

Spatio-Temporal Hierarchical Causal Models

Xintong Li¹, Haoran Zhang¹, Xiao Zhou^{2,3,4*}

¹School of Statistics, Renmin University of China Beijing, China

²Gaoling School of Artificial Intelligence, Renmin University of China Beijing, China

³Beijing Key Laboratory of Research on Large Models and Intelligent Governance

⁴Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
{lixintong, zhanghaoran1015, xiaozhou}@ruc.edu.cn

Abstract

The abundance of fine-grained spatio-temporal data, such as traffic sensor networks, offers vast opportunities for scientific discovery. However, inferring causal relationships from such observational data remains challenging, particularly due to unobserved confounders that are specific to units (e.g., geographical locations) yet influence outcomes over time. Most existing methods for spatio-temporal causal inference assume that all confounders are observed, an assumption that is often violated in practice. In this paper, we introduce *Spatio-Temporal Hierarchical Causal Models (ST-HCMs)*, a novel graphical framework that extends hierarchical causal modeling to the spatio-temporal domain. At the core of our approach is the *Spatio-Temporal Collapse Theorem*, which shows that a complex ST-HCM converges to a simpler flat causal model as the amount of subunit data increases. This theoretical result enables a general procedure for *causal identification*, allowing ST-HCMs to recover causal effects even in the presence of unobserved, time-invariant unit-level confounders, a scenario where standard non-hierarchical models fail. We validate the effectiveness of our framework on both synthetic and real-world datasets, demonstrating its potential for robust causal inference in complex dynamic systems.

Code — <https://github.com/CAMELLIAxt/ST-HCMs>

1 Introduction

Fine-grained spatio-temporal data are becoming ubiquitous, generated by sources ranging from satellite imagery in environmental science to sensor networks in urban planning (Wang et al. 2018; Xu and Zhou 2024; Ali et al. 2024; Dong et al. 2024; Li, Li, and Zhou 2025). Such data inherently possess a hierarchical structure, offering unprecedented opportunities to uncover causal mechanisms in complex dynamic systems. However, reliable causal inference from observational data remains challenging (Hernán and Robins 2010; Reich et al. 2021), obscuring mechanistic insights and impeding accurate policy intervention predictions in complex spatio-temporal systems like urban operations (Zhang and Ning 2023; Dong, Zhang, and Zhang 2025; Xie et al. 2025a).

Consider the urban traffic system in Figure 1, where a city is divided into several regions (units), and within each region, numerous sensors monitor traffic flow (outcome) on

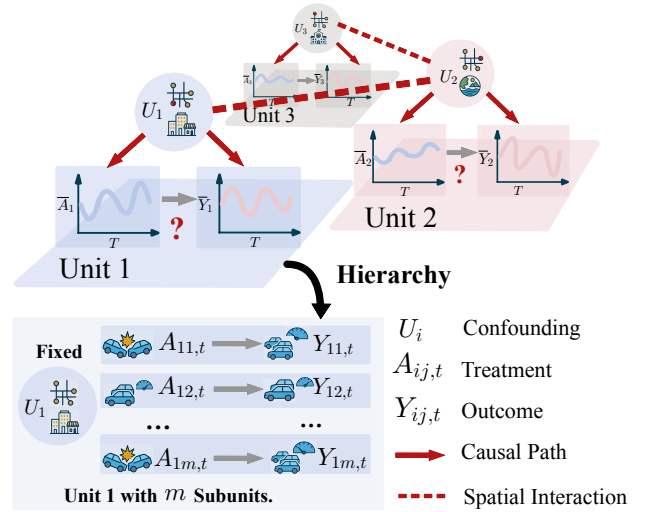


Figure 1: An example of hierarchical causality in traffic.

specific road segments (subunits). A traffic accident (treatment) can disrupt the flow, but a region’s unique characteristics, such as its geographical layout and road network structure, serve as an **unobserved confounder**. This confounder affects both the likelihood of accidents and typical traffic patterns, making it difficult to disentangle the true causal effect of an accident from the region’s inherent characteristics (Weinstein and Blei 2024; Clarke and Polselli 2025). The problem is further compounded by temporal dynamics and spatial spillovers from neighboring regions, which introduces additional layers of complexity (Robins 1986; Aronow and Samii 2017; Reich et al. 2021).

This hierarchical structure offers a powerful, yet often underutilized, path forward. Standard analytical methods often bypass this structure, instead relying on the strong “no unmeasured confounding” assumption (Hernán and Robins 2006; Ali, Faruque, and Wang 2024; Oprescu et al. 2025), which is untenable for complex systems like the one in Figure 1. While instrumental variables (IV) provide a theoretical alternative (Martinussen et al. 2017; Cheng et al. 2024), finding a valid instrument that is both relevant and exogenous is notoriously difficult in practice. Methods like fixed-effect (FE) models do leverage within-unit variation (An-

*Corresponding author.

grist and Pischke 2009; Jensen, Burroni, and Rattigan 2020), but they typically impose restrictive parametric assumptions, such as linearity, limiting their applicability.

In contrast, our approach builds on the key insight illustrated in Figure 1: we treat each unit as a local natural experiment, enabling non-parametric estimation of region-specific causal effects by analyzing the relationship between treatments and subunit-level outcomes. This paper extends this powerful principle to spatio-temporal settings, where such effects are confounded not only by static heterogeneity but also by dynamic evolution and spatial interactions.

While the idea of leveraging hierarchical structure for causal inference has been explored in static settings (Weinstein and Blei 2024), its extension to the spatio-temporal domain presents non-trivial. The introduction of time and space creates two fundamental challenges: (1) **dynamic confounding**, where the history of the system acts as a time-varying confounder for future treatments and outcomes, and (2) **spatial confounding**, where the states of neighboring units interfere with each other’s causal processes. To address these challenges, we introduce **Spatio-Temporal Hierarchical Causal Models (ST-HCMs)**, a novel graphical framework that explicitly models these dependencies. The core of our solution is a **Spatio-Temporal Collapse Theorem**, which provides a theoretical guarantee that the complex, dynamic ST-HCM mathematically converges to a simpler, equivalent representation. This theorem is the key that enables causal identification, allowing us to disentangle effects from the web of spatio-temporal confounding.

The contributions of this paper are as follows:

- **Conceptually**, we propose the first graphical framework (ST-HCMs) for causal modeling of nested spatio-temporal data, providing a formal formulation to reason about causality within spatio-temporal systems.
- **Theoretically**, we prove a Spatio-Temporal Collapse Theorem (Theorem 2) that guarantees a complex ST-HCM converges to a simpler, equivalent model, which in turn enables causal identification (Theorem 3,4) despite unobserved confounders.
- **Experimentally**, we validate our framework through extensive simulations, showing that it robustly recovers causal effects despite unobserved confounders and spatial spillovers. On a real-world dataset, we further demonstrate that it substantially reduces the biases inherent to traditional non-hierarchical models.

2 Preliminaries

Notation Consider a system composed of N units, indexed by $i \in \{1, \dots, N\}$, observed over T discrete time steps, indexed by $t \in \{1, \dots, T\}$. Each unit i contains m_i subunits, indexed by $j \in \{1, \dots, m_i\}$. For simplicity, we assume $m_i = m$ for all i . Let V be a finite index set for endogenous variable types. This set is partitioned into subunit-level variables S and unit-level variables $U = V \setminus S$.

We define the key variables as follows:

- $X_{ij,t}^v$: The value of a subunit-level variable $v \in S$ for subunit j in unit i at time t .

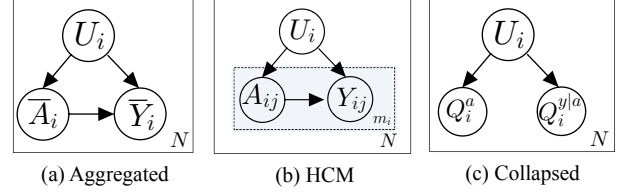


Figure 2: Hierarchy and Collapsing in HCMs.

- $X_{i,t}^w$: The value of a unit-level variable $w \in U$ for unit i at time t .
- $\mathbf{X}_{i,<t}$: The full history of all variables associated with unit i prior to time t .
- $\mathcal{N}(i)$: The set of spatial neighbors of unit i .
- $\mathcal{X}$: A generic state space. $\mathcal{P}(\mathcal{X})$ denotes the space of all probability measures on $\mathcal{X}$.
- U_i : A time-invariant, unobserved random variable representing the static latent properties of unit i .
- $\gamma_{i,t}^v, \epsilon_{ij,t}^v$: Time-varying exogenous noise variables at the unit and subunit levels, respectively.

Causal Graphical Models and HCMs A Causal Graphical Model (CGM) represents causal relationships as a Directed Acyclic Graph (DAG), where nodes are variables and directed edges denote causal effects. Interventions are denoted by the do-operator (Pearl 2009).

The static Hierarchical Causal Model (HCM) framework (Weinstein and Blei 2024) is designed to overcome confounding in nested data. As illustrated in Figure 2(a), in a standard aggregated view, an unobserved confounder U_i biases the estimated effect of treatment A_i on outcome Y_i . The key insight of HCM is to leverage the fine-grained subunit data within each unit, as depicted in Figure 2(b). This hierarchical structure enables a powerful identification strategy known as collapsing. The procedure transforms the HCM into an equivalent, flat causal graph (Figure 2(c)), where the subunit-level causal mechanisms are succinctly captured by new random variables, termed Q-variables (e.g., $Q_i^a, Q_i^{y|a}$). However, the static HCM is limited to cross-sectional settings and does not account for temporal dynamics or spatial dependencies, which are essential in spatio-temporal domains. Our work extends this collapsing principle to the dynamic spatio-temporal domain.

3 Spatio-Temporal Hierarchical Causal Models

In this section, we introduce our Spatio-Temporal Hierarchical Causal Model (ST-HCM) framework. To build intuition, we first present a temporal extension of the static HCM, the T-HCM. We then incorporate spatial dependencies to formulate the ST-HCM and present the convergence theorem.

3.1 A Temporal Extension: T-HCMs

Definition 1 (Temporal Hierarchical Structural Causal Model (T-HSCM)). A T-HSCM, $\mathcal{M}_{t-hscm}$, is defined by: (1) a summary causal graph $\mathcal{G}_T$ whose unrolling over

time yields a DAG; (2) a set of static exogenous variables $U_i \sim P(U)$; (3) sets of time-varying exogenous noises $\{\gamma_{i,t}^v\}, \{\epsilon_{ij,t}^v\}$; and (4) a set of deterministic mechanism functions $\{f^v\}_{v \in V}$. The value of each endogenous variable is generated as follows:

- For a subunit-level variable $v \in S$:

$$x_{ij,t}^v = f^v(U_i, pa_U(x_{i,t}^v), pa_S(x_{ij,t}^v), \mathbf{X}_{i,<t}, \gamma_{i,t}^v, \epsilon_{ij,t}^v).$$

- For a unit-level variable $w \in U$:

$$x_{i,t}^w = f^w(U_i, pa_U(x_{i,t}^w), \{\{x_{ij,t}^{v'}\}_{v' \in pa_S(w,t)}\}_{j=1}^m, \mathbf{X}_{i,<t}, \gamma_{i,t}^w),$$

where the mechanism f^w is permutation-invariant with respect to the subunit index j .

