag of $k > 0$ for $i = j$ (*autodependencies*) and $k \geq 0$ for $i \neq j$, and analogously for $\mathbf{R}$ and $\mathbf{A}$.

We do not require *causal stationarity* in $\mathcal{G}$ (also known as the *consistency throughout time* assumption), which states that any link $X_{t-k}^i \to X_t^j$ with lag k recurs for all t (Runge, 2018). Importantly, we assume the *causal Markov condition*.

Definition 3.3 (Causal Markov condition)**.** Let $\mathcal{G} = (\mathbf{X} \cup \mathbf{A} \cup \mathbf{R}, \mathbf{E})$ be a full time graph. For node $X_t^i \in \mathbf{X}$, let $\mathbf{nd}_{X_t^i}$ denote its nondescendants and $\mathbf{pa}_{X_t^i}$ its parents. For every $X_t^i \in \mathbf{X}$,

$$X_t^i \perp\!\!\!\perp \mathbf{nd}_{X_t^i} \mid \mathbf{pa}_{X_t^i}. \tag{2}$$

We do not make explicit assumptions over $\mathbf{A}, \mathbf{R}$. Definition 3.3 is important because with this, checking conditional independence between $X_{t_1}^{i_1}, X_{t_2}^{i_2}$ conditioned on $X_{t_3}^{i_3}$ is equivalent to check whether the $X_{t_3}^{i_3}$ blocks all path from $X_{t_1}^{i_1}$ to $X_{t_2}^{i_2}$ in the DAG $\mathcal{G}$ (Pearl, 1999). Though the *time series model with independent noise* (TiMINo) (Peters et al., 2013) is a popular extension of the NPSEM-IE for which causal Markov is known to hold, we find this setting overly restrictive for RL. In many RL applications like healthcare, the noise often has seasonal trends which breaks the i.i.d. assumption. Further, although there always exists an equivalent SCM describing the data when the underlying SCM is an MDP and the value space of $\mathbf{X}_t$ is countable (detailed discussion in Appendix D), this does not always hold when the space is not countable. Therefore, we relax the strong assumption of jointly independent noise to conditional independence as stated below.

Assumption 3.4 (Conditional independence of exogenous variables)**.** Given an SCM denoted by $\mathcal{M}$, we assume the exogenous variables $\mathbf{U}$ over $\mathbf{X}$ satisfy $U_t^i \perp\!\!\!\perp \left\{U_h^j : X_h^j \in \mathbf{nd}_{X_t^i}\right\} \mid \mathbf{pa}_{X_t^i}$.

Assumption 3.4 states that all the influence of past noise to the current node X_t^i are fully mediated through the parents of X_t^i. Assumption 3.4 ensures that the causal markov condition holds. Formally,

Proposition 3.5. *Under Assumption 3.4, for node $X_t^i \in \mathbf{X}$, let $\mathbf{nd}_{X_t^i}$ denote its nondescendants and $\mathbf{pa}_{X_t^i}$ its parents. For every $X_t^i \in \mathbf{X}$, $X_t^i \perp\!\!\!\perp \mathbf{nd}_{X_t^i} \mid \mathbf{pa}_{X_t^i}$.*

Proposition 3.5 can be proven using the same technique as Peters et al. (2013) and we provide proof in Appendix F. When Assumption 3.4 fails, the causal Markov condition might also fail and it becomes unclear whether it is possible to recover the Markov state from the DAG alone. See Appendix E for discussion.

Remark 3.6 (Relaxation of Assumption 3.4). We only need *causal markov condition* for Theorem 4.1-4.2. Though Assumption 3.4 is still a restrictive assumption to ensure causal markov condition, it already relaxes the independent noise assumption used in literature (Peters et al., 2013). In this work, we focus on constructing markovian state from DAG under causal markov condition. Thus, finding the exact necessary and sufficient condition for causal markov condition is important future work.

W$^{\text{th}}$ Order MDP We consider an episodic finite-horizon problem with horizon T. At each time step $t \in \{0, \ldots, T\}$, the learner observes m scalar variables, represented by $\mathbf{X}_t = (X_t^0, \ldots, X_t^m) \in \mathcal{X}$, where $\mathcal{X}$ denotes the observation space. After choosing an action $A_t \in \mathcal{A}$, the learner receives a reward r_t that satisfies $\mathbb{E}[r_t \mid \mathbf{X}_t, A_t] = r(\mathbf{X}_t, A_t)$, where $r(\mathbf{X}_t, A_t)$ is the expected reward function. We say the system is a W^{th} order MDP if, after defining the state

$$Z_t = (\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{\max\{t-W,0\}}), \tag{3}$$

the process $\{Z_t\}_{t=0}^T$ satisfies the Markov property, namely,

$$\mathbb{P}(Z_{t+1} \mid O_0, \ldots, Z_t, A_0, \ldots, A_t) = \mathbb{P}(Z_{t+1} \mid Z_t, A_t). \tag{4}$$

For any policy $\boldsymbol{\pi} = \{\pi_t\}_{t=0}^T$, we define the value function and Q-function at time t by

$$Q_t^{\boldsymbol{\pi}}(z_t, a) \coloneqq \mathbb{E}^{\boldsymbol{\pi}}\left[\sum_{h=t}^{T} r(Z_h, A_h) \,|\, Z_t = z_t,\ A_t = a\right], \quad V_t^{\boldsymbol{\pi}}(Z_t) \coloneqq \mathbb{E}^{\boldsymbol{\pi}}\left[\sum_{h=t}^{T} r(Z_h, A_h) \,|\, Z_t = z_t\right].$$

By optimal Bellman equation (Sutton & Barto, 1998), we have $Q_t^*(z_t, a) = r(z_t, a) + \mathbb{E}\left[V_{t+1}^*(Z_{t+1}) \mid Z_t = z_t, A_t = a\right], V_t^*(z_t) = \max_a Q_t^*(z_t, a)$, where $Q_t^*(z_t, a) = \max_{\boldsymbol{\pi}} Q_t^{\boldsymbol{\pi}}(z_t, a)$ and $V_t^*(z_t) = \max_{\boldsymbol{\pi}} V_t^{\boldsymbol{\pi}}(z_t)$. We further have the optimal policy $\boldsymbol{\pi}^*$ is given by $\pi_t^*(z_t) = \arg\max_a Q_t^*(z_t, a)$.

In the online setting, at each time t of episode k, the learner observes $\mathbf{X}_t$ and then chooses an action A_t based on the history $h_t = (\mathbf{X}_0, \ldots, \mathbf{X}_t,\ A_0, \ldots, A_{t-1},\ R_0, \ldots, R_{t-1})$. Then the learner observes a reward $r_t^{(k)}$. Over K episodes, the goal is to minimize the regret of online policy $\boldsymbol{\pi}^{(k)}$: $\mathbb{E}\left[\sum_{k=1}^K V_0^*(z_0) - V_0^{\boldsymbol{\pi}^{(k)}}(z_0)\right]$. However, directly taking all variables in Z_t as the state can substantially slow learning. In tabular RL, the regret scales polynomially with the number of states (Auer et al., 2008), so when the state is formed by stacking all m coordinates across $W+1$ time steps, the effective state space can grow rapidly with both m and W.

We therefore seek a lower-dimensional state process that can preserve the Markov property and optimality. This leads to the following definition of a *valid state*.

Definition 3.7 (Valid States)**.** Consider a W^{th} order MDP, and let Z_t be the stacked observations from window $t-W$ to t, as defined in Eq. (3). A sequence of random variable sets $\{S_t\}_{t=0}^T$ is called a *valid state* if, for each t, S_t is a subset of the observed variables contained in $\{\mathbf{X}_i\}_{i\in[t]}$ and the following two conditions hold:

1. **Markov Property.** For every t and any sequence of actions, $\mathbb{P}\big(S_{t+1} \mid S_t, A_t, \ldots, S_0, A_0\big) = \mathbb{P}\big(S_{t+1} \mid S_t, A_t\big)$.
2. **Optimality preservation.** For every t and every realization $s_t \in Z_t, s_t \in S_t, z_t \in Z_t$, $\arg\max_a Q_t^*(s_t, a) = \arg\max_a Q_t^*(z_t, a)$.

4 Minimal Markovian State & Multi-Order State Exposure

Section 4.1 presents the minimal Markovian state based on a known causal DAG. Section 4.2 introduces *Multi-Order State Exposure* (MOSE), which constructs the state that can be fed to *any deep RL algorithm* as input when the DAG is unknown.

4.1 From DAG to MDP: Minimal Markovian State Construction

We introduce a graphical criterion for constructing valid states (Definition 3.7) in Theorem 4.1, and then show that Algorithm G.1 constructs a minimal set satisfying this criterion (Theorem 4.2, Lemma G.1). Complete proofs are provided in Appendix G.

Theorem 4.1 (Graphical criterion for valid state construction)**.** *Let $\mathcal{G}$ be a full time graph as defined in Definition 3.2. For each t, we are given $S_t \subset \cup_{j\in[t]} O_j$, which is a subset of state variables observed up to the current time. Let $\mathbf{A}_t \coloneqq \{A_j\}_{j\in[t]}$ be the set of actions taken so far. We say $\{S_t\}_{t\in[T]}$ are valid states (Definition 3.7) for all possible DAG if and only if the following conditions hold:*

1. ***Reward parent inclusion.*** *For every t, the parents of the reward R_t (as defined in the graph $\mathcal{G}$) are contained in S_t: $\mathbf{pa}_{R_t} \setminus \mathbf{A}_t \subseteq S_t$.*
2. ***Inclusion of extended next state parents.*** *For every t, let $\mathcal{C}_{t+1}^*(S_{t+1})$ be the closure of S_{t+1} under same-time parents, constructed iteratively as follows (by replacing S_{t+1} with Z):*

$$\mathcal{C}_{t+1}^0(Z) = Z,\ \mathcal{C}_{t+1}^{k+1}(Z) = \mathcal{C}_{t+1}^k(Z) \cup \Big(\bigcup_{Y\in\mathcal{C}_{t+1}^k(Z)} \big(\mathbf{pa}_Y \cap \mathbf{X}_{t+1}\big)\Big),\ \mathcal{C}_{t+1}^*(Z) = \bigcup_{k\geq 0} \mathcal{C}_{t+1}^k(Z).$$

We require that every parent *of any $X \in \mathcal{C}_{t+1}^*(S_{t+1} \setminus S_t)$ that lies in the past (times $\leq t$) and is not an action must be included in S_t: $\bigcup_{X\in\mathcal{C}_{t+1}^*(S_{t+1}\setminus S_t)} \big(\mathbf{pa}_X \setminus (\mathbf{X}_{t+1} \cup \mathbf{A}_t)\big) \subseteq S_t$.*

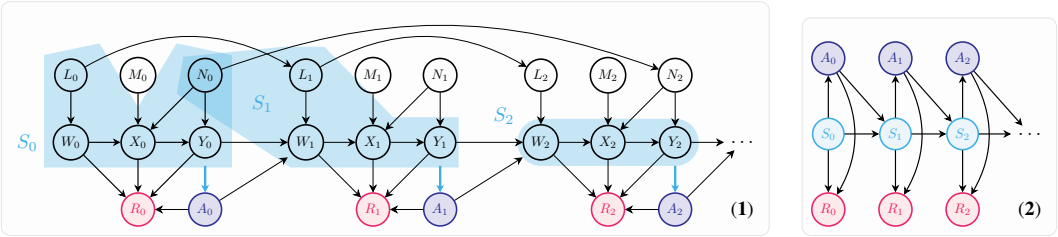


Figure 1: From DAG to MDP. (**1**) Time-series causal DAG shown at time $t \in [0, 2]$, where S_0, S_1, and S_2 are valid states selected by Algorithm G.1. (**2**) Causal DAG representation of the corresponding MDP (Bareinboim et al., 2025), a subgraph of (**1**).

