

A Organization

A schematic of our framework and its logical flow is illustrated in Fig. 7. The diagram visually maps out how the paper develops, starting from the formulation of the non-stationary SCM-MAB problem in §3 and introducing the temporal model along with key graphical assumptions about the nature of non-stationarity. §5 establishes that in a temporal model, partial assignments of predetermined variables can affect optimal interventions at future time steps, motivating the definition of POMIS^+ as a non-myopic intervention. This leads into the graphical characterization in §6 which defines key structures (IB^+ , QIB) that construct POMIS^+ . Finally, §7 presents an algorithm that enumerates optimal intervention sequences based on these graphical elements. The structure clarifies the dependencies between the paper’s components and highlights how our contributions build upon one another.

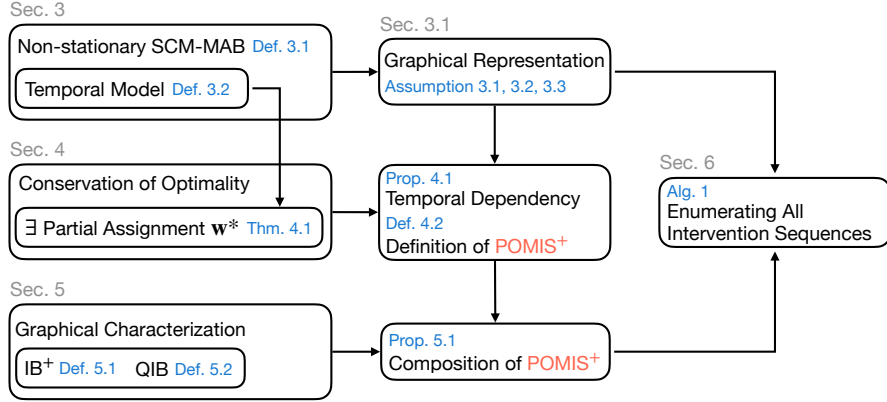


Figure 7: An overall schema of this paper.

B Preliminaries

Here we present definitions of MIS, POMIS, MUCT, and IB.

Definition B.1. (Minimal Intervention Set (MIS) [Lee and Bareinboim, 2018]). A set of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$ is said to be a *minimal intervention set* relative to $[\mathcal{G}, Y]$ if there is no $\mathbf{X}' \subset \mathbf{X}$ such that $\mu_{\mathbf{x}[\mathbf{X}']} = \mu_{\mathbf{x}}$ for every SCM conforming to the \mathcal{G} .

Definition B.2. (Possibly-Optimal Minimal Intervention Set (POMIS) [Lee and Bareinboim, 2018]). Given information $[\mathcal{G}, Y]$, let \mathbf{X} be a MIS. If there exists an SCM conforming to \mathcal{G} such that $\mu_{\mathbf{x}^*} > \forall_{\mathbf{Z} \in \mathbb{Z} \setminus \{\mathbf{X}\}} \mu_{\mathbf{Z}^*}$, where \mathbb{Z} is the set of MISs with respect to \mathcal{G} and Y , then \mathbf{X} is a *possibly-optimal minimal intervention set* with respect to the information $[\mathcal{G}, Y]$.

Definition B.3. (*C*-component [Tian and Pearl, 2003]). In a causal diagram \mathcal{G} , two variables are said to be in the same confounded component (for short, *C*-component or $\text{CC}(\cdot)_{\mathcal{G}}$) if and only if they are connected by a bi-directed edge (i.e., a path composed solely of $V_i \leftrightarrow V_j$).

In this paper, we denote $\text{CC}(\mathbf{X})_{\mathcal{G}} = \bigcup_{X \in \mathbf{X}} \text{CC}(X)_{\mathcal{G}}$.

Definition B.4. (Unobserved-Confounders’ Territory [Lee and Bareinboim, 2018]). Given information $[\mathcal{G}, Y]$, let H be $\mathcal{G}[An(Y)_{\mathcal{G}}]$. A set of variables $\mathbf{T} \subseteq \mathbf{V}(H)$ containing Y , where $\mathbf{V}(H)$ is the set of variables in H , is called a *UC-territory* on \mathcal{G} with respect to Y if $\text{De}(\mathbf{T})_H = \mathbf{T}$ and $\text{CC}(\mathbf{T})_H = \mathbf{T}$.

A UC-territory \mathbf{T} is said to be *minimal* if no $\mathbf{T}' \subset \mathbf{T}$ is a UC-territory. A minimal UC-Territory (MUCT) for \mathcal{G} and Y can be constructed by extending a set of variables, starting from $\{Y\}$, alternatively updating the set with the *c*-component and descendants of the set.

Definition B.5. (Interventional Border [Lee and Bareinboim, 2018]). Let \mathbf{T} be a minimal UC-territory on \mathcal{G} with respect to Y . Then, $\mathbf{X} = \text{pa}(\mathbf{T})_{\mathcal{G}} \setminus \mathbf{T}$ is called an *interventional border* for \mathcal{G} with respect to Y .

Algorithm 2 Minimal unobserved confounders' territory

```
1: function MUCT( $\mathcal{G}, \mathbb{Y}$ )
2:  $H = \mathcal{G}[An(Y)_{\mathcal{G}}]$ 
3:  $\mathbf{Q} = \{Y\}; \mathbf{T} = \{Y\}$ 
4: while  $\mathbf{Q} \neq \emptyset$  do
5:   remove an element  $Q_1$  from  $\mathbf{Q}$ 
6:    $\mathbf{W} = CC(Q_1)_H; \mathbf{T} = \mathbf{T} \cup \mathbf{W}; \mathbf{Q} = (\mathbf{Q} \cup de(\mathbf{W})_H) \setminus \mathbf{T}$ 
7: return  $\mathbf{T}$ 
```

C Nomenclature

NS	Non-stationary
SCM	Structural causal model
RL	Reinforcement learning
MAB	Multi-armed bandit
MIS	Minimal intervention set
UC	Unobserved confounder
POMIS	Possibly-optimal minimal intervention set
MUCT	Minimal unobserved confounders' territory
IB	Interventional border
QIB	Qualified interventional border
DUC	Dynamically unobserved confounder
CR	Cumulative regret
OAP	Optimal arm-selection probability
KL-UCB	Kullback-Leibler upper confidence bound

D Stationary SCM-MAB

The cumulative regret of the stationary setting is given by:

$$R_T \triangleq \sum_{t=1}^T \left(\mu_{\mathbf{x}^\dagger} - \mathbb{E}[\mu_{\mathbf{x}_t}] \right), \quad (4)$$

where \mathbf{x}^\dagger denotes the globally optimal arm, defined as

$$\mathbf{x}^\dagger \triangleq \arg \max_{\mathbf{x} \in \mathcal{D}(\mathbf{X}), \mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}} \mathbb{E}[Y \mid \text{do}(\mathbf{X} = \mathbf{x})].$$

where $\mu_{\mathbf{x}_t}$ is the arm played at round t . The SCM-MAB setting assumes that the agent has full access to the causal graph \mathcal{G} of \mathcal{M} , although its parametrization remains unknown—i.e., the agent knows the structure \mathcal{G} , but not the structural functions \mathcal{F} or the distribution over exogenous variables $P(\mathbf{U})$. Furthermore, the causal graph \mathcal{G} is assumed to be static, meaning that the underlying causal structure of the domain does not change over time. As a result, the agent interacts with the same causal model in each round.

In this setting, there is no confounding across time slices, and thus, no information propagates between rounds. Consequently, the reward distribution remains fixed across rounds.

In practice, under the stationary setting, the agent effectively observes, in each round, a causal diagram \mathcal{G} composed of temporally disconnected slices, as illustrated in Fig. 1(b). This stationary (and also non-stationary) formulation typically assumes that the number of arms is constant throughout the interaction. The graphical structure and topology of each time slice are identical across all rounds.

The SCM-MAB framework inherently introduces dependencies between arms, stemming from the underlying causal relationships among endogenous and exogenous variables. Lee and Bareinboim [2018, 2019] identified two structural properties that can be derived from any SCM-MAB framework:

1. Arm equivalence: a characterization of arms that share identical reward distributions, determined using constraints from do-calculus, and
2. Partial-orders among arms: under what topological conditions one arm can be optimal.

Leveraging these properties, one can identify minimal intervention sets (MIS) that constitute a non-redundant collection of informative interventions. In addition, Lee and Bareinboim [2019] identified both MIS and POMIS for the stationary setting with non-manipulative variables. However, these characterizations rest on the assumption that the causal graph \mathcal{G} remains static and stationary—meaning that no information is carried over from previous decisions. §3 extends beyond the stationary assumption to present the SCM-MAB framework in the non-stationary setting.