By integrating out the exogenous noises from the structural model, we obtain the T-HCGM. This graphical model is described by stochastic mechanisms, centered on the dynamic evolution of Q-variables, which represent the conditional distributions governing subunit-level phenomena.

Definition 2 (Temporal Hierarchical Causal Graphical Model (T-HCGM)). A T-HCGM, $\mathcal{M}_{t-hcgm}$, corresponding to a T-HSCM, is characterized by a two-stage generative process at each time step t :

1. **Q-variable Evolution:** A vector of Q-variables $\mathbf{Q}_{i,t}$ is generated for each unit i , conditional on its macro-history: $\mathbf{Q}_{i,t} \sim p(\mathbf{q}_t | U_i, \mathbf{S}_{i,<t})$ where $\mathbf{S}_{i,<t} = (X_{i,<t}^U, \mathbf{Q}_{i,<t})$ is the macro-history.
2. **Variable Generation:** Endogenous variables are drawn conditional on the current Q-variables and their history.

To prove the convergence of this model to a simpler, collapsed representation, we require the following assumptions.

Assumption 1 (Mechanism Convergence). For any $w \in U$, its conditional distribution's generating mechanism is continuous with respect to the empirical measure of its subunit-level parent histories. Let $\mathcal{P}_{i,<t}$ be the true probability measure on the subunit history space, determined by the macro-history $\text{Hist}_{i,<t}^{\text{col}} = (U_i, \mathbf{S}_{i,<t})$. Let $\hat{\mathcal{P}}_{i,m,<t}$ be the empirical measure from m subunits. We assume:

$$\lim_{m \rightarrow \infty} \mathbb{E}_{\hat{\mathcal{P}}_{i,m,<t}} [\text{KL}(p(X_{i,t}^w | \mathbf{S}_{i,<t}, \mathcal{P}_{i,<t}) || p(X_{i,t}^w | \mathbf{S}_{i,<t}, \hat{\mathcal{P}}_{i,m,<t}))] = 0. \quad (1)$$

This assumption formalizes the idea that with enough subunit data, the uncertainty from sampling subunit histories becomes negligible for determining unit-level outcomes.

Assumption 2 (Regularity of State Spaces). The state spaces for all variables are Polish spaces. Furthermore, the class of functions defining subunit histories forms a Glivenko-Cantelli class, ensuring the uniform convergence of empirical measures.

This is a standard technical condition for ensuring the regularity of the underlying probability spaces (Hernán and Robins 2006; Ali, Faruque, and Wang 2024).

Theorem 1 (Convergence of T-HCMs). Under Assumptions 1 and 2, as the number of subunits $m \rightarrow \infty$, the joint distribution of the macro-state history from a T-HCGM, marginalized over subunit variables ($P_{m,T}^{\text{marg}}$), converges in Kullback-Leibler divergence to the distribution of its corresponding Dynamic Collapsed Model ($P_{\text{col},T}$). For any finite time T :

$$\lim_{m \rightarrow \infty} \text{KL}(P_{\text{col},T} || P_{m,T}^{\text{marg}}) = 0. \quad (2)$$

Proof Sketch. The proof proceeds by mathematical induction on the time step t , using the chain rule: $\text{KL}(P_t || Q_t) = \text{KL}(P_{t-1} || Q_{t-1}) + \mathbb{E}_{P_{t-1}} [\text{KL}(P(S_t | H_{t-1}) || Q(S_t | H_{t-1}))]$. The inductive step shows that the expected KL-divergence of the one-step transition kernels converges to zero. We decompose this inner KL term into two parts: one for the Q-variable evolution and one for the unit-level variable generation. The KL-divergence for the Q-variable part is identically zero by definition, as the collapsed model preserves this mechanism exactly. The KL-divergence for the unit-level variable part converges to zero due to a continuous mapping argument: Assumption 2 ensures the convergence of the input, and Assumption 1 ensures the continuity of the function. Thus, the output distributions converge. The full proof is provided in Appendix A. $\square$

3.2 The Framework: ST-HCMs

We now incorporate spatial dependencies to construct the full Spatio-Temporal Hierarchical Causal Model. This introduces new challenges, primarily the symmetric, contemporaneous interactions between units, which can create cycles in the causal graph. We address this by imposing a causal ordering on spatial interactions.

Definition 3 (Spatio-Temporal Hierarchical Causal Graphical Model (ST-HCGM)). An ST-HCGM, $\mathcal{M}_{st-hcgm}$, extends the T-HCGM by allowing the generating mechanisms to depend on the states of spatial neighbors $\mathcal{N}(i)$.

- The Q-variable evolution for unit i now depends on its own macro-history and that of its neighbors:

$$\mathbf{Q}_{i,t} \sim p(\mathbf{q}_t | U_i, \mathbf{S}_{i,<t}, \{\mathbf{S}_{k,<t}\}_{k \in \mathcal{N}(i)}). \quad (3)$$

- The generation of unit-level variables $X_{i,t}^w$ similarly depends on the history of unit i and its neighbors $\mathcal{N}(i)$.

To ensure the model is well-defined, we introduce the following assumptions regarding its spatial structure.

Assumption 3 (Spatial Markov Property (S1)). The state of any unit i at time t is conditionally independent of all non-neighboring units, given the history of unit i and its designated neighbors $\mathcal{N}(i)$.

Assumption 4 (Spatial Causal Ordering (S2)). There exists a known, global ordering $<$ on the set of all units $\{1, \dots, N\}$. For any two neighbors $i, j \in \mathcal{N}(i)$, contemporaneous causal influence is unidirectional: unit j can influence unit i at time t only if $j < i$.

Assumption 5 (Spatial Homogeneity (S3)). All units share the same set of mechanism functions $\{f^v\}$. The functional form of spatial influence is invariant across space, depending only on the relative relationship between units, not their absolute identities.

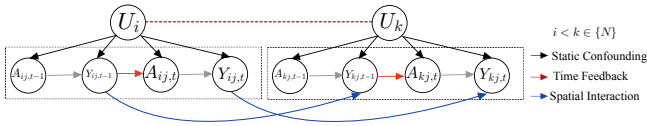


Figure 3: The challenge of identifiability.

With these assumptions in place, we can state the main convergence theorem for our full framework.

Theorem 2 (Convergence of ST-HCMs). *Under Assumptions 1, 2, and 3-5, as the number of subunits $m \rightarrow \infty$, the joint distribution of the macro-state history from an ST-HCGM, marginalized over subunit variables ($P_{m,T,N}^{marg}$), converges in Kullback-Leibler divergence to the distribution of its corresponding Spatio-Temporal Dynamic Collapsed Model ($P_{col,T,N}$). For any finite N and T :*

$$\lim_{m \rightarrow \infty} \text{KL}(P_{col,T,N} \parallel P_{m,T,N}^{marg}) = 0. \quad (4)$$

Proof Sketch. We construct a single “Super-Unit” model, $\mathcal{M}_{super}$, whose state at time t is the concatenated vector of states of all N spatial units, i.e., $\mathbf{S}_{super,t} = (\mathbf{S}_{1,t}, \dots, \mathbf{S}_{N,t})$. Assumption 4 guarantees that the internal dependencies within this Super-Unit’s state vector at any time t are acyclic. Consequently, $\mathcal{M}_{super}$ is a valid (though high-dimensional) T-HCM that satisfies all premises of Theorem 1. Since the collapsing process for $\mathcal{M}_{super}$ is mathematically equivalent to the collapsing process for the original ST-HCM, the convergence guaranteed by Theorem 1 for $\mathcal{M}_{super}$ directly implies the convergence for the ST-HCM. The full proof is provided in Appendix B. $\square$

4 Identification and Estimation

Having established definition and convergence properties for ST-HCMs, we now address its primary purpose: enabling causal inference from spatio-temporal data. This section tackles identifiability, determining when causal effects can be uniquely computed from observational data. We develop sufficient conditions for identifiability in ST-HCMs.

4.1 The Identifiability Problem in Dynamic HCMs

The fundamental challenge of causal inference in dynamic hierarchical systems is threefold (Figure 3). First, we face unobserved unit-level confounding, represented by variable U_i . Second, the system’s history acts as a confounder for current treatments and future outcomes. Third, neighboring units’ states can confound effects, creating spatial dependencies that traditional methods cannot address.

Our goal is to identify the full post-intervention distribution $P(Y_{k,t} | \text{do}(A_{i,t'} = a^*))$, which describes the outcome at a location k and time t under an intervention on a treatment at location i and time t' . Thanks to the convergence theorems (Theorems 1 and 2), this problem is equivalent to identifying the effect in the corresponding Dynamic Collapsed Model (DCM). Our identifiability results are derived by analyzing the graphical structure of this DCM.

4.2 Identifiability via Sequential Adjustability

Our first identification strategy relies on the principle of adjusting for a sufficiently rich set of observed covariates to block all backdoor paths between treatment and outcome. We begin by stating the result for the purely temporal case, which forms the foundation for the more general theorem.

Proposition 1 (Identifiability in T-HCMs via Adjustability). *In a T-HCGM satisfying the premises of Theorem 1, the causal effect of a subunit-level treatment $A_{ij,t'}$ on a future outcome $Y_{ij,t}$ ($t \geq t'$) is identifiable if, in its corresponding DCM, all backdoor paths from the treatment node $Q_{i,t'}^A$ to the outcome node $Q_{i,t}^Y$ are blocked by conditioning on the observable macro-history $\mathbf{S}_{i,<t'}$.*

Proof Sketch. The proof operates on the DCM. Conditioning on the observable history $\mathbf{S}_{i,<t'}$ blocks all confounding paths originating from the system’s dynamic evolution. The only remaining backdoor path is the static confounding path $Q_{i,t'}^A \leftarrow U_i \rightarrow Q_{i,t}^Y$. This path is resolved by the hierarchical structure itself: we first estimate the conditional effect within each unit i (where U_i is fixed), and then average these estimates across all units. This two-step procedure of sequential adjustment followed by hierarchical averaging allows for identification. The full proof is in Appendix C.1. $\square$

This principle naturally extends to the full spatio-temporal setting, where the adjustment set must also include the history of spatial neighbors.

Theorem 3 (Identifiability in ST-HCMs via Adjustability). *In an ST-HCGM satisfying the premises of Theorem 2, the causal effect $P(Y_{k,t} | \text{do}(A_{i,t'} = a^*))$ is identifiable if, in its ST-DCM, all backdoor paths from the treatment $Q_{i,t'}^A$ to the outcome $Q_{k,t}^Y$ are blocked by conditioning on the observable spatio-temporal history. This history comprises the macro-states of unit i and its neighbors $\mathcal{N}(i)$ prior to time t' .*

Proof Sketch. The logic directly parallels that of Proposition 1. The adjustment set is expanded to include the histories of neighbors, $\{\mathbf{S}_{j,<t'}\}_{j \in \mathcal{N}(i)}$, to block confounding paths that traverse space. After this spatio-temporal adjustment, the remaining confounding is purely from the static, unit-level variables $\{U_k\}$. As before, this static confounding is handled by leveraging the subunit-level variation within each unit to estimate conditional effects, which are then averaged. The full proof is in Appendix C.3. $\square$

4.3 Identifiability via Dynamic Instruments

When a sufficiently rich history is not observable, we can achieve identification by leveraging an exogenous source of variation—an instrumental variable.