3. ***Retension of Extended Neighbours*** *For any t, let $N_{t+1} := \{X_h^i : X_h^i \in S_{t+1} \setminus S_t, h \leq t\}$ represent the variables that is new in the next time state whose index is smaller than t. For any $s \leq t$, let $A_{t,s}$ to be the extended parents of variables indexed by s that are in historic state, i.e. $A_{t,s} := C_s^* \left(\bigcup_{h \leq t+1} S_h \cap \boldsymbol{X}_s\right)$. Then let M_t be the moral graph for $\mathcal{G}$ constrained on variables X_h^i in $A_{t,s}$ and $\boldsymbol{pa}_{X_h^i}$ that are not in $\cup_{h \leq t} S_h$. For every maximal clique K of M_t such that $K \cap N_{t+1} \neq \emptyset$, the following must hold: $\bigcup_{X_h^i \in A_{t,s}, \{X_h^i\} \cup \boldsymbol{pa}_{X_h^i} \cap K \neq \emptyset} \left(\{X_h^i\} \cup \boldsymbol{pa}_{X_h^i}\right) \bigcap \left(\bigcup_{h \leq t} S_h\right) \subset S_t$.*
4. ***Persistence of included variables.*** *If a variable appears in the state at two different times, it must also appear at all intermediate times: $\forall\, t_1 \leq t_2 \leq t_3,\ S_{t_1} \cap S_{t_3} \subseteq S_{t_2}$.*

Condition 2 ensures that S_t captures every variable at times $\leq t$ that directly influences the extended next-state set $\mathcal{C}_{t+1}^*(S_{t+1})$. Any causal path from the earlier history to (S_{t+1}) must pass through one of these parents. Condition 3 guarantees that markov blanket of S_{t+1} when restricted on vertices indexed before t is included in S_t. Condition 4 ensures that when some variable is present in the state of two time steps, this will not cause the Markov property to fail because that variable will be in the state of every time step in between.

In this work, we assume for simplicity, the action A_t can only affect $\mathbf{X}_{t+1}$. When this assumption does not hold, we need to modify Theorem 4.1 to also include the parents of actions to the state to avoid edge cases like $X_{t-2}^i \to A_{t-2} \to X_t^i$. We note that for a W^{th} order MDP, stacking all variables within the window length W satisfies all four criteria and is a valid state.

However, a state construction $\{S_t\}_{t \in [T]}$ that satisfies the criteria under Theorem 4.1 may contain redundant variables that are not needed to maintain validity. To remove redundant variables, we propose a backward construction procedure in Algorithm G.1. Algorithm G.1 constructs the state sets $\{S_t\}_{t \in [T]}$ backwards, starting from the empty set at time $T + 1$. At step t, we are given S_{t+1} and an auxiliary set $\mathcal{C}_{t+1}$; together, $S_{t+1} \cup \mathcal{C}_{t+1}$ forms the extended next-state set $\mathcal{C}_{t+1}^*(S_{t+1})$ required by Condition (2). We construct S_t as follows. First, any variable in S_{t+1} with time index $\leq t$ is included in S_t. This avoids recursively expanding their own parents, which would needlessly increase the state size. Second, to enforce Condition 2, S_t is augmented with all all parents of the extended next-state set $S_{t+1} \cup \mathcal{C}_{t+1}$ that lie at times $\leq t$, together with parents of the reward $\mathbf{pa}_{R_t}$. To illustrate the construction, we provide an example in Figure 1. In Theorem 4.2, we prove that Algorithm G.1 returns a *minimal* set satisfying Theorem 4.1 such that removal of any variable from $\mathbf{S}$ is impossible.

Theorem 4.2 (Minimality of the state space returned by Algorithm G.1)**.** *Given a full time DAG $\mathcal{G}$, let $\{S_t\}_{t=0}^T$ be the state sequence returned by Algorithm G.1. Then $\{S_t\}_{t=0}^T$ is valid (satisfies Definition 3.7). Moreover, it is* minimal *in the following sense: for any sequence $\{\widetilde{S}_t\}_{t=0}^T$ such that $\widetilde{S}_t \subseteq S_t,\ \forall t,$ and $\widetilde{S}_{t_0} \subsetneq S_{t_0}$ for at least one time step t_0. There exists a transition probability model admissible by the DAG such that $\{\widetilde{S}_t\}_{t=0}^T$ is invalid.*

The proof of Theorem 4.2 proceeds in two steps 1) verifying the criterion in Theorem 4.1 are satisfied naturally by construction and 2) to preserve optimality, the causal parents of the parents of reward has to be involved in the state as established by Lemma G.1.

To illustrate the scale of reduction of the state space size, we evaluate the number of variables contained in S_t for randomly generated graphs. In Table 1 we report the 95% confidence interval for the average number of variables contained over all

	Nodes Per State
$W = 2$	$[16.77, 16.90]$
$W = 5$	$[28.06, 28.41]$
$W = 10$	$[35.17, 35.93]$

Table 1: Nodes per state (95% CI).

horizons. The confidence interval is calculated based on 100 randomly generated graphs with 10 nodes per time step. The exact generation process is the same as the one shown in Section 5. This illustrates that we can often scale down the state space size relative to the full window state.

4.2 Multi-Order State Exposure

Deep RL is known to suffer from decreasing expressivity and generalizability (Sokar et al., 2023) due to overfitting to early samples (Cobbe et al., 2019). Feeding a neural network the minimal Markovian representation can further worsen this behavior, as it removes many variables from observation that can provide additional information to help the network adapt to new samples. This is consistent with our observations in Figure 4, where the minimal Markovian representation did not improve over window policies when $W = 5$. To confirm this, we plot the difference between training and testing loss when the state is minimal Markovian state. The curve is averaged over H horizons using linear ANM detailed in Appendix H.1. As shown in Figure 2, the DAG state starts with good generalization in the beginning and fluctuates toward negative as more samples come in. This indicates that the model using DAG state overfits to early samples and cannot adjust to later samples, as the testing loss consistently becomes larger. Meanwhile, though the magnitude of stacked window state is smaller than DAG state, it still fluctuates. In contrast, MOSE and Causal MOSE, which will be introduced later, consistently have near-zero loss difference, showing the benefits of controlled redundancy. We include the per-horizon loss curve in Appendix H.3.

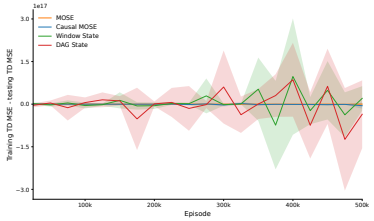


Figure 2: Loss Difference Averaged over Horizons.

To address this issue, we introduce *Multi-Order State Exposure* (MOSE; Algorithm G.2). Based on the idea that selecting features jointly with the model that uses them will improve model performance (Sutton, 2019; Balın et al., 2019), MOSE explicitly adds a selection layer before the final MLP that estimates the value function. As shown in Figure 4.2, MOSE operates on a history that is either the raw observation or the belief history learned by a sequential architecture like LSTM. We first embed each dimension into a higher-dimensional space by $E_t^i = \tanh\left(W^i X_t^i + b_t^i\right)$. We use a separate weight W for different dimensions of X (Gorishniy et al., 2021) because each dimension of X describes a different aspect of the environment. We let the same dimension of X across different time steps share the same weight because they should lie in the same direction, with the effect of time introduced by a bias term. To choose which X_h^i in the history Z_t should be part of the state, we need to calculate the importance of this dimension to estimating value functions. Attention provides a natural way to parameterize this dependence on the value-estimation task (Vaswani et al., 2017). Accordingly, we introduce a single learned query q_0, shared across candidate coordinates and trained jointly with the Q-network. Its association with value estimation comes from the learning objective: gradients from value prediction determine which information the query learns to retrieve.

The selector uses this query to obtain a context for evaluating the current history. Let i index the available variable–lag pairs, i.e., $i = (h, j)$ for $t - W \le h \le t, 0 \le j \le m$, and let $k_{t,i}, v_{t,i} \in \mathbb{R}^d$ denote the key and value obtained from the embedding of coordinate i at time t. The attention read is

$$\alpha_{t,i} = \frac{\exp\left(q_0^\top k_{t,i}/\sqrt{d}\right)}{\sum_j \exp\left(q_0^\top k_{t,j}/\sqrt{d}\right)}, \qquad c_t = \sum_i \alpha_{t,i} v_{t,i}. \tag{5}$$

We then use c_t to score the same candidate coordinates by $r_{t,i} = \frac{c_t^\top k_{t,i}}{\sqrt{d}}$. This second step allows selection to reflect the information retrieved from the current history while retaining a single independently learned query. The selection layer converts these scores by comparing the sigmoid of score and a learnable threshold τ. We learn τ using a straight-through estimator (Bengio et al., 2013) with hard gates $\mathbf{1}\{\sigma(z) \ge \tau\}$ in the forward pass and the differentiable surrogate $\sigma(\sigma(z) - \tau)$ in the backward pass. When $\sigma(r_{t,i}) \ge \tau$, then the corresponding X_h^j will be retained. Otherwise the value will be replaced by 0. The Q-network receives the selected original observations together with their mask. Preserving the observations retains their numerical values, while including the mask distinguishes an excluded coordinate from an observed zero. This mask can be seen as an artificial missingness because the corresponding element is not helping the value function estimation. Therefore, we also include the mask as input to the value function following the common structure when dealing with missingness in MLP (Le Morvan et al., 2020).

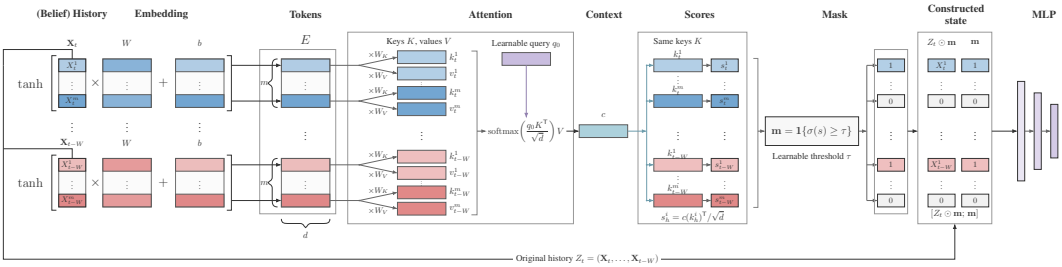


Figure 3: **MOSE architecture.** A learned query q_0 attends to embedded history coordinates to form a context for assessing their relevance to Q-value estimation. This context scores each coordinate through its key to determine a selection mask. The masked observations and mask form the state passed to the MLP Q-network.