E Comparison with Conventional Non-Stationary Bandit Algorithms

We contrast our approach to non-stationarity in the SCM-MAB with traditional non-stationary bandit settings. Our focus is on how causal modeling provides a structured explanation for reward distribution shifts over time, as opposed to treating them as statistical artifacts. These build upon early foundational work on dynamic allocation and index policies [Whittle, 1988, Gittins, 1979].

Conventional NS-bandit formulations Conventional non-stationary bandit algorithms aim to maintain *low regret* with respect to a comparator (or competitor) class—a predefined set of benchmark policies that may adapt over time to account for changing environments (e.g., policies that allow a limited number of switches between arms, comparators that track the best-performing arm over recent time windows, or strategies that assume bounded changes in the underlying reward distribution). These algorithms are typically categorized into two regimes: adversarial and stochastic bandits, illustrated in Table 1.

Table 1: Representative non-stationary bandit settings and algorithms (adapted from Lattimore and Szepesvári [2020]).

Regime	Setting	Description	Representative Algorithms
Adversarial	L -switching	The identity of the optimal arm may change abruptly up to L times	Exp3.S, AdaHedge
Adversarial	Variation budget	Total variation in reward sequences is bounded by V	Rexp3, Adapt-EvE
Stochastic	Piecewise-stationary	Rewards are stationary within intervals, with occasional change-points	Sliding-window UCB, Change-point Thompson Sampling
Stochastic	Drifting	Expected rewards evolve smoothly over time	Discounted UCB, Sliding-window UCB, SW-TS

Each setting assumes non-stationarity as a statistical property: either abrupt shifts, or slowly drifting rewards. Some algorithms require knowledge of the number of switches L , while others are designed to be adaptive. In the variation budget setting, the cumulative amount of change in reward distributions is bounded by a budget V , offering a finer-grained control of non-stationarity than simple change-point models.

Causal non-stationarity (our perspective) Our framework models non-stationarity as a consequence of causal information propagation across time. Specifically, transition edges in the causal graph \mathcal{G} (e.g., $(X_t \rightarrow Z_{t'})$) induce changes in the downstream reward variables (e.g. $Y_{t'}$), illustrated in Fig. 2. This structure directly captures *why* the reward distribution changes.

For instance, whereas traditional settings might treat $\mathbb{E}[Y_t \mid \text{do}(x)]$ as shifting arbitrarily with t , our approach identifies structural causes: $P(Y_t \mid \text{do}(X_t))$ is influenced by information propagation

from previous variables (e.g. X_{t-1} , Z_{t-1}). This allows us to model the *mechanism* behind reward distribution shifts.

Moreover, the presence of arcs (or edges) between time slices in \mathcal{G} determines where and how changes occur. This is closely aligned with the “mean payoff drift” interpretation in traditional models [Lattimore and Szepesvári, 2020, Chapter 31], but we interpret it in terms of explicit graphical information in \mathcal{G} .

Summary Most traditional algorithms detect or adapt to change, but do not explain it. Our approach, by contrast, offers a mechanism-based explanation of non-stationarity—one grounded in a causal understanding of the system under investigation, which is itself a strong assumption—via the SCM structure. This allows for:

- Identification of reward-relevant intervention targets (POMIS⁺)
- Intervention sequence planning backed by theoretical guarantees derived from the underlying causal structure

Ultimately, our method treats non-stationarity as a structured phenomenon emergent from a dynamic causal model, rather than as an arbitrary change in observed statistics.

F Semi Time-Slice Markovian Non-Stationary SCM-MAB

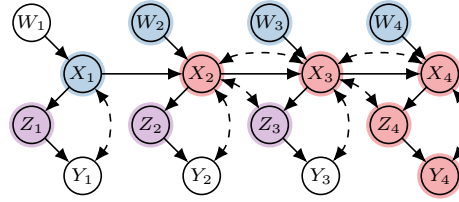


Figure 8: IB^+ ranges over four time steps.

We begin by formalizing dynamic unobserved confounders (DUCs), which represent exogenous variables inducing dependencies across time.

Definition F.1. (Dynamic unobserved confounders (DUCs)). Given $[\mathcal{G}, \mathbf{Y}]$, let \mathbf{U}_t^* denote the set of exogenous variables that induce dependencies between variables in \mathbf{V}_t and $\mathbf{V}_{t'}$ for some $t < t'$. If $U_j \in \mathbf{U}_t^*$, then we refer to U_j as a dynamic unobserved confounder (DUC), indicating that U_j introduces confounding effects that persist across time.

When such DUCs are permitted in the graphical structure, we say the temporal graph satisfies the semi-time-slice Markov property.

Definition F.2 (Semi-time-slice Markov). A temporal graph \mathcal{G} satisfies the semi-time-slice Markov property if it allows the presence of dynamic unobserved confounders (DUCs) across time. That is, the graph permits bidirected edges between variables $V_t \in \mathbf{V}_t$ and $V_{t'} \in \mathbf{V}_{t'}$ for $t \neq t'$, representing confounding induced by exogenous variables $U_j \in \mathbf{U}_t^*$ that simultaneously affect both time slices.

Under the semi-time-slice Markov assumption induced by dynamic unobserved confounders (Def. F.1), we need to perform a time-expanded search for valid intervention targets across multiple time steps. Consider the example in Fig. 8. When we calculate IB^+ for $\text{MUCT}(\mathcal{G}[\bigcup_{i=1}^4 \mathbf{V}_i], Y_4)$, the result consists of four sets: $\text{IB}_{1,4}^+ = \{X_1\}$, $\text{IB}_{2,4}^+ = \{W_2\}$, $\text{IB}_{3,4}^+ = \{W_3\}$ and $\text{IB}_{4,4}^+ = \{W_4\}$ (blue shaded). Given those IB^+ at each time step, we can determine each QIB_t according to Def. 6.2. For example, when $\text{IB}_{1,4}^+ = \{X_1\}$ is given, the IB corresponding to the $\text{MUCT}(\mathcal{G}[\mathbf{V}_1]_{\overline{X_1 \cup Z_1}}, Y_1) = \{Y_1\}$ is Z_1 , hence $\text{QIB}_1 = \{Z_1\}$. Similarly, for $\text{MUCT}(\mathcal{G}[\mathbf{V}_2]_{\overline{X_2 \cup W_2}}, Y_2)$, we can select $\text{QIB}_2 = \{Z_2\}$, and for $\text{MUCT}(\mathcal{G}[\mathbf{V}_3]_{\overline{X_3 \cup W_3}}, Y_3)$, we can choose $\text{QIB}_3 = \{Z_3\}$ (purple shaded). For now, by Prop. 6.1, we can obtain POMIS^+ s by taking the union of each $\text{IB}_{t,t'}^+$ and QIB_t — $\text{POMIS}_{1,4}^+ = \{X_1, Z_1\}$, $\text{POMIS}_{2,4}^+ = \{W_2, Z_2\}$, $\text{POMIS}_{3,4}^+ = \{W_3, Z_3\}$ and $\text{POMIS}_{4,4}^+ = \{W_4\}$. These POMIS^+ sets can then be used as components of an intervention sequence in the NS-SCM-MAB setting.

G Proofs

Theorem 5.1. (Existence of Optimal Partial Assignment). Given information $\llbracket \mathcal{G}, \mathbf{Y} \rrbracket$, let $\mathbf{X}_{t'}$ be a POMIS with respect to $\llbracket \mathcal{G}[\mathbf{V}_{t'}], Y_{t'} \rrbracket$ for the subsequent two time steps $t, t' \in [T]$ with $t < t'$. Then, there exists an assignment \mathbf{w}^* for a subset of variables $\mathbf{W} \subseteq Pa(\mathbf{V}_{t'}) = \mathbf{V}_{t'}^*$ such that

$$\mathbb{E}^{\mathcal{M}_{t'} | \mathbf{v}_{t'}^*} [Y_{t'} \mid \text{do}(\mathbf{x}_{t'}^*)]$$

achieves the maximum expected reward for any $\mathbf{v}_{t'}^*$ with $\mathbf{w}^* = \mathbf{v}_{t'}^*[\mathbf{W}]$.

Proof. Let $t, t' \in [T]$ with $t < t'$ be two time steps. Given the information $\llbracket \mathcal{G}, \mathbf{Y} \rrbracket$, we fix an arbitrary target variable $Y_{t'} \in \mathbf{Y}$ located in the time slice $\mathcal{G}[\mathbf{V}_{t'}]$.

We now consider any temporal model $\mathcal{M}_{t'}$ that conforms to the time slice $\mathcal{G}[\mathbf{V}_{t'}]$. Let $\mathbf{X}_{t'}$ be a POMIS for $Y_{t'}$ with respect to the time slice $\mathcal{G}[\mathbf{V}_{t'}]$. The goal is to show that under any temporal model conforming to $\mathcal{G}[\mathbf{V}_{t'}]$, there exists a partial assignment \mathbf{w}^* that preserves the optimality of $\mathbf{X}_{t'}$.