Proposition 2 (Identifiability in T-HCMs via Instruments). *In a T-HCGM, the causal effect of $A_{ij,t'}$ on $Y_{ij,t}$ is identifiable if there exists a valid subunit-level instrumental variable $Z_{ij,t'}$. A valid instrument must (1) be a cause of $A_{ij,t'}$ (relevance) and (2) have no unblocked backdoor path to $Y_{ij,t}$ in the DCM that does not pass through $A_{ij,t'}$.*

Proof Sketch. A valid instrument $Z_{ij,t'}$ in the DCM, represented by $Q_{i,t'}^Z$, is d-separated from all confounders, including the static U_i and any unobserved history. This allows us to formulate a Fredholm integral equation of the first kind: $P(Y_t|Z_{t'} = z) = \int P(Y_t|\text{do}(A_{t'} = a))P(A_{t'} = a|Z_{t'} = z)da$. Since the terms involving the observational distribution are identifiable, the causal effect $P(Y_t|\text{do}(A_{t'} = a))$ can be recovered by solving this integral equation. The full proof is in Appendix C.2. $\square$

This result also generalizes to the spatio-temporal framework, requiring a correspondingly stronger instrument.

Theorem 4 (Identifiability in ST-HCMs via Instruments). *In an ST-HCGM, the causal effect $P(Y_{k,t}|\text{do}(A_{i,t'} = a^*))$ is identifiable if there exists a valid spatio-temporal subunit-level instrumental variable $Z_{ij,t'}$. A valid instrument must satisfy the exclusion restriction with respect to the entire system: its node $Q_{i,t'}^Z$ in the ST-DCM must be d-separated from all static confounders $\{U_k\}$ and the complete, unobserved history of the entire spatio-temporal system.*

Proof Sketch. The proof is identical in structure to that of Proposition 2. The instrument’s exogenous variation allows us to establish the same integral equation relating the observational distribution to the desired causal effect. The key difference lies in the strength of the required assumption: the instrument must be exogenous not just to the history of its own unit, but to the history of the entire interacting system. Given such an instrument, identification follows. The full proof is in Appendix C.4. $\square$

4.4 ATE with Time-Invariant Spatial Confounding

To demonstrate our framework’s practical power, we focus on a canonical yet challenging problem in real-world applications: **estimating average treatment effects with unobserved, time-invariant spatial confounding**. This scenario is particularly insightful because traditional spatio-temporal methods often fail, while our hierarchical structure provides a unique solution.

Problem Formulation We consider an ST-HCM where the primary challenge is an unobserved, time-invariant, unit-level confounder, U_i . This confounder affects the treatment assignment, outcomes, and potentially the dynamics of neighboring units through spatial spillover, creating complex confounding across both time and space.

Our goal is to identify and estimate the **Average Treatment Effect (ATE)** of a global, static treatment policy. Specifically, we are interested in the effect on the outcome at a future time horizon T , resulting from setting the subunit-level treatment A to a constant value a^* for all units i and all time steps $t \leq T$. The formal estimand is:

$$\text{ATE}_T = \mathbb{E}[Y_T|\text{do}(A_t = 1)] - \mathbb{E}[Y_T|\text{do}(A_t = 0)], \quad (5)$$

where $\mathbb{E}[Y_T|\text{do}(A_t = a^*)]$ denotes the expected global average outcome at time T under the specified policy.

Identification Strategy The identification of this ATE is made possible by our theoretical framework, specifically by applying Theorem 3. The strategy involves three conceptual steps: (1) collapsing the ST-HCM to its equivalent ST-DCM, as guaranteed by Theorem 2; (2) applying a conditional version of sequential adjustment on the ST-DCM; and (3) leveraging the hierarchical structure to average out the unobserved confounder U_i .

Assuming that the dynamic and spatial confounding paths are blockable by the observable history (comprising the history of unit i and its neighbors $\mathcal{N}(i)$), Theorem 3 ensures that the ATE is identifiable. The identification formula can be expressed as an expectation over the distribution of the unobserved confounders $\mathbf{U} = (U_1, \dots, U_N)$:

$$\mathbb{E}[Y_T|\text{do}(A_t = a^*)] = \mathbb{E}_{\mathbf{U}} [\mathbb{E}_{\text{G-Comp}}[Y_T|\text{do}(A_t = a^*), \mathbf{U}]].$$

The inner expectation, $\mathbb{E}_{\text{G-Comp}}[\cdot]$, represents the conditional causal effect for units with fixed static properties $\mathbf{U}$, which can be computed via a recursive procedure analogous to Robins’ G-Computation formula. The outer expectation, $\mathbb{E}_{\mathbf{U}}[\cdot]$, is then computed by averaging these conditional effects over the population of units.

Estimation Algorithm We translate the identification strategy into a practical, two-stage estimation algorithm.

Stage 1: Learning Conditional Dynamics. For each unit i , we estimate a unit-specific model, $\mathcal{M}_i$, that captures the full spatio-temporal dynamics conditional on its latent static properties U_i . This model is trained on the subunit-level panel data centered at unit i , including the histories of its neighbors. The choice of model for $\mathcal{M}_i$ is flexible. Any method capable of capturing the relevant temporal and spatial dependencies can be employed, provided it can yield an estimate of the conditional expectation of the outcome. As our core contribution lies in the causal identification framework itself, rather than a specific predictive model, we prioritize methods that are well-established and interpretable. Suitable candidates range from semi-parametric models, such as Linear Mixed-Effects Models (LMM), to non-parametric approaches like Gradient Boosting Machines (GBM). The specific implementation of $\mathcal{M}_i$ and further details are provided in Appendix D.

Stage 2: Conditional G-Computation via Simulation. With the trained models $\{\mathcal{M}_i\}_{i=1}^N$, we simulate the effect of the intervention for each unit and then aggregate the results. The procedure is detailed in Algorithm 1. For each unit i , we recursively simulate the system’s evolution under a fixed treatment policy a^* . At each time step t , the model $\mathcal{M}_i$ predicts the outcome $\hat{y}_{i,t}(a^*)$ based on the simulated history of unit i and the observed history of its neighbors. The final outcome at time T , $\hat{\mu}_i(T; a^*)$, represents an estimate of the conditional causal effect for that unit. Finally, we average these estimates across all units to obtain the ATE.

5 Experiments

We conduct experiments on both a synthetic setting and a real-world case to evaluate ST-HCMs. We design 4 research questions to guide the evaluation:

Algorithm 1: ATE Estimation in ST-HCM

```

1: Input: Spatio-temporal data  $\mathcal{D}$ , time horizon  $T$ .
2: Output: Estimated Average Treatment Effect  $\widehat{\text{ATE}}_T$ .
3: Stage 1: Learning Conditional Dynamics
4: Train a set of unit-specific models  $\{\mathcal{M}_i\}_{i=1}^N$  from  $\mathcal{D}$ .
5: Stage 2: Simulation and Aggregation
6: Initialize result sets  $\mathcal{R}_0 \leftarrow \emptyset, \mathcal{R}_1 \leftarrow \emptyset$ .
7: for  $a^* \in \{0, 1\}$  do
8:   for  $i = 1$  to  $N$  do
9:     Initialize simulation history  $\mathcal{H}_i$ .
10:    for  $t = 1$  to  $T$  do
11:      Get neighbors' history  $\mathcal{H}_{\mathcal{N}(i), < t}$  from  $\mathcal{D}$ .
12:      Predict outcome  $\hat{y}_{i,t}^{(a^*)} \leftarrow \mathcal{M}_i(\text{do}(A_t = a^*), \mathcal{H}_i, \mathcal{H}_{\mathcal{N}(i), < t})$ .
13:      Update  $\mathcal{H}_i \leftarrow \mathcal{H}_i \cup \{(a^*, \hat{y}_{i,t}^{(a^*)})\}$ .
14:    end for
15:    Add final outcome  $\hat{y}_{i,T}^{(a^*)}$  to result set  $\mathcal{R}_{a^*}$ .
16:  end for
17: end for
18: Final Calculation
19:  $\hat{\mathbb{E}}[Y_T | \text{do}(A_t = 1)] \leftarrow \text{mean}(\mathcal{R}_1)$ .
20:  $\hat{\mathbb{E}}[Y_T | \text{do}(A_t = 0)] \leftarrow \text{mean}(\mathcal{R}_0)$ .
21:  $\widehat{\text{ATE}}_T \leftarrow \hat{\mathbb{E}}[Y_T | \text{do}(A_t = 1)] - \hat{\mathbb{E}}[Y_T | \text{do}(A_t = 0)]$ .
22: return  $\widehat{\text{ATE}}_T$ 

```

- **RQ1 (Correctness)** Can our method recover known causal effects in an ideal synthetic setting?
- **RQ2 (Superiority)** How does our framework perform compared to baselines that neglect either the hierarchical structure or spatio-temporal dependencies of the data?
- **RQ3 (Robustness)** How robust is our method to violations of its key assumptions?
- **RQ4 (Applicability)** How sensitive are causal estimates to underlying structural assumptions in a real-world system, and how does ST-HCM address this challenge?

We implement our estimator with both parametric Linear Mixed Models (LMM) (Faraway 2006) and non-parametric Gradient Boosting Machines (GBM) models (Natekin and Knoll 2013) to serve as the unit-specific conditional mechanism $\mathcal{M}_i$. Full details of our experimental setup and data generation processes are provided in Appendix E.

5.1 Simulation Experiments

Correctness To verify the fundamental correctness of our proposed estimation procedure, we first assess its ability to produce unbiased and consistent estimates of the Average Treatment Effect (ATE). We generate data from a canonical spatio-temporal hierarchical process with a known true ATE.

Figure 4 demonstrates the statistical properties of our ST-HCM estimator. The left panel shows the distribution of ATE estimates over multiple independent trials. The mean of this distribution (4.91) is nearly identical to the true ATE (5.0), demonstrating that our estimator is **unbiased**. The

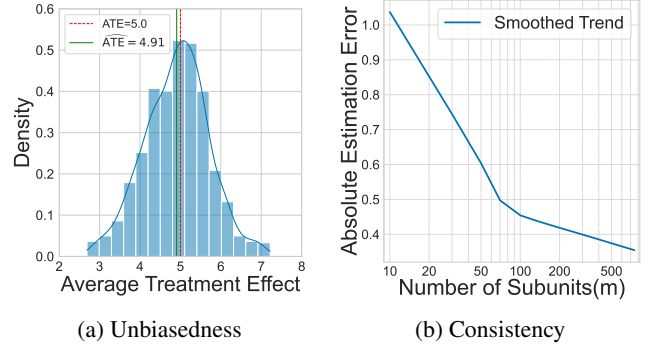


Figure 4: Validation of ST-HCM estimator.

right panel shows that the absolute estimation error converges towards zero as the number of subunits per unit (m) increases. This confirms the **consistency** of our estimator, a key property guaranteed by our Collapse Theorem 2. These results provide strong evidence that our framework is correctly specified and theoretically sound (**answering RQ1**).