We initialize the threshold so that all available coordinates are selected. The selector and Q-network are then trained jointly using the temporal-difference objective. Starting with full selection gives the network access to potentially useful redundancy before learning which inputs it can dispense with. Because the gates remain trainable and the selector continues to inspect the available history, these decisions can be revised as learning proceeds. The aim is thus to learn a compact input representation suited to the Q-network, with room to retain redundant coordinates when they make value estimation easier. We have done ablations to replace the attention mechanism with a simple weight to directly represent the score and the results are deferred to Appendix H.2. When DAG information is available, we use the minimal Markovian state to specify inputs that MOSE must retain. We initialize the corresponding input gates to 1 and keep them fixed at 1 throughout training. The remaining gates are learned as before. Therefore the DAG state determines which inputs must remain available, and end-to-end selection determines which additional inputs are useful. To further emphasize DAG state, we additionally add an gradient update on the Q network using only DAG state as input (Algorithm G.3).

5 Experiments

We tested the benefits of MOSE (Algorithm G.2) and Causal MOSE (Algorithm G.3) compared with naively stacking Window State: Defining all observations within a time window w as state (Mnih et al., 2015; Hessel et al., 2018) and DAG-State: minimal valid state representation constructed by Algorithm G.1. We evaluate the performance on synthetic data generating processes (DGPs) and Atari games (Machado et al., 2018). We use a variant of Double Q Learning (van Hasselt et al., 2015) as the RL algorithm for synthetic environments and R2D2 (Kapturowski et al., 2018) for Atari games.

Synthetic DGPs We first randomly generate a graph with 10 nodes per time step and the causal parents of each node X_t^i can only lie in $\{t-W, t-1, t\}$ where W denotes the order of MDP and can take value in $\{5, 10\}$ in our experiment. Given the causal graph, data were sampled from additive noise models (ANMs) (Hoyer et al., 2008; Peters et al., 2014). We describe the details of the DGPs in Appendix H.1. We compare two length of horizon per episode as 25 and 30 and we feed our state to a variant of Double Q Learning (van Hasselt et al., 2015) for episodic finite RL. Linear ANM (14) are included in main body and nonlinear ANM results are deferred to Appendix H.

Atari Games We additionally evaluate MOSE on the Atari game. In this environment, episode lengths are policy-dependent, since an episode terminates once the agent loses all lives. For this reason, we formulate the problem as an infinite-horizon discounted W^{th} order stationary MDP; the formal definition is deferred to Appendix C. As the underlying RL backbone, we use R2D2 (Kapturowski et al., 2018), a strong value-based deep RL algorithm. We compare MOSE against vanilla R2D2, where vanilla R2D2 directly uses the hidden states learned by LSTM as its input state. We claim that hidden states at current time might not be sufficient and it is beneficial to choose which dimensions of hidden states are enough. Therefore, we use MOSE to operate on the belief history learned by LSTM. Since the ground truth causal graph of the game environment is unavailable, we do not include Causal-MOSE in this experiment. For both methods, we train the agent for 2 million environment interactions. Every 50,000 environment steps, we evaluate the current policy over 10 episodes and report the average episodic reward across these evaluation episodes.

5.1 Experimental Results

Results Using Eq. (14) Figure 4 shows a clear and consistent pattern across all three choices of the order W: MOSE outperforms all competing benchmarks throughout the full training horizon

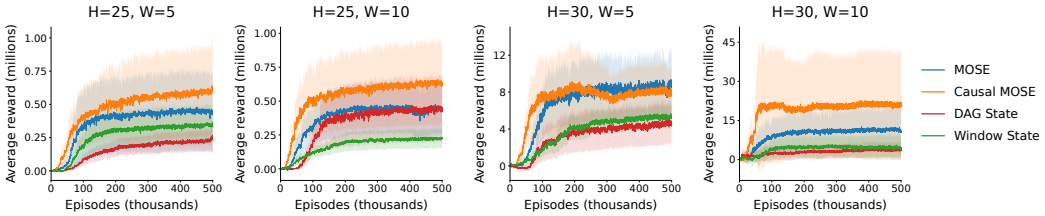


Figure 4: Average per-episode reward averaged over 18 instances. From left to right: 5^{th}-order MDP, Horizon 25, 10^{th}-order MDP, Horizon 25, 5^{th}-order MDP, Horizon 30,10^{th}-order MDP, Horizon 25. Linear ANM (Eq. (14))

of 500K episodes for Linear ANM. The fact that this advantage appears for every tested W and H suggests that the gain from MOSE is not tied to a particular order and horizon length. By selectively constructing the state representation, MOSE reduces the burden of learning from irrelevant or weakly informative variables for Q-function, leading to better performance across training.

A particularly notable observation from Figure 4 is that the benchmark based only on the minimal Markovian state does *not* improve performance, and in fact remains worse to window policies across 3 (W, H) of 4 experiments. This is an important finding. From a theoretical perspective, the minimal Markovian state is sufficient to preserve the optimal control problem. However, the figure suggests that such minimality alone does not necessarily translate into the best empirical performance when the value function is approximated by a neural network. This indicates that although some variables are redundant for satisfying the Markov property, they may still contain auxiliary statistical information that helps the network estimate the Q-function more accurately in finite samples. Removing all such variables may therefore make the representation too sparse from the standpoint of function approximation, even if it remains sufficient in principle for optimal control.

This interpretation is further supported by the comparison between MOSE and Causal MOSE in Figure 4. When the minimal Markovian state is provided as prior structural information, Causal MOSE consistently outperforms MOSE over the entire 500K-episode training period. This shows that by fixing DAG state to be always input to the model, Causal MOSE can selectively choose which redundant information is needed for statistical efficiency and build state around DAG state. Therefore, the best performance is achieved by combining a principled Markovian core with a mechanism that can incorporate extra informative variables.

Finally, we plot the false negative and positive rate of the state constructed by MOSE compared with DAG state to check what state MOSE is trying to construct. As we can see from Figure H.5 the false positive rate of both MOSE and Causal MOSE is dropping and Causal MOSE dropped faster than MOSE. This shows that MOSE are slowly learning the DAG state and with DAG state hard encoded, Causal MOSE can drop truly redundant information more quickly. Though the false negative rate of MOSE is increasing, this is due to under finite horizon, MOSE may falsely drop important information and need more samples to collect back. Detailed discussions are in Appendix H.5.

Results on Atari Games Table 2 shows that MOSE outperforms vanilla R2D2 after 2M training samples, improving mean scores by **25.8% on BattleZone** and **34.7% on Krull**. Specific learning curves are deferred to Appendix H.6. A possible explanation is that learned belief states are not necessarily Markovian: the LSTM may forget information relevant to control, making past belief states useful for current Q-value estimation. MOSE selects relevant components across the belief-state history, providing access to this information while avoiding the high-dimensional input produced by concatenating full past belief states. Furthermore, dimension-wise selection allows gradients to propagate back to the LSTM, guiding it toward learning belief representations that better retain control-relevant information.

Game	R2D2	MOSE
BattleZone	$12{,}435.0 \pm 1{,}009.4$	$15{,}640.0 \pm 5{,}409.6$ **(+25.8%)**
Krull	$5{,}556.9 \pm 888.6$	$7{,}483.2 \pm 77.7$ **(+34.7%)**

Table 2: Final Average results over past 10 evaluations after 2M training samples.

AI USE STATEMENT

In this work, we used generative AI tools for 1) Generating synthetic data sets; 2) implementing methods. We have not used generative AI tools for 1) helping develop theoretical models or conceptual frameworks; 2) formulating mathematical claims; 3) providing critical ingredients for proving mathematical claims and assisting in the writing of proofs; 4) proposing or refine hypotheses; 5) designing or provide feedback on research methodology or experiments; 6) supporting qualitative and thematic data analysis; 7) interpreting results and assist with translation, clean and reformat dataset are not applicable to this work.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International conference on machine learning*, pp. 10–19. PMLR, 2018.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

David Andre, Stuart J Russell, et al. State abstraction for programmable reinforcement learning agents. In *Annual AAAI Conference on Artificial Intelligence*, 2002.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International conference on machine learning*, pp. 507–517. PMLR, 2020.

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pp. 444–453. PMLR, 2019.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in neural information processing systems*, 28, 2015.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

Elias Bareinboim, Sanghack Lee, and Junzhe Zhang. An introduction to causal reinforcement learning. *arXiv preprint arXiv:2101.06498*, 2025.

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2022.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bin Chen and Yongmiao Hong. Testing for the markov property in timeseries. *Econometric Theory*, 28(1):130–178, 2012.

Boyuan Chen, Pieter Abbeel, and Deepak Pathak. Unsupervised learning of visual 3d keypoints for control. In *International Conference on Machine Learning*, pp. 1539–1549. PMLR, 2021a.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021b.

Nhat Chung, Taisei Hanyu, Toan Nguyen, Huy Le, Frederick Bumgarner, Duy Minh Ho Nguyen, Khoa Vo, Kashu Yamazaki, Chase Rainwater, Tung Kieu, et al. Rethinking progression of memory state in robotic manipulation: An object-centric perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3407–3415, 2026.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pp. 1282–1289. PMLR, 2019.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, 16, 2010.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

M Fortunato, MG Azar, B Piot, J Menick, M Hessel, I Osband, A Graves, V Mnih, R Munos, D Hassabis, O Pietquin, and S Blundell, C Legg. Noisy networks for exploration. In *International Conference on Learning Representations (ICLR)*, 2018.

Daigo Fujiwara, Kazuki Koyama, Keisuke Kiritoshi, Tomomi Okawachi, Tomonori Izumitani, and Shohei Shimizu. Causal discovery for non-stationary non-linear time series data using just-in-time modeling. In *Conference on Causal Learning and Reasoning*, pp. 880–894. PMLR, 2023.

Daiqi Gao, Hsin-Yu Lai, Predrag Klasnja, and Susan Murphy. Harnessing causality in reinforcement learning with bagged decision times. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378):304–313, 1982.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2023.

Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for iid and time series data. *Transactions on Machine Learning Research*, 2023.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, volume 45, pp. 141, 2015.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

HyeongJoo Hwang, Seokin Seo, Youngsoo Jang, Sungyoon Kim, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Information-theoretic state space model for multi-view reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 14249–14282, 2023.

Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

Mateo Juliani, Mingxuan Li, and Elias Bareinboim. Confounding robust continuous control via automatic reward shaping. *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026.

Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020.

Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*, 2018.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pp. 5639–5650. PMLR, 2020a.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, 35:34478–34491, 2022.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in Neural Information Processing Systems*, 33: 19884–19895, 2020b.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33:5980–5990, 2020.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4164–4172, 2019.

Timothée Lesort, Natalia Díaz-Rodríguez, Jean-Franois Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Automatic reward shaping from confounded offline data. *arXiv preprint arXiv:2505.11478*, 2025a.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Confounding robust deep reinforcement learning: A causal approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Minne Li, Lisheng Wu, Jun Wang, and Haitham Bou Ammar. Multi-view reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mingzhou Liu, Xinwei Sun, and Yizhou Wang. Conditional local independence testing for it\^ o processes with applications to dynamic causal discovery. *arXiv preprint arXiv:2506.07844*, 2025.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.

Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pp. 141–150. PMLR, 2020.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. *Advances in Neural Information Processing Systems*, 34:24817–24828, 2021.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: evaluation protocols and open problems for general agents (extended abstract). In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 5573–5577, 2018. ISBN 9780999241127.

Ričards Marcinkevičs and Julia E Vogt. Interpretable models for granger causality using self-explaining neural networks. *arXiv preprint arXiv:2101.07600*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

George E Monahan. State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management science*, 28(1):1–16, 1982.

Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2017–2025. PMLR, 2021.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive rl. *arXiv preprint arXiv:2401.08898*, 2024.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International conference on machine learning*, pp. 16828–16847. PMLR, 2022.

James R Norris. *Markov chains*. Cambridge university press, 1998.

Kei Ota, Tomoaki Oiki, Devesh Jha, Toshisada Mariyama, and Daniel Nikovski. Can increasing input dimensionality improve deep reinforcement learning? In *International conference on machine learning*, pp. 7424–7433. PMLR, 2020.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. PMLR, 2020.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Judea Pearl. Probabilities of Causation: Three Counterfactual Interpretations and Their Identification. *Synthese*, 121:93–149, 1999.

Judea Pearl. Causality: Models, reasoning, and inference. *Cambridge, UK: Cambridge University Press*, 19(2):3, 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7), 2018.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 1388–1397, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pp. 3057–3066. PMLR, 2017.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Chengchun Shi, Runzhe Wan, Rui Song, Wenbin Lu, and Ling Leng. Does the markov decision process fit the data: Testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pp. 8807–8817. PMLR, 2020.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 32145–32168. PMLR, 2023.

Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, et al. Robust reward modeling via causal rubrics. In *ICML 2025 Workshop on Models of Human Feedback for AI Alignment*, 2025.

Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In *International conference on machine learning*, pp. 9870–9879. PMLR, 2021.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. In *International Conference on Artificial Intelligence and Statistics*, pp. 1942–1964. PMLR, 2023.

Richard Sutton. The bitter lesson. 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022.

Geraud Nangue Tasse, Matthew Riemer, Benjamin Rosman, and Tim Klinger. Finding the framestack: Learning what to remember for non-markovian reinforcement learning. In *Finding the Frame Workshop at RLC 2025*, 2025.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning, 2015. URL https://arxiv.org/abs/1509.06461.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.

Linbo Wang, Thomas Richardson, and James Robins. Causal inference: A tale of three frameworks. *arXiv preprint arXiv:2511.21516*, 2025.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. In *International Conference on Machine Learning*, pp. 23151–23180. PMLR, 2022.

Zizhao Wang, Caroline Wang, Xuesu Xiao, Yuke Zhu, and Peter Stone. Building minimal and reusable causal state abstractions for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15778–15786, 2024.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations (ICLR)*, 2021.

Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. Deep reinforcement learning with relational inductive biases. In *International conference on learning representations*, 2019.

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

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017.

Junzhe Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*. PMLR, 2020.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: a general setting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 60057–60075, 2024.

Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In *international conference on machine learning*, pp. 27042–27059. PMLR, 2022.

Yunzhe Zhou, Chengchun Shi, Lexin Li, and Qiwei Yao. Testing for the markov property in time series via deep conditional generative learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1204–1222, 2023.

Lutong Zou, Ziping Xu, Daiqi Gao, and Susan Murphy. Causal directed acyclic graph-informed reward design. In *The Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2025.

Appendix

Table of Contents

A Limitations **17**

B Extended Related Works **17**

C Extended Preliminaries **18**

D Validity of I.I.D Noise **18**

E When is the Causal DAG Helpful? **18**

F Proof of Proposition 3.5 **18**

G Technical Details in Section 4.1 **19**
- G.1 Proof of Theorem 4.1 . . . 19
- G.2 Proof of Theorem 4.2 . . . 21
- G.3 Proof of Lemma G.1 . . . 22

H Experimental Details **22**
- H.1 Synthetic DGP Details . . . 22
- H.2 Ablations on Attention . . . 24
- H.3 Per Horizon Loss Difference Curve . . . 24
- H.4 Nonlinear ANM (Eq. 15) . . . 25
- H.5 What Is MOSE Learning? . . . 25
- H.6 Learning Curve of Atari Games . . . 29
- H.7 Choices of Hyperparameters . . . 29

A Limitations

In this work, we established a formal construction of minimal markovian state and design MOSE and Causal MOSE that can leverage the minimality of the state we constructed in deep RL. Several limitations remain. First, our approach assumes a known full time graph; learning such a DAG from high-dimensional longitudinal data, particularly with temporally correlated noise, remains an open challenge. Though state constructed by MOSE may be converging to minimal markovian state, its convergence speed has not been established and would be interesting future work. Second, integrating our framework with other causal RL techniques could further improve performance. These would be interesting future research directions.

B Extended Related Works

Longitudinal Causal Discovery Despite extensive research addressing causal discovery in time series (Assaad et al., 2022; Hasan et al., 2023), longitudinal causal discovery in RL remains an open problem. Existing longitudinal discovery approaches include Granger causality (Granger, 1969; Geweke, 1982), neural Granger causality (Marcinkevičs & Vogt, 2021; Tank et al., 2022; Löwe et al., 2022), methods that assume functional causal models (Fujiwara et al., 2023), constraint-based algorithms (Entner & Hoyer, 2010; Runge et al., 2019; Runge, 2020; Günther et al., 2023), and other deep learning approaches (Pamfil et al., 2020; Bellot et al., 2022; Sun et al., 2023). These methods typically make stationarity assumptions and sometimes deploy classical CITs that assume i.i.d. data. For RL, two challenges arise: (1) in many applications, the noise associated with a given variable is

correlated across time, and hence no longer i.i.d.; (2) data are generated under a behavioral policy, potentially introducing distributional shifts and selection bias. Despite recent advances in CITs for stochastic processes (Liu et al., 2025), Markovian tests, and causal RL (Bareinboim et al., 2025), open questions surround the adaptation of CITs to policy-intervened settings and, more generally, designing discovery algorithms that are robust in RL contexts. In this paper, we assume access to a ground truth longitudinal causal graph and explore (1) how it can be used to build a Markovian state and (2) whether that representation can be effectively leveraged by modern deep RL algorithms.

Causal Bandits and Causal Representation Learning Causal bandits (Lattimore et al., 2016; Bareinboim et al., 2015; Zhang & Bareinboim, 2017) focus on regret minimization in bandits, where actions causally affect the rewards via a (often known) causal graph (Sen et al., 2017; Lee & Bareinboim, 2018; 2019; Lu et al., 2020; Nair et al., 2021; Lu et al., 2021), and the environment resets after each round. Causal representation learning (Schölkopf et al., 2021; Zhang et al., 2024; Ahuja et al., 2023; von Kügelgen et al., 2023; Varici et al., 2024) aims to recover latent causal variables and their structural relationships from high-dimensional variables. Both address distinct settings orthogonal to the objective of this paper.

C Extended Preliminaries

Infinite Horizon Discounted Reward W^{th} Order MDP The only difference compared with episodic finite horizon W^{th} order MDP lies in the objective. In infinite horizon discounted reward setting, we have an discount factor γ and the objective is maximizing

$$\mathbb{E}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(Z_h, A_h) \mid Z_t = z_t, A_t = a\right]$$

D Validity of I.I.D Noise

In this section, we establish that when the space of X_t^i is countable, then if the SCM is describing an MDP, there exists a function f_t^i such that $x_t^i = f_t^i(\mathbf{pa}_{X_t^i}, u_t^i)$ where u_t^i are independent uniform $[0, 1]$. Without loss of generality, we assume there exists only one node per time step and the underlying MDP is 0-th order MDP, i.e. $\mathbb{P}(x_t|x_{t-1}, \ldots, x_0, a_{t-1}) = \mathbb{P}(x_t|x_{t-1}, a_{t-1})$. As stated in Norris (1998), for any markov chain with countable state space, there exists an representation such that $x_t = f(x_{t-1}, a_{t-1}, u_t)$ where U_t are i.i.d uniform random variable. Therefore, suppose the data forms an MDP, there must exist an causal function such that the noise is independent for countable state space.

E When is the Causal DAG Helpful?

As discussed in Section 3 and shown in Theorem 4.1, given DAG that satisfies Definition 3.3 we can construct state such that it defines an MDP. However, when causal Markovian condition does not hold because of dependent noise, we can not construct state directly from DAG alone. For instance, suppose the DAG is just a single chain having $X_0 \to X_1 \to X_2 \to X_3$ but the same noise U controls all four nodes, we can not get that information from DAG and thus do not know whether just taking current observation state satisfies Markov property. We think it is an important future work to study how to construct valid state given DAG and other further information when causal Markovian condition fails. It is also important to study exact conditions on exogenous noise so that an SCM can admit an MDP.

F Proof of Proposition 3.5

Before we prove Theorem 3.5, we introduce the notion of topological ordering which resorts the $\{\mathbf{X}_t\}_{t=0}^T$ so that the parents of $X^{(i)}$ can only be in $X^{(j)}, j < i$. Formally,

Definition F.1 (Topological Ordering)**.** Given a DAG $\mathcal{G} = (\mathbf{X}, E)$, we say a mapping $g : [T] \times [m] \to \{0, 1, \ldots, (T+1)(m+1)\}$ is a linear topological sort iff $\forall X_t^i$, whenever $X_h^j \in \boldsymbol{pa}_{X_t^i}$, $g(h, i) \geq g(t, j)$.

For any X_t^i, assume $\left|\mathbf{nd}_{X_t^i}\right| = n$. Then we can topologically sort these nodes into $X^{(0)}, X^{(1)}, \ldots, X^{(n)}$. Since X_t^i does not have descendants among this set, without loss of generality, assume $X^{(n)} = X_t^i$. Then we have

$$\begin{aligned}
&\mathbb{P}\left(x^{(n)} \mid x^{(0)}, x^{(1)}, x^{(2)}, \ldots, x^{(n-1)}\right) \\
&= \mathbb{P}\left(f^{(n)}\left(\mathbf{pa}_{X^{(n)}}, U^{(n)}\right) \mid \mathbf{pa}_{X^{(n)}}, U^{(0)}, f^{(1)}\left(U^{(0)}, U^{(1)}\right), \ldots, f^{(n-1)}\left(U^{(0)}, \ldots, U^{(n-1)}\right)\right) \\
&= \mathbb{P}(x^{(n)} \mid \mathbf{pa}_{X^{(n)}}),
\end{aligned}$$

where the second equality holds because the variables are sorted by topological ordering and the last equality holds because of Assumption 3.4.

