Global Average Treatment Effects for Individualized Randomization Experiments with Aggregate Data

Anonymous Author(s)
Anonymous Institution
Anonymous Address
anonymous@email.com

Abstract

Individualized randomized experiments are central to online platforms for optimizing personalized decisions in complex environments. In two-sided markets, however, standard treatment effect estimation is often invalid due to strong temporal and cross-unit interference, a challenge compounded when only aggregated data are available because of privacy or system constraints. To address these issues, we identify the Global Average Treatment Effect (GATE) using only group-level data from treatment and control groups. We first establish identification conditions based on aggregated observations, and then propose the Individualized Randomized Experiment Varying Coefficient Decision Process (IRE-VCDP) model, which accounts for interference through supply–demand dynamics. Building on this framework, we develop a complete procedure for estimation and statistical inference of the GATE, along with theoretical guarantees for the proposed test. Extensive simulations and real-world experiments using data from a leading ridesharing platform demonstrate the effectiveness of our approach.

1 Introduction

Individualized randomized experiments (IREs) assign treatments at the individual level over time, with treatment status held fixed within each experimental period. Policy evaluation under IREs is of particular importance in two-sided markets — such as ridesharing, food delivery, and freelance platforms — where interventions simultaneously affect both sides of the market and induce complex interactions and interference patterns across participants [Johari et al., 2022, Shi et al., 2023]. The primary objective is to introduce an intervention to a subset of the market and use the observed outcomes to estimate its effect if deployed at full scale. We refer to this estimand as the Global Average Treatment Effect (GATE, formally defined in (1)). Accurate estimation of GATE is central to platform decisions about whether to roll out a new policy to the entire market.

Our study is motivated by a concrete ridesharing marketplace setting, in which a city-scale online experiment is conducted to evaluate a new passenger subsidy policy. Passengers are randomly assigned at the outset to either a treatment group, receiving the new subsidy, or a control group, remaining under the existing policy, over a two- to four-week experimental period. In practice, individual-level data are often unavailable due to privacy constraints. Instead, only aggregated group-level metrics are observed in one-hour intervals, including Gross Merchandise Value (GMV), passenger demographic summaries, and key supply and demand variables such as order frequencies and driver online hours. The platform's goal is to quantify the causal impact of this policy on GMV under full deployment using only these aggregate observations.

Challenges. Despite the widespread adoption of IREs in two-sided markets, reliable treatment effect estimation in this dynamic setting remains fundamentally difficult, for three reasons. The first challenge is cross-unit interference: interventions applied to one side of the market, for example, a

Submitted to STAI-X 2026 (Full Paper Track). Do not distribute.

passenger subsidy, propagate through supply–demand interactions and affect outcomes on the other side, thereby violating the standard Stable Unit Treatment Value Assumption (SUTVA) [Imbens and Rubin, 2015]. The second challenge is temporal interference: treatment effects evolve over time through dynamic market adjustments, so that an intervention at one time point can influence both contemporaneous and future outcomes [Wang et al., 2023, Li et al., 2024, 2025]. Ignoring either form of interference leads to biased treatment effect estimates [Li et al., 2023]. The third challenge is data availability: privacy and system constraints restrict access to individual-level records, leaving only aggregated group-level observations for analysis — a regime that remains largely underexplored in the causal inference literature. Taken together, these challenges render classical causal inference methods inadequate and motivate the development of new approaches tailored to two-sided markets.

Contributions. The primary goal of this paper is to study the GATE in individualized randomized experiments within two-sided markets using aggregated data, while explicitly accounting for both cross-unit and temporal interference. Our contributions are fourfold.

(i) **Identification.** We establish identification conditions for the GATE using only group-level data from the treatment and control groups, together with market-level features. This result enables valid GATE estimation without access to individual-level records. To the best of our knowledge, this is the first work to provide such identification based solely on aggregated data.

(ii) **Modeling framework.** We propose a two-layer modeling framework, the Varying Coefficient Decision Process (VCDP), that separates the outcome process from the underlying supply–demand dynamics. Interference is encoded through the supply–demand mechanism, reflecting the intrinsic structure of two-sided markets in which units interact via matching. The VCDP accommodates time-varying treatment effects and dynamic interference, with supply and demand variables serving as mediators that jointly capture cross-unit and temporal interference.

(iii) **Estimation.** We show that the GATE can be expressed as an explicit function of the VCDP model coefficients, and we develop a two-step estimation procedure that combines least squares estimation with kernel smoothing to improve efficiency and stability under weak signals and limited sample sizes. The GATE is then recovered via plug-in estimation from the fitted coefficients.

(iv) **Inference and empirical validation.** We develop a bootstrap-based testing procedure, adapted from the high-dimensional Gaussian approximation framework of Chernozhukov et al. [2013], to assess whether deploying the new treatment across the full market improves platform revenue relative to the baseline policy in the setting of aggregate-level observations and supply–demand interference.

1.1 Related Work

Causal inference in A/B testing. The classical rationale for A/B testing rests on the SUTVA [Rubin, 1980], which requires that the treatment assigned to one unit does not affect the outcomes of others. In two-sided marketplaces, however, interference is typically unavoidable: interventions on one side of the market propagate through supply–demand interactions to affect outcomes on the other side, rendering standard A/B analyses causally invalid [Halloran and Hudgens, 2016, Sävje et al., 2021].

Within the switchback design literature, there is a rich body of work on policy evaluation under interference [Hu and Wager, 2022, Bojinov et al., 2023, Farias et al., 2023, Luo et al., 2024, Li et al., 2024]. By contrast, individualized randomized experiments (IREs) have received considerably less attention. The most closely related work is Liu and Hou [2023], who proposed UniTE to estimate individual-level conditional average treatment effects in two-sided markets via Robinson decomposition.

While effective in principle, UniTE requires granular transaction-level data that are often severely sparse in practice – in a two-week experiment, a single passenger may complete only one ride. A related line of work studies off-policy evaluation in contextual bandits, including doubly robust and adaptive weighting estimators for evaluating target policies from logged individual-level data [Dudik et al., 2011, Wang et al., 2017, Zhan et al., 2021]. These methods provide useful individual-level policy evaluation tools, but they do not explicitly model the dynamic demand–supply feedback that drives interference in two-sided markets. Similarly, recent approaches that leverage aggregate statistics of state evolution for causal inference under interference, such as those by Shirani et al. [2025] and Shirani and Bayati [2025], still fundamentally require access to individual-level observations to construct subpopulations or learn state dynamics. Practitioners constrained to aggregated group-level

data are thus largely limited to simple mean differences or difference-in-differences estimators, which are well known to be biased in the presence of interference [Imbens and Rubin, 2015]. Bridging this gap between data availability and methodological requirements is the central motivation of this paper. However, if individual-level data or empirical distribution data are accessible, methods such as the Distribution-Preserving Network Bootstrap (DPNB) [Shirani et al., 2025] and the Experimental State Evolution (ESE) framework [Shirani and Bayati, 2025] can be employed to diagnose the form of interference. Once the interference structure is characterized, aggregate-level data can be released for GATE estimation using our method.

Varying coefficient models. Varying coefficient models, introduced by Hastie and Tibshirani [1993], allow regression coefficients to vary smoothly as functions of an index variable such as time. A rich estimation and inference theory has since developed, spanning kernel and local polynomial methods [Hoover et al., 1998, Fan and Zhang, 1999], spline-based approaches for longitudinal data [Chiang et al., 2001, Huang et al., 2004], and broad applications in spatial, epidemiological, and longitudinal settings [Sun et al., 2023, Zhang et al., 2025, Sadovnichiy et al., 2022]; see Park et al. [2015] for a comprehensive survey. More recently, varying coefficient models have been applied to online A/B testing, though existing work is confined to switchback designs [Luo et al., 2024, Li et al., 2024]. To the best of our knowledge, the present paper is the first to extend this framework to individualized randomized experiments, where interference must be handled through the supply–demand structure rather than the experimental design itself.

2 Problem Formulation

Data. Consider an experiment conducted over n consecutive days, each divided into m time intervals. At the outset, $N = N_0 + N_1$ subjects are randomly assigned to either a treatment group of size N_1 or a control group of size N_0, with assignments held fixed throughout the experiment. For day $d \in \{1, \dots, n\}$ and interval $t \in \{1, \dots, m\}$, subjects in the treatment group receive the new policy ($A_t^d = 1$) and subjects in the control group receive the baseline policy ($A_t^d = 0$). Let $Y_d^i(t) \in \mathbb{R}$ denote the outcome for subject i at interval t on day d.

Due to privacy constraints, individual-level outcomes are not observed. Instead, only group-level aggregates are available for each day d and interval t. Specifically, for each group C_k ($k \in \{0, 1\}$), we observe three types of summaries: the average outcome $Y_{d,C_k}(t) = N_k^{-1} \sum_{i \in C_k} Y_d^i(t)$, the average covariate vector $X_{d,C_k}(t) \in \mathbb{R}^p$, and the group-level market variables $S_{d,C_k}(t) \in \mathbb{R}^q$, which capture supply and demand conditions faced by each group. Observations are assumed to be independent across days, while temporal dependence within each day is explicitly permitted.

Objective. We adopt the potential outcomes framework [Rubin, 2005] to formalize the problem. Let $\bar{a}_t^i = (a_1^i, \dots, a_t^i)^\top \in \{0, 1\}^t$ denote the treatment history of subject i up to time t, and let $\bar{\boldsymbol{a}}_{N,t} = (\bar{a}_t^1, \dots, \bar{a}_t^N)$ denote the joint treatment history for all N subjects. Define $S(t, \bar{\boldsymbol{a}}_{N,t-1})$ and $Y^i(t, \bar{\boldsymbol{a}}_{N,t})$ as the counterfactual market state and outcome for subject i, respectively, under treatment history $\bar{\boldsymbol{a}}_{N,t}$. Our primary objective is to quantify the Global Average Treatment Effect (GATE), defined as the difference in expected cumulative outcomes between deploying the new policy to the entire market and maintaining the baseline policy:

$$\text{GATE} = \frac{1}{N}\mathbb{E}\{\sum_{t=1}^{m}\sum_{i=1}^{N} Y^i\big(t, \mathbf{1}_{N,t}\big)\} - \frac{1}{N}\mathbb{E}\{\sum_{t=1}^{m}\sum_{i=1}^{N} Y^i\big(t, \mathbf{0}_{N,t}\big)\}, \tag{1}$$

where $\mathbf{1}_{N,t}$ (respectively, $\mathbf{0}_{N,t}$) denotes that all entries in $\bar{\boldsymbol{a}}_{N,t}$ are equal to 1 (respectively, 0).