Superiority over Baselines Having established the correctness of our framework, we now conduct a rigorous stress test to evaluate its superiority over several strong baselines in the presence of unobserved confounding and spatial spillovers. To isolate the core causal identification capabilities of each method, we instantiate our framework and the hierarchical baselines using Linear Mixed Models (LMMs). This creates a challenging but well-understood linear setting where performance differences can be clearly attributed to the structural assumptions of each model. We compare our ST-HCM (LMM) against two primary classes of baselines:

1. **Aggregated Models:** A standard spatio-temporal LMM applied to data aggregated at the unit-time level, which ignores the hierarchical structure.
2. **Non-Spatial Models:** A Temporal HCM that models the hierarchy but ignores spatial dependencies, equivalent to a panel data model with fixed effects.

Table 1 presents the ATE absolute error for these models across a wide grid of confounding strengths (γ) and spatial spillover strengths (ρ). The results reveal two clear patterns.

First, hierarchy is essential for handling confounding. Across all levels of spatial spillover (all row groups), the error of the Aggregated model increases dramatically with the confounding strength. In contrast, both hierarchical models (T-HCM and ST-HCM) maintain low error, demonstrating their robustness to the unobserved unit-level confounder U_i . For instance, at $\rho = 2.0$ and $\gamma = 4.0$, the Aggregated model's error is 7.94, whereas the hierarchical models' errors are an order of magnitude smaller.

Second, spatial modeling is crucial under spillover. While the T-HCM performs well when spillover is mild ($\rho = 0.0, 0.5$), its error systematically increases as ρ grows. Our proposed ST-HCM, which explicitly models these spatial dependencies, consistently maintains the lowest error across nearly all settings. At $\gamma = 4.0$, as ρ increases from 0.0 to 2.0, the error of the T-HCM explodes from 0.11 to 2.35,

ρ	Model	Confounding Strength (γ)				
		0.0	1.0	2.0	3.0	4.0
0.0	Aggregated	0.173	1.909	2.886	3.009	3.086
	T-HCM	0.054	0.070	0.084	0.126	0.114
	ST-HCM	0.057	0.073	0.086	0.118	0.117
0.5	Aggregated	0.164	1.854	3.038	3.387	3.785
	T-HCM	0.097	0.068	0.072	0.128	0.080
	ST-HCM	0.056	0.069	0.088	0.126	0.123
1.0	Aggregated	0.189	2.261	3.624	4.013	4.260
	T-HCM	0.132	0.108	0.161	0.251	0.224
	ST-HCM	0.056	0.066	0.076	0.127	0.121
1.5	Aggregated	0.211	3.420	5.051	5.950	6.734
	T-HCM	0.174	0.135	0.335	0.507	0.606
	ST-HCM	0.058	0.066	0.064	0.123	0.113
2.0	Aggregated	0.228	3.962	5.620	6.819	7.942
	T-HCM	0.329	0.271	0.894	1.320	2.354
	ST-HCM	0.056	0.072	0.070	0.124	0.110

Table 1: Mean ATE Absolute Error for different levels of confounding strength (γ) and spatial spillover (ρ). Both T-HCM (no spatial modeling) and ST-HCM (full model) are proposed hierarchical estimators. T-HCM is optimal when $\rho = 0$, while ST-HCM excels when $\rho \geq 1$, demonstrating the importance of correctly specifying spatial dependencies.

while our ST-HCM’s error remains stable at approximately 0.11. These findings indicate that our ST-HCM framework, by correctly modeling both the hierarchical structure and spatio-temporal dependencies, provides substantially more accurate causal estimates than methods that ignore either dimension of the data’s complexity (**answering RQ2**).

Robustness We assess the robustness of our ST-HCM (LMM) estimator by systematically violating two key assumptions: the time-invariance of the unobserved confounder and the causal ordering of spatial effects (Assumption 4). Our simulations are based on a challenging linear setting ($N = 16, m = 50, T = 8$) with fixed baseline confounding ($\gamma = 2.0$) and spatial spillover ($\rho = 1.5$). See Appendix E for data generation details.

As illustrated in Figure 5, our proposed model demonstrates considerable robustness. When the time-invariance assumption is relaxed by introducing a temporal drift to the confounder (Figure 5a), the error of our ST-HCM (LMM) increases mildly but remains significantly lower than that of the non-spatial T-HCM baseline. More impressively, when the spatial ordering assumption is violated by introducing simultaneous feedback loops (Figure 5b), the performance of our ST-HCM (LMM) remains nearly unaffected, maintaining a consistently low error and substantially outperforming both the T-HCM and Aggregated baselines across all violation intensities. This suggests our framework is a robust tool for practical applications where these assumptions may not strictly hold (**answering RQ3**).

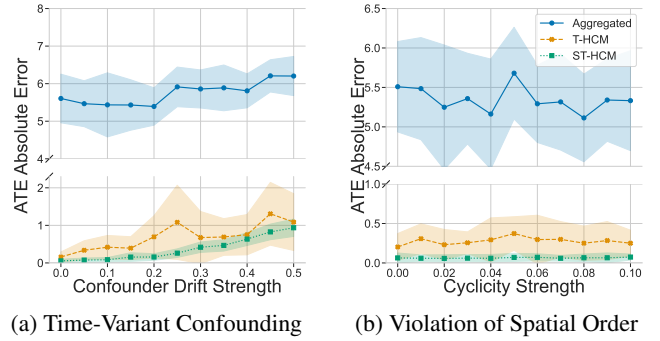


Figure 5: Robustness analysis of estimators.

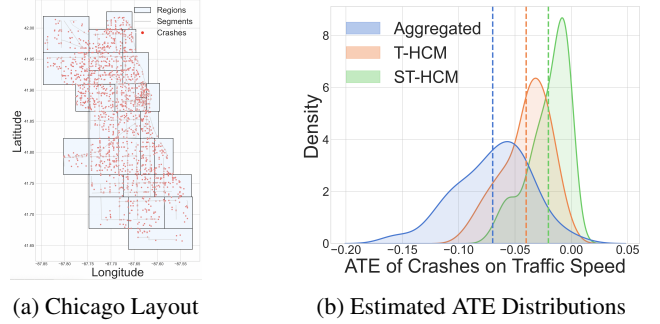


Figure 6: Real-world application.

5.2 Real-World Application: Urban Traffic

We demonstrate the applicability of our framework in real-world using a Chicago traffic data set. Our analysis uses traffic speed data from 1,025 road segments (subunits) across 29 traffic regions (units) over a one-week period, comprising over 1.2 million observations and 1943 treatment events (Figure 6(a)).

To capture the complex non-linearities inherent in traffic systems, we employ Gradient Boosting Machines (GBM) as the implementation for the unit-specific conditional mechanism $\mathcal{M}_i$. We compare our full ST-HCM against a T-HCM that ignores spatial effects, and an Aggregated model that ignores the hierarchy entirely.

Results summarized in Figure 6(b) reveal that causal estimates are sensitive to structural assumptions. Progressively incorporating the hierarchical and then the spatial structure systematically attenuates the estimated ATE. Notably, this process also sharpens the posterior distribution, demonstrating that a more correctly specified model yields not only a less biased but also a more precise estimate by disentangling the treatment effect from unobserved confounders and spatial spillovers (**answering RQ4**).

6 Related Work

Hierarchical causal models The hierarchical causal model framework (Weinstein and Blei 2024), exploits sub-unit variation to address unobserved unit-level confounding, generalizing fixed-effects and related econometric methods (Gelman and Hill 2007; Wooldridge 2010). However, they

are designed for static i.i.d. settings and cannot capture temporal or spatial dependencies. Recent advances such as DML-panel methods with correlated random effects and dynamic multivariate panel models suggest possible extensions toward temporal settings (Clarke and Polselli 2025; Tikka and Helske 2024). Our ST-HCMs extend HCMs by explicitly modeling temporal dynamics and spatial interactions while still addressing latent unit-level confounders.

Causal inference for spatio-temporal data Existing approaches fall into two categories: (1) causal discovery methods that learn dynamic causal graphs from observational data (Pamfil et al. 2020; Zhao et al. 2023; Gong et al. 2024; Wang et al. 2025) (2) causal effect estimators that adapt frameworks like G-computation, marginal structural models, or propensity scores to dynamic or spatial settings, with recent extensions using recurrent or transformer architectures (Bica, Alaa, and van der Schaar 2020; Melnychuk, Frauen, and Feuerriegel 2022; Bhattacharya and Sen 2024; Li et al. 2025; Opreescu et al. 2025). However, most assume no latent unit-level confounders, which is unrealistic in domains like environment or transportation. In contrast, ST-HCM address this gap by structurally incorporating unobserved, static confounders, enabling more credible causal inference.

Panel data and IV-based causal inference Panel methods such as fixed effects, difference-in-differences, and synthetic control estimate effects via within-unit variation or counterfactuals (Wooldridge 2010; Abadie, Diamond, and Hainmueller 2010), but their i.i.d. and linearity assumptions limit use in spatio-temporal contexts (Millimet and Bellemare 2023; Arkhangelsky and Imbens 2024). IV-based approaches mitigate confounding, for example through DeepIV and semiparametric IVs (Hartford et al. 2017; Xie et al. 2025b; Cui et al. 2025); and the recent Time-dependent Instrumental Factor Model (TIFM) addresses time-varying confounders with learned substitutes (Cheng et al. 2024). Still, IV methods hinge on valid instruments, which are often scarce in practice. Our ST-HCM offers a flexible alternative by combining hierarchical modeling with graphical reasoning and spatio-temporal analysis.

7 Conclusion

In this paper, we introduced Spatio-Temporal Hierarchical Causal Models (ST-HCMs), a novel framework enabling reliable causal inference from nested spatio-temporal data. Our theoretical foundation rests on a Spatio-Temporal Collapse Theorem, guaranteeing that complex hierarchical models converge to simpler, tractable counterparts, providing rigorous causal identification. Extensive experiments verified this framework, demonstrating theoretical correctness, practical necessity over baselines ignoring hierarchy or spatial effects, and robustness to assumption violations.

8 Acknowledgements

We thank the reviewers for their valuable feedback. This work was supported by Public Computing Cloud, Renmin University of China and the fund for building world-class universities (disciplines) of Renmin University of China.