By the definition of POMIS, there exists an intervention assignment $\mathbf{x}_{t'}^* \in \mathcal{D}(\mathbf{X}_{t'})$ such that:

$$\mathbb{E}^{\mathcal{M}_{t'}} [Y_{t'} \mid \text{do}(\mathbf{x}_{t'}^*)] > \mathbb{E}^{\mathcal{M}_{t'}} [Y_{t'} \mid \text{do}(\mathbf{z})] \quad \text{for all } \mathbf{z} \in \mathbb{Z} \setminus \{\mathbf{x}_{t'}^*\},$$

where \mathbb{Z} is the set of all MISs for $Y_{t'}$ in $\mathcal{G}[\mathbf{V}_{t'}]$.

Now consider how expectations are evaluated in the temporal model $\mathcal{M}_{t'}$: they are conditioned on predetermined variables $\mathbf{v}_{t'}^* \in \mathcal{D}(Pa(\mathbf{V}_{t'}))$. Hence, there must exist at least one such $\mathbf{v}_{t'}^*$ under which the above inequality holds.

Fix any such $\mathbf{v}_{t'}^*$. Since the domain $\mathcal{D}(Pa(\mathbf{V}_{t'}))$ is finite, and the conditional expectation $\mathbb{E}^{\mathcal{M}_{t'} | \mathbf{v}_{t'}^*} [Y_{t'} \mid \text{do}(\cdot)]$ is a deterministic function of $\mathbf{v}_{t'}^*$ and the do-intervention, there exists a minimal subset $\mathbf{W} \subseteq Pa(\mathbf{V}_{t'})$ such that $\mathbf{w}^* = \mathbf{v}_{t'}^*[\mathbf{W}]$ satisfies:

$$\mathbb{E}^{\mathcal{M}_{t'} | \mathbf{v}_{t'}^*} [Y_{t'} \mid \text{do}(\mathbf{x}_{t'}^*)] > \mathbb{E}^{\mathcal{M}_{t'} | \mathbf{v}_{t'}'} [Y_{t'} \mid \text{do}(\mathbf{z})], \quad \text{for all } \mathbf{v}_{t'}' \text{ such that } \mathbf{v}_{t'}'[\mathbf{W}] = \mathbf{w}^*.$$

That is, fixing the partial assignment \mathbf{w}^* suffices to preserve the optimality of the POMIS $\mathbf{X}_{t'}$ regardless of how the remaining variables in $Pa(\mathbf{V}_{t'}) \setminus \mathbf{W}$ are instantiated.

Since the choice of $\mathcal{M}_{t'}$ was arbitrary (subject to conforming to $\mathcal{G}[\mathbf{V}_{t'}]$), this concludes that such a subset \mathbf{W} and partial assignment \mathbf{w}^* must exist under any temporal model consistent with the time-slice causal structure. \square

Proposition 5.1. (Temporal Dependency). Given causal diagram \mathcal{G} and a collection of time-specific variables \mathbb{V} , $\text{POMIS}_{t,t+1}^+ \subseteq \mathbf{V}_t \setminus \{Y_t\}$ for every $t \in [T]$ under Assumption 3.1.

Proof. For the sake of contradiction, suppose that the proposition does not hold. Then, there exists some $t \in [T]$ such that:

$$\nexists \text{POMIS}_{t,t+1}^+ \subseteq \mathbf{V}_t \setminus \{Y_t\}$$

That is, there exists a POMIS^+ set $\mathbf{X}_{t,t+1}^+$ such that:

$$\exists X \in \mathbf{X}_{t,t+1}^+ \text{ where } X \notin \mathbf{V}_t \setminus \{Y_t\}$$

By this existence, X must either lie in a future time slice ($X \in \mathbf{V}_{t'}$ with $t < t'$), or $X = Y_t$.

We now argue that such an X cannot be part of any valid POMIS^+ set under Assumption 3.1. Recall that under the assumption, the only causal influences on Y_{t+1} must flow through variables in \mathbf{V}_t (i.e., no time cycles and no backward edges from $\mathbf{V}_{t'}$ with $t < t'$).

Furthermore, $\text{POMIS}_{t,t+1}^+$ is defined as the minimal set in \mathbf{V}_t such that the intervention at time t maximizes the expected reward at time $t + 1$. Including any $X \notin \mathbf{V}_t \setminus \{Y_t\}$ violates both this minimality and the validity of the do-intervention within time t .

This leads to a contradiction. Hence, for all $t \in [T]$, we must have:

$$\text{POMIS}_{t,t+1}^+ \subseteq \mathbf{V}_t \setminus \{Y_t\}$$

\square

We now prove the composition property of $\text{POMIS}_{t,t'}^+$, which guarantees that interventions on $\text{IB}_{t,t'}^+$ do not block the causal effect of QIB_t on Y_t . Before proceeding, we formalize the fact that the set constructed by $\text{IB}_{t,t'}^+$ satisfies the partial assignment condition required by Thm. 5.1.

Proposition G.1 (IB^+ satisfies the condition of Thm. 5.1). *Let $t < t'$ and let $\mathbf{X}_{t'}$ be a POMIS for $Y_{t'}$ in $\mathcal{G}[\mathbf{V}_{t'}]$. Then, the set $\text{IB}_{t,t'}^+(\mathcal{G}[\bigcup_{i=t}^{t'} \mathbf{V}_i], Y_{t'})$ identifies a subset of $\text{Pa}(\mathbf{V}_{t'})$ such that there exists a partial assignment \mathbf{w}^* over this set which satisfies the condition of Thm. 5.1*

Proof. Let $t < t'$ and let $\mathbf{X}_{t'}$ be a POMIS for $Y_{t'}$ in $\mathcal{G}[\mathbf{V}_{t'}]$.

From Thm. 5.1, this implies the existence of a subset $\mathbf{W} \subseteq \text{Pa}(\mathbf{V}_{t'})$ and a partial assignment \mathbf{w}^* such that for all $\mathbf{v}_{t'}^*$ with $\mathbf{v}_{t'}^*[\mathbf{W}] = \mathbf{w}^*$, the reward under $\text{do}(\mathbf{x}_{t'}^*)$ can be maximized.

Now consider the construction of $\text{IB}_{t,t'}^+(\mathcal{G}[\bigcup_{i=t}^{t'} \mathbf{V}_i], Y_{t'})$. By definition, this set is formed by computing the interventional border IB in the full unrolled graph \mathcal{G} , and selecting from it only those variables that reside in \mathbf{V}_t .

Since the IB of $Y_{t'}$ is known to be a POMIS, and since IB^+ is a subset of this IB restricted to variables available at time t , the values assigned to IB^+ in any temporal model appear as part of some $\mathbf{v}_{t'}^* \in \mathcal{D}(\text{Pa}(\mathbf{V}_{t'}))$.

Note that although Thm. 5.1 is stated in terms of a temporal model $\mathcal{M}_{t'}$, the structure of $\mathcal{M}_{t'}$ —including its structural functions and the set of predetermined variables $\text{Pa}(\mathbf{V}_{t'})$ —is dependent on the \mathcal{M} . Since $\text{IB}_{t,t'}^+$ is computed from the global graph $\mathcal{G}[\bigcup_{i=t}^{t'} \mathbf{V}_i]$, it selects variables that reflect this SCM-induced structure, and thus corresponds to the \mathbf{W} required in Thm. 5.1.

Therefore, $\text{IB}_{t,t'}^+$ fulfills the role of \mathbf{W} in Thm. 5.1, ensuring that fixing its values preserves the optimality of $\mathbf{X}_{t'}$ consistent with \mathbf{w}^* . \square

Proposition 6.1. (*Composition of POMIS^+*). *Given $\llbracket \mathcal{G}, \mathbf{Y} \rrbracket$, $\text{IB}_{t,t'}^+(\mathcal{G}[\bigcup_{i=t}^{t'} \mathbf{V}_i], Y_{t'}) \cup \text{QIB}_t(\mathcal{G}[\mathbf{V}_t], Y_t)$ is a $\text{POMIS}_{t,t'}^+$ for $t, t' \in [T]$ and $t < t'$.*

Proof. Let $\mathbf{X}_t := \text{QIB}_t(\mathcal{G}[\mathbf{V}_t], Y_t)$ and $\mathbf{W}_t := \text{IB}_{t,t'}^+(\mathcal{G}[\bigcup_{i=t}^{t'} \mathbf{V}_i], Y_{t'})$. By Def. 6.2, \mathbf{X}_t is a Qualified Interventional Border (QIB) for Y_t with respect to $\mathcal{G}[\mathbf{V}_t]$, which implies that \mathbf{X}_t is a POMIS for Y_t . Thus, the first condition of Def. 5.2 is satisfied.