Note that the GATE in (1) is defined at the daily level: it measures the expected difference in cumulative outcomes over the m time intervals within a single day. In estimation, we pool observations across all n experimental days to estimate this common day-level estimand.

Our goal is to estimate the GATE using aggregated data and to test the hypothesis

$$H_0 : \text{GATE} \leq 0 \quad \text{versus} \quad H_1 : \text{GATE} > 0, \tag{2}$$

which evaluates whether deploying the new policy leads to an improvement over the baseline in terms of overall market outcomes.

3 Identification

The SUTVA [Imbens and Rubin, 2015] is violated in our setting for two reasons. First, a subject's outcome depends not only on their own current treatment, but also on their prior treatment history. Second, the treatment trajectories of other subjects affect individual outcomes through shared supply–demand dynamics, inducing both cross-unit and temporal interference. Without additional structure on the nature of this interference, identifying the treatment effect may be impossible [Basse and Airoldi, 2018, Liu et al., 2024]. We therefore impose the following four assumptions to identify the GATE from the observed aggregate data.

Assumption 3.1 (Dynamic local interference)**.** For any $\bar{\boldsymbol{a}}_{N,t}$ and $\bar{\boldsymbol{a}}'_{N,t}$, if for each t, $a_t^i = a_t'^i$ and there exists a known function $g(\cdot)$ such that $g(a_t^1, \ldots, a_t^N) = g(a_t'^1, \ldots, a_t'^N)$, then $Y^i(t, \bar{\boldsymbol{a}}_{N,t}) = Y^i(t, \bar{\boldsymbol{a}}'_{N,t})$, and $S(t, \bar{\boldsymbol{a}}_{N,t-1}) = S(t, \bar{\boldsymbol{a}}'_{N,t-1})$.

Assumption 3.2 (Consistency)**.** If the treatment policy satisfies $\bar{\boldsymbol{A}}_{N,t} = \bar{\boldsymbol{a}}_{N,t}$, then $S(t) = S(t, \bar{\boldsymbol{a}}_{N,t-1})$ and $Y^i(t) = Y^i(t, \bar{\boldsymbol{a}}_{N,t})$.

Assumption 3.3 (Sequential ignorability)**.** The treatment $\boldsymbol{A}^i(t)$ assigned to subject i at time t is conditionally independent of all potential variables given the observed data history.

Assumption 3.4 (Bernoulli design)**.** Each subject is randomly assigned to treatment with probability p and to control with probability $1-p$, where $0 < p < 1$. That is, $\mathbb{P}\{A^i(t) = 1\} = p$ for all i and t.

We now discuss the role and justification of each assumption. Assumption 3.1 restricts the interference structure: it requires that a subject's outcome and the subsequent market state depend on other subjects' treatments only through a summary statistic of treatment assignments, rather than through individual identities or specific assignments. For example, the function $g(\cdot)$ can be taken as the mean function or a quantile function, similar to forms used in the exposure mapping literature [Yang et al., 2018, Li and Wager, 2022, Shi et al., 2023]. This structure generalizes the static local interference assumption of Johari et al. [2022] and Masoero et al. [2026] to a dynamic setting, and extends the classical SUTVA by accounting for interference among subjects. If individual-level data are available, methods like DPNB [Shirani et al., 2025] or ESE [Shirani and Bayati, 2025] can be used to diagnose the interference form and determine the appropriate $g(\cdot)$. Our identification result holds under this identified structure, which need not be mean-field. It is particularly natural in two-sided markets, where individual outcomes are driven by market-level supply and demand conditions rather than by the identity of any particular treated peer.

Assumption 3.2 is a standard consistency condition: it requires that potential outcomes and states, under the observed treatment history, coincide with the actually observed outcomes and states. Assumption 3.3 is a sequential ignorability condition requiring that, given the accumulated data history, the treatment at each period is conditionally independent of all remaining potential variables; this is a standard assumption in the dynamic treatment literature [Wang et al., 2018]. Assumption 3.4 specifies a Bernoulli randomized design at the outset of the experiment, which guarantees that both treatment and control groups have positive probability and is necessary for identification of the GATE from aggregate data. We note that the identification result continues to hold in non-randomized settings as long as treatment assignment probabilities are bounded away from zero and one.

Proposition 3.5. *Suppose Assumptions 3.1–3.4 hold, and that there exists a function $R(\cdot)$ such that*

$$\mathbb{E}\Big(Y^i(t, \bar{\boldsymbol{a}}_{N,t}) \,\Big|\, \bar{\boldsymbol{a}}_{N,t}, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j \le t}\Big) = R\Big(a_t^i,\ \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j \le t}\Big), \tag{3}$$

where $f(j)$ denotes the fraction of treated subjects at time j. Then the following results hold.

(i) Identification of $R(\cdot)$ from observed data. *Under the full-treatment policy $\bar{\boldsymbol{a}}_{N,t} = \mathbf{1}_{N,t}$ and the full-control policy $\bar{\boldsymbol{a}}_{N,t} = \mathbf{0}_{N,t}$, the function $R(\cdot)$ can be learned from the observed data as*

$$R\Big(a,\ \{S(j), f(j)\}_{j \le t}\Big) = \mathbb{E}\Big(N_a^{-1} \sum_{i \in C_a} Y^i(t) \,\Big|\, \{S(j), f(j)\}_{j \le t}\Big), \tag{4}$$

where N_a denotes the number of subjects with $a_t^i = a$, for $a \in \{0, 1\}$.

(ii) Identification of GATE from observed data. *The GATE defined in* (1) *is identifiable from the observed summary data. Specifically,*

$$\begin{aligned} & \mathbb{E}\Big(Y^i(t, \bar{\boldsymbol{a}}_{N,t})\Big) = \mathbb{E}\Big[R\Big(a,\, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}\Big)\Big] \qquad (5) \\ = \;& \mathbb{E}\Big[\mathbb{E}\Big\{R\Big(a,\, \{S(j), f(j)\}_{j\le t}\Big) \;\Big|\; \bar{\boldsymbol{A}}_{N,t} = \bar{\boldsymbol{a}}_{N,t},\, \{S(j), f(j)\}_{j\le t},\, \{Y(j)\}_{j<t}\Big\}\Big]. \end{aligned}$$

Proposition 3.5 shows that the GATE can be presented as a function of the observed summary data, making it identifiable without access to the individual-level data. When $\bar{\boldsymbol{a}}_{N,t} = \mathbf{1}_{N,t}$, the corresponding ratio is $f(t) = 1$; when $\bar{\boldsymbol{a}}_{N,t} = \mathbf{0}_{N,t}$, we have $f(t) = 0$.

4 Models for the outcome and state variables

We introduce the IRE-VCDP model to jointly characterize outcome and state dynamics in two-sided markets. The state variables explicitly encode the platform's demand–supply system. For each group $C \in \{C_0, C_1\}$, let $D_{d,C}(t)$ denote group-specific demand and $S_d(t)$ denote platform-wide supply, which is shared across groups since drivers serve all users regardless of group membership. We define the group-level state vector as $S_{d,C}(t+1) = (D_{d,C}(t+1), S_d(t+1))^\top$.

The outcome and state evolution are modeled as

$$\begin{aligned} Y_{d,C}(t) &= g_C\big(t,\, X_{d,C}(t),\, S_{d,C}(t)\big) + \epsilon_{d,C}(t), \\ S_{d,C}(t+1) &= G_C\big(t,\, \widetilde{X}_{d,C}(t),\, S_{d,C}(t),\, f_d(t)\big) + E_{d,C}(t+1), \end{aligned}$$

where $g_C(\cdot)$ and $G_C(\cdot)$ are unknown regression functions. The outcome is driven by group characteristics $X_{d,C}(t)$ and the current demand–supply state $S_{d,C}(t)$, while the state transition depends additionally on the fraction of treated subjects $f_d(t)$, which serves as the channel through which the treatment propagates across groups. The error terms $\epsilon_{d,C}(t)$ and $E_{d,C}(t+1)$ are mutually independent; however, we allow correlation between $\epsilon_{d,C_0}(t)$ and $\epsilon_{d,C_1}(t)$ to accommodate shared latent shocks at the platform level. The inclusion of covariates $X_{d,C}(t)$ and $\widetilde{X}_{d,C}(t)$ improves estimation efficiency [Su and Ding, 2021].

To enable tractable estimation and interpretation, we adopt the following linear approximations:

$$Y_{d,C}(t) = \alpha_{0,C}(t) + \alpha_{1,C}^\top(t)\, X_{d,C}(t) + \alpha_{2,C}^\top(t)\, S_{d,C}(t) + \epsilon_{d,C}(t), \qquad (6)$$

$$S_{d,C}(t+1) = \gamma_{0,C}(t) + f_d(t)\, \gamma_{1,C}(t) + \Phi_{0,C}(t)\, \widetilde{X}_{d,C}(t) + \Phi_{1,C}(t)\, S_{d,C}(t) + E_{d,C}(t+1). \qquad (7)$$

This specification is motivated by both structural and empirical considerations. Structurally, it reflects the platform's operational asymmetry: demand is group-specific, arising from user-level randomization, while supply is shared across groups. This asymmetry is the mechanism through which cross-group interference operates — a demand shock in one group affects service availability for the other through the common supply pool, and consequently affects outcomes. Note that the treatment fraction $f_d(t)$ does not appear directly in the outcome model; its influence operates indirectly through the state variable $S_{d,C}(t)$. This reflects the ridesharing setting where the treatment effect is mediated by supply–demand dynamics. In other experimental settings where $f_{d,t}$ may directly affect Y independent of the state variables, the outcome model can be straightforwardly extended to include $f(t)$ as an additional regressor.

A key feature of two-sided markets is that the imbalance between supply and demand is a primary determinant of outcomes [Chen et al., 2024]. When supply is sufficient, demand and GMV exhibit an approximately linear relationship. Under excess demand, however, a fraction of ride requests cannot be fulfilled, and this congestion breaks down the linear relationship. To capture this nonlinearity, we incorporate the supply–demand gap

$$G_d(t) = \max\{\mathrm{Demand}_d(t) - \mathrm{Supply}_d(t),\, 0\} \qquad (8)$$

as a component of $X_{d,C}(t)$, where $\mathrm{Demand}_d(t)$ is the number of ride requests and $\mathrm{Supply}_d(t)$ is the number of rides that can be fulfilled given total available driver hours, both in city d at time t. The gap $G_d(t)$ captures congestion effects that are particularly pronounced during peak periods and play a crucial role in explaining GMV dynamics over the experimental horizon, as confirmed by our empirical analysis.