References

- Abadie, A.; Diamond, A.; and Hainmueller, J. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American statistical Association*, 105(490): 493–505.
- Ali, S.; Faruque, O.; and Wang, J. 2024. Estimating direct and indirect causal effects of spatiotemporal interventions in presence of spatial interference. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 213–230. Springer.
- Ali, S.; Hasan, U.; Li, X.; Faruque, O.; Sampath, A.; Huang, Y.; Gani, M. O.; and Wang, J. 2024. Causality for Earth Science – A Review on Time-series and Spatiotemporal Causality Methods. arXiv:2404.05746.
- Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press.
- Arkhangelsky, D.; and Imbens, G. 2024. Causal models for longitudinal and panel data: a survey. *The Econometrics Journal*, 27(3): C1–C61.
- Aronow, P. M.; and Samii, C. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4): 1912 – 1947.
- Bhattacharya, S.; and Sen, S. 2024. Causal effect estimation under network interference with mean-field methods. arXiv:2407.19613.
- Bica, I.; Alaa, A. M.; and van der Schaar, M. 2020. Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders. arXiv:1902.00450.
- Billingsley, P. 1995. *Probability and measure*. A Wiley-Interscience publication. New York [u.a.]: Wiley, 3. ed edition. ISBN 0471007102.
- Cheng, D.; Xu, Z.; Li, J.; Liu, L.; Liu, J.; Gao, W.; and Le, T. D. 2024. Instrumental variable estimation for causal inference in longitudinal data with time-dependent latent confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11480–11488.
- Clarke, P. S.; and Polselli, A. 2025. Double machine learning for static panel models with fixed effects. *The Econometrics Journal*, utaf011.
- Cover, T. M.; and Thomas, J. A. 1991. *Elements of information theory*. Wiley series in telecommunications. New York, NY [u.a.]: Wiley.
- Cressie, N.; and Wikle, C. K. 2011. *Statistics for Spatio-Temporal Data*. Wiley.
- Cui, E. H.; Lu, X.; Zhou, J.; Zhou, H.; and Li, G. 2025. A Semiparametric Bayesian Method for Instrumental Variable Analysis with Partly Interval-Censored Time-to-Event Outcome. arXiv:2501.14837.
- Dong, P.; Wang, X.-L.; Bose, I.; Ng, K. K. H.; Zhang, X.; and Zhang, X. 2024. Causally-Aware Spatio-Temporal Multi-Graph Convolution Network for Accurate and Reliable Traffic Prediction. arXiv:2408.13293.

- Dong, P.; Zhang, X.; and Zhang, X. 2025. A data-driven approach for spatio-temporal causal analysis in large-scale urban traffic networks. *Transportation Research Part E: Logistics and Transportation Review*, 202: 104244.
- Dudley, R. M. 1999. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.
- Duvenaud, D. 2014. *Automatic model construction with Gaussian processes*. Ph.D. thesis, Apollo - University of Cambridge Repository.
- Faraway, J. J. 2006. *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton, FL: Chapman & Hall/CRC. ISBN 1-584-88424-X.
- Folland, G. 2013. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley. ISBN 9781118626399.
- Gelfand, A.; Fuentes, M.; Guttorp, P.; and Diggle, P. 2010. *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis. ISBN 9781420072877.
- Gelman, A.; and Hill, J. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
- Gong, C.; Zhang, C.; Yao, D.; Bi, J.; Li, W.; and Xu, Y. 2024. Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 57(4): 1–38.
- Hartford, J.; Lewis, G.; Leyton-Brown, K.; and Taddy, M. 2017. Deep IV: A Flexible Approach for Counterfactual Prediction. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1414–1423. PMLR.
- Hernán, M. A.; and Robins, J. M. 2006. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7): 578–586.
- Hernán, M. A.; and Robins, J. M. 2010. Causal inference.
- Jensen, D.; Burroni, J.; and Rattigan, M. 2020. Object Conditioning for Causal Inference. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 1072–1082. PMLR.
- Kechris, A. 2012. *Classical descriptive set theory*, volume 156. Springer Science & Business Media.
- Li, J.; Li, X.; and Zhou, X. 2025. FAP-CD: Fairness-Driven Age-Friendly Community Planning via Conditional Diffusion Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27): 28168–28176.
- Li, S.; Li, H.; Li, X.; Xu, Y.; Lin, Z.; and Jiang, H. 2025. Causal Intervention Is What Large Language Models Need for Spatio-Temporal Forecasting. *IEEE Transactions on Cybernetics*, 55(8): 3825–3837.
- Martinussen, T.; Vansteelandt, S.; Tchetgen Tchetgen, E. J.; and Zucker, D. M. 2017. Instrumental variables estimation of exposure effects on a time-to-event endpoint using structural cumulative survival models. *Biometrics*, 73(4): 1140–1149.
- Melnichuk, V.; Frauen, D.; and Feuerriegel, S. 2022. Causal Transformer for Estimating Counterfactual Outcomes. arXiv:2204.07258.
- Millimet, D.; and Bellemare, M. F. 2023. Fixed effects and causal inference.
- Natekin, A.; and Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurobotics*, Volume 7 - 2013.
- Newey, W. K.; and Powell, J. L. 2003. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5): 1565–1578.
- Opreescu, M.; Park, D. K.; Luo, X.; Yoo, S.; and Kallus, N. 2025. GST-UNet: Spatiotemporal Causal Inference with Time-Varying Confounders. arXiv:2502.05295.
- Pamfil, R.; Sriwattanaworachai, N.; Desai, S.; Pilgerstorfer, P.; Beaumont, P.; Georgatzis, K.; and Aragam, B. 2020. DYNOTEARS: Structure Learning from Time-Series Data. arXiv:2002.00498.
- Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none): 96 – 146.
- Rasmussen, C. E.; and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. MIT Press.
- Reich, B. J.; Yang, S.; Guan, Y.; Giffin, A. B.; Miller, M. J.; and Rappold, A. 2021. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, 89(3): 605–634.
- Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9): 1393–1512.
- Tikka, S.; and Helske, J. 2024. dynamite: An R Package for Dynamic Multivariate Panel Models. arXiv:2302.01607.
- Wang, K.; Varambally, S.; Watson-Parris, D.; Ma, Y.-A.; and Yu, R. 2025. Discovering Latent Causal Graphs from Spatiotemporal Data. arXiv:2411.05331.
- Wang, Y.; Zhou, X.; Mascolo, C.; Noulas, A.; Xie, X.; and Liu, Q. 2018. Predicting the Spatio-Temporal Evolution of Chronic Diseases in Population with Human Mobility Data.
- Weinstein, E. N.; and Blei, D. M. 2024. Hierarchical Causal Models. arXiv:2401.05330.
- Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. MIT press.
- Xie, Z.; Hong, Z.; Lyu, W.; Wang, H.; Wang, G.; and Zhang, D. 2025a. CoAlign: Uncertainty Calibration of LLM for Geospatial Repartition. In Rehm, G.; and Li, Y., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, 254–263. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-288-6.
- Xie, Z.; Lyu, W.; Song, Y.; Wang, H.; Yang, G.; Liu, Y.; He, T.; Zhang, D.; and Wang, G. 2025b. Scalable Area Difficulty Assessment with Knowledge-enhanced AI for Nationwide Logistics Systems. In *Proceedings of the 31st ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining V.1, KDD '25, 2713–2724. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712456.

Xu, Z.; and Zhou, X. 2024. CGAP: Urban Region Representation Learning with Coarsened Graph Attention Pooling. arXiv:2407.02074.

Zhang, W.; and Ning, K. 2023. Spatiotemporal heterogeneities in the causal effects of mobility intervention policies during the COVID-19 outbreak: A spatially interrupted time-series (SITS) analysis. *Annals of the American Association of Geographers*, 113(5): 1112–1134.

Zhao, W.; Zhang, S.; Wang, B.; and Zhou, B. 2023. Spatio-temporal causal graph attention network for traffic flow prediction in intelligent transportation systems. *PeerJ Computer Science*, 9: e1484.

A Proof of Theorem 1

This section provides the complete proof for the convergence of a Temporal Hierarchical Causal Model (T-HCGM) to its corresponding Dynamic Collapsed Model (DCM).

Proof. The proof proceeds by mathematical induction on the time horizon T . Let $P_{col,t}$ and $P_{m,t}^{marg}$ denote the joint distributions of the macro-state history $(S_1, \dots, S_t)$.

Base Case : For $T = 1$, we need to show that $\lim_{m \rightarrow \infty} \text{KL}(P_{col,1} \parallel P_{m,1}^{marg}) = 0$. The macro-state at $t = 1$ is $S_1 = (X_1^U, Q_1)$. Using the chain rule for KL-divergence (Cover and Thomas 1991), we have:

$$\begin{aligned} \text{KL}(P_{col,1} \parallel P_{m,1}^{marg}) &= \text{KL}(P_{col}(Q_1) \parallel P_m^{marg}(Q_1)) \\ &+ \mathbb{E}_{P_{col}(Q_1)} [\text{KL}(P_{col}(X_1^U | Q_1) \parallel P_m^{marg}(X_1^U | Q_1))] \quad (6) \end{aligned}$$

The first term, $\text{KL}(P_{col}(Q_1) \parallel P_m^{marg}(Q_1))$, is identically zero. This is because the generation of Q_1 is, by definition, a unit-level process that does not depend on any subunit realizations, and the DCM is constructed to preserve this mechanism exactly. Thus, $P_{col}(Q_1) = P_m^{marg}(Q_1)$.

For the second term, we consider the inner KL-divergence, $\text{KL}(P_{col}(X_1^U | Q_1) \parallel P_m^{marg}(X_1^U | Q_1))$. The distribution $P_{col}(X_1^U | Q_1)$ is determined by the true (but empty) subunit history measure, $\mathcal{P}_{<1}$, while $P_m^{marg}(X_1^U | Q_1)$ is determined by the empirical measure from m subunits, $\hat{\mathcal{P}}_{m,<1}$. By Assumption 1 (applied to an empty history), this KL-divergence converges to zero as $m \rightarrow \infty$. Since the integrand converges to zero for every Q_1 and is non-negative, the expectation also converges to zero by the Dominated Convergence Theorem. Thus, the base case holds.

Inductive Step: Inductive Hypothesis: Assume that for a time horizon $t - 1$, the theorem holds:

$$\lim_{m \rightarrow \infty} \text{KL}(P_{col,t-1} \parallel P_{m,t-1}^{marg}) = 0 \quad (7)$$

We must prove that the theorem also holds for the time horizon t . We apply the chain rule for KL-divergence to the joint distribution up to time t :

$$\text{KL}(P_{col,t} \parallel P_{m,t}^{marg}) = \text{KL}(P_{col,t-1} \parallel P_{m,t-1}^{marg}) \quad (8)$$

$$+ \mathbb{E}_{P_{col,t-1}} [\text{KL}(P_{col}(S_t | S_{<t}) \parallel P_m^{marg}(S_t | S_{<t}))] \quad (9)$$

By the Inductive Hypothesis, Term (8) converges to zero as $m \rightarrow \infty$. The proof thus reduces to showing that Term (9), the expected KL-divergence of the one-step transition kernels, also converges to zero. We will establish this by showing that the integrand converges to zero pointwise for any given macro-history $S_{<t}$.