Next, as established in Prop. G.1, the set \mathbf{W}_t constructed via $\text{IB}_{t,t'}^+$ satisfies the condition of Thm. 5.1: there exists a partial assignment \mathbf{w}_t over \mathbf{W}_t such that the expected reward for $Y_{t'}$ under $\text{do}(\mathbf{x}_{t'})$ is maximized.

Moreover, due to the construction of QIB_t , which requires that $\text{IB}(\mathcal{G}[\mathbf{V}_t]_{\overline{\mathbf{Z} \cup \mathbf{W}_t}}, Y_t) = \mathbf{X}_t$ for some $\mathbf{Z} \subseteq \mathbf{V}_t \setminus \{Y_t\}$, it follows that interventions on \mathbf{W}_t do not interfere with the causal effect of \mathbf{X}_t on Y_t within $\mathcal{G}[\mathbf{V}_t]$.

Therefore, the expected reward at time t remains unchanged $\mu_{\mathbf{x}_t} = \mu_{\mathbf{x}_t \cup \mathbf{w}_t}$. Hence, by Def. 5.2, the combined set $\mathbf{X}_t \cup \mathbf{W}_t$ is a $\text{POMIS}_{t,t'}^+$. \square

Theorem 7.1. (*Soundness and Completeness*). *Given information $\llbracket \mathcal{G}, \mathbf{Y} \rrbracket$, Alg. 7 returns all intervention sequences composed solely of POMIS^+ sets.*

Proof. We prove the theorem by showing (i) soundness: every sequence returned by Alg. 7 is composed solely of POMIS^+ , and (ii) completeness: every valid sequence of POMIS^+ is returned.

(Soundness) Fix the final time step T . Let $\mathcal{G}' = \mathcal{G}[\text{An}(Y_T)_{\mathcal{G}}]$ and suppose the algorithm selects a valid MUCT/IB pair (\mathbf{X}, \mathbf{T}) under the \mathcal{G}' . $\text{IB}^+(\mathbb{V}, \mathbf{X}, \mathcal{I}^+)$ updates the IB^+ map by assigning each $X \in \mathbf{X}$ to its respective time step t_X . For each $t \in \text{keys}(\mathcal{I}^+)$, the function $\text{QIB}(\mathcal{G}, \mathbb{V}, \mathbf{T}, \mathcal{Q}, \mathcal{I}^+)$ constructs a mutilated graph $\mathcal{G}[\mathbf{V}_t]_{\overline{\mathcal{I}^+[t]}}$, and from it computes all valid POMISs (i.e., \mathbf{q}_t) that do not intersect with \mathbf{T} . In the base case (i.e., when the earliest $t_e = 0$), the Cartesian product over all t gives sequences $\{\mathcal{I}^+[t] \cup \mathbf{q}_t \mid \mathbf{q}_t \in \mathcal{Q}[t]\}$. By Prop. 6.1, each $\mathcal{I}^+[t] \cup \mathbf{q}_t$ is a valid POMIS^+ set at time t . Thus, every sequence in \mathbb{S} consists of only POMIS^+ sets.

(Completeness) The algorithm recursively traces backward from $t = T$ to earlier time steps, guided by the smallest time index in the current IB^+ map, denoted t_e , and recurses with horizon $t_e - 1$. At each step, the algorithm considers all valid MUCT/IB pairs for Y_t . The correctness of this step is guaranteed by the soundness and completeness of the algorithm POMISs from Lee and Bareinboim [2018, Theorem 9]. Each recursive call updates the maps \mathcal{I}^+ and \mathcal{Q} by accumulating IB^+ and QIB sets across all considered time steps. Once the earliest time index in IB^+ reaches $t = 0$, the recursion terminates. At this point, the algorithm forms the Cartesian product $\prod_{t \in \text{keys}(\mathcal{I}^+) \cap \text{keys}(\mathcal{Q})} \{\mathcal{I}^+[t] \cup \mathbf{q}_t \mid \mathbf{q}_t \in \mathcal{Q}[t]\}$, enumerating every possible combination across time. Since every possible IB^+ and QIB configuration is explored and retained, the final set \mathbb{S} contains all valid sequences composed solely of POMIS⁺ sets.

□

H Algorithmic Characterization of POMIS⁺

Algorithm 3 Update \mathcal{I}^+ and \mathcal{Q} from \mathcal{I}^+

```

1: function  $\text{IB}^+(\mathbb{V}, \mathbf{X}, \mathcal{I}^+)$ 
2: for each  $X \in \mathbf{X}$  do
3:   Identify the time step  $t$  such that  $X \in \mathbf{V}_t$ 
4:   if  $t \notin \mathcal{I}^+$  then
5:      $\mathcal{I}^+[t] \leftarrow []$ 
6:   append  $X$  to  $\mathcal{I}^+[t]$ 
7: return  $\mathcal{I}^+$ 

8: function  $\text{QIB}(\mathcal{G}, \mathbb{V}, \mathbf{T}, \mathcal{Q}, \mathcal{I}^+)$ 
9: for each  $t \in \text{keys}(\mathcal{I}^+)$  do
10:  if  $t \notin \mathcal{Q}$  then
11:     $G_t, \mathcal{Q}[t] \leftarrow \mathcal{G}[\mathbf{V}_t]_{\overline{\mathcal{I}^+[t]}}, []$ 
12:    for each  $\mathbf{X} \in \text{POMISs}(G_t, Y_t)$  do
13:      if  $\mathbf{X} \cap \mathbf{T} = \emptyset$  then
14:        append  $\mathbf{X}$  to  $\mathcal{Q}[t]$ 
15: return  $\mathcal{Q}$ 

```

In this appendix, we provide a detailed explanation of the algorithmic components used in the enumeration of POMIS⁺ intervention sequences, as introduced in §7. At first, we define and clarify the role of the two internal maps, \mathcal{I}^+ and \mathcal{Q} , which correspond to the Interventional border for the subsequent time steps (IB^+) and qualified interventional borders (QIB), respectively.

Map of interventional border for the subsequent time steps (\mathcal{I}^+) Given a POMIS candidate \mathbf{X} obtained from the MUCT-IB procedure, we identify the time step t associated with each variable $X \in \mathbf{X}$, and store it in the map $\mathcal{I}^+[t]$. The map \mathcal{I}^+ thus organizes variables in \mathbf{X} by their corresponding time step, preserving the temporal alignment of possible interventions. This time-indexed representation allows the algorithm to recursively evaluate which time steps still require backward expansion, and is central to the structure of Alg. 1. This structure ensures that only the earliest relevant time step is explored recursively, avoiding redundant expansion into irrelevant subgraphs.

Map of qualified interventional Border (\mathcal{Q}) Once \mathcal{I}^+ is populated, the QIB map \mathcal{Q} is computed by evaluating each subgraph $\mathcal{G}[\mathbf{V}_t]$ after mutilating it with the intervention set $\mathcal{I}^+[t]$. From this mutilated subgraph, we re-run the POMISs algorithm on the local reward variable Y_t to identify any remaining minimal intervention sets. These are stored in $\mathcal{Q}[t]$ only if they do not overlap with the current MUCT set \mathbf{T} , ensuring independence across temporal intervention levels. The QIB thus captures any residual variables at time t that are necessary to optimize the local reward, complementing the already selected IB^+ .

The recursive design of Alg. 1 ensures that the enumeration of intervention sequences terminates when all relevant intervention variables are assigned by time $t = 0$. At that point, the algorithm takes the Cartesian product of all IB^+ and QIB sets across time steps to generate complete intervention

sequences. This backward recursive approach avoids exhaustive enumeration by pruning irrelevant branches early and leveraging graph locality in the causal diagram.

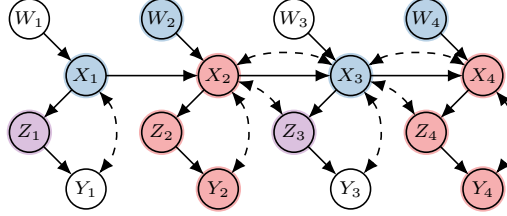


Figure 9: Illustration of the intermediate stage of the algorithm after two recursive calls.

Example traces of recursive enumeration. Fig. 9 illustrates an intermediate stage in the execution of our recursive POMIS⁺ enumeration algorithm. The color-coded graph shows how MUCT, IB⁺, and QIB evolve across time steps as the algorithm proceeds backward.

At the final time step $t = 3$ (corresponding to Y_4 in the figure), the algorithm computes the MUCT and IB from the subgraph $\mathcal{G}[\mathbf{V}_3 \cup \mathbf{V}_4]$, resulting in a MUCT set $\{X_4, Z_4, Y_4\}$ (highlighted in red). The corresponding IB⁺ consists of the set $\{X_3, W_4\}$, indicated in blue. Next, the QIB is computed by evaluating the mutilated subgraph $\mathcal{G}[\mathbf{V}_3]_{\overline{X_3, W_4}}$, yielding $\{Z_3\}$ as an intervention set (highlighted in purple).