G Technical Details in Section 4.1

G.1 Proof of Theorem 4.1

We will first prove the sufficiency. First, we prove that any state space satisfying the criterion satisfies the Markov property. It suffices to show that $\mathbb{P}(C_{t+1}^* = c_{t+1}, L_{t+1} = l_{t+1} \mid A_t = a, S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0) = \mathbb{P}(C_{t+1}^* = c_{t+1} \mid A_t = a, S_t = s_t)$ where $\mathcal{L}_{t+1} = S_{t+1} \cap S_t$. This holds since $S_{t+1} \setminus S_t \subset \mathcal{C}_{t+1}^*$. Since $\mathcal{L}_{t+1} \subset S_t$, we have

$$\begin{aligned}
&\mathbb{P}(C_{t+1}^* = c_{t+1}, L_{t+1} = l_{t+1} \mid A_t = a, S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0) \\
&= \mathbb{I}(l_{t+1} = g(s_t))\mathbb{P}(C_{t+1}^* = c_{t+1} \mid A_t = a, S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0),
\end{aligned}$$

where g is the projection from S_t to $\mathcal{L}_\sqcup$. By construction of $\mathcal{C}_{t+1}^*$ and Criterion (2), for any $X \in \mathcal{C}_{t+1}^*$, X's parent can only be in $\mathcal{C}_{t+1}^* \cup S_t$. Let $X_{t+1}^{(0)}, \ldots, X_{t+1}^{(i)}, \ldots$ be the variables in $\mathcal{C}_{t+1}^*$ that is ordered according to topological ordering. Then we have the parents of X_{t+1}^i can only be $X_{t+1}^{(j)}, j > i$ and S_t. Since by Criterion (4), if $S_{t+1} \cap S_k$ for $k < t$ is non empty, then those value are also contained in S_t, without losing generality, assume $S_{t+1} \cap S_k = \emptyset$ for $k < t$. Then we have

$$\begin{aligned}
&\mathbb{P}(C_{t+1}^* = c_{t+1} \mid A_t = a, S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0) \\
&= \prod_{i=1}^{|\mathcal{C}_{t+1}^*|} \mathbb{P}\left(x_{t+1}^{(i)} \mid x_{t+1}^{(i-1)}, \ldots, x_{t+1}^{(1)}, S_t = s_t, A_t = a, \ldots, S_0 = s_0\right) \\
&= \prod_{i=1}^{|\mathcal{C}_{t+1}^*|} \mathbb{P}\left(x_{t+1}^{(i)} \mid x_{t+1}^{(i-1)}, \ldots, x_{t+1}^{(1)}, S_t = s_t, A_t = a\right),
\end{aligned}$$

where the last equality holds because Criterion (3) ensures no path will point to S_{t+1} without passing S_t.

Then we prove that such state space also preserves the optimal value. We prove by induction. The induction hypothesis is

$$\forall\, t, s_t, z_t, a, \quad Q_t^*(s_t, a) = Q_t^*(z_t, a). \tag{6}$$

When $t = T$, since by Criterion (1), $\mathbf{pa}_{R_T} \subset \mathcal{S}_T$, we have $Q_T^*(s_T, a) = R(\mathbf{pa}_{R_T}, a) = Q_T^*(z_t, a)$. Suppose this holds for $t = h + 1$, when $t = h$, we have

$$\begin{aligned}
Q_h^*(s_h, a) &= R(s_h, a) + \mathbb{E}\left[\max_a Q_{h+1}^*(s_{h+1}, a) \mid S_h = s_h, A_h = a\right] \\
&= R(\mathbf{pa}_{R_h}, a) + \mathbb{E}\left[\max_a Q_{h+1}^*(s_{h+1}, a) \mid S_h = s_h, A_h = a\right].
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
Q_h^*(z_h, a) &= R(z_h, a) + \mathbb{E}\left[\max_a Q_{h+1}^*(z_{h+1}, a) \mid Z_h = z_h, A_h = a\right] \\
&= R(\mathbf{pa}_{R_h}, a) + \mathbb{E}\left[\max_a Q_{h+1}^*(s_{h+1}, a) \mid Z_h = z_h, A_h = a\right] \\
&= R(\mathbf{pa}_{R_h}, a) + \mathbb{E}\left[\max_a Q_{h+1}^*(s_{h+1}, a) \mid S_h = s_h, A_h = a\right].
\end{aligned}$$

where the second equality holds by Eq. (6) and the last equality holds because of Criterion (4) and 2. Therefore, we have $Q_h^*(s_h, a) = Q_h^*(z_h, a)$.

Next, we prove the necessity. To prove Criterion 1,2 are necessary, we introduce Lemma G.1. The proof will be deferred to the end of this section.

Lemma G.1 (Necessity of parent inclusion)**.** *Fix $t \in \{0, \dots, T-1\}$. Suppose that the state variables $\{\mathcal{S}_\tau\}_{\tau=t+1}^T$ have been constructed such that the following holds: there exists a transition probability model consistent with the times series DAG under which, for any $\tau \in \{t+1, \dots, T\}$, removing any variable from $\mathcal{S}_\tau$ strictly reduces the optimal value.*

Then, if there exists a variable $X \in \mathcal{S}_{t+1} \cup \{R_t\}$ such that

$$X \notin \mathcal{S}_t \quad \textit{and} \quad \mathbf{pa}(C_{t+1}^*(X)) \cap \{\textit{variables at time } \le t\} \nsubseteq \mathcal{S}_t,$$

then there exists a transition probability model consistent with the DAG under which the optimal value at time t under $\mathcal{S}_t$ is strictly smaller than that under the full construction.

Lemma G.1 immediately shows that for any valid state, it must satisfy the first two criteria.

Suppose Criterion 4 fails, there exists X_h^i and $t_1 < t_2 < t_3$ such that $X_h^i \in S_{t_1}, S_{t_3}, X_h^i \notin S_{t_2}$. Then we have S_{t_3} is dependent of S_{t_1} given S_{t_2} since even conditioned on S_{t_2}, X_h^i in S_{t_1} determines the value of X_h^i in S_{t_3}. This breaks the Markovian property.

Since Criterion 4 must hold, it implies $N_{t+1} \cap \left(\bigcup_{h=0}^t S_h\right) = \varnothing$. Indeed, if a variable in N_{t+1} had appeared in some S_h with $h < t$, then its appearances in S_h and S_{t+1} would imply its inclusion in S_t by persistence. This contradicts the definition of N_{t+1}.

Suppose Criterion 3 fails, there exists t and $s \le t$, maximal clique K of M_t satisfying $K \cap N_{t+1} \neq \varnothing$, X_s^k satisfying $X_s^k \in A_{t,s}$ and a variable X_r^j such that

$$X_r^j \in \left(\{X_s^k\} \cup \mathbf{pa}(X_s^k)\right) \cap \left[\left(\bigcup_{h=0}^t S_h\right) \setminus S_t\right],$$

and

$$\left(\{X_s^k\} \cup \mathbf{pa}(X_s^k)\right) \cap K \neq \varnothing.$$

Choose $X_h^i \in K \cap N_{t+1}$ and $X_\ell^m \in \left(\{X_s^k\} \cup \mathbf{pa}(X_s^k)\right) \cap K$. Thus X_r^j is a previously included variable that is absent from S_t, and it shares a family with a variable in K.

Consider the subgraph of $\mathcal{G}$ induced by $N_{t+1} \cup \left(\bigcup_{h=0}^t S_h\right)$ and all ancestors of these variables, and form its moral graph. Recall that moralization connects every pair of parents sharing a child and then removes all arrowheads.

Every variable in $A_{t,s}$ belongs to this ancestral subgraph. To see this, the variables used to define $A_{t,s}$ are indexed by $s \le t$ and belong to states up to time $t+1$. Each such variable therefore either belongs to $\bigcup_{h=0}^t S_h$ or belongs to N_{t+1}. The same-time ancestors added by C_s^*, and all their parents, also belong to the ancestral subgraph.

Consequently, every edge of M_t is also present in this ancestral moral graph. Furthermore,

$$\{X_s^k\} \cup \mathbf{pa}(X_s^k)$$

forms a clique there: the child is adjacent to each parent, and every pair of parents is joined by moralization. In particular, the ancestral moral graph contains an edge between X_ℓ^m and X_r^j.

Because K is a clique, it is connected. Hence there is a path inside K from X_h^i to X_ℓ^m; if these variables coincide, this path has length zero. Appending the edge from X_ℓ^m to X_r^j gives a path from X_h^i to X_r^j in the ancestral moral graph.

Every vertex of K lies outside $\bigcup_{h=0}^t S_h$, by the construction of M_t. Therefore, every vertex of the path inside K lies outside S_t. Its final vertex X_r^j also lies outside S_t by its choice. Thus this path avoids S_t entirely.

It follows that S_t does not separate N_{t+1} from

$$\left(\bigcup_{h=0}^{t} S_h\right) \setminus S_t$$

in their ancestral moral graph. By the ancestral moral graph criterion for d-separation,

$$N_{t+1} \not\perp_d \left[\left(\bigcup_{h=0}^{t} S_h\right) \setminus S_t\right] \mid S_t.$$

By completeness of d-separation, there exists a probability model compatible with $\mathcal{G}$ under which

$$N_{t+1} \not\perp\!\!\!\perp \left[\left(\bigcup_{h=0}^{t} S_h\right) \setminus S_t\right] \mid S_t.$$

However, $N_{t+1} \subseteq S_{t+1}$, whereas the Markov property requires

$$S_{t+1} \perp\!\!\!\perp (S_0, \ldots, S_{t-1}) \mid S_t.$$

The latter would imply the conditional independence just shown to fail. This contradicts Markovianity for every compatible model. Therefore, the required inclusion holds.