The covariate vector $X_{d,C}(t)$ also includes the average per-subject non-experimental subsidy received by subjects in group C, i.e., subsidies not controlled by the experimental policy. The state vector $S_{d,C}(t)$ has two components: $D_{d,C}(t)$, the total ride requests from group C_k at interval t, and $S_d(t)$, the total driver online hours citywide, shared across groups. The covariate vector $\widetilde{X}_{d,C}(t)$ contains city-level weather and calendar variables: temperature, precipitation, and a holiday indicator.

Based on models (6)–(7), the GATE in (1) admits a closed-form representation as a function of the model coefficients, which greatly facilitates estimation and inference. The coefficients in (6)–(7) can be consistently estimated using the observed aggregate data. The GATE estimator is then obtained by plugging these estimates into (5), with $f(t) = 1$ corresponding to full-scale treatment and $f(t) = 0$ corresponding to the control.

Proposition 4.1. *Suppose the conditions in Proposition 3.5 holds, and the potential outcomes admits the form of models in* (6)-(7)*, then we have*

$$
\begin{aligned}
GATE = &\sum_{t=1}^{m}(\alpha_{0,C_1}(t) - \alpha_{0,C_0}(t)) + \sum_{t=1}^{m}\Big[\bar{X}_N(\alpha_{1,C_1}(t) - \alpha_{1,C_0}(t))\Big] + \qquad (9)\\
&\sum_{t=1}^{m}\Big[\alpha_{2,C_1}(t)^\top\Big(\prod_{l=1}^{t-1}\Phi_{1,C_1}(l)\mathbb{E}(S_{C_1}(1)) + \sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}\Phi_{1,C_1}(l)\cdot[\gamma_{0,C_1}(k) + \gamma_{1,C_1}(k) + \Phi_{0,C_1}(k)\bar{\tilde{X}}_N(k)]\}\Big)\\
&- \alpha_{2,C_0}(t)^\top\Big(\prod_{l=1}^{t-1}\Phi_{1,C_0}(l)\mathbb{E}(S_{C_0}(1)) + \sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}\Phi_{1,C_0}(l)\cdot[\gamma_{0,C_0}(k) + \Phi_{0,C_0}(k)\bar{\tilde{X}}_N(k)]\}\Big)\Big],
\end{aligned}
$$

where $\bar{X}_N = \mathbb{E}(N^{-1}\sum_{i=1}^{N}X^i(t))$ *and* $\bar{\tilde{X}}_N(t) = \mathbb{E}(N^{-1}\sum_{i=1}^{N}\tilde{X}^i(t))$, $\mathbb{E}(S(1))$ *denotes the demand and supply of all subjects, and* $\prod_{l=k+1}^{t-1}\Phi_{1,C_0}(l) = \prod_{l=k+1}^{t-1}\Phi_{1,C_1}(l) = 1$*, if* $t-1 < k+1$.

From Proposition 4.1, the GATE can be estimated by plugging the estimated coefficients from models (6)–(7) into the closed-form expression. The resulting GATE admits a natural decomposition into five interpretable components. See Appendix E for details.

If the predictive covariates are centered to have zero mean, the expression of the GATE in Proposition 4.1 further simplifies. In practice, this can be achieved by centering the covariates across all subjects (rather than within each group). While such centering simplifies the expression, the inclusion of these covariates also improves estimation efficiency by reducing variance.

To better illustrate the role of interference, we consider a naive treatment effect defined as the difference in mean outcomes between the two groups: $\tau = \mathbb{E}\Big(\sum_{t=1}^{m}Y_{C_1}(t)\Big) - \mathbb{E}\Big(\sum_{t=1}^{m}Y_{C_0}(t)\Big)$. This estimand ignores interference across subjects. Under models (6)–(7), τ can also be expressed in terms of the model coefficients, capturing both the direct treatment effect and its interactions with predictive and state variables. It can be viewed as a dynamic extension of the regression-adjusted estimator of Li and Ding [2020]. However, τ fails to account for temporal and cross-subject interference. Consequently, it coincides with the GATE only when the expected demand and supply processes are identical under full-scale treatment and under control—an assumption that is typically violated in practice. See Appendix F.1 for detailed derivations and discussion.

5 Estimation and Inference

This section presents the estimation procedure and statistical inference framework for the hypothesis (2), based on models (6)–(7).

Estimation. Estimation of the GATE proceeds in two steps. In the first step, the time-varying coefficients in models (6)–(7) are estimated by ordinary least squares at each time point. Specifically, for $t = 1, \ldots, m$ (and $t = 1, \ldots, m-1$ for the state equation), define

$$
\widehat{\theta}_C(t) = \arg\min_{\theta_C(t)}\sum_{d=1}^{n}\Big(Y_{d,C}(t) - Z_{d,C}(t)^\top\theta_C(t)\Big)^2,
$$

$$
\widehat{\Theta}_C^{(\nu)}(t) = \arg\min_{\Theta^{(\nu)}(t)}\sum_{d=1}^{n}\Big(S_{d,C}^{(\nu)}(t+1) - \widetilde{Z}_{d,C}(t)^\top\Theta_C^{(\nu)}(t)\Big)^2,
$$

where $C \in \{C_0, C_1\}$, $Z_{d,C}(t) = [1,\ X_{d,C}(t)^\top,\ S_{d,C}(t)^\top]^\top$, $S^{(\nu)}_{d,C}(t{+}1)$ denotes the ν-th component of $S_{d,C}(t+1)$, and $\widetilde{Z}_{d,C}(t) = [1,\ f(t),\ \widetilde{X}_{d,C}(t)^\top,\ S_{d,C}(t)^\top]^\top$.

In the second step, kernel smoothing is applied to the initial estimates to reduce variance, enforce temporal smoothness, and improve signal detection under weak effect sizes [Zhu et al., 2014]. For a kernel function $K(\cdot)$ with bandwidth h, the smoothed estimators are defined as

$$\widetilde{\theta}_C(t) = \sum_{j=1}^{m} \omega_{j,h}(t)\, \widehat{\theta}_C(j) \qquad \text{and} \qquad \widetilde{\Theta}^{(\nu)}_C(t) = \sum_{j=1}^{m-1} \omega_{j,h}(t)\, \widehat{\Theta}^{(\nu)}_C(j), \tag{10}$$

where $\omega_{j,h}(t) = K((j-t)/(mh)) \,/\, \sum_{k=1}^{m} K((k-t)/(mh))$ are normalized kernel weights. We adopt a truncated Gaussian kernel $K(u) = c^{-1}\exp(-u^2)\cdot \mathbb{I}\{|u| \le 1\}$, where $c = \int_{-1}^{1} \exp(-s^2)\, ds$ is the normalizing constant. The smoothed estimators are weighted averages of the initial pointwise estimates, with weights decaying as a function of temporal distance. This yields continuous, stable coefficient trajectories that are particularly well-suited for settings in which the signal is weak or the underlying parameters evolve slowly over time.

The GATE estimator $\widehat{\text{GATE}}$ is then obtained by substituting $\widetilde{\theta}_C(t)$ and $\widetilde{\Theta}^{(\nu)}_C(t)$ into the closed-form expression of Proposition 4.1, with the population means of the predictive covariates and the initial state variable replaced by their empirical counterparts across all subjects (see Appendix F.2).

Inference. To test the hypothesis (2), we use the test statistic $T = \widehat{\text{GATE}}$. Under the null hypothesis, T is expected to be non-positive or close to zero, so we reject H_0 when T exceeds a sufficiently large positive threshold. Deriving the exact limiting distribution of T is analytically intractable due to its complex functional dependence on the estimated model parameters. We therefore approximate the null distribution of T via a multiplier bootstrap procedure adapted from Chernozhukov et al. [2013].

Specifically, after estimating the outcome and state models, fitted values and residuals are computed within each group. The residuals are then perturbed using independent standard normal multipliers to generate bootstrap pseudo-data, thereby mimicking the variability of the original observations. For each bootstrap sample, the full estimation procedure is repeated using the pseudo-data to obtain a bootstrap GATE estimate $\widehat{\text{GATE}}^{\,b}$. The bootstrap statistic is defined as $T^b = \widehat{\text{GATE}}^{\,b} - \widehat{\text{GATE}}$. Repeating this procedure over B replications provides an empirical approximation of the null limit distribution. Then we reject H_0 at significance level α if T excedds the $1-\alpha$ quantile of $\{T_b\}_{b=1}^{B}$.

The complete estimation and inference procedure is summarized in Algorithm 1 in Appendix A. The bootstrap-based critical value enables valid inference without requiring a closed-form asymptotic distribution. The theoretical validity of Algorithm 1 is established below.

Theorem 5.1. *Suppose Assumptions G.1–G.4 in the supplement hold. If $h = o(n^{-1/4})$, $m \asymp n^{c_2}$ for some $\frac{1}{2} < c_2 < \frac{3}{2}$, and $mh \to \infty$ as $n \to \infty$, then*

$$\sup_z \left|\mathbb{P}(T - \text{GATE} \le z) - \mathbb{P}\big(T^b - T \le z \mid \text{Data}\big)\right| \;\le\; \widetilde{C}\Big(\sqrt{n}\,h^2 + \sqrt{n}\,m^{-1} + n^{-1/8}\Big)$$

holds with probability approaching 1*, for some positive constant $\widetilde{C}$.*

Theorem 5.1 establishes that the bootstrap distribution of $T^b - T$ consistently approximates the sampling distribution of $T - \text{GATE}$, with an explicit convergence rate. The bound comprises three terms: a smoothing bias $O(\sqrt{n}\,h^2)$ from kernel estimation, a discretization error $O(\sqrt{n}\,m^{-1})$ due to the finite number of time intervals, and the Gaussian approximation error $O(n^{-1/8})$ inherited from the theory of Chernozhukov et al. [2013]. The bandwidth conditions $h = o(n^{-1/4})$ and $mh \to \infty$ balance these three sources of error and ensure that the overall bound vanishes as $n \to \infty$. The proof is given in Appendix H.3.