Let us analyze the inner KL-divergence by decomposing it for the components of $S_t = (X_t^U, Q_t)$:

$$\begin{aligned} &\text{KL}(P_{col}(S_t | S_{<t}) \parallel P_m^{marg}(S_t | S_{<t})) \\ &= \text{KL}(P_{col}(Q_t | S_{<t}) \parallel P_m^{marg}(Q_t | S_{<t})) \\ &+ \mathbb{E}_{P_{col}(Q_t | S_{<t})} [\text{KL}(P_{col}(X_t^U | Q_t, S_{<t}) \parallel P_m^{marg}(X_t^U | Q_t, S_{<t}))] \\ &= \text{Part A} + \text{Part B} \quad (10) \end{aligned}$$

Part A is identically zero because the generating mechanism for the Q-variables, $p(\mathbf{q}_t|U_i, \mathbf{S}_{i,<t})$, is a unit-level process preserved exactly in the DCM, making the two conditional distributions identical.

For **Part B**, we focus on the inner KL-divergence term inside the expectation. The two conditional distributions for X_t^U differ only in their dependence on the measure over the subunit history space: $P_{col}(X_t^U|\mathbf{Q}_t, \mathbf{S}_{<t})$ depends on the true measure $\mathcal{P}_{<t}$, while $P_m^{marg}(X_t^U|\mathbf{Q}_t, \mathbf{S}_{<t})$ depends on the empirical measure $\hat{\mathcal{P}}_{m,<t}$. The convergence of this KL term to zero follows from a continuous mapping argument.

By Assumption A2, the class of subunit histories is a Glivenko-Cantelli class. This guarantees the uniform convergence of the empirical measure to the true measure (a result from empirical process theory, see, e.g., (Dudley 1999)), i.e., $\hat{\mathcal{P}}_{m,<t} \rightarrow \mathcal{P}_{<t}$ almost surely in a suitable metric as $m \rightarrow \infty$. By Assumption A1, the mapping from this measure to the output distribution of X_t^U is continuous in the sense of KL-divergence. The Continuous Mapping Theorem (Billingsley 1995) states that a continuous function of a convergent sequence of random variables converges to the function of the limit. Applying this here, since the input measures converge and the mapping is continuous, the output distributions must also converge. The mode of convergence guaranteed by Assumption A1 is precisely what we need: the KL-divergence between the output distributions converges to zero. Since the integrand in Part B converges to zero for every $\mathbf{Q}_t$, and is non-negative and appropriately bounded, the entire expectation in **Part B converges to zero** by the Dominated Convergence Theorem (Folland 2013).

Since both Part A and Part B are zero in the limit, the KL-divergence of the one-step transition kernel converges to zero. Consequently, Term (9) converges to zero. This completes the inductive step.

By the principle of mathematical induction, the theorem holds for any finite time horizon T . $\square$

B Proof of Theorem 2

This section provides the proof for the convergence of a Spatio-Temporal Hierarchical Causal Model (ST-HCGM) to its corresponding Spatio-Temporal Dynamic Collapsed Model (ST-DCM).

Proof. The core of the proof is to construct a ‘‘Super-Unit’’ model, $\mathcal{M}_{super}$, that is mathematically equivalent to the ST-HCGM but formally satisfies the definition of a T-HCGM. We will then show that the convergence of $\mathcal{M}_{super}$ guaranteed by Theorem 1 implies the convergence stated in Theorem 2.

Step 1: Construction of the Super-Unit Model

We define a single Super-Unit by aggregating all N spatial units. The state variables and mechanisms of this Super-Unit are defined as concatenations of the corresponding elements from the individual spatial units.

- **Super-Unit State Variables:**

- Static Exogenous Variable:

$$U_{super} := (U_1, U_2, \dots, U_N).$$

- Unit-Level Variables at time t :

$$X_{super,t}^U := (X_{1,t}^U, X_{2,t}^U, \dots, X_{N,t}^U).$$

- Q-Variables at time t :

$$\mathbf{Q}_{super,t} := (\mathbf{Q}_{1,t}, \mathbf{Q}_{2,t}, \dots, \mathbf{Q}_{N,t}).$$

- Macro-State at time t :

$$\mathbf{S}_{super,t} := (X_{super,t}^U, \mathbf{Q}_{super,t}).$$

- **Super-Unit Subunits:** The set of subunits for the Super-Unit is the union of all subunits from all spatial units, totaling $N \times m$. The mechanism for each unit-level variable $X_{i,t}^w$ within the Super-Unit’s state depends on the empirical measure constructed from this entire collection of subunits.
- **Super-Unit Mechanism:** The mechanism $\mathbf{f}_{super}$ for the Super-Unit is a vector-valued function where the i -th component is the mechanism f_i from the original ST-HCGM. The evolution of the Super-Unit’s state is thus described by:

$$\mathbf{S}_{super,t} = \mathbf{f}_{super}(\mathbf{S}_{super,<t}, U_{super}, \text{Noise}_{super,t}, \{\text{Subunit Info}\}_{super}) \quad (11)$$

Step 2: Verify $\mathcal{M}_{super}$ is a Valid T-HCGM

We must verify that $\mathcal{M}_{super}$ adheres to the definition of a T-HCGM. The most critical requirement is that its underlying causal graph, when unrolled over time, must be a Directed Acyclic Graph (DAG). The structure of $\mathcal{M}_{super}$ is inherently acyclic across time steps, as any state at time t can only depend on states at times $t' \leq t$. The crucial step is to establish acyclicity within a single time step t . The dependencies among the components of the state vector $\mathbf{S}_{super,t} = (\mathbf{S}_{1,t}, \dots, \mathbf{S}_{N,t})$ are governed by the spatial interactions of the original ST-HCGM. Assumption 4 (Spatial Causal Ordering) imposes a strict, global ordering on these contemporaneous interactions. Specifically, any component $\mathbf{S}_{i,t}$ can be influenced by another component $\mathbf{S}_{j,t}$ only if $j < i$. This enforces a directed, acyclic structure on the dependencies within any given time slice. Since the graph is acyclic both across and within time steps, the unrolled graph for $\mathcal{M}_{super}$ is a DAG. The other definitional aspects of a T-HCGM, such as the presence of a static confounder and time-varying noise, are satisfied by construction. Thus, $\mathcal{M}_{super}$ is a valid T-HCGM.

Step 3: Verify $\mathcal{M}_{super}$ Satisfies Premises of Theorem 1

We now verify that $\mathcal{M}_{super}$ satisfies the assumptions required by the temporal collapsing theorem (Theorem 1). The state space of the Super-Unit is a finite Cartesian product of Polish spaces, which is itself a Polish space (Kechris 2012), satisfying the regularity condition. The mechanism convergence assumption for $\mathcal{M}_{super}$ is also met, as it is a direct consequence of the corresponding assumption on the original ST-HCGM, combined with the uniform convergence of the aggregated empirical measure over all subunits, which is guaranteed by the Glivenko-Cantelli class assumption.

Step 4: Invoking Theorem 1 Since $\mathcal{M}_{super}$ is a valid T-HCGM that satisfies all the necessary premises, we can directly invoke Theorem 1. This theorem guarantees that the

marginalized distribution of the Super-Unit model converges to its collapsed distribution in KL-divergence:

$$\lim_{m \rightarrow \infty} \text{KL}(P_{\text{super}, \text{col}, T} || P_{\text{super}, m, T}^{\text{marg}}) = 0$$

where $P_{\text{super}, \text{col}, T}$ is the distribution of the collapsed Super-Unit model and $P_{\text{super}, m, T}^{\text{marg}}$ is its marginalized distribution.

The final step is to establish the equivalence between the distributions of the Super-Unit model and the original ST-HCGM. By construction, the macro-state space of the ST-HCGM is identical to the state space of the Super-Unit model. Consequently, the marginalized distribution $P_{m, T, N}^{\text{marg}}$ of the ST-HCGM is the same as $P_{\text{super}, m, T}^{\text{marg}}$. Similarly, the Spatio-Temporal Dynamic Collapsed Model (ST-DCM) describes the exact same limiting process as the collapsed model of the Super-Unit, meaning $P_{\text{col}, T, N} = P_{\text{super}, \text{col}, T}$. Given these equivalences, the convergence result for the Super-Unit model directly implies the convergence for the Spatio-Temporal model. $\square$

C Proofs for Identifiability

This section provides proofs for the identifiability theorems for Temporal Hierarchical Causal Models (T-HCMs) and Spatio-Temporal Hierarchical Causal Models (ST-HCMs).

C.1 Proof of Proposition 1

We first restate the propositions concerning identifiability in the purely temporal setting. The proofs operate on the Dynamic Collapsed Model (DCM), whose existence and convergence properties are guaranteed by the temporal collapsing theorem.

Proof. The problem of identifying the causal effect is equivalent to identifying the post-intervention distribution $P(Q_{i,t}^Y | \text{do}(Q_{i,t'}^A = \delta_{a*}))$ in the DCM. The primary obstacles are the backdoor paths connecting $Q_{i,t'}^A$ and $Q_{i,t}^Y$. These paths arise from two sources: (1) dynamic confounding from the system's history, and (2) static confounding from the time-invariant latent variable U_i .

By the premise of the proposition, all dynamic backdoor paths are intercepted by the observable macro-history $\mathcal{S}_{i, < t'}$. We can therefore apply the backdoor adjustment formula (Pearl 2009), integrating over the distribution of this history to block these paths:

$$P(Q_{i,t}^Y | \text{do}(Q_{i,t'}^A)) = \int P(Q_{i,t}^Y | \text{do}(Q_{i,t'}^A), \mathbf{s}_{< t'}) P(\mathbf{s}_{< t'}) d\mathbf{s}_{< t'}$$

After conditioning on a specific history realization $\mathbf{s}_{< t'}$, the only remaining unblocked backdoor path is the one mediated by the static confounder, $Q_{i,t'}^A \leftarrow U_i \rightarrow Q_{i,t}^Y$. At this stage, we leverage the hierarchical structure. The law of total probability allows us to express the conditional interventional distribution as an expectation over U_i :

$$\begin{aligned} & P(Q_{i,t}^Y | \text{do}(Q_{i,t'}^A), \mathbf{s}_{< t'}) \\ &= \mathbb{E}_{U_i | \mathbf{s}_{< t'}} [P(Q_{i,t}^Y | \text{do}(Q_{i,t'}^A), \mathbf{s}_{< t'}, U_i)] \end{aligned} \quad (12)$$

Crucially, once we condition on U_i , all backdoor paths are blocked. Thus, the do operator can be removed, and the inner term becomes an estimable quantity from the observational distribution of a specific unit i (for which U_i is fixed). While U_i is unobserved, we can estimate this expectation by first estimating the unit-specific effect from its subunit-level panel data, and then averaging these estimates across all units. Since all components of the final expression are computable from observational data, the causal effect is identifiable. $\square$