Since the current earliest time step in IB⁺ is $t = 3 > 1$, the algorithm proceeds recursively with $T \leftarrow 2$. In this second round, the algorithm again computes the MUCT/IB pair for Y_2 , resulting in MUCT $\{X_2, Z_2, Y_2\}$ (red), IB⁺ = $\{X_1, W_2\}$ (blue), and QIB = $\{Z_1\}$ (purple), after mutilating $\mathcal{G}[\mathbf{V}_1]$ with $\{X_1, W_2\}$.

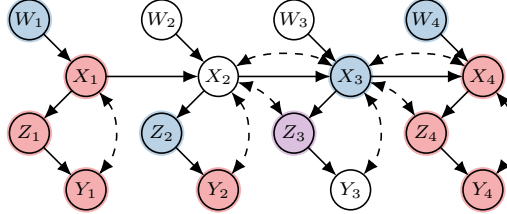


Figure 10: Illustration of the intermediate stage of the algorithm after three recursive calls.

Fig. 10 also provides one of the recursive traces of our POMIS⁺ enumeration algorithm over steps, starting from the final reward Y_4 at time $t = 4$ and proceeding backward to Y_1 at $t = 1$. As before, red nodes indicate variables in the MUCT set, blue nodes indicate IB⁺ variables and purple nodes denote QIB elements at each stage.

At time $t = 4$, the algorithm identifies the MUCT $\{Y_4, Z_4, X_4\}$ (highlighted in red), computed from the subgraph $\mathcal{G}[\mathbf{V}_3 \cup \mathbf{V}_4]$. The IB⁺ map is then updated to include $\{X_3, W_4\}$ (blue), and since $\{Z_3\}$ influences to Y_3 in $\mathcal{G}[\mathbf{V}_3]_{\overline{X_3, W_4}}$, the QIB is $\{Z_3\}$ at this step.

Proceeding to time $t = 2$, the algorithm observes that a single node Y_3 can be a MUCT. The corresponding IB⁺ is updated to $\{Z_2\}$ (purple), with no further QIB identified. At time $t = 1$, the MUCT set becomes $\{Y_1, Z_1, X_1\}$ (red). The IB⁺ is now $\{Z_1\}$ (blue), as Z_1 is needed to activate paths to Y_2 under the current mutilated graph. Again, no QIB is generated at this time. Finally, at $t = 0$, the algorithm computes MUCT as $\{X_1, Z_1, Y_1\}$ (red), and identifies IB⁺ = $\{W_1\}$ (blue), completing the backward exploration.

This example highlights how our algorithm incrementally expands the maps \mathcal{I}^+ and \mathcal{Q} over time by tracing from Y_T back to earlier rewards. Each recursive call explores a new MUCT/IB combination while updating time-specific intervention targets, eventually enabling full enumeration when the earliest IB⁺ reaches time $t = 1$.

Time Complexity of Alg. 1 The recursive construction of POMIS⁺ sequences explores possible intervention combinations over a time horizon T . In the worst case, each time step requires enumer-

ating subsets of ancestors of Y_t to compute MUCT/IB pairs, followed by QIB exploration—both involving searches over subsets of the variable set \mathbf{V} (i.e., $\mathcal{O}(2^{|\mathbf{V}|})$). Each of these steps incurs an exponential cost of up to $2^{|\mathbf{V}|}$, resulting in $2^{2^{|\mathbf{V}|}}$ complexity per time step. After T recursion, the total worst-case time complexity becomes $\mathcal{O}(2^{2^{|\mathbf{V}| \cdot T}})$.

I Experiment Details

In this section, we provide detailed specifications of the SCMs used in our experiments to ensure reproducibility. For each experiment, the simulation was repeated 200 times using the corresponding SCM. All experiments were run on a dual-socket Intel Xeon Gold 5317 system with 24 physical cores (48 logical threads) at 3.0GHz.

Task 1. We implement an SCM composed of three time steps $t = 1, 2, 3$ with variable sets $\{X_t, Z_t, Y_t\}$ and structural equations defined to model the propagation of intervention effects over time. The probability distributions over exogenous variables \mathbf{U} are defined as: We use the following probabilities over exogenous variables:

$$\begin{aligned} P(U_{X_1} = 1) &= 0.85, & P(U_{X_1 Y_1} = 1) &= 0.47, & P(U_{Z_1} = 1) &= 0.14, & P(U_{Y_1} = 1) &= 0.02, \\ P(U_{X_2} = 1) &= 0.80, & P(U_{X_2 Y_2} = 1) &= 0.55, & P(U_{Z_2} = 1) &= 0.14, & P(U_{Y_2} = 1) &= 0.01, \\ P(U_{X_3} = 1) &= 0.01, & P(U_{X_3 Y_3} = 1) &= 0.51, & P(U_{Z_3} = 1) &= 0.13, & P(U_{Y_3} = 1) &= 0.05. \end{aligned}$$

The structural functions f_V for each endogenous variable V are as follows (where \oplus denotes binary XOR and v is the valuation of its parents):

$$\begin{array}{lll} f_{X_1} = U_{X_1} \oplus U_{X_1 Y_1} & f_{X_2} = X_1 \oplus U_{X_2} \oplus U_{X_2 Y_2} & f_{X_3} = X_2 \oplus U_{X_3} \oplus U_{X_3 Y_3} \\ f_{Z_1} = X_1 \oplus U_{Z_1} & f_{Z_2} = X_2 \oplus U_{Z_2} & f_{Z_3} = X_3 \oplus U_{Z_3} \\ f_{Y_1} = Z_1 \oplus U_{Y_1} \oplus U_{X_1 Y_1} & f_{Y_2} = Z_2 \oplus U_{Y_2} \oplus U_{X_2 Y_2} & f_{Y_3} = Z_3 \oplus U_{Y_3} \oplus U_{X_3 Y_3} \\ \text{(a) } t = 1 & \text{(b) } t = 2 & \text{(c) } t = 3 \end{array}$$

Figure 11: SCM definition for Task 1.

Task 2. We conducted the experiment with a two-step SCM defined over variables $\{W_t, X_t, Z_t, Y_t\}$ for $t = 1, 2$, as illustrated in Fig. 4. The structure explicitly models how intervention effects on X_t propagate across time via X_{t+1} and influence downstream outcomes Y_t and Y_{t+1} through intermediary variables Z_t .

The exogenous distribution $P(\mathbf{U})$ is parameterized to highlight the long-term influence of early interventions. The assigned probabilities are as follows:

$$\begin{aligned} P(U_{X_1} = 1) &= 0.15, & P(U_{Z_1} = 1) &= 0.15, & P(U_{Y_1} = 1) &= 0.02, & P(U_{X_1 Y_1} = 1) &= 0.47 \\ P(U_{W_1} = 1) &= 0.15, & P(U_{X_2} = 1) &= 0.02, & P(U_{Z_2} = 1) &= 0.12, & P(U_{X_2 Y_2} = 1) &= 0.55 \\ P(U_{Y_2} = 1) &= 0.01, & P(U_{W_2} = 1) &= 0.12. \end{aligned}$$

We define the structural equations in Fig. 12 (where \oplus denotes binary XOR). Notably, the structural

$$\begin{array}{ll} f_{W_0} = U_{W_0} & f_{W_1} = U_{W_1} \\ f_{X_0} = U_{X_0} \oplus W_0 \oplus U_{X_0 Y_0} & f_{X_1} = U_{X_1} \oplus W_1 \oplus U_{X_1 Y_1} \oplus X_0 \\ f_{Z_0} = X_0 \oplus U_{Z_0} & f_{Z_1} = X_1 \oplus U_{Z_1} \\ f_{Y_0} = Z_0 \oplus U_{Y_0} \oplus U_{X_0 Y_0} & f_{Y_1} = Z_1 \oplus U_{Y_1} \oplus U_{X_1 Y_1} \\ \text{(a) } t = 0 & \text{(b) } t = 1 \end{array}$$

Figure 12: SCM definition for Task 2.

functions for X_2 and Y_2 include X_1 as a parent, creating a direct dependency between early decisions and future rewards. This design allows us to test the effectiveness of DUC-based POMIS⁺ strategies in capturing delayed causal influences that are missed by myopic approaches.

Task 3. To clearly illustrate how information from early IB^+ and QIB components can propagate across multiple time steps in the absence of the time-slice Markov assumption (Assumption 3.1), we design an SCM inducing Fig. 6. This structure focuses on exposing the long-range effect of early interventions. The SCM contains three time steps $t = 1, 2, 3$ over variables $\{W_t, X_t, Y_t\}$, with unobserved confounding between X_t and Y_t at each step. A bidirected edge between X_2 and X_3 further emphasizes the departure from the time-slice Markov assumption. This configuration enables us to test whether $POMIS^+$ can successfully capture reward-relevant information originating from earlier steps.