Algorithm G.1 Backward State-Space Construction from a Time Series DAG

Require: Time horizon T; times series DAG $\mathcal{G}$

1: Initialize $S_{T+1} \leftarrow \emptyset, \mathcal{C}_{T+1} \leftarrow \emptyset$
2: **for** $t = T, T-1, \ldots, 0$ **do**
3: Construct state S_t and auxiliary variable set $\mathcal{C}_t$:

$$S_t \leftarrow \left\{X_s^i : X_s^i \in S_{t+1}, s \le t\right\} \cup \left\{X_s^i : X_s^i \in \mathbf{pa}_{X_{t+1}^j}, X_{t+1}^j \in S_{t+1} \cup \mathcal{C}_{t+1}, s \le t\right\} \cup \mathbf{pa}_{R_t}$$

4: Initialize $\mathcal{C}_t \leftarrow \left\{X_t^i \ : \ X_t^i \in \mathbf{pa}_{X_t^j},\ X_t^j \in S_t,\ X_t^i \notin S_t\right\}$.
5: **while** $\left\{X_t^i \ : \ X_t^i \in \mathbf{pa}_{X_t^j},\ X_t^j \in \mathcal{C}_t,\ X_t^i \notin \mathcal{C}_t\right\} \neq \emptyset$ **do**
6: $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \left\{X_t^i \ : \ X_t^i \in \mathbf{pa}_{X_t^j},\ X_t^j \in \mathcal{C}_t,\ X_t^i \notin \mathcal{C}_t\right\}$
7: **end while**
8: **end for**
9: **Output:** $\{\mathcal{S}_t\}_{t=0}^{T}$

G.2 Proof of Theorem 4.2

We first show that this construction gives a valid state space by verifying the four criteria of Theorem 4.1. Criteria (1) and 2 is satisfied by Line 3 of Algorithm G.1 since $\mathcal{C}_{t+1} \cup \mathcal{S}_{t+1}$ is the same as $\mathcal{C}^*_{t+1}$. For Criterion (3), again by Line 3, we have for all $k \le t_2$ if $X_k^i \in S_{t_1}$, $X_k^i \in S_{t_2}$. Therefore, for every $X_k^i \in S_{t_1} \setminus S_{t_2}$, time index k must be larger than t_2. Since the variables in S_{t_2} can only have ancestors from time smaller or equal to t_2, Criterion (3) is satisfied. Criterion (4) holds because of Line 3 similarly.

We will use induction to show that we can't move any variable from the state constructed by Algorithm G.1. At $t = T$, by Lemma G.1, we must have the parents of reward to preserve optimality, which is exactly the same as the construction of Algorithm G.1. Suppose for $t \ge h+1$, we can't remove any variable from the constructed state. When $t = h$, by Lemma G.1, for any $X \in \mathcal{S}_{h+1}$, either itself or $\mathbf{pa}(C^*_{h+1}(X))$ needs to be in $\mathcal{S}_h$. If X is indexed by $k \le h$, X can be in the state at time h. Since this will only increase the dimension of state space by 1 which is smaller than the cardinality of $\mathbf{pa}(C^*_{h+1}(X))$, we should put X to $\mathcal{S}_t$. If X indexed by $h+1$, then $\mathbf{pa}(C^*_{h+1}(X))$ must be in $\mathcal{S}_t$.

Since this exactly matches the output of Algorithm G.1, every variable retained by the construction is indispensable: deleting any one of them causes failure of at least one of the two desired properties for some transition instance.

Therefore, no strictly smaller state sequence can simultaneously preserve Markovianity and preserve the optimal policy for all transition instances. This proves the minimality of $\{\mathcal{S}_t\}_{t=0}^{T}$.

G.3 Proof of Lemma G.1

It suffices to consider the case where exactly one necessary variable X is omitted from the state at time t. Suppose X is either a causal parent of some variable in S_{t+1} with time index at most t, or a variable in S_{t+1} itself, but $X \notin S_t$. Fix some realization $S_t = s$, and let $X \in \{0, 1\}$ with

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p, \tag{7}$$

for some $p \in (0, 1)$. Consider two actions a_0 and a_1, and define a transition model consistent with the DAG as follows: the next state contains a binary variable $Y_{t+1} \in S_{t+1}$, and

$$\mathbb{P}(Y_{t+1} = 1 \mid S_t = s, X = x, A_t = a) = \mathbb{I}(a = a_x), \tag{8}$$

where $a_x = a_0$ if $x = 0$ and $a_x = a_1$ if $x = 1$. Let the continuation value be

$$V_{t+1}(Y_{t+1}) = \mathbb{I}(Y_{t+1} = 1). \tag{9}$$

Then if X were observed as part of the state at time t, the controller could choose a_x after observing x, and hence attain value

$$\max_a \sum_y \mathbb{P}(Y_{t+1} = y \mid S_t = s, X = x, A_t = a) V_{t+1}(y) = 1 \tag{10}$$

for both $x = 0$ and $x = 1$. In contrast, if X is not included in the state, then the policy must choose the same action for both values of X, so the value is obtained by averaging over the conditional law of the unobserved variable:

$$Q_t(s, a) = \sum_{x \in \{0,1\}} \mathbb{P}(X = x \mid S_t = s) \sum_y \mathbb{P}(Y_{t+1} = y \mid S_t = s, X = x, A_t = a) V_{t+1}(y). \tag{11}$$

By construction,

$$Q_t(s, a_0) = 1 - p, \qquad Q_t(s, a_1) = p, \tag{12}$$

and therefore

$$\max_a Q_t(s, a) = \max\{p, 1 - p\} < 1 \tag{13}$$

whenever $p \in (0, 1)$. Thus, the optimal value under the reduced state representation is strictly smaller than under the full state representation. Moreover, any policy based only on the reduced state $S_t = s$ must choose the same action regardless of the realized value of X, whereas the optimal full-state policy chooses a_0 when $X = 0$ and a_1 when $X = 1$. Hence the policy induced by the reduced state representation is not optimal under the full state representation. This proves that omitting even a single necessary variable can strictly reduce the optimal value.

H Experimental Details

Compute Resources Experiments were run on an AWS EC2 c6a.4xlarge instance with 16 vCPUs.

H.1 Synthetic DGP Details

In the linear additive case, each variable takes the form

$$X_t^i = b^i + \omega_{t,a}^i a + \sum_{x_h^j \in \mathbf{pa}(X_t^i)} \omega_{t,x_{h,j},a}^i x_h^j + \epsilon_t^i, \tag{14}$$

where ω is randomly generated, b^i follows normal distribution with mean 0 and ϵ_t^i is uniform over $[-\sqrt{3}, \sqrt{3}]$. And the reward follows the same DGP.

Additionally, we have implemented a non-linear ANM. In the nonlinear additive case, we standardize the parent values as $z_h^j = (x_h^j - \mu_h^j)/s_h^j$ and $z_a = (a - \mu_a)/s_a$, where μ and $s > 0$ are randomly

Algorithm G.2 Multi-Order State Exposure (MOSE)

Require: **Update-Value-Function**; order W; batch size B; head dimension d

Initialize separate parameters θ_h for each $h = 1, \dots, H$, including the attention module and Q-network
Initialize thresholds $\tau_h = 0$ and replay buffers $\mathcal{D}_h = \emptyset$
for $k = 1, 2, \dots, K$ **do**
  Reset the environment
  for $h = 1, 2, \dots, H$ **do**
    Observe $\mathbf{X}_h^k$ and construct $Z_h^k = (\mathbf{X}_h^k, \dots, \mathbf{X}_{h-W}^k)$
    Construct $\hat{O}_h^k$ using the step-h attention and selection operations defined below
    Take action $a_h^k = \arg\max_a Q_h^{\theta_h}(\hat{O}_h^k, a)$
    Observe r_h^k and O_{h+1}^k
    Store $(Z_h^k, a_h^k, r_h^k, O_{h+1}^k)$ in $\mathcal{D}_h$
  end for
  for $h = H, H-1, \dots, 1$ **do**
    Sample B transitions $\{(Z_h^{k_b}, a_h^{k_b}, r_h^{k_b}, O_{h+1}^{k_b})\}_{b=1}^B$ from $\mathcal{D}_h$
    for $b = 1, \dots, B$ **do**
      Let $\mathbf{A}_h$ denote the availability mask and $\mathcal{I}_h$ the available coordinates
      $E_{h;t,i}^b = \tanh\Big(W_h^i X_t^{k_b,i} + b_{h;h-t,i}\Big)$
      $\mathbf{E}_h^b := (E_{h;t,i}^b)$
      Compute keys and values: $\mathbf{K}_h^b = \mathbf{E}_h^b W_{K,h}$, $\mathbf{V}_h^{b,j} = \mathbf{E}_h^b W_{V,h}$
      Compute attention weights: $\boldsymbol{\alpha}_h^b = \mathrm{softmax}_{\mathcal{I}_h}\Big(\frac{\mathbf{q}_{0,h}(\mathbf{K}_h^b)^\top}{\sqrt{d}}\Big)$
      Compute context: $\mathbf{c}_h^b = \boldsymbol{\alpha}_h^b \mathbf{V}_h^b$
      Compute coordinate scores: $\mathbf{z}_h^b = \dfrac{1}{\sqrt{d}} \mathbf{c}_h^b (\mathbf{K}_h^b)^\top$
      Select coordinates: $\mathbf{m}_h^b = \mathbf{A}_h \odot \mathbf{1}\big\{\sigma(\mathbf{z}_h^b) \geq \tau_h\big\}$
      Construct the exposed state: $\hat{O}_h^{k_b} = [\, Z_h^{k_b} \odot \mathbf{m}_h^b ; \mathbf{m}_h^b \,]$
    end for
    $\theta_h \leftarrow$ **Update-Value-Function**$\Big(\theta_h, \Big\{(\hat{O}_h^{k_b}, a_h^{k_b}, r_h^{k_b}, O_{h+1}^{k_b})\Big\}_{b=1}^B\Big)$
  end for
end for

Algorithm G.3 Causal MOSE

Require: **Update-Value-Function**; W, B, d; minimal Markovian state masks $\{\mathbf{m}_h^\star\}_{h=1}^H$

Initialize as in MOSE, with $\theta_h = (\phi_h, \psi_h)$: ϕ_h contains attention and selection parameters; ψ_h contains Q-network weights
for $k = 1, \dots, K$ **do**
  Collect transitions as in MOSE
  for $h = H, \dots, 1$ **do**
    Sample B transitions $\mathcal{B}_h$ from $\mathcal{D}_h$
    $(\theta_h, \mathbf{y}_h) \leftarrow$ **MOSE**$(h, \theta_h, \mathcal{B}_h)$ {One minibatch update; retain detached TD targets}
    $\psi_h \leftarrow$ **DAG-State-Update**$(\psi_h, \mathcal{B}_h, \mathbf{m}_h^\star, \mathbf{y}_h)$
  end for
end for

generated parameters whose dependence on the target node X_t^i is suppressed. The action modifies each state-parent coefficient as

$$\omega_{t,x_{h,j},a}^i = \omega_{t,x_{h,j}}^i + \Delta_{t,x_{h,j},a}^i.$$

Define $\rho(z) = z/(1+|z|)$ and the weighted nonlinear functions

$$\varphi_1(z,w) = w\tanh(z), \qquad \varphi_2(z,w) = w\max\{0, \mathrm{sgn}(w)z\},$$
$$\varphi_3(z,w) = w[\sin(z+\theta_t^i) + \tfrac{1}{2}\cos(\nu_t^i z)], \quad \varphi_4(z,w) = \rho(wz).$$

Algorithm G.4 DAG-State-Update

Require: Q-network weights ψ_h; batch $\mathcal{B}_h$; minimal-state mask $\mathbf{m}_h^\star$; detached TD targets $\mathbf{y}_h$