C.2 Proof of Proposition 2

Proof. A valid instrumental variable $Z_{ij, t'}$, represented by the node $Q_{i, t'}^Z$ in the DCM, is by definition d-separated from all confounding sources, including the static U_i and any unobserved history. This exogeneity allows us to relate the desired causal effect to the observational distribution. Let $g(a) := P(Y_t | \text{do}(A_{t'} = a))$ be the target causal effect distribution. The exclusion restriction of the instrument implies that the observed association between the instrument $Z_{t'}$ and the outcome Y_t is mediated entirely through the treatment $A_{t'}$. This relationship can be formalized by the following integral equation:

$$P(Y_t | Z_{t'} = z) = \int g(a) \cdot P(A_{t'} = a | Z_{t'} = z) da \quad (13)$$

This is a Fredholm integral equation of the first kind. The left-hand side, $P(Y_t | Z_{t'} = z)$, and the kernel of the integral, $P(A_{t'} = a | Z_{t'} = z)$, are both identifiable from the observational joint distribution of variables in the DCM. Under standard regularity conditions that ensure the invertibility of the integral operator defined by the kernel (a completeness condition, see (Newey and Powell 2003)), this equation can be solved for the unknown function $g(a)$. Since $g(a)$ can be uniquely determined from observational quantities, the causal effect is identifiable. $\square$

C.3 Proof of Theorem 3

Proof. The proof directly generalizes that of Proposition 1. The problem is first translated to the ST-DCM via the spatio-temporal collapsing theorem. The set of confounding backdoor paths now includes those that traverse through the histories of spatial neighbors. By the premise of the theorem, this entire set of dynamic and spatial backdoor paths is blocked by conditioning on the observable spatio-temporal history, $\text{Hist}_{ST, < t'}^{\text{obs}}$. Applying the backdoor adjustment formula yields:

$$P(Q_{k,t}^Y | \text{do}(Q_{i,t'}^A)) = \int P(Q_{k,t}^Y | \text{do}(Q_{i,t'}^A), h_{st}) P(h_{st}) dh_{st} \quad (14)$$

After this adjustment, the only remaining backdoor paths are those mediated by the set of all static confounders $\{U_j\}_{j=1}^N$. As in the temporal case, the hierarchical structure allows us to overcome this. By conditioning on the full set of static confounders $U_{\text{all}} = (U_1, \dots, U_N)$, all backdoor paths become blocked. The resulting conditional interventional distribution can be estimated from the observational data within

unit-groups that share the same (unobserved) values of U_{all} . The final estimand is obtained by averaging over the population of units. Therefore, the causal effect is identifiable. $\square$

C.4 Proof of Theorem 4

Proof. The proof structure is identical to that of Proposition 2. A valid spatio-temporal instrument $Z_{ij,t'}$, represented by $Q_{i,t'}^Z$ in the ST-DCM, must be exogenous with respect to the entire confounding system. This means it must be separated from the set of all static confounders $\{U_k\}$ and the complete unobserved history of the entire spatio-temporal system.

Given such an instrument, we can again formulate the Fredholm integral equation relating the causal effect distribution $g_k(a) := P(Y_{k,t} | \text{do}(A_{i,t'} = a))$ to observational quantities:

$$P(Y_{k,t} | Z_{i,t'} = z) = \int g_k(a) \cdot P(A_{i,t'} = a | Z_{i,t'} = z) da \quad (15)$$

The terms on the left and the kernel of the integral are identifiable from the observational distribution of the ST-DCM. Provided the completeness condition holds for the instrument, this equation has a unique solution for $g_k(a)$. Thus, the full causal effect distribution is identifiable. $\square$

D Details for ATE Estimation

D.1 Linear Mixed-Effects Model Specification

This section details the implementation of our two-stage estimation algorithm when using a semi-parametric Linear Mixed-Effects Model (LMM) as the unit-specific conditional mechanism, $\mathcal{M}_i$. LMMs, also known as panel data models with fixed effects in econometrics, are a natural choice for this problem as they are explicitly designed to handle unobserved time-invariant heterogeneity.

Stage 1: Learning Conditional Dynamics with LMMs

The first stage aims to estimate the conditional expectation of the outcome, $\mathbb{E}[Y_{ij,t} | \text{History}_{i,<t}, U_i]$, under the assumption of linear relationships. Unlike the non-parametric approach where an independent model is trained for each unit, the LMM leverages a more statistically efficient **pooled regression** across all units and subunits.

The core idea is to model the static, unobserved confounder U_i as a **unit-specific fixed effect**. This is achieved by including a dummy variable for each unit in the regression model. This approach effectively absorbs all time-invariant differences between units, including the effect of U_i , into these unit-specific intercepts.

We fit a single, global LMM to the entire subunit-level dataset. The model is specified by the following linear formula:

$$Y_{ij,t} = \alpha_i + \beta_A A_{ij,t} + \beta_{\text{temp}} \bar{Y}_{i,t-1} + \rho \bar{Y}_{N(i),t-1} + \epsilon_{ij,t} \quad (16)$$

where α_i represents the fixed effect for unit i , and the coefficients β_A , β_{temp} , and ρ are shared across all units. The treatment effect, β_A , is the primary parameter of interest estimated in this stage.

Stage 2: Conditional G-Computation via Simulation

With the fitted LMM, we perform a conditional G-computation procedure as in Algorithm 1 to estimate the average treatment effect. This stage remains conceptually identical to the non-parametric case, but the prediction at each step uses the fitted linear model instead of a GBM.

The simulation for each unit i under a fixed policy a^* proceeds recursively. At each time step t , we predict the potential outcome $\hat{y}_{i,t}(a^*)$ using the fitted coefficients from Equation 16:

- The fixed treatment assignment, a^* , is used for the term $\beta_A a^*$.
- The unit-specific intercept, $\hat{\alpha}_i$, learned during Stage 1, is used. This is the crucial step that conditions the prediction on the unit's unique characteristics (i.e., on U_i).
- The temporal lag is taken from the *simulated* history of unit i .
- The spatial lag is taken from the *observed* history of unit i 's neighbors.

This process is repeated until we obtain the final potential outcome $\hat{y}_{i,T}(a^*)$, which represents the estimate of $\mathbb{E}_{G\text{-Comp}}[Y_T | \text{do}(A_t = a^*), U = u_i]$.

Final ATE Calculation The final ATE is computed by averaging the simulated potential outcomes across all units, exactly as in the GBM case:

$$\widehat{\text{ATE}}_T = \frac{1}{N} \sum_{i=1}^N \hat{y}_{i,T}(1) - \frac{1}{N} \sum_{i=1}^N \hat{y}_{i,T}(0). \quad (17)$$

D.2 Gradient Boosting Machines Specification

This section provides a description of the two-stage estimation algorithm presented in the main text, specifically for the non-parametric Gradient Boosting Machine (GBM) implementation of the unit-specific conditional mechanism, $\mathcal{M}_i$.

Stage 1: Learning Conditional Dynamics with GBMs

The first stage involves learning the conditional dynamics for each unit. The goal is to estimate the conditional expectation of the outcome, $\mathbb{E}[Y_{ij,t} | \text{History}_{i,<t}, U_i]$, where the conditioning on the unobserved, time-invariant confounder U_i is crucial.

To achieve this, we train an **independent GBM for each unit** i , which we denote as $\mathcal{M}_i$. Training a separate model for each unit is a standard strategy to implicitly condition on unit-specific fixed effects. In this way, each model $\mathcal{M}_i$ learns the unique dynamic patterns of its corresponding region, effectively accounting for the influence of that region's unique time-invariant characteristics, U_i .

The target variable for each model $\mathcal{M}_i$ is the subunit-level speed, $Y_{ij,t}$. The feature set consists of the relevant temporal and spatial history, including lagged speeds, lagged treatments, and the historical states of neighboring units, as specified in our experimental setup.

Stage 2: Conditional G-Computation via Simulation

With the set of trained models $\{\mathcal{M}_i\}_{i=1}^N$, we perform a conditional G-computation procedure as outlined in Algorithm

1 to estimate the potential outcomes under different global treatment policies. This stage involves two parallel simulations for each unit i , one for the policy $do(A_t = 0)$ and another for $do(A_t = 1)$.

The simulation for a single unit i under a fixed policy a^* proceeds recursively from $t = 1$ to T . At each time step t , we predict the potential outcome $\hat{y}_{i,t}(a^*)$ using the trained model $\mathcal{M}_i$. The input features for this prediction are constructed from:

- The fixed treatment assignment, a^* .
- The *simulated* history of unit i up to time $t - 1$, which is built using the outcomes predicted in previous steps of the simulation.
- The *observed* history of unit i 's neighbors, $\{S_{k,<t}\}_{k \in \mathcal{N}(i)}$.

This recursive process is repeated until we obtain the final potential outcome for unit i at the time horizon T , denoted $\hat{y}_{i,T}(a^*)$. This value serves as an estimate of the inner expectation in our identification formula, $\mathbb{E}_{G\text{-Comp}}[Y_T | do(A_t = a^*), U = u_i]$.

Final ATE Calculation After completing the simulations for all N units under both treatment policies ($a^* = 0$ and $a^* = 1$), we obtain two sets of potential outcomes: $\{\hat{y}_{i,T}(0)\}_{i=1}^N$ and $\{\hat{y}_{i,T}(1)\}_{i=1}^N$. The final ATE is computed by averaging across these unit-specific estimates, which corresponds to the outer expectation over the distribution of U :

$$\widehat{\text{ATE}}_T = \frac{1}{N} \sum_{i=1}^N \hat{y}_{i,T}(1) - \frac{1}{N} \sum_{i=1}^N \hat{y}_{i,T}(0). \quad (18)$$

This non-parametric, simulation-based approach allows us to estimate the causal effect while accommodating the complex, non-linear dynamics learned by the GBMs.

D.3 Gaussian Process Model Specification

In our case study, we can also employ a unit-specific Gaussian Process (GP) model (Rasmussen and Williams 2006), denoted as $\mathcal{M}_i$, to estimate the conditional dynamics for each unit i . This appendix provides the detailed specification of this GP model.