The structural functions f_V for each endogenous variable V are defined as follows:

$$\begin{array}{lll}
f_{W_1} = U_{W_1} & f_{W_2} = U_{W_2} & f_{W_3} = U_{W_3} \\
f_{X_1} = U_{X_1} \oplus W_1 \oplus U_{X_1 Y_1} & f_{X_2} = U_{X_2} \oplus W_2 \oplus U_{X_2 X_3} & f_{X_3} = U_{X_3} \oplus W_3 \oplus U_{X_2 X_3} \\
f_{Y_1} = X_0 \oplus U_{Y_1} \oplus U_{X_1 Y_1} & \oplus U_{X_2 Y_2} \oplus X_1 & \oplus U_{X_3 Y_3} \oplus X_2 \\
& f_{Y_2} = X_2 \oplus U_{Y_2} \oplus U_{X_2 Y_2} & f_{Y_3} = X_3 \oplus U_{Y_3} \oplus U_{X_3 Y_3}
\end{array}
\begin{array}{l}
(a) \ t = 1 \\
(b) \ t = 2 \\
(c) \ t = 3
\end{array}$$

Figure 13: SCM definition for Task 3.

The probability distributions over exogenous variables \mathbf{U} are defined as:

$$\begin{aligned}
P(U_{X_1} = 1) &= 0.82, & P(U_{W_1} = 1) &= 0.15, & P(U_{X_1 Y_1} = 1) &= 0.52, & P(U_{X_2} = 1) &= 0.92, \\
P(U_{Y_2} = 1) &= 0.02, & P(U_{X_2 Y_2} = 1) &= 0.44, & P(U_{X_2 X_3} = 1) &= 0.43, & P(U_{Y_2} = 1) &= 0.01, \\
P(U_{W_2} = 1) &= 0.42, & P(U_{X_3} = 1) &= 0.20, & P(U_{X_3 Y_3} = 1) &= 0.48, & P(U_{Y_3} = 1) &= 0.05, \\
P(U_{W_3} = 1) &= 0.41.
\end{aligned}$$

J Limited Alternative Methods to Approach on NS-SCM-MAB

J.1 Projection-based approach on NS-SCM-MAB

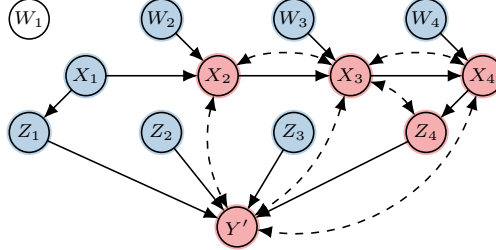


Figure 14: MUCT and IB in the $\mathcal{G}_{[V \setminus Y]}$

Lee and Bareinboim [2019] builds on the framework of Lee and Bareinboim [2018] by relaxing the assumption that every variable in the system must be manipulable, while still operating under a stationary setting. Their method uses latent projections (Verma and Pearl [1990], Verma [1992]) to identify possibly-optimal intervention targets and infer dependency relations among them, even when only partial causal information is available.

In our setting, although the underlying process is non-stationary, the complete causal graph is revealed after all rounds. This allows the latent projection technique to be applied post hoc for identifying possibly-optimal arms. Nevertheless, such an approach overlooks the temporal dynamics inherent to the non-stationary process, thereby failing to capture how specific interventions affect rewards across different time steps. In what follows, we compare this projection-based approach with our method.

First of all, we introduce the projection-based approach. In a stationary setting with non-manipulable variables, Lee and Bareinboim [2019] characterize possibly-optimal arms using the latent projection technique (Verma and Pearl [1990], Verma [1992]). They begin with a causal diagram $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, and define a set of non-manipulative variables $\mathbf{N} \subset \mathbf{V} \setminus \{Y\}$, where Y is the target variable. To construct

the latent projection onto the manipulative variables $\mathbf{V} \setminus \mathbf{N}$, they consider an augmented graph $\hat{\mathcal{G}}$ that explicitly represents unobserved confounders. They initialize a graph $\mathcal{H} = \langle \mathbf{V} \setminus \mathbf{N}, \emptyset \rangle$, then add edges as follows:

1. A directed edge between V_i and V_j if $V_i \rightarrow V_j \in \mathcal{G}$ or there exists a directed path from V_i to V_j where all non-end vertices in the path between there are in \mathbf{N} .
2. A bi-directed edge between V_i and V_j if $V_i \leftrightarrow V_j \in \mathcal{G}$; or there exists directed paths from an unobserved confounder to V_i and V_j in $\hat{\mathcal{G}}$ where all non-end vertices are in \mathbf{N} .

Let $\mathcal{G}_{[\mathbf{V}']}$ denote the causal diagram resulting from projecting \mathcal{G} onto \mathbf{V}' . They prove that $\mathbb{P}_{\mathcal{G}, Y}^{\mathbf{N}} = \mathbb{P}_{\mathcal{H}, Y}$ (i.e., POMISs given $\langle \mathcal{G}, Y, \mathbf{N} \rangle = \text{POMISs given } \langle \mathcal{H}, Y \rangle$) via two propositions, which ensures that the optimality of an arm remains under (i) projection from \mathcal{G} to \mathcal{H} and (ii) the reverse projection.

Proposition J.1 (Causal Identification without Non-manipulative Variables [Lee and Bareinboim, 2019]). *Given an SCM $\mathcal{M}^1 = \mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$, there exists an SCM $\mathcal{M}^2 = \mathcal{M}_{\mathbf{V} \setminus \mathbf{N}} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}', P(\mathbf{U}) \rangle$ such that $P_{\mathbf{x}}^1(\mathbf{y}) = P_{\mathbf{x}}^2(\mathbf{y})$ for any $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{N}$ and $\mathbf{Y} \neq \emptyset$.*

Proposition J.2 (Causal Identification under the Projected Graph [Lee and Bareinboim, 2019]). *Given a causal diagram \mathcal{G} , let $\mathcal{H} = \mathcal{G}_{[\mathbf{V} \setminus \mathbf{N}]}$. For a SCM $\mathcal{M}^1 = \mathcal{M}_{[\mathbf{V} \setminus \mathbf{N}]} = \langle \mathbf{V} \setminus \mathbf{N}, \mathbf{U}, \mathcal{F}', P(\mathbf{U}) \rangle$ conforming to \mathcal{H} , there exists a SCM $\mathcal{M}^2 = \mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ that conforms to \mathcal{G} such that $P_{\mathbf{x}}^1(\mathbf{y}) = P_{\mathbf{x}}^2(\mathbf{y})$, for any $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{N}$ and $\mathbf{Y} \neq \emptyset$.*

Theorem J.1 (POMIS Invariance under Projection). *Given a causal diagram $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, a reward variable $Y \in \mathbf{V}$, and a set of non-manipulable variables $\mathbf{N} \subseteq \mathbf{V} \setminus \{Y\}$, let \mathcal{H} be the projection of \mathcal{G} onto $\mathbf{V} \setminus \mathbf{N}$. Then,*

$$\mathbb{P}_{\mathcal{G}, Y}^{\mathbf{N}} = \mathbb{P}_{\mathcal{H}, Y}$$

where $\mathbb{P}_{\mathcal{G}, Y}^{\mathbf{N}}$ denotes a set of POMISs given $\langle \mathcal{G}, Y, \mathbf{N} \rangle$.

Based on the projection-based approach introduced above, we can derive Fig. 8 via the latent projection. In Fig. 14, we obtain $\text{POMIS} = \{Z_1, X_1, Z_2, W_2, Z_3, W_3, W_4\}$. This result coincides with the outcome of our proposed POMIS^+ -based method (Fig. 8) in terms of maximizing the cumulative reward $Y' = \sum_{t=1}^4 Y_t$. However, since the graph in Fig. 14 is constructed by projecting all temporal models into a single static, stationary structure, it lacks temporal interpretability. We argue that such a projection is insufficient for modeling and analyzing non-stationary bandit problems, where the notion of time is inseparable from the causal dynamics.

Non-stationarity in bandits inherently involves modeling how reward distributions evolve over time, and our method is specifically designed to capture and exploit this temporal evolution. The significant advantage of our framework lies in its ability to identify time-specific intervention sets for each reward variable Y_t . This makes POMIS^+ not only optimal in terms of cumulative reward but also more interpretable and practically applicable to real-world scenarios where intervention constraints or objectives vary over time. Below, we present one example that highlights this property.

Illustrative case: Sequential treatment with risk constraint Suppose each time-indexed intervention variable Z_t, X_t, W_t represents a distinct medication administered at time t :

- Z_t : fast-acting drug that strongly affects immediate health (Y_t) but may induce side effects.
- X_t : slow-acting drug that influences future outcomes (e.g., Y_{t+1}, Y_{t+2}).
- W_t : a supportive drug that amplifies or stabilizes drug effects in the future.