1: **for** $b = 1, \dots, B$ **do**
2: $\quad O_h^{\star,k_b} = [\, Z_h^{k_b} \odot \mathbf{m}_h^\star ; \mathbf{m}_h^\star \,]$
3: **end for**
4: Bypass attention and selection; compute

$$\mathcal{L}_h^{\mathrm{DAG}} = \frac{1}{B} \sum_{b=1}^{B} \left(Q_h^{\psi_h}(O_h^{\star,k_b}, a_h^{k_b}) - y_h^b \right)^2$$

5: Take one optimizer step on ψ_h using $\nabla_{\psi_h} \mathcal{L}_h^{\mathrm{DAG}}$; keep ϕ_h fixed
6: **return** ψ_h

Then define

$$f_{m,t,a}^i = \varphi_m(z_a, \omega_{t,a}^i) + \sum_{x_h^j \in \mathbf{pa}_{X_t^i}} \varphi_m(z_h^j, \omega_{t,x_{h,j},a}^i), \qquad m = 1, \dots, 4.$$

Pairwise interactions and a sigmoid gate are defined by

$$f_{5,t,a}^i = \sum_{\{x_h^j, x_u^k\} \subseteq \mathbf{pa}_X(X_t^i)} \eta_{t,h,u}^{i,j,k} \rho(z_h^j z_u^k),$$

$$f_{6,t,a}^i = \sigma_{20}\left(\gamma_{t,a}^i z_a + \sum_{x_h^j \in \mathbf{pa}_{X_t^i}} \gamma_{t,h,j}^i z_h^j \right), \qquad \sigma_{20}(z) = \frac{1}{1 + \exp\{-\operatorname{clip}_{[-20,20]}(z)\}}.$$

The pairwise sum counts each distinct unordered pair once. Each node randomly selects a coefficient vector c_t^i from its eligible mixtures in the fixed library, giving

$$f_{t,a}^i = \sum_{m=1}^{5} c_{m,t}^i f_{m,t,a}^i + c_{6,t}^i f_{6,t,a}^i f_{5,t,a}^i + c_{7,t}^i \rho(f_{5,t,a}^i) + c_{8,t}^i f_{6,t,a}^i f_{4,t,a}^i.$$

Each variable then takes the form

$$X_t^i = \operatorname{clip}_{[-10^6, 10^6]} \left[b^i + g_t^i f_{t,a}^i + \alpha_t^i \sin(\nu_t^i t + \theta_t^i) \right] + \epsilon_t^i. \tag{15}$$

The weights ω, action adjustments Δ, interaction coefficients η, gate coefficients γ, and parameters g, α, ν, θ are randomly generated. The intercept b^i follows a normal distribution with mean zero, and ϵ_t^i is independently drawn from $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$. Reward nodes follow the same construction, with the clipped component divided by 1000 and an action-dependent offset added before the noise.

H.2 Ablations on Attention

We compared MOSE with a version without attention. Specifically, instead of using attention to calculate s and then compare with learnable threshold τ, we directly set s as learnable parameter for each dimension. As shown in Figure H.1, MOSE without attention is worse than MOSE with attention. This indicates that with the learnable query, it can more directly calculates the relationship between the estimation error and each dimension of history.

H.3 Per Horizon Loss Difference Curve

As shown in Figures H.2-H.3, many curves are initially flat and close to zero, indicating little discrepancy between training and testing losses on the early samples. Their subsequent departures from zero are consistent with our interpretation of overfitting to early samples: a model may generalize well within the early sample distribution but become overly adapted to it, limiting its ability to accommodate samples encountered later in training. For Figure H.2, this deterioration is particularly visible for DAG State at $h = 20$–25, whereas MOSE and Causal MOSE remain comparatively close

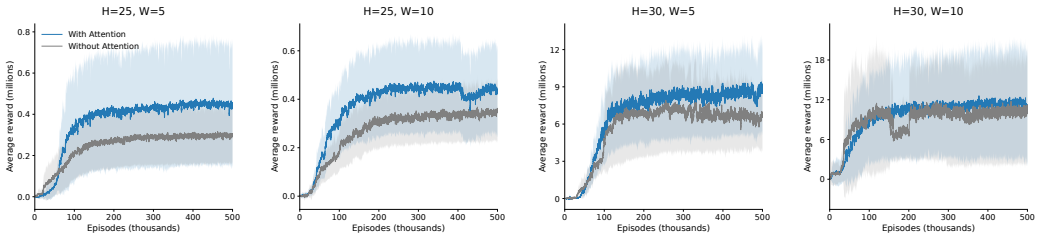


Figure H.1: Performance Comparison Between Using Attention to Score and Directly Learn Score. Averaged over 18 instances.

to zero. Window State also develops substantial fluctuations at $h = 26$ and $h = 28$, suggesting that a fixed history representation can exhibit similar difficulties. For Figure H.3, the contrast is especially pronounced at $h = 22$–25 and $h = 29$–30, where DAG State exhibits large fluctuations while Causal MOSE remains comparatively stable. MOSE also performs well across many of these steps, although it develops late-training deviations at $h = 26$ and $h = 28$. At the earliest horizon steps, the advantage is less consistent, as the MOSE variants also depart from zero later in training. These patterns suggest that learned state selection can help mitigate excessive adaptation to early samples, particularly at intermediate and late horizon steps.

H.4 Nonlinear ANM (Eq. 15)

As shown in the top row of Figure H.4, the closely clustered reward curves and small absolute performance gaps suggest that obtaining a clear advantage in this complex-transition setting is challenging. Nevertheless, MOSE eventually achieves higher mean rewards than both the Window and DAG state benchmarks, as highlighted in the upper-right panel. This result suggests that MOSE can capture useful information for decision making even under complicated transition dynamics. The second row shows that its false-positive rate decreases during training, although its false-negative rate increases. Thus, the performance gains support the usefulness of MOSE's learned representation, while the gate errors indicate that it does not exactly recover the DAG-defined state due to finite samples.

H.5 What Is MOSE Learning?

As shown in the Figure H.5 and bottom row of Figure H.4, MOSE progressively eliminates some historical coordinates and retains a smaller subset, as indicated by its decreasing false-positive rate. However, its increasing false-negative rate raises the question of whether this selection recovers a minimal Markovian state. First, the minimal state constructed by Algorithm G.1 does not imply uniqueness without additional assumptions. For a simple counterexample, let X_t be a nondegenerate binary Markov chain, observe both X_t and $Y_t = 1 - X_t$, and define the reward as $R_t = \mathbf{1}\{A_t = X_t\}$. Both $\{X_t\}$ and $\{Y_t\}$ are minimal Markovian states: either determines the transition and reward distributions, whereas the empty state cannot preserve the optimal action. Thus, omitting a coordinate from one particular minimal state does not necessarily imply losing essential information.

Second, Causal MOSE consistently outperforms MOSE in the available state-selection comparisons and eliminates unnecessary coordinates more rapidly, supporting the usefulness of the DAG state constructed by our algorithm in guiding selection. In these runs, DAG coordinates are retained by construction and this results in faster reduction in false positives, demonstrating improved elimination of additional information by having minimal markovian state information. Together, these observations are consistent with finite-sample estimation and optimization errors causing MOSE to discard coordinates belonging to the algorithm's minimal Markovian state. They motivate the conjecture that, with sufficient data and suitable learning conditions, MOSE may recover the DAG state or an informationally equivalent minimal state asymptotically.