6 Experiments

In this section, we conduct real-data–based simulation studies to assess the finite-sample performance of the proposed test for (2). We then apply the method to real data from a leading ridesharing company. We compare our procedure against three alternatives: (i) a two-sample t-test, which is currently

used within the company; (ii) a Difference-in-Differences (DiD) estimator, which compares pre-post outcome changes between treatment and control groups [Ashenfelter, 1978, Callaway and Sant'Anna, 2021]; and (iii) a Direct Effect (DE) estimator, which estimates the treatment effect by extending the regression adjustment approach in Li and Ding [2020] to a dynamic setting. The outcome is the GMV, the state variables are the number of order requests and the drivers's total online time within each t. Detailed data-generating process and additional results are provided in Appendix B. To further assess the efficiency loss due to aggregation, we also conduct an individual-level benchmark simulation based on order-level records, with details reported in Appendix D.

Example 6.1 (Real-data-based simulations within cities). We construct a simulation environment based on real A/A experimental data from a ridesharing platform across three cities of different sizes: a large city (5–10 million residents), a medium city (1–5 million), and a small city (0.5–1 million). In an A/A experiment, identical policies are applied to both groups, so the true GATE is zero by construction. Each day is divided into $m = 24$ time intervals, and the number of experimental days is set to $n \in \{14, 28\}$. To assess power, we introduce a tunable parameter $\eta \in [0, 12]$ that scales the treatment effect, with $\eta = 0$ corresponding to the null and larger values of η yielding stronger effects. All results are based on 1,000 simulation replications, each using 500 bootstrap samples.

Figure 1 summarizes the results. The left panel shows that the bootstrap distribution closely tracks the empirical null distribution of $\widehat{\text{GATE}}$ across replications, confirming the accuracy of the bootstrap approximation established in Theorem 5.1. The right panel reports the empirical rejection rates of the proposed IRE-VCDP-based test, the two-sample t-test, the DiD estimator, and the DE estimator. All methods control the type I error at the nominal 5% level; however, the proposed test demonstrates superior power, with particularly notable gains for small effect sizes. Figure 3 in the Appendix further confirms that the empirical p-value distributions under the null are approximately uniform across all city sizes and observation windows, indicating reliable Type I error control. Additional experiments under nonlinear data-generating processes, estimated using B-spline basis functions, further demonstrate the robustness of our procedure beyond the linear specification (see Appendix C).

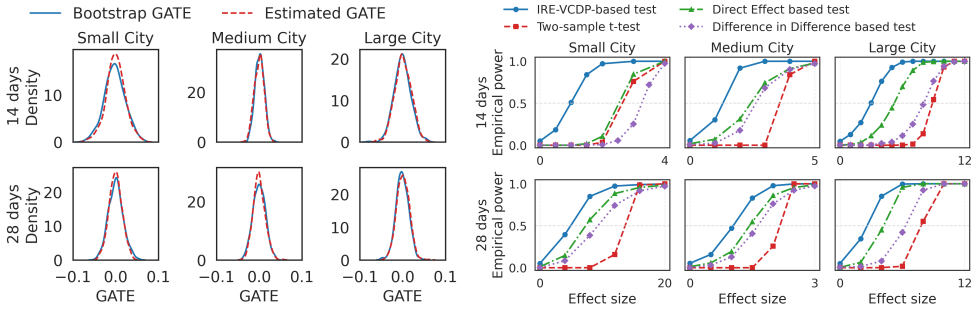

Figure 1: Left panel: empirical and bootstrap sampling distributions of the estimated GATE under null; Right panel: empirical rejection rates of the proposed IRE-VCDP-based test, the t-test, the DiD estimator, and the DE estimator.

Example 6.2 (Individual-level benchmark simulation). We further conduct an individual-level benchmark simulation to quantify the efficiency loss due to aggregation. Using order-level records, we construct simulated individual panels and compare the aggregate IRE-VCDP with two individual-level benchmarks: a micro-cohort individual-data VCDP oracle and a contextual-bandit doubly robust off-policy estimator. All methods are evaluated on the same simulated individual panels and target the same 24-hour aggregate-hourly GATE.

In the positive scenario, the micro-cohort individual-data VCDP achieves the lowest RMSE. The average true GATE is 178.999, and the micro-cohort VCDP estimates 178.972 on average, with bias -0.028 and RMSE 1.182. In comparison, the aggregate IRE-VCDP has RMSE 2.462, and the individual-level DR benchmark has RMSE 8.635. These results indicate that individual-level information can improve estimation when combined with explicit demand–supply dynamics, and the aggregate estimator incurs about 2.08 times the RMSE of the micro-cohort oracle. Details are reported in Appendix D.

Example 6.3 (Real-data-based simulations across all cities). Beyond individual cities, our framework also accommodates joint evaluation across multiple cities. In this example, we define the GATE

as the overall difference in GMV across the three cities between the treatment and control groups. The findings are qualitatively similar to those obtained for individual cities: the bootstrap approximation remains accurate under the null and the proposed test maintains a clear power advantage over all three alternative methods. Detailed results are deferred to Appendix B.

Example 6.4 (Real data analysis). We apply our method to a real-world A/B experiment conducted by Didi Chuxing from November 1 to November 21, 2025 (21 days), in which users were randomly assigned to treatment and control groups receiving different subsidy policies. The new policy was designed to increase GMV relative to the incumbent one.

Before fitting the models, we examine the raw data. Figure 5 in the Appendix displays the temporal evolution of GMV, order requests, and drivers' total online time across cities over the experimental period. GMV closely tracks order volume under normal conditions, but visibly decouples on days with abnormally high demand, when requests exceed available supply and cannot be fully converted into completed rides. This pattern motivates the inclusion of the supply–demand imbalance term (8) and confirms the suitability of models (6)–(7) for this setting.

Figure 2 presents, for the medium city, the fitted values of GMV, order requests, and drivers' total online time against their observed counterparts, together with the corresponding residuals. Figures 6 and 7 in the Appendix show analogous results for the large and small cities, respectively. Across all cities, the proposed models achieve a close fit to both the outcome and state variables, with residuals that fluctuate randomly and exhibit no systematic structure.

Table 1 in Appendix B reports the p-values from all four methods for both A/A and A/B analyses across cities. The A/A results serve as a sanity check: all methods yield p-values above 5%, confirming proper Type I error control. For the A/B analysis, the proposed test shows a meaningful drop in p-values from the A/A period, with the small city reaching 0.078 (significant at the 10% level); no significant effects are detected for the medium or large cities. In contrast, the alternative methods yield similar p-values in both A/A and A/B periods, failing to distinguish between them, suggesting limited sensitivity to treatment effects transmitted through supply–demand dynamics.

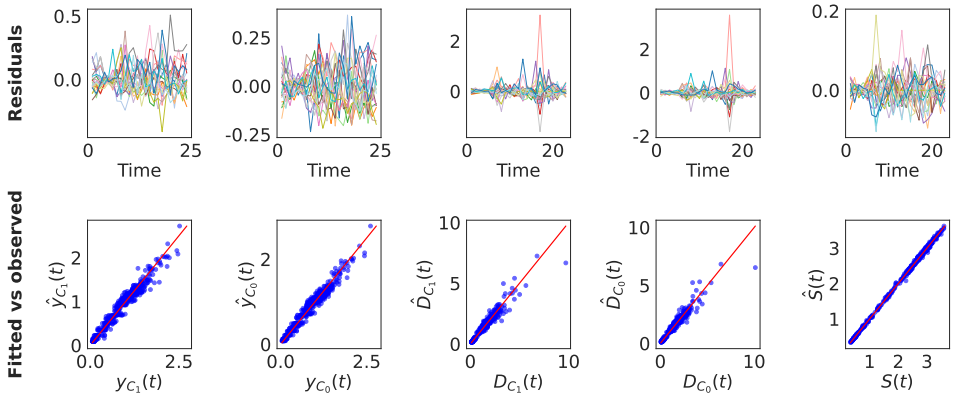

Figure 2: Medium city: residuals (top row) and fitted vs. observed values (bottom row). Columns (left to right): GMV (treatment), GMV (control), number of requests (treatment), number of requests (control), and drivers' total online time.

References

Orley Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, pages 47–57, 1978.

Guillaume W Basse and Edoardo M Airoldi. Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, 48(1):136–151, 2018.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230, 2021.

Q Chen, Y Lei, and Stefanus Jasin. Real-time spatial-intertemporal dynamic pricing for balancing supply and demand in a ride-hailing network: near-optimal policies and the value of dynamic pricing. *Operations Research*, 72(5):2097–2118, 2024.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6): 2786–2819, 2013.

Chin-Tsang Chiang, John A Rice, and Colin O Wu. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96(454):605–619, 2001.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *The annals of Statistics*, 27(5):1491–1518, 1999.

Vivek Farias, Hao Li, Tianyi Peng, Xinyuyang Ren, Huawei Zhang, and Andrew Zheng. Correcting for interference in experiments: A case study at douyin. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 455–466, 2023.

M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: a recent methodological review. *Current epidemiology reports*, 3(4):297–305, 2016.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779, 1993.

Donald R Hoover, John A Rice, Colin O Wu, and Li-Ping Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998.

Yuchen Hu and Stefan Wager. Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*, 2022.

Jianhua Z Huang, Colin O Wu, and Lan Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, pages 763–788, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 68(10):7069–7089, 2022.

Mengbing Li, Chengchun Shi, Zhenke Wu, and Piotr Fryzlewicz. Testing stationarity and change point detection in reinforcement learning. *The Annals of Statistics*, 53(3):1230–1256, 2025.

Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358, 2022.

Ting Li, Chengchun Shi, Jianing Wang, Fan Zhou, et al. Optimal treatment allocation for efficient policy evaluation in sequential decision making. *Advances in Neural Information Processing Systems*, 36:48890–48905, 2023.

Ting Li, Chengchun Shi, Zhaohua Lu, Yi Li, and Hongtu Zhu. Evaluating dynamic conditional quantile treatment effects with applications in ridesharing. *Journal of the American Statistical Association*, 119(547):1736–1750, 2024.

Xinran Li and Peng Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):241–268, 2020.

Runshi Liu and Zhipeng Hou. Unite: A unified treatment effect estimation method for one-sided and two-sided marketing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1472–1481, 2023.

Yang Liu, Yifan Zhou, Ping Li, and Feifang Hu. Cluster-adaptive network a/b testing: From randomization to estimation. *Journal of Machine Learning Research*, 25(170):1–48, 2024.

Shikai Luo, Ying Yang, Chengchun Shi, Fang Yao, Jieping Ye, and Hongtu Zhu. Policy evaluation for temporal and/or spatial dependent experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):623–649, 2024.