The GP framework offers a principled, non-parametric approach to learn the complex, non-linear function governing the spatio-temporal dynamics, conditional on the unit's latent static properties U_i . For each unit i , we model the subunit-level outcome $Y_{ij,t}$ as being generated by a latent function $g_i(\cdot)$ corrupted by i.i.d. Gaussian noise:

$$Y_{ij,t} = g_i(\mathbf{x}_{ij,t}) + \eta_{ij,t}, \quad \eta_{ij,t} \sim \mathcal{N}(0, \sigma_n^2) \quad (19)$$

where $\mathbf{x}_{ij,t}$ is the input feature vector for subunit j in unit i at time t , and σ_n^2 is the noise variance. We place a GP prior on the latent function g_i :

$$g_i(\cdot) \sim \mathcal{GP}(m_i(\cdot), k_i(\cdot, \cdot)) \quad (20)$$

where $m_i(\cdot)$ is the mean function and $k_i(\cdot, \cdot)$ is the covariance (kernel) function. The use of GPs for spatio-temporal modeling is a well-established practice in (Gelfand et al. 2010; Cressie and Wikle 2011)

Input Feature Vector The input feature vector $\mathbf{x}_{ij,t}$ is designed to encode all relevant information for predicting the outcome. For our specific problem, it comprises:

- **Treatment:** The current treatment assignment $A_{ij,t}$.
- **Time Stamp:** The current time index t .
- **Lagged Outcome:** The subunit's own outcome from the previous time step, $Y_{ij,t-1}$.
- **Spatial Lag:** A summary of neighbors' outcomes from the previous time step, defined as the average outcome in the neighborhood $\mathcal{N}(i)$:

$$\bar{Y}_{\mathcal{N}(i),t-1} = \frac{1}{|\mathcal{N}(i)|} \sum_{k \in \mathcal{N}(i)} \bar{Y}_{k,t-1} \quad (21)$$

where $\bar{Y}_{k,t-1}$ is the average subunit outcome in unit k at time $t - 1$.

The full input vector is thus:

$$\mathbf{x}_{ij,t} = [A_{ij,t}, t, Y_{ij,t-1}, \bar{Y}_{\mathcal{N}(i),t-1}]^T. \quad (22)$$

Mean and Covariance Functions For simplicity and flexibility, we set the prior mean function to zero, $m_i(\mathbf{x}) = 0$, allowing the kernel to capture the full complexity of the dynamics.

The soul of our GP model is the composite kernel function $k_i(\mathbf{x}, \mathbf{x}')$, which we construct as a sum of kernels to model different sources of variation (Duvenaud 2014). For two input vectors $\mathbf{x}$ and $\mathbf{x}'$, the kernel is:

$$k_i(\mathbf{x}, \mathbf{x}') = k_{\text{time}}(t, t') \times k_{\text{cov}}(\mathbf{z}, \mathbf{z}') \quad (23)$$

where $\mathbf{z} = [A, Y_{\text{lag}}, \bar{Y}_{\text{spatial_lag}}]$ represents the non-temporal features. This product structure allows for a non-stationary model where the covariance structure of the covariates can evolve over time.

Temporal Kernel: We use the Matérn 5/2 kernel to model temporal dependencies, which is a common choice for physical processes as it assumes the underlying function is twice-differentiable.

$$k_{\text{Matern52}}(t, t') = \sigma_t^2 \left(1 + \frac{\sqrt{5}d}{l_t} + \frac{5d^2}{3l_t^2} \right) \exp \left(-\frac{\sqrt{5}d}{l_t} \right) \quad (24)$$

where $d = |t - t'|$, and σ_t^2 and l_t are the variance and length-scale hyperparameters, respectively.

Covariate Kernel: We use the standard Radial Basis Function (RBF) kernel, also known as the squared exponential kernel, for the remaining covariates. We employ Automatic Relevance Determination (ARD), which assigns a separate lengthscale parameter l_d to each dimension d of the input $\mathbf{z}$.

$$k_{\text{RBF-ARD}}(\mathbf{z}, \mathbf{z}') = \sigma_z^2 \exp \left(-\frac{1}{2} \sum_{d=1}^D \frac{(z_d - z'_d)^2}{l_d^2} \right) \quad (25)$$

The ARD mechanism allows the model to learn the relative importance of each feature.

Model Fitting For each unit i , the hyperparameters $\theta_i = \{\sigma_n^2, \sigma_t^2, l_t, \sigma_z^2, \{l_d\}\}$ of the model $\mathcal{M}_i$ are learned by maximizing the log marginal likelihood of the observed data $\mathcal{D}_i$. This is typically done using gradient-based optimizers. The resulting trained models $\{\mathcal{M}_i\}_{i=1}^N$ are then used in Stage 2 of our estimation algorithm (Algorithm 1) to perform counterfactual simulations.

Crucially, while $\mathcal{M}_i$ is trained as a predictive model, its role in our framework is to serve as a non-parametric estimator for the unit-specific conditional dynamics. Because $\mathcal{M}_i$ is trained on data centered at unit i , it implicitly conditions on the latent static properties U_i . Therefore, under the identification guarantees of Theorem 3, using this trained model $\mathcal{M}_i$ to simulate outcomes under a do-intervention (as performed in Stage 2 of Algorithm 1) yields a valid estimate of the conditional causal effect $E[\cdot | \text{do}(\cdot), U_i]$.

E Experiment Details

This section provides supplementary details for the experiments presented in the main paper. We elaborate on the synthetic data generation process, model implementations, and provide a description of the real-world dataset.

E.1 Details of the simulation experiment

Data Generation Process We generate synthetic data from a Spatio-Temporal Hierarchical Causal Model (ST-HCM) designed to test the core challenges of unobserved confounding and spatial spillover. The process involves N units, each with m subunits, observed over T time steps.

A time-invariant, unobserved confounder U_i is drawn for each unit i from a standard normal distribution, $U_i \sim \mathcal{N}(0, 1)$. This confounder influences both treatment assignment and outcomes throughout the simulation.

The results presented in the main paper’s Table 1 are based on a data generating process with linear relationships, providing a clear testbed for the causal identification capabilities of different model structures. The generation process for each subunit (i, j) at time t is as follows:

1. Treatment Assignment: The treatment $A_{ij,t} \in \{0, 1\}$ is a binary variable drawn from a Bernoulli distribution whose probability is a non-linear function of the confounder U_i :

$$A_{ij,t} \sim \text{Bernoulli}(\sigma(\gamma \cdot U_i - 0.5)) \quad (26)$$

where $\sigma(\cdot)$ is the sigmoid function and γ is the **confounding strength**. This ensures that treatment assignment is endogenously determined by the unobserved confounder.

2. Outcome Generation: The outcome $Y_{ij,t}$ is generated as a linear combination of the confounder’s main effect, the treatment effect, temporal and spatial lags, and i.i.d. noise:

$$Y_{ij,t} = \gamma \cdot U_i + \beta_A \cdot A_{ij,t} + \beta_{\text{temp}} \cdot \bar{Y}_{i,t-1} + \rho \cdot \bar{Y}_{\mathcal{N}(i),t-1} + \epsilon_{ij,t} \quad (27)$$

where:

- β_A is the true Average Treatment Effect (ATE).
- $\bar{Y}_{i,t-1}$ is the average outcome of unit i at time $t - 1$.

- $\bar{Y}_{\mathcal{N}(i),t-1}$ is the average outcome of unit i ’s neighbors $\mathcal{N}(i)$ at time $t - 1$.
- ρ is the spatial spillover strength.
- $\epsilon_{ij,t} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the subunit-level noise.

Simulation Parameters: For the experiments in Table 1, we used the following parameters: $N = 16$, $m = 50$, $T = 8$, $\beta_A = 5.0$, $\beta_{\text{temp}} = 0.5$, $\sigma_\epsilon^2 = 4$. The values for γ and ρ are varied as indicated in the table.

E.2 Robustness Analysis Data Generation

This section details the data generation processes used for the robustness analysis. Our aim is to systematically introduce controlled violations of key assumptions underlying hierarchical linear models (LMMs), particularly the time-invariance of the unobserved confounder and the acyclic nature of contemporaneous spatial effects. All experiments are conducted with $N = 16$ units, $m = 50$ subunits, and $T = 8$ time steps. The base confounding strength is $\gamma = 2.0$ and spatial spillover strength is $\rho = 1.5$.

Violating Time-Invariance of the Confounder The standard ST-HCM (LMM) with fixed effects assumes that the unobserved unit-level confounder U_i is time-invariant. To violate this assumption, we allow U_i to evolve over time, transforming it into a time-varying confounder U_{it} . Specifically, U_{it} follows a simple random walk:

$$U_{it} = U_{i,t-1} + \delta \cdot \epsilon_{it}^U$$

where $U_{i,0}$ is drawn from a standard normal distribution, ϵ_{it}^U is i.i.d. standard normal noise, and δ is the confounder drift strength parameter. A higher δ implies a more rapidly changing confounder, making it harder for fixed effects to control. The treatment assignment and outcome generation mechanisms then depend on this time-varying U_{it} instead of U_i . When $\delta = 0$, the confounder is static, corresponding to the assumption holding.

Violating Spatial Ordering Our framework assumes a causal ordering among units at each time step, implying that contemporaneous spatial influences are unidirectional (e.g., from unit j to unit i only if $j < i$). To violate this strict ordering and introduce simultaneous, bi-directional spatial effects, we modify the noise term of the outcome generation. Instead of being independent, the subunit-level noise $\epsilon_{ij,t}$ is influenced by the average noise of its neighbors at the same time step. Specifically, the observed noise for subunit j in unit i at time t becomes:

$$\tilde{\epsilon}_{ij,t} = \epsilon_{ij,t} + \kappa \cdot \left(\frac{1}{|\mathcal{N}(i)|} \sum_{k \in \mathcal{N}(i)} \epsilon_{kj,t} \right)$$

where $\epsilon_{ij,t}$ are i.i.d. standard normal noise, $\mathcal{N}(i)$ denotes the set of all spatial neighbors of unit i , and κ is the cyclicity strength parameter. A non-zero κ introduces contemporaneous spatial correlation in the unobserved error terms. This violation cannot be fully controlled by standard linear models, as it introduces endogenous error terms that are correlated across space within the same time step. When $\kappa = 0$, the spatial ordering assumption holds, and the noise is independent across units.

E.3 Chicago Traffic Dataset

Our real-world analysis uses three publicly available datasets from the City of Chicago data portal¹. We focus on a one-week period in January 2025 (from January 5th to January 11, 2025) for our case study.

Data Sources The core components of our dataset are:

- **Regions:** The 29 official traffic regions (our *units*) provide the spatial partitioning of the city. Each region is composed of two to three community areas with comparable traffic patterns.
- **Segments:** Arterial traffic segments (our *subunits*) provide real-time speed estimates every 10 minutes. These estimates are derived from the GPS traces of Chicago Transit Authority (CTA) buses. As noted by the data provider, individual segment speeds can be highly volatile due to factors like traffic signals and intersections.
- **Crashes:** Records of traffic accidents provide the precise time and location of crash events, which form the basis of our treatment variable.

Preprocessing and Feature Engineering We first performed a spatial join to assign each of the 1,025 unique road segments to its corresponding traffic region(s). The crash events were then assigned to a region and subsequently assigned to the nearest road segment within the boundaries of that region.

The key step in our analysis was defining the treatment variable at the unit level. A region was considered treated at a given 10-minute time step if one or more crashes occurred within it. This *unit-level treatment* allows its effect to be estimated across all subunits in the region. The final feature set for our full ST-HCM model included:

- **Temporal lags:** The subunit's own speed at $t - 1$ (`speed_lag1`) and the unit's treatment status at $t - 1$ (`unit_treatment_lag1`).
- **Spatial lags:** The average speed and treatment status of neighboring regions at $t - 1$.
- **Time features:** The hour of the day and the day of the week, treated as categorical variables.

¹<https://data.cityofchicago.org/>