Now consider a scenario where we impose a safety constraint on early reward:

Let $\mathbf{A} = (\{Z_1, X_1\}, \{Z_2, W_2\}, \{Z_3, W_3\}, \{W_4\})$ denote the intervention sequence obtained from the POMIS^+ method (or projection-based method) across all time steps, and let $\mathbf{a}_0 \in \mathcal{D}(\mathbf{A}[0])$ be a joint intervention assignment to the first intervention set.

$$\mathbb{E}[Y_1 \mid \text{do}(\mathbf{a}_0)] \leq \epsilon,$$

to ensure that early treatments do not induce excessive physiological stress. A projection-based method (Fig. 14) optimizes $\mathbb{E}[Y']$ and may select $(Z_1 = 1)$ if it increases Y' overall—despite the fact that Z_1 directly affects Y_1 and may violate the safety constraint.

In contrast, our method (Fig. 8) distinguishes that:

- Z_1 is QIB_1 influencing primarily Y_1 , while
- X_1 is $\text{IB}_{1,4}^+$ that contributes to Y_4 through the path $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots \rightarrow Y_4$.

By leveraging this temporal decomposition, we can selectively intervene on X_1 while avoiding Z_1 , satisfying the constraint on Y_1 and still improving long-term outcomes. Formally, we solve:

$$\mathbf{a}^* = \underset{\substack{\mathbf{a} \in \mathcal{D}(\mathbf{A}) \\ \text{s.t. } \mathbb{E}[Y_1 | \text{do}(\mathbf{a}_0)] \leq \epsilon}}{\arg \max} \mathbb{E} \left[\sum_{t=1}^4 Y_t \mid \text{do}(\mathbf{a}) \right]$$

Summary This example highlights the limitations of projection in non-stationary bandits. While projection-based methods may yield high cumulative rewards, they are blind to when and how specific interventions act. Our POMIS^+ framework enables temporally structured intervention planning, allowing for interpretable, constraint-aware, and sequentially optimal policies. In domains such as medicine or education—where interventions at each stage must consider safety or ethical constraints—such temporal disentanglement is essential.

J.2 Forced stationary approach on NS-SCM-MAB

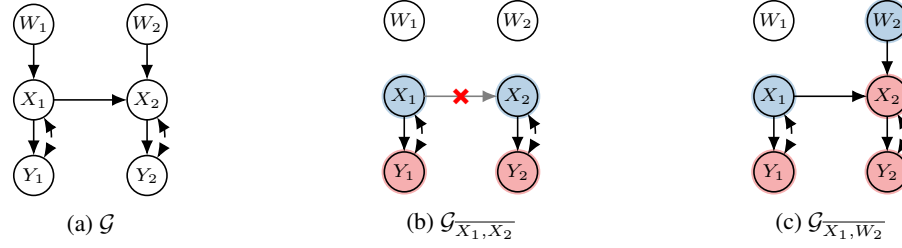


Figure 15: (a) Non-stationary causal diagram \mathcal{G} and (b, c) mutilated graphs

In the main text, we analyzed the non-stationary SCM-MAB setting in which the reward distribution changes over time due to the causal influence of past interventions (see Def. 3.1). While our main algorithmic framework leverages time-specific causal structure, it is also possible to consider an alternative approach: *forced stationarity* via direct intervention. The idea is to identify intervention sets that block the information propagation (see §3) across time slices—thus preventing reward distribution shifts. This allows the agent to reuse previously learned interventional effects, reducing the need to re-identify POMIS sets or re-calculate expected rewards at every time step. Before formalizing this concept, we introduce the notion of transition edge blocking.

Definition J.1 (Block). Let \mathcal{G} be a causal diagram and \mathbf{V}_t the set of endogenous variables at time t . An intervention $\text{do}(\mathbf{X}_t = \mathbf{x}_t)$ is said to block transition edges if, in the mutilated graph $\mathcal{G}_{\mathbf{X}_t}$, all incoming edges from variables in previous time slices $\mathbf{V}_{<t}$ to \mathbf{X}_t are removed. This operation prevents the information propagation from $\mathbf{V}_{<t}$ to \mathbf{V}_t via \mathbf{X}_t .

When such a blocking intervention is applied, the reward variable Y_t becomes conditionally isolated from its historical influences. This motivates the following definition of forced stationarity, grounded in the interventional reward distributions formalized in Def. 3.1

Definition J.2 (Forced Stationarity). Given a non-stationary SCM-MAB $\langle \mathcal{M}, \mathbf{Y} \rangle$, we say that an intervention set $\mathbf{F} \subseteq \mathbf{V}_t$ induces forced stationarity over reward variable Y_t if for every $t < t'$, and for every $\mathbf{f} \in \mathcal{D}(\mathbf{F})$,

$$P(Y_t \mid \text{do}(\mathbf{f}), \mathbb{1}_{t>1} \cdot I_{1:t-1}) = P(Y_{t'} \mid \text{do}(\mathbf{f}), \mathbb{1}_{t'>1} \cdot (I_{1:t-1} \cup I_{t:t'-1})),$$

In other words, the reward distribution remains invariant over time under repeated interventions on \mathbf{F} .

This formalization allows us to distinguish between *actual* stationarity (inherent in the SCM) and stationarity that is induced by intervention. The following lemma clarifies the relationship between blocking transition edges and enforcing reward stationarity.

Lemma J.1 (Blocking transition edges induces forced stationarity). *Let \mathcal{G} be a causal diagram, and suppose an intervention on $\{Z_t\}$ blocks the transition edge $Z_{t-1} \rightarrow Z_t$. Then, we have:*

$$P(Y_{t-1} \mid \text{do}(Z_{t-1}), \mathbb{1}_{t>1} \cdot I_{1:t-2}) = P(Y_t \mid \text{do}(Z_t), \mathbb{1}_{t>1} \cdot I_{1:t-1}).$$

Therefore, the intervention enforces reward stationarity across time.

Proof. By intervening on Z_t , we remove the transition edge from Z_{t-1} to Z_t (Def. J.1), thereby preventing any causal effect from prior interventions from propagating to Z_t . As a result, Y_t under $\text{do}(Z_t)$ becomes conditionally independent of earlier actions, and its interventional distribution same as that of Y_{t-1} under $\text{do}(Z_{t-1})$. \square

The idea of forced stationarity provides a mechanism to simplify learning in non-stationary SCM-MABs: if one can construct a policy that always intervenes on such blocking variables (e.g., Z_t), the resulting reward distribution no longer shifts over time. This allows the agent to reuse previously learned information without recalculating POMISs at each step.

However, such interventions may reduce the optimality of the resulting policy. Although reward distributions may appear stationary under intervention \mathbf{F} , the values themselves may not be maximized. Blocking information from previous time slices may prevent the agent from exploiting long-term causal pathways that yield higher rewards.

To illustrate this trade-off, consider the causal graphs in Fig. 15. In Fig. 15(b), we apply an intervention on X_2 , which removes the incoming edge from X_1 to X_2 , thereby blocking temporal information transfer and enforcing stationarity. While this simplifies the learning problem, it overlooks alternative intervention options. In Fig. 15(c), we do not intervene on X_2 , but instead on W_2 —a variable that is structurally valid and potentially in maximizing reward, even though it does not block the transition path.

This example reveals an important limitation of the forced stationarity paradigm: sets such as $\{X_2\}$ and $\{W_2\}$ both are interventional border (IB), yet only X_2 induces stationarity when intervened. If one naively prioritizes only those IB variables that block transition edges, the policy may become overly myopic and ignore possibly higher-rewarding options. In contrast, our framework (i.e., POMIS⁺) retains the full temporal structure and dynamically adapts to changes in reward distribution without forcibly blocking information. This enables both fine-grained interpretability and optimal intervention selection over time.

While the limitations of forced stationarity are evident in scenarios where blocking discards useful long-term causal dependencies, we emphasize that this strategy is not inherently flawed. In fact, under certain structural conditions, forced stationarity may serve as a practical approach to simplifying learning in temporally complex environments. For example, if the reward variable Y_t consistently depends on a repeated set of parent variables over time—such as through recurring transition edges—then blocking those transitions can yield invariant reward distributions without sacrificing performance.

In such cases, domain knowledge or structural priors can inform targeted interventions that induce stationarity while still capturing the most influential causal paths. This highlights a broader point: rather than rejecting forced stationarity entirely, it can be selectively applied when supported by sufficient knowledge about the system’s temporal regularities.