Lorenzo Masoero, Suhas Vijaykumar, Thomas S Richardson, James McQueen, Ido Rosen, Brian Burdick, Pat Bajari, and Guido Imbens. Multiple randomization designs: Estimation and inference with interference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf073, 2026.

Byeong U Park, Enno Mammen, Young K Lee, and Eun Ryung Lee. Varying coefficient regression models: a review and new developments. *International Statistical Review*, 83(1):36–64, 2015.

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

VA Sadovnichiy, Askar Akayevich Akaev, AI Zvyagintsev, and Askar Islamovich Sarygulov. Mathematical modeling of overcoming the covid-19 pandemic and restoring economic growth. In *Doklady Mathematics*, volume 106, pages 230–235. Springer, 2022.

Fredrik Sävje, Peter Aronow, and Michael Hudgens. Average treatment effects in the presence of unknown interference. *Annals of statistics*, 49(2):673, 2021.

Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Hongtu Zhu, and Rui Song. A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *The Annals of Applied Statistics*, 17(4):2701–2722, 2023.

S. Shirani and M. Bayati. On evolution-based models for experimentation under interference. *arXiv preprint arXiv:2511.21675*, 2025.

S. Shirani, Y. Luo, W. Overman, R. Xiong, and M. Bayati. Can we validate counterfactual estimations in the presence of general network interference? *arXiv preprint arXiv:2502.01106*, 2025.

Rober H Shumway. Time series analysis and its applications: With r examples, 2006.

Fangzhou Su and Peng Ding. Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):994–1015, 2021.

Yanqing Sun, Qiong Shou, Peter B Gilbert, Fei Heng, and Xiyuan Qian. Semiparametric additive time-varying coefficients model for longitudinal data with censored time origin. *Biometrics*, 79(2): 695–710, 2023.

Jitao Wang, Chengchun Shi, and Zhenke Wu. A robust test for the stationarity assumption in sequential decision making. In *International Conference on Machine Learning*, pages 36355–36379. PMLR, 2023.

Lan Wang, Yu Zhou, Rui Song, and Ben Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3589–3597, 2017.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.

Ruohan Zhan, Vitor Hadad, David A. Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2125–2135, 2021.

Yuzi Zhang, Howard H Chang, Angela D Iuliano, and Carrie Reed. A bayesian spatial–temporal varying coefficients model for estimating excess deaths associated with respiratory infections. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188(3):843–858, 2025.

Hongtu Zhu, Jianqing Fan, and Linglong Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098, 2014.

Appendix

A The proposed algorithm

Algorithm 1 Estimation and Inference for GATE

Require: Observed aggregate data $\{Y_{d,C}(t),\, S_{d,C}(t),\, X_{d,C}(t),\, \widetilde{X}_{d,C}(t),\, f(t)\}$ for $d = 1, \ldots, n$, $t = 1, \ldots, m$, $C \in \{C_0, C_1\}$; number of bootstrap samples B; significance level α.

1: Estimate the time-varying coefficients in the outcome model (6) and the state model (7) via two-step kernel smoothing, yielding $\widetilde{\theta}_C(t)$ and $\widetilde{\Theta}_C$ as in (10). Substitute these into (9) to obtain $\widehat{\text{GATE}}$.

2: For each $d = 1, \ldots, n$ and $C \in \{C_0, C_1\}$, compute the fitted values $\widetilde{S}_{d,C}(t+1)$ and $\widetilde{Y}_{d,C}(t)$, along with the residuals:

$$\widehat{\epsilon}_{d,C}(t) = Y_{d,C}(t) - Z_{d,C}(t)^\top \widetilde{\theta}_C(t), \quad t = 1, \ldots, m,$$
$$\widehat{E}_{d,C}(t+1) = S_{d,C}(t+1) - \widetilde{\Theta}_C\, \widetilde{Z}_{d,C}(t), \quad t = 1, \ldots, m-1.$$

3: **for** $b = 1$ **to** B **do**

4: Draw i.i.d. standard normal multipliers $\left\{\xi^b_{d,Y,C_0}, \xi^b_{d,Y,C_1}, \xi^b_{d,S,C_0}, \xi^b_{d,S,C_1}\right\}_{d=1}^{n}$.

5: Construct bootstrap pseudo-responses by perturbing the fitted values with the multiplier-weighted residuals:

$$\widehat{S}^b_{d,C}(t+1) = \widetilde{S}_{d,C}(t+1) + \xi^b_{d,S,C}\, \widehat{E}_{d,C}(t+1),$$
$$\widehat{Y}^b_{d,C}(t) = \widetilde{Y}_{d,C}(t) + \xi^b_{d,Y,C}\, \widehat{\epsilon}_{d,C}(t).$$

6: Re-estimate the model using the pseudo-data $\{\widehat{S}^b_{d,C}(t),\, \widehat{Y}^b_{d,C}(t)\}$ to obtain $\widetilde{\theta}^b_C(t)$ and $\widetilde{\Theta}^b_C$.

7: Compute the centered bootstrap statistic:

$$T^b = \widehat{\text{GATE}}^{\,b} - \widehat{\text{GATE}}.$$

8: **end for**

9: Reject H_0 at significance level α if

$$T \;=\; \widehat{\text{GATE}} \;>\; \text{the } (1-\alpha)\text{-quantile of } \left\{T^b\right\}_{b=1}^{B}.$$

Ensure: Estimate $\widehat{\text{GATE}}$ and the corresponding decision rule for testing H_0.

B Additional Experiment Results

In this section, we describe the data-generating process for the real-data–based simulation and report additional experimental results.

Example 6.1 (continued). The A/A experiment spans two weeks, from October 18 to October 31, 2025, with each day divided into $m = 24$ time intervals. For the GMV outcome model, in addition to the supply and demand state variables, the predictive covariates include the other subsidy amount and the supply–demand gap $G_d(t)$. The state-transition model uses temperature, precipitation, and a holiday indicator (equal to 1 if the time interval falls on a public holiday, and 0 otherwise) as predictors. We first fit models (6)–(7) using the full A/A dataset to obtain coefficient estimates $\{\widetilde{\alpha}_0, \widetilde{\alpha}_1, \widetilde{\alpha}_2\}$ and $\{\widetilde{\gamma}_0(t), \widetilde{\Phi}_0(t), \widetilde{\Phi}_1(t)\}$, along with estimated error processes $\widetilde{e}_d(t)$ and $\widetilde{E}_d(t)$.

To generate simulated data, we adopt a bootstrap procedure. In each run, we sample $n \in \{14, 28\}$ initial state observations and n estimated error trajectories with replacement. For each group $C \in$

$\{C_0, C_1\}$, we sample the initial demand $\widetilde{D}_{d,C}(1)$ within the group and the initial supply $\widetilde{S}_d(1)$ from platform-level supply. For each city and time interval t, predictive covariates are generated as follows. In the outcome model, the other subsidy amount and the supply–demand gap $G_d(t)$ are independently drawn from uniform distributions over their empirical ranges at time t during the A/A period; the subsidy is sampled separately for treatment and control groups, while $G_d(t)$ is shared. In the state-transition model, temperature and precipitation are sampled analogously and shared across groups, while the holiday indicator follows the calendar. We then generate n days of data under the VCDP model:

$$\begin{aligned} \widetilde{Y}_{d,C}(t) &= \alpha_{0,C}(t) + \alpha_{1,C}^{\top}(t) X_{d,C}(t) + \alpha_{2,C}^{\top}(t) \widetilde{S}_{d,C}(t) + \widetilde{\epsilon}_{d,C}(t), \\ \widetilde{S}_{d,C}(t+1) &= \gamma_{0,C}(t) + \Phi_{0,C}(t) \widetilde{X}_{d,C}(t) + \Phi_{1,C}(t) \widetilde{S}_{d,C}(t) + \widetilde{E}_{d,C}(t+1). \end{aligned}$$

We set $\{\alpha_{0,C}, \alpha_{1,C}, \alpha_{2,C}\} = \{\widetilde{\alpha}_0, \widetilde{\alpha}_1, \widetilde{\alpha}_2\}$ for both groups $C \in \{C_0, C_1\}$. The parameters $\gamma_{0,C}(t)$ and $\Phi_{1,C}(t)$ are shared across groups and set equal to $\{\widetilde{\gamma}_0(t), \widetilde{\Phi}_1(t)\}$. To control the magnitude of the GATE, we let $\Phi_{0,C_0}(t) = \widetilde{\Phi}_0(t)$ and $\Phi_{0,C_1}(t) = \eta \widetilde{\Phi}_0(t)$ for a constant $\eta \geq 0$. When $\eta = 0$, there is no treatment effect; when $\eta > 0$, the new policy is beneficial. For each city and each value of η, we generate 1000 independent datasets and apply the proposed test, the two-sample t-test, the DiD estimator, and the DE estimator at the 5% significance level.

Figure 3 shows that the empirical distributions of bootstrap p-values under the null across the three cities closely approximate the uniform distribution, confirming the validity of the proposed testing procedure.

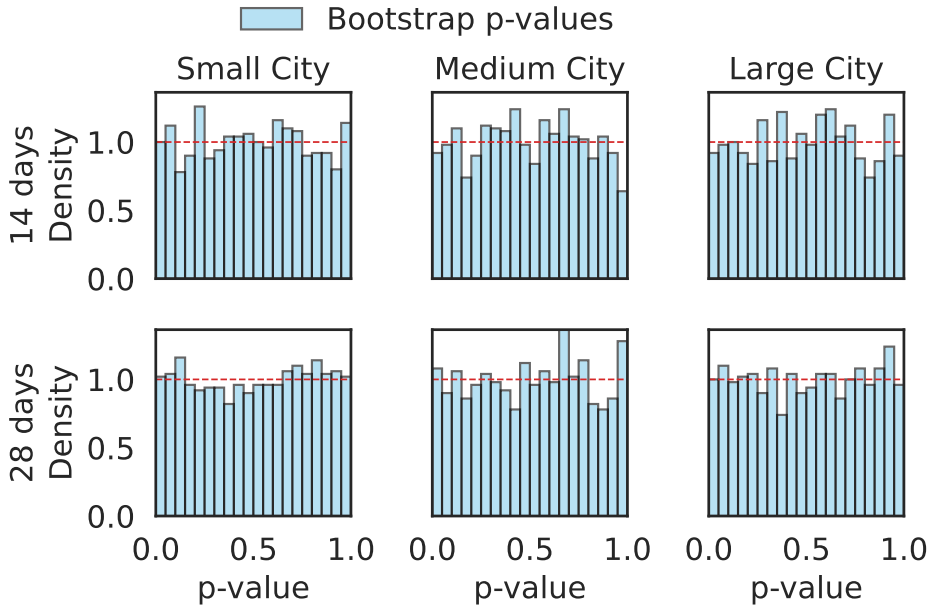

Figure 3: Empirical distributions of bootstrap p-values under the null for three cities.