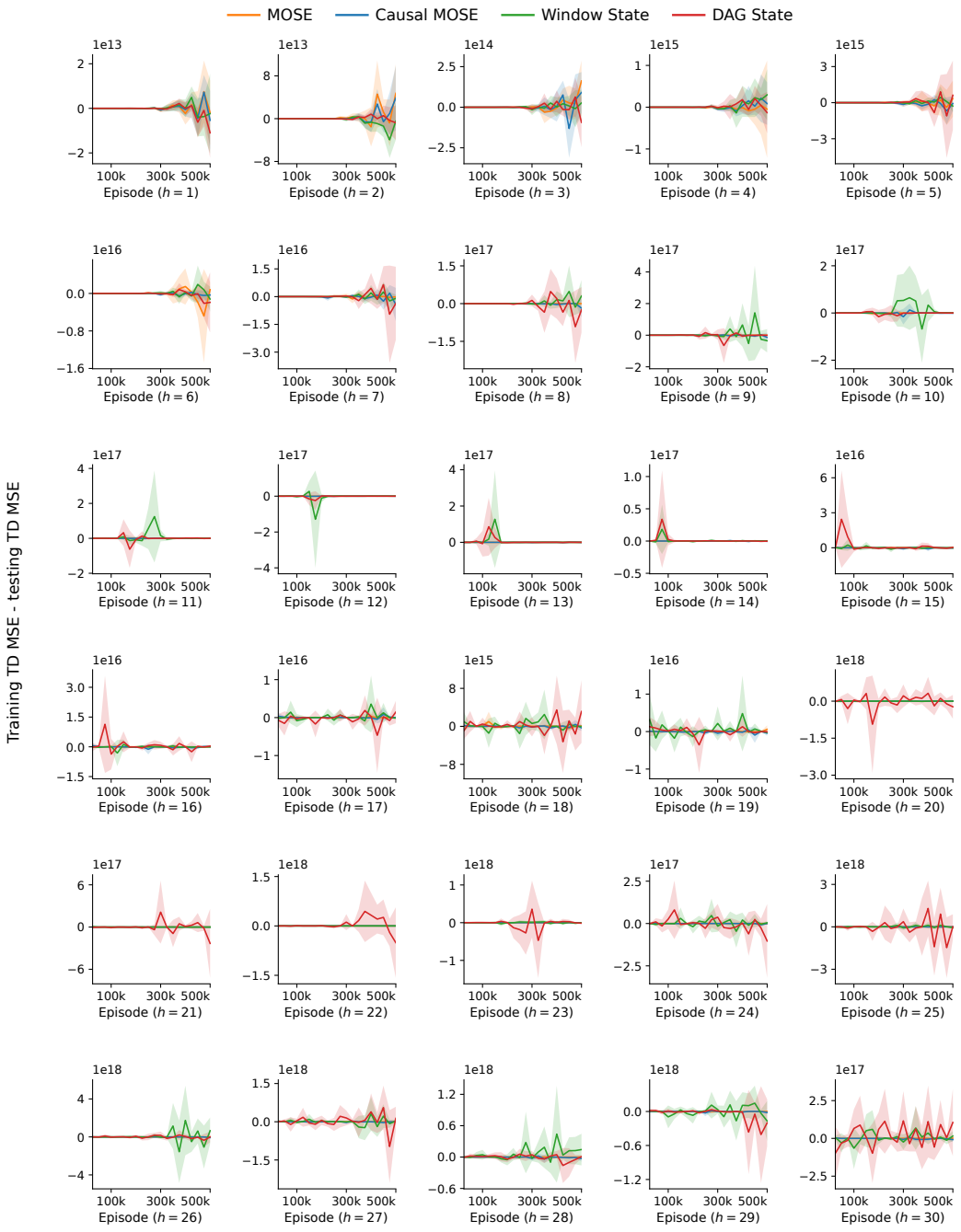


Figure H.2: Difference of Training and Testing Loss for Horizon 30 and $W = 5$

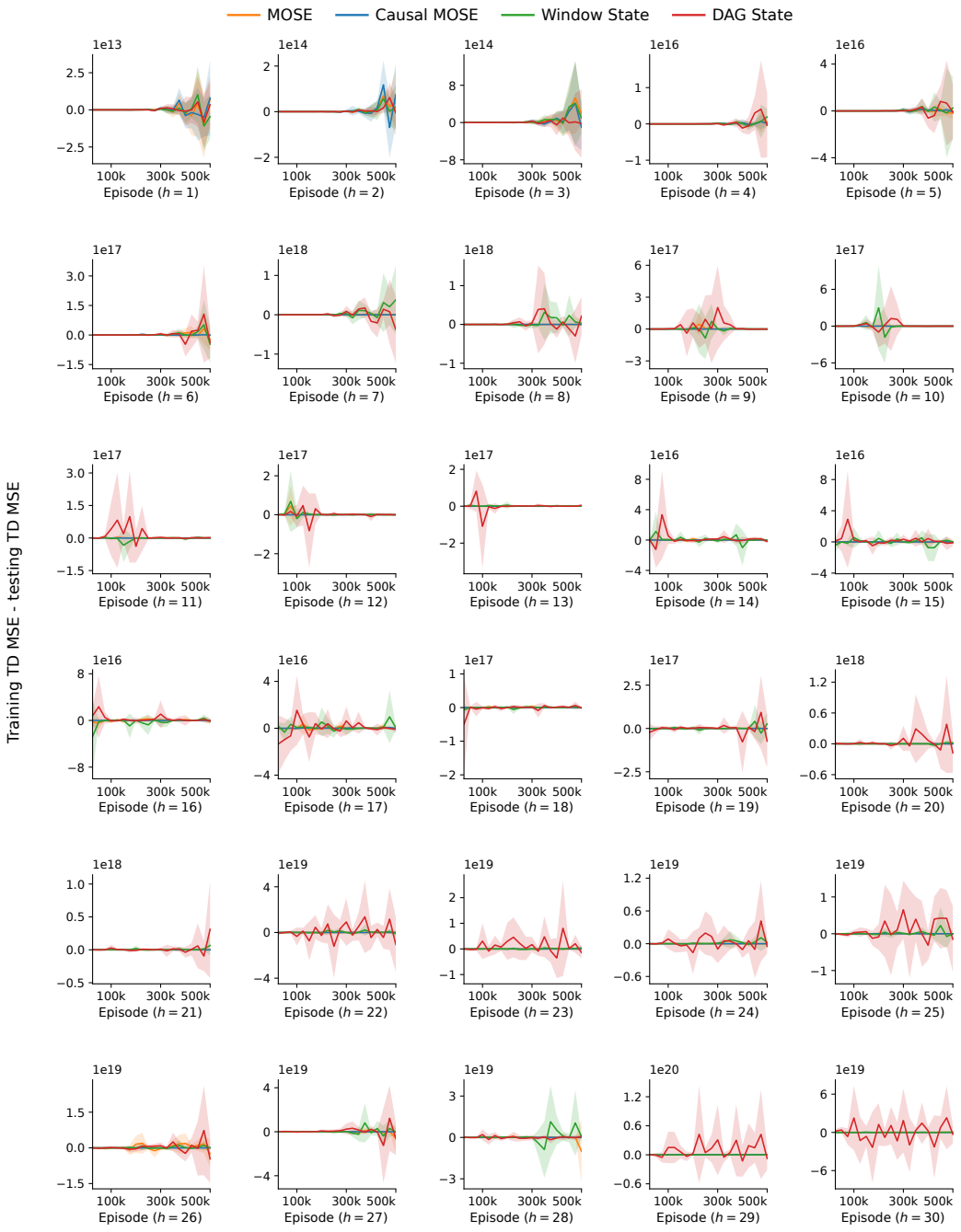


Figure H.3: Difference of Training and Testing Loss for Horizon 30 and $W = 10$

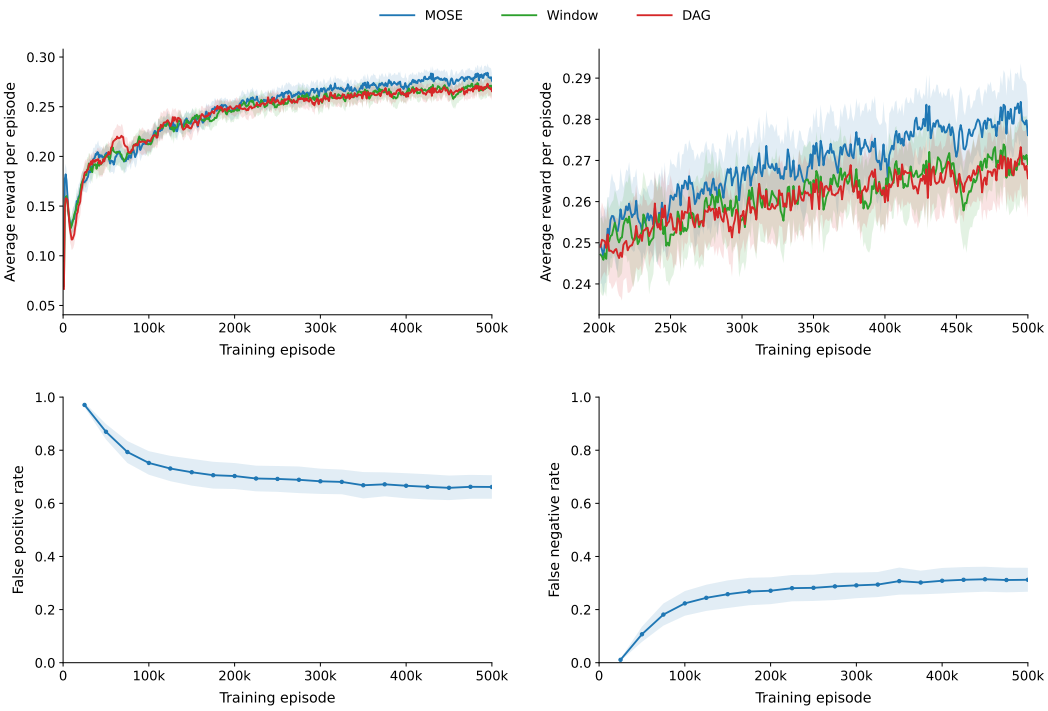


Figure H.4: Avergaed over 18 instances of nonlinear ANM. Horizon is 30 and $W = 10$. 20 nodes per time step. Top: Performance comparison; Bottom: False positive/negative rate of MOSE selected state.

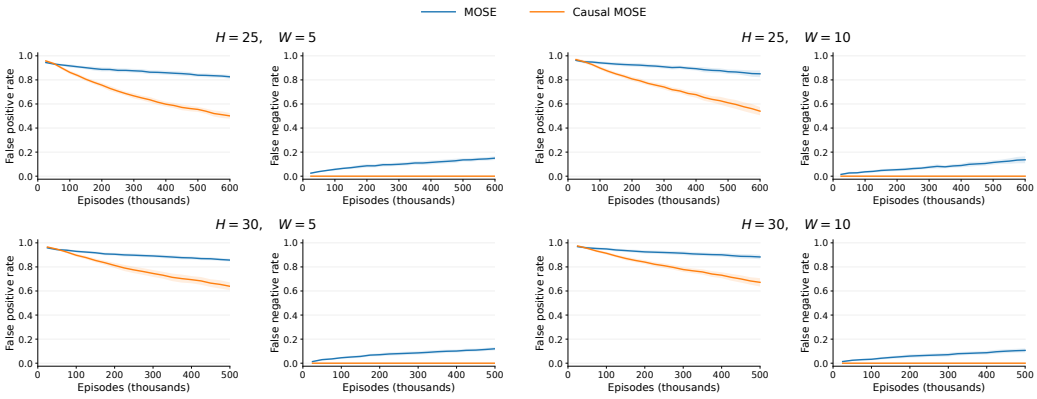


Figure H.5: False positive/negative rate of MOSE selected state and Causal MOSE selected state.

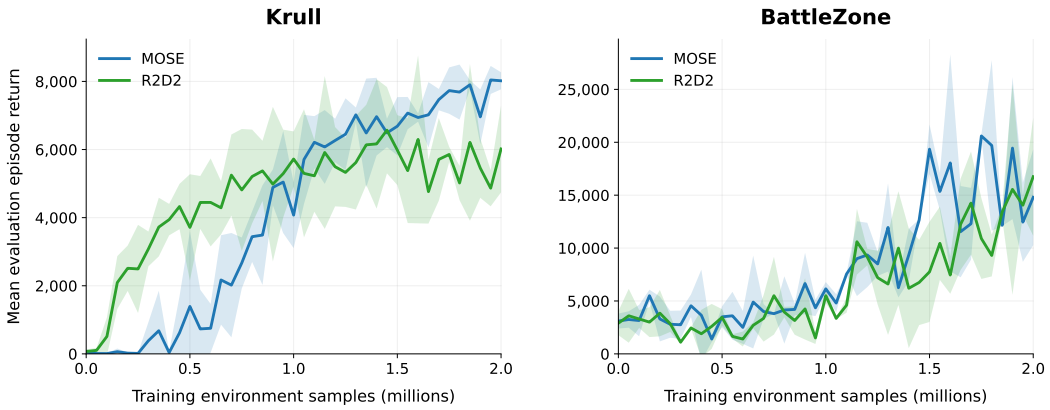


Figure H.6: Performance Comparison on Krull and Battlezone

Table H.1: Hyperparameters of Network for Synthetic

Hyperparameter	Value
Hidden layers	2
Hidden size	128
Learning rate	10^{-3}
Adam ϵ	10^{-8}
Batch size	64
Replay capacity	50,000
Target update interval	50
Replay frequency	1
Gradient clipping	1000.0
Learning starts	32

H.6 Learning Curve of Atari Games

As shown in Figure H.6, MOSE frequently outperforms R2D2 on BattleZone, particularly during later training. On Krull, MOSE requires a longer initial learning period but subsequently overtakes R2D2 and sustains higher returns toward the end of training. This delayed yet sustained improvement suggests that explicit selection may help MOSE eventually learn a more informative belief state.

H.7 Choices of Hyperparameters

We report the hyperparameters of Q-netwrok for synthetic and the attention related hyperparameters for Atari games respectively in Table H.1 and H.7. For the choices of hyperparameters of R2D2, we match the ones reported in Kapturowski et al. (2018) except we only have one learner that collects sample and learns from it. We did not have multiple actors as in R2D2 due to compute constraint.

Table H.2: MOSE attention selection-layer hyperparameters.

Hyperparameter	Value
History length	4
Attention dimension	32