K More Discussions

K.1 Relaxing Natural Characteristics on Temporal Structure

Our main theoretical development naturally assumes both an identical graphical structure across time slices and the existence of transition edges between them. These assumptions are introduced to reflect natural characteristics of non-stationary environments and to facilitate tractable inference. Importantly, the assumption of identical structure is not essential to the correctness of our algorithm: even when the assumption is relaxed, our method remains sound and complete, provided that the full temporal causal graph is accessible.

In contrast, the existence of transition edges plays a more critical role. Without such edges, causal information cannot propagate across time, and the algorithm may fail to identify relevant variables at earlier steps, limiting its ability to construct non-myopic intervention strategies.

- The property of identical time slices reflects the natural structure of stationary bandit models (see, Fig. 1(b)), where variables and their dependencies are replicated across time for each time slice. While this natural assumption simplifies the search space by enabling repeated reasoning patterns, which is a characteristic of bandit settings, it is not required for the correctness of the graphical analysis or the definitions of IB^+ and QIB.
- The existence of transition edges between time slices enables information to propagate forward in time, allowing the algorithm to identify earlier variables that support future interventions. Without transition edges, information cannot propagate across time, preventing the algorithm from identifying long-range dependencies or constructing effective non-myopic strategies. More concretely, certain paths to earlier supporting variables (i.e., IB^+) become unreachable, limiting the algorithm’s ability to exploit temporal structure for long-term optimization.

This flexibility ensures that our framework is broadly applicable, even in domains where temporal structure varies over time. We emphasize that these assumptions serve as modeling conveniences that facilitate analysis and interpretation, rather than strict algorithmic requirements.

K.2 Non-Stationarity in Causal Bandits vs. Causality in Non-Stationary Bandits

It is useful to distinguish our formulation from alternative approaches that combine causality and non-stationarity in opposite directions. Our work builds upon structural causal bandits (SCM-MABs) [Lee and Bareinboim, 2018, 2019], which assume a fixed underlying causal graph and utilize it to guide interventions. We extend this line of work by introducing non-stationarity at the level of the causal structure itself—modeling how causal dependencies and reward mechanisms evolve over time. In this sense, we develop a causal bandit framework that internalizes non-stationarity as a structural property.

By contrast, recent efforts that introduce causal reasoning into non-stationary bandit settings typically begin with traditional statistical formulations—such as sliding-window or change-point models—and incorporate causal estimators to correct for confounding or adapt to changing environments (details are available in Appendix E). For instance, Huang and Wu [2024] propose a method for handling confounded and selection-biased offline data by deriving robust causal bounds for each arm. Similarly, Nourani-Koliji et al. [2023] address piecewise-stationary bandits with causally related rewards by detecting changes in both reward distributions and causal structure using a Generalized Likelihood Ratio (GLR)-based change-point detector. In these settings, causality is layered onto a fundamentally statistical treatment of temporal change, often without structural assumptions about the underlying dynamics.

The distinction is more than philosophical. Our model-based approach enables:

- Reasoning about how non-stationarity arises (via temporal models and transition edges),
- Derivation of non-myopic intervention strategies through temporally-aware graphical structures ($POMIS^+$), and
- Theoretical guarantees grounded in structural assumptions.

In contrast, approaches based on statistical modeling typically emphasize adaptation—e.g., through sliding windows, discounting, or change-point detection—without modeling the underlying structural mechanisms that generate non-stationarity. These methods act as black-box in the sense that they detect shifts or trends in the data but do not explain them in terms of causal relationships or system dynamics.

K.3 $POMIS^+$ Sequences as Composition of Possibly Optimal Intervention Sets within Temporal Models

An intervention sequence and the causal responses of each time-indexed variable determine the expected cumulative reward in a sequential decision problem. In our framework, this is formal-

ized by an SCM \mathcal{M} together with a sequence of possibly-optimal intervention sets—namely, POMIS^+ —identified across time.

While SCM-level definitions of POMIS (cf. Def. B.2) evaluate optimality based on the existence of at least one SCM where a given intervention set performs optimally, temporal decision-making imposes a stricter constraint: each reward Y_t must be evaluated in the context of prior interventions. This context naturally gives rise to a family of temporal models $\mathcal{M}_t \mid \mathbf{v}_t^*$, where \mathbf{v}_t^* represents the predetermined values induced by earlier actions. Thus, the optimal intervention at each time step may be determined within this temporal model, reflecting only the information propagation available at time t .

In this subsection, we show that the expected cumulative reward under a full SCM can be equivalently decomposed into a sum of expected rewards under temporal models. Each such term corresponds to an intervention on a POMIS^+ set at time t , selected with respect to $\mathcal{M}_t \mid \mathbf{v}_t^*$. The result is formalized in Prop. K.1 which clarifies how the temporal roll-out of POMIS^+ constructs a compositionally valid and possibly-optimal global intervention plan.

Proposition K.1 (Composition of Temporal Optima). *Let \mathcal{M} be an SCM conforming to \mathcal{G} , and let $(\mathbf{X}_1, \dots, \mathbf{X}_T)$ be an optimal intervention sequence with realizations $\mathbf{x}_1^*, \dots, \mathbf{x}_T^*$. Let \mathbf{v}_t^* denote the predetermined variables at time t given past interventions. Suppose each intervention set \mathbf{X}_t is constructed as $\text{POMIS}_t \cup \mathbf{W}_t$, where $\mathbf{W}_t \subseteq \text{Pa}(\mathbf{V}_t)$ satisfies the condition in Thm. 5.1. Then:*

$$\mathbb{E}^{\mathcal{M}} \left[\sum_{t=1}^T Y_t \mid \text{do}(\mathbf{x}_1^*, \dots, \mathbf{x}_T^*) \right] = \sum_{t=1}^T \mathbb{E}^{\mathcal{M}_t \mid \mathbf{v}_t^*} [Y_t \mid \text{do}(\mathbf{x}_t^*)]$$

Proof. We begin by expanding the expected cumulative reward under the full SCM \mathcal{M} :

$$\mathbb{E}^{\mathcal{M}} \left[\sum_{t=1}^T Y_t \mid \text{do}(\mathbf{x}_1^*, \dots, \mathbf{x}_T^*) \right] = \sum_{t=1}^T \mathbb{E}^{\mathcal{M}} [Y_t \mid \text{do}(\mathbf{x}_1^*, \dots, \mathbf{x}_T^*)].$$

For each $t \in [T]$, let \mathbf{v}_t^* denote the predetermined assignment to $\text{Pa}(\mathbf{V}_t)$ induced by prior interventions $\text{do}(\mathbf{x}_1^*, \dots, \mathbf{x}_{t-1}^*)$. Then the reward at time t can be equivalently computed in the conditioned temporal model $\mathcal{M}_t \mid \mathbf{v}_t^*$ as:

$$\mathbb{E}^{\mathcal{M}} [Y_t \mid \text{do}(\mathbf{x}_1^*, \dots, \mathbf{x}_T^*)] = \mathbb{E}^{\mathcal{M}_t \mid \mathbf{v}_t^*} [Y_t \mid \text{do}(\mathbf{x}_t^*)].$$

Applying this to each t and summing, we obtain:

$$\mathbb{E}^{\mathcal{M}} \left[\sum_{t=1}^T Y_t \mid \text{do}(\mathbf{x}_1^*, \dots, \mathbf{x}_T^*) \right] = \sum_{t=1}^T \mathbb{E}^{\mathcal{M}_t \mid \mathbf{v}_t^*} [Y_t \mid \text{do}(\mathbf{x}_t^*)],$$

as claimed. \square

This proposition formalizes how the cumulative reward in a full SCM can be decomposed into a sequence of rewards, each governed by a temporal model conditioned on prior interventions (i.e., predetermined values). It supports the view that the POMIS^+ sequence acts as a possibly-optimal plan across time, even though the optimality of individual interventions may depend on specific prior values.

While each temporal model $\mathcal{M}_t \mid \mathbf{v}_t^*$ determines the optimal intervention at time t under a fixed history of prior interventions, the globally optimal sequence across the entire SCM must be selected jointly, since each temporal model depends on the conditioning induced by earlier decisions.

Corollary K.1 (Local Optimality Does Not Imply Global Optimality). *There exists an SCM \mathcal{M} and intervention sequences $(\mathbf{X}_1, \dots, \mathbf{X}_T)$ such that for each t , \mathbf{X}_t is a possibly-optimal intervention set in the temporal model $\mathcal{M}_t \mid \mathbf{v}_t^*$, but the joint sequence $(\mathbf{X}_1, \dots, \mathbf{X}_T)$ is not globally optimal for $\mathbb{E}^{\mathcal{M}} \left[\sum_{t=1}^T Y_t \right]$.*

This corollary illustrates that local optimality under temporal models does not guarantee global optimality in the full SCM. Due to the interdependencies between time steps, early interventions may have long-term consequences that are not captured by myopic optimization. Therefore, joint planning across time is necessary to achieve globally optimal decisions.