Example 6.3 (Continued). This example mimics a setting where the experiment is conducted simultaneously across R cities. In each city, treatment and control groups receive different subsidy amounts under a common budget constraint. The objective is to evaluate whether the new subsidy allocation policy improves the overall GMV across the R cities. Participants within each city are randomly selected to participate in the experiment. Let $\{N_r,\ r = 1, \ldots, R\}$ denote the number of participants in the r-th city, and let $N = \sum_{r=1}^{R} N_r$. The GATE is defined as

$$\text{GATE} = N^{-1} \sum_{r=1}^{R} \Big[\mathbb{E}\Big(\sum_{t=1}^{m} \sum_{i=1}^{N_r} Y^i(t, \mathbf{1}_{N,t}) \Big) - \mathbb{E}\Big(\sum_{t=1}^{m} \sum_{i=1}^{N_r} Y^i(t, \mathbf{0}_{N,t}) \Big) \Big]. \tag{11}$$

The data-generating process is the same as in Example 6.1: data are generated independently within each city following the same procedure. However, the GATE is evaluated jointly across all cities.

Specifically, we first estimate the GATE within each city, and then aggregate these estimates to obtain the overall GATE. The bootstrap-based test is constructed accordingly.

As shown in Figure 4, the distribution of the bootstrap statistic closely approximates the empirical sampling distribution of the estimated GATE. Under the null hypothesis, the empirical distribution of bootstrap p-values remains approximately uniform on $[0, 1]$, indicating good calibration. Moreover, the proposed test demonstrates a clear power advantage over all three alternative methods, particularly for small effect sizes.

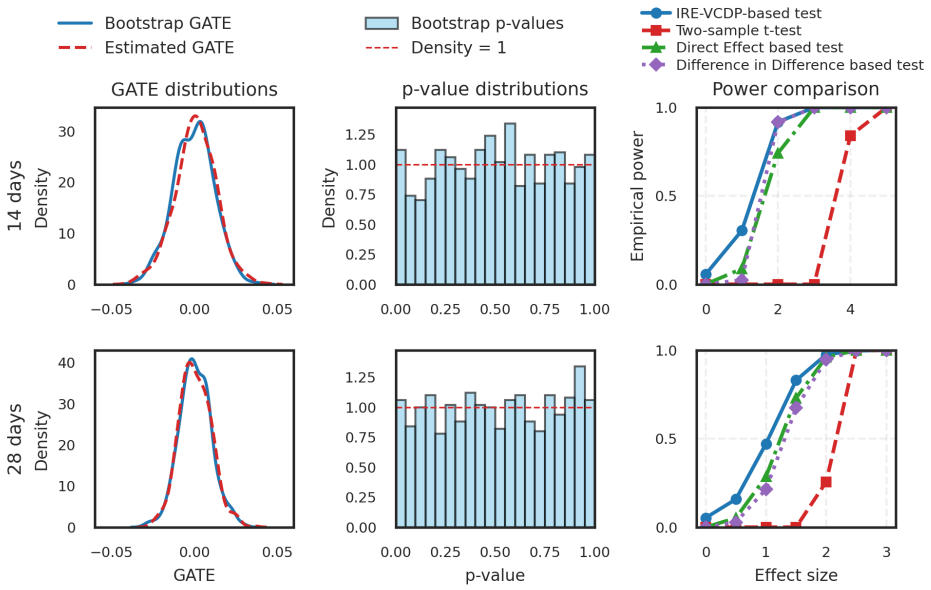

Figure 4: Joint simulation results pooling all three cities. Columns report GATE sampling distributions, bootstrap p-value distributions under the null, and power curves of the proposed test versus the t-test, the DiD estimator, and the DE estimator as a function of the true GATE; rows correspond to the two observation horizons ($n = 14$ and $n = 28$ days).

Example 6.4 (Continued). Figure 5 illustrates the temporal evolution of GMV and key state variables capturing supply–demand dynamics (number of requests and drivers' total online time) across cities over the experimental period. Figures 6–7 present the fitted values of GMV, number of requests, and drivers' total online time against their observed counterparts, together with the corresponding residuals over time for both the small and large cities. These results demonstrate that the proposed VCDP model provides a good fit to both the outcome and state variables.

Table 1 reports the p-values from all four methods for A/A and A/B analyses across cities. Detailed discussion is provided in Section 6.

Table 1: P-value results for different methods and cities under A/A and A/B periods.

Method	Small City		Medium City		Large City		All Cities	
	AA	AB	AA	AB	AA	AB	AA	AB
IRE-VCDP	0.382	0.078	0.544	0.244	0.488	0.208	0.446	0.12
Two-sample t-test	0.485	0.476	0.397	0.525	0.405	0.415	0.391	0.553
DE	0.5	0.472	0.248	0.54	0.252	0.306	0.158	0.534
DiD	0.515	0.362	0.099	0.631	0.532	0.253	0.366	0.429

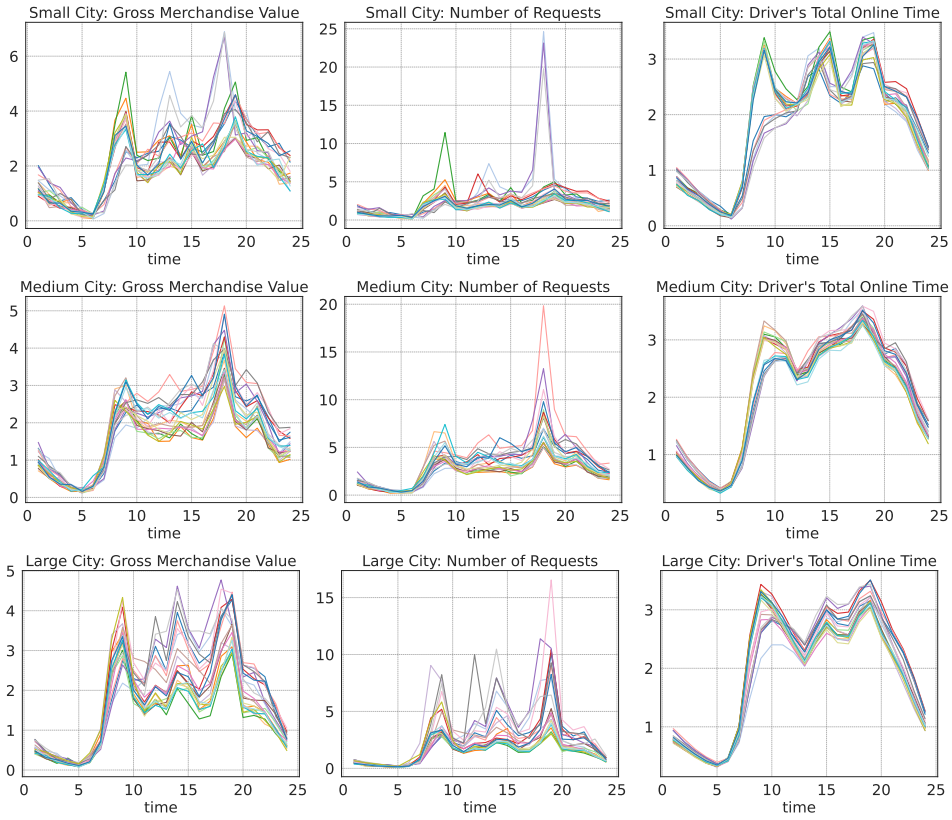

Figure 5: Scaled business metrics from the small, medium and large cities across 21 days.

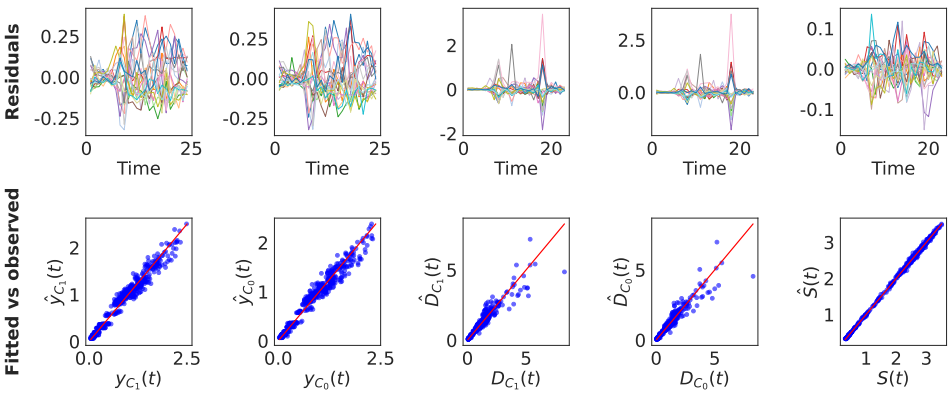

Figure 6: Large city: residuals (top row) and fitted vs. observed values (bottom row). Columns (left to right): GMV (treatment), GMV (control), number of requests (treatment), number of requests (control), and drivers' total online time.

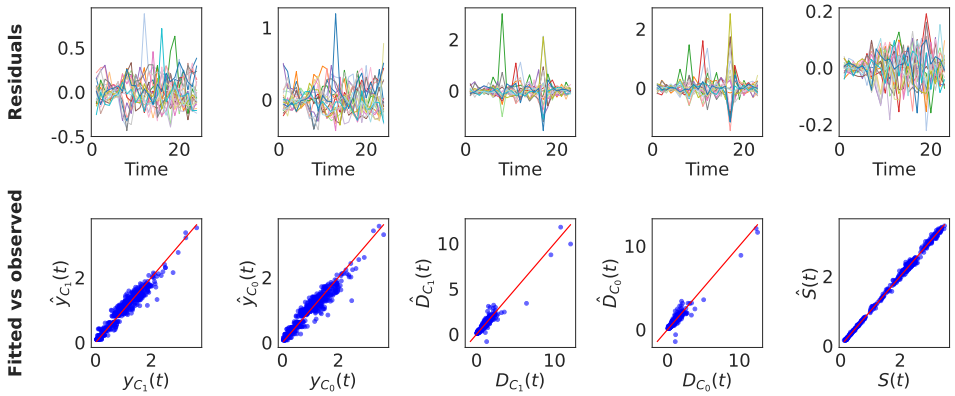

Figure 7: Small city: residuals (top row) and fitted vs. observed values (bottom row). Columns (left to right): GMV (treatment), GMV (control), number of requests (treatment), number of requests (control), and drivers' total online time.

C Additional Results under Nonlinear Data-Generating Processes

In the main text, the simulation data are generated from the linear VCDP model in equations (4)–(5). To assess the robustness of our procedure beyond the linear specification, we conduct additional experiments under nonlinear data-generating processes. Specifically, we introduce quadratic and sinusoidal terms with heterogeneous treatment effects between groups to capture nonlinear dynamics, and we estimate the model using B-spline basis functions.

Nonlinear data-generating process. The outcome model for the treatment group is specified as:

$$\begin{aligned} Y_{d,C}(t) &= \mu_C\big(t, X_{d,C}(t), S_{d,C}(t)\big) + \epsilon_{d,C}(t) \\ &= \beta_{0,C}(t) + \beta_{1,C}^\top(t) X_{d,C}(t) + \beta_{2,C}^\top(t) S_{d,C}(t) \\ &\quad + \underbrace{\lambda_{1,C} S_{d,C}(t)^2 + \lambda_{2,C} \sin(S_{d,C}(t)/\text{scale})}_{\text{quadratic + sinusoidal}} + \epsilon_{d,C}(t), \end{aligned} \tag{12}$$

and similarly for the control group with different coefficient values to capture heterogeneous effects.

The state equation for the treatment group is:

$$\begin{aligned} S_{d,C}(t+1) &= \psi_C\big(t, \widetilde{X}_{d,C}(t), S_{d,C}(t), f_d(t)\big) + E_{d,C}(t+1) \\ &= \gamma_{0,C}(t) + f_d(t)\gamma_{1,C}(t) + \Phi_{0,C}(t)\widetilde{X}_{d,C}(t) \\ &\quad + \Phi_{1,C}(t) S_{d,C}(t) + \underbrace{\Phi_{2,C}(t) S_{d,C}^2(t) + \Phi_{3,C}(t) \sin(S_{d,C}(t)/\text{scale})}_{\text{quadratic + sinusoidal}} + \eta_{d,C}(t+1). \end{aligned} \tag{13}$$

The B-spline basis expansion is used to estimate both $\mu_C(\cdot)$ and $\psi_C(\cdot)$:

$$\mu_C(t, x, s) = \sum_{k=1}^{K} \beta_k^{(g)} B_k^{(d)}(t), \tag{14}$$

$$\psi_C(t, \tilde{x}, s, f) = \sum_{k=1}^{K} \beta_k^{(G)} B_k^{(d)}(t), \tag{15}$$

where $B_k^{(d)}(t)$ are degree-d B-spline basis functions with knots placed at each time interval, and K is the total number of basis functions. This allows the conditional expectation functions to take arbitrary nonlinear forms.

Figure 8 shows that the bootstrap distribution (blue) closely aligns with the simulated ATE distribution (red) under nonlinear DGP, demonstrating that our test maintains calibration and accurate Type I

error control. Figure 9 shows that under the null hypothesis, our method produces p-values that are near-uniform, while the t-test, DiD, and DE estimators fail to control Type I error. Among valid methods that control Type I error, our B-spline estimator exhibits the fastest power growth with effect size (Figure 10), indicating superior sensitivity to treatment effects under nonlinearity. Figure **??** provides the full power comparison across all methods and settings.

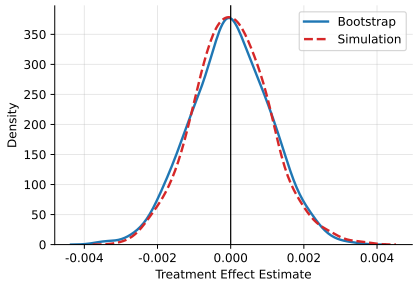

Figure 8: IRE-VCDP: Bootstrap vs. Simulation Distribution. Comparison of the sampling distribution of the treatment effect estimate obtained from 500 bootstrap resamples (blue solid line) and 500 independent simulations (red dashed line) under the null hypothesis.

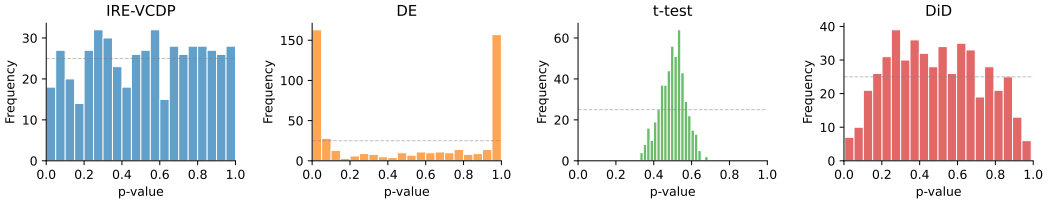

Figure 9: P-value Distribution under the Null Hypothesis. Histograms of p-values from 500 simulations under H_0 (effect size = 0) for four methods: IRE-VCDP, DE, two-sample t-test, and difference-in-differences.

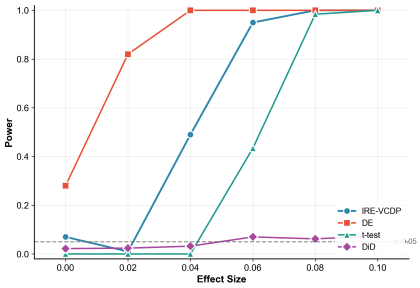

Figure 10: Power comparison under nonlinear effect setting.

D Additional Individual-Level Benchmark Simulation

This section reports an additional simulation study based on order-level records. The goal is to assess the efficiency loss from aggregating individual records into group-level summaries. All estimators are evaluated on the same simulated individual panel and target the same 24-hour aggregate-hourly GATE as in the main experiments.

D.1 Micro-Cohort Individual-Data VCDP

A direct passenger-level dynamic model is not well suited to ridesharing data because passenger demand is highly sparse: within a two-week experiment, a passenger may only generate one or a few order records. We therefore construct micro-cohorts from individual records. Each micro-cohort can be viewed as a synthetic individual formed by grouping passengers with the same treatment status, which preserves individual-level information while producing a denser demand trajectory.

Within each city and treatment arm, passengers are randomly partitioned into micro-cohorts of size K_{cohort}. In the reported experiment, $K_{\text{cohort}} = 200$. Let g denote a micro-cohort and let $\mathcal{I}_g$ be the corresponding passenger set. We define

$$R_{g,d,t} = \sum_{i \in \mathcal{I}_g} R_{i,d,t}, \qquad Y_{g,d,t} = \frac{1}{|\mathcal{I}^{\text{act}}_{g,d,t}|} \sum_{i \in \mathcal{I}_g} Y_{i,d,t}, \qquad Q_{g,d,t} = \frac{1}{|\mathcal{I}^{\text{act}}_{g,d,t}|} \sum_{i \in \mathcal{I}_g} Q_{i,d,t},$$

where $R_{i,d,t}$, $Y_{i,d,t}$, and $Q_{i,d,t}$ are individual demand, GMV, and subsidy rate, respectively.

The micro-cohort VCDP keeps the same supply–demand structure as the aggregate IRE-VCDP. For each hour t, we fit

$$R_{g,d,t} = \alpha^R_t + \tau^R_t A_g + \beta^R_{t,Q} Q_{g,d,t} + \beta^R_{t,D} D_{d,t-1} + \beta^R_{t,M} M_{d,t-1} + \beta^R_{t,G} G_{d,t-1} + \beta^{R\top}_{t,W} W_{d,t} + \varepsilon^R_{g,d,t},$$

$$M_{d,t+1} = \alpha^M_t + \gamma_{t,D} D_{d,t} + \gamma_{t,M} M_{d,t} + \gamma_{t,G} G_{d,t} + \gamma^{M\top}_{t,W} W_{d,t+1} + \eta^M_{d,t},$$

and

$$Y_{g,d,t} = \alpha^Y_t + \tau^Y_t A_g + \theta^Y_{t,R} R_{g,d,t} + \theta^Y_{t,Q} Q_{g,d,t} + \theta^Y_{t,D} D_{d,t} + \theta^Y_{t,M} M_{d,t} + \theta^Y_{t,G} G_{d,t} + \theta^{Y\top}_{t,W} W_{d,t} + \varepsilon^Y_{g,d,t}.$$

Here A_g is the cohort treatment indicator, $D_{d,t}$ is market demand, $M_{d,t}$ is market supply, $G_{d,t}$ is the demand–supply gap, and $W_{d,t}$ contains exogenous covariates. The supply equation remains market-level because drivers serve the whole market.

To evaluate full treatment and full control, we set all cohorts to $A_g = 1$ or $A_g = 0$ and recursively predict cohort demand, market demand, supply, and cohort outcomes over the 24-hour horizon. The estimator is reported on the same aggregate-hourly GATE scale:

$$\widehat{\tau}_{\text{mcVCDP}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sum_{t=0}^{23} \left[\overline{\widehat{Y}}^{(1)}_{d,t} - \overline{\widehat{Y}}^{(0)}_{d,t} \right].$$

Since the input records are sparse active order records, we use the randomized observed contrast as an anchor and add a clipped recursive model adjustment to avoid excessive extrapolation.

D.2 Contextual-Bandit Doubly Robust Benchmark

We also compare with a contextual-bandit doubly robust off-policy evaluation benchmark [Dudik et al., 2011, Wang et al., 2017, Zhan et al., 2021]. This benchmark treats each logged interaction as a one-step decision problem and does not explicitly model the dynamic feedback from individual demand to market demand and supply.

Each passenger-hour observation is written as

$$O_{i,d,t} = (X_{i,d,t}, A_i, Y_{i,d,t}),$$

where $X_{i,d,t}$ contains market covariates, subsidy rate, lagged demand and supply variables, time indicators, and weather/event variables. Since treatment is randomized at the passenger level, the propensity score is known:

$$e(X_{i,d,t}) = \mathbb{P}(A_i = 1 \mid X_{i,d,t}) = 0.5.$$

For each hour t, we fit outcome regressions $\widehat{\mu}_{1,t}(X)$ and $\widehat{\mu}_{0,t}(X)$ for the treatment and control arms. The hour-specific doubly robust contrast is

$$\widehat{\tau}_{\text{DR},t} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{1}{N_{d,t}} \sum_{i \in \mathcal{I}_{d,t}} \left[\psi^{(1)}_{i,d,t} - \psi^{(0)}_{i,d,t} \right],$$

where

$$\psi_{i,d,t}^{(a)} = \widehat{\mu}_{a,t}(X_{i,d,t}) + \frac{\mathbf{1}\{A_i = a\}}{0.5}\left[Y_{i,d,t} - \widehat{\mu}_{a,t}(X_{i,d,t})\right], \qquad a \in \{0,1\}.$$

The final estimate is

$$\widehat{\tau}_{\mathrm{DR}} = \sum_{t=0}^{23} \widehat{\tau}_{\mathrm{DR},t}.$$

We use a conservative feature set that excludes current demand $R_{i,d,t}$, since it is a post-treatment mediator, but includes lagged demand and lagged market states.

D.3 Computational Considerations

The aggregate IRE-VCDP operates on city-day-hour group summaries. With m time intervals, n experimental days, and p covariates, its main cost comes from hour-specific least-squares estimation and kernel smoothing, with complexity on the order of $O(mnp^2)$. This is substantially cheaper than methods that operate on individual records.

The micro-cohort individual-data VCDP uses a finer unit of analysis. If G denotes the number of micro-cohorts and p_{mc} denotes the number of cohort-level covariates, its hour-specific demand and outcome regressions scale approximately as $O(mnGp_{\mathrm{mc}}^2)$, in addition to the market-level supply transition and the recursive full-treatment/full-control rollout. Thus, the micro-cohort estimator is computationally more expensive, but it can exploit individual-level demand and outcome information that is unavailable to the aggregate estimator.

The contextual-bandit DR benchmark also uses individual-level observations. For a sampled panel with roughly N_s active passenger-hour observations per hour and p_{DR} covariates, its hourly outcome regressions scale approximately as $O(mN_s p_{\mathrm{DR}}^2)$. Unlike VCDP-based methods, however, it does not recursively model the demand–supply feedback. These complexity differences reflect the main trade-off in this benchmark: individual-level data can improve statistical accuracy when combined with the market feedback structure, but this comes at a higher computational cost.

D.4 Simulation Design

The individual-level benchmark follows the real-data-based simulation design in Example 6.1. We use the same large city, the same 14-day horizon, the same hourly time intervals, and the same market-level covariates. The main difference is that the baseline data-generating mechanism is estimated from order-level records instead of city-hour aggregate observations.

We first aggregate the order-level records to the passenger-day-hour level. Let $R_{i,d,t}^{\mathrm{obs}}$, $Y_{i,d,t}^{\mathrm{obs}}$, and $Q_{i,d,t}^{\mathrm{obs}}$ denote the observed individual demand, GMV, and subsidy rate, respectively. These individual records are merged with the market covariates used in the aggregate simulation. For each hour t, we fit baseline individual demand and outcome models,

$$R_{i,d,t}^{\mathrm{obs}} = m_t^R(Q_{i,d,t}^{\mathrm{obs}}, S_{d,t}) + \varepsilon_{i,d,t}^R, \qquad Y_{i,d,t}^{\mathrm{obs}} = m_t^Y(R_{i,d,t}^{\mathrm{obs}}, Q_{i,d,t}^{\mathrm{obs}}, S_{d,t}) + \varepsilon_{i,d,t}^Y,$$

where $S_{d,t}$ denotes the market state and covariates. The fitted residuals are retained for bootstrap simulation.

In each Monte Carlo replication, we generate treatment and control potential outcomes at the individual level. Under treatment arm $a \in \{0,1\}$, the subsidy rate is shifted as

$$Q_{i,d,t}^{(a)} = Q_{i,d,t}^{\mathrm{base}} + ca.$$

Individual demand is generated by

$$R_{i,d,t}^{(a)} = \max\left\{\widehat{m}_t^R(Q_{i,d,t}^{(a)}, S_{d,t}^{(a)}) + \varepsilon_{i,d,t}^{R,(a)} + \mathbf{1}\{a=1\}\Delta_{i,d,t}^R, 0\right\},$$

where $\varepsilon_{i,d,t}^{R,(a)}$ is sampled from the fitted demand residuals and $\Delta_{i,d,t}^R$ controls the treatment-induced demand enhancement in the positive scenario. The individual demands are aggregated to market demand,

$$D_{d,t}^{(a)} = \sum_i R_{i,d,t}^{(a)},$$

which is then used to update the market supply path through the same supply transition as in the aggregate simulation. Given the simulated individual demand and market state, individual outcomes are generated by

$$Y_{i,d,t}^{(a)} = \max\left\{\widehat{m}_t^Y(R_{i,d,t}^{(a)}, Q_{i,d,t}^{(a)}, S_{d,t}^{(a)}) + \varepsilon_{i,d,t}^{Y,(a)}, 0\right\},$$

where $\varepsilon_{i,d,t}^{Y,(a)}$ is sampled from the fitted outcome residuals.

Treatment is assigned at the passenger level, $A_i \sim \text{Bernoulli}(0.5)$, and the observed simulated panel is obtained by selecting the corresponding potential outcomes. All estimators are evaluated on the same sampled simulated individual panel. To control computation, we use stratified sampling within city-day-hour-treatment cells, preserving the distributions of demand, GMV, and subsidy rate. The reported experiment uses 100 Monte Carlo replications and at most 1000 sampled active users per hour.

D.5 Results

Table 2 reports the null-scenario results. The true GATE is zero, and all methods produce estimates close to zero. The micro-cohort VCDP is not the most stable estimator in this null setting, which is expected: when there is no treatment-induced demand or supply response, simpler estimators that rely less on recursive state modeling can have slightly smaller finite-sample variance. This null scenario mainly serves as a sanity check that the simulation does not mechanically generate a positive effect.

Table 2: Individual-level benchmark simulation under the null scenario.

Method	Bias	SD	RMSE
Micro-cohort individual VCDP	0.031	1.195	1.189
DR-OPE individual	0.105	0.912	0.914
Aggregate IRE-VCDP	0.096	0.971	0.971
DR-OPE aggregate	0.094	0.928	0.928

Table 3 reports the positive-scenario results, where the treatment effect is generated through demand and market-state responses. In this setting, explicitly modeling the demand–supply feedback is important. The micro-cohort individual VCDP achieves the smallest RMSE, reducing the RMSE from 2.462 for the aggregate IRE-VCDP to 1.182. Thus, the aggregate estimator incurs about 2.08 times the RMSE of the micro-cohort individual-data oracle. The contextual-bandit DR benchmarks are less accurate because they use logged individual or aggregate observations but do not model the dynamic feedback from individual demand to market demand, supply, and outcomes.

Table 3: Individual-level benchmark simulation under the positive scenario.

Method	Mean estimate	Bias	SD	RMSE
Micro-cohort individual VCDP	178.972	-0.028	1.304	1.182
Aggregate IRE-VCDP	176.714	-2.285	1.118	2.462
DR-OPE aggregate	173.125	-5.874	1.956	6.133
DR-OPE individual	170.412	-8.588	1.090	8.635

Table 4 summarizes the average runtime per Monte Carlo replication. The aggregate IRE-VCDP is the fastest because it models the demand–supply feedback on city-date-hour aggregate arrays, so the recursive rollout is performed only at the group-state level. The micro-cohort individual VCDP applies the same feedback structure at a finer cohort level; it must construct micro-cohorts and recursively predict cohort demand, market demand, supply, and cohort outcomes under both full treatment and full control, which increases computation. The individual DR-OPE benchmark is faster than the micro-cohort VCDP because it treats each passenger-hour as a one-step contextual-bandit observation and does not propagate treatment effects through future market states. Thus, the results reflect a trade-off among data granularity, dynamic feedback modeling, statistical accuracy, and computational cost.

Overall, the results highlight two points. First, explicitly modeling market feedback is valuable: both VCDP-based methods outperform the contextual-bandit DR benchmarks in the positive scenario.

Table 4: Average runtime per replication in the individual-level benchmark simulation.

Component	Null scenario	Positive scenario
Micro-cohort individual VCDP	47.56s	44.95s
DR-OPE individual	1.30s	1.19s
DR-OPE aggregate	0.66s	0.63s
Aggregate IRE-VCDP	0.26s	0.25s

Second, if individual-level records are available, the micro-cohort VCDP can further improve estimation accuracy by using finer demand and outcome information, but this improvement comes with substantially higher computational cost. This comparison quantifies the efficiency loss due to aggregation while also explaining why the aggregate IRE-VCDP is practically attractive when only group-level data are available.

E GATE Decomposition

The GATE in Proposition 4.1 admits the following decomposition:

1. **Direct effect on outcome.**
$$\sum_{t=1}^{m}\big[\alpha_{0,C_1}(t)-\alpha_{0,C_0}(t)\big],$$
the immediate effect of the treatment on the outcome intercept.
2. **Covariate effect.**
$$\sum_{t=1}^{m}\bar{X}_N\big[\alpha_{1,C_1}(t)-\alpha_{1,C_0}(t)\big],$$
the interaction between treatment and predictive covariates.
3. **Initial state propagation.**
$$\sum_{t=1}^{m}\Big[\alpha_{2,C_1}(t)^\top\prod_{l=1}^{t-1}\Phi_{1,C_1}(l)\,\mathbb{E}(S_{C_1}(1))-\alpha_{2,C_0}(t)^\top\prod_{l=1}^{t-1}\Phi_{1,C_0}(l)\,\mathbb{E}(S_{C_0}(1))\Big],$$
capturing how differences in initial states propagate over time through the state dynamics.
4. **Cumulative state-level constant and treatment effects.**
$$\sum_{t=1}^{m}\Bigg[\alpha_{2,C_1}(t)^\top\sum_{k=1}^{t-1}\prod_{l=k+1}^{t-1}\Phi_{1,C_1}(l)\big(\gamma_{0,C_1}(k)+\gamma_{1,C_1}(k)\big)-\alpha_{2,C_0}(t)^\top\sum_{k=1}^{t-1}\prod_{l=k+1}^{t-1}\Phi_{1,C_0}(l)\,\gamma_{0,C_0}(k)\Bigg],$$
where $\gamma_{1,C_1}(k)$ represents the direct effect of treatment on the state equation, and $\gamma_{0,C}(k)$ captures the baseline state evolution.
5. **Covariate effect via state propagation.**
$$\sum_{t=1}^{m}\Bigg[\alpha_{2,C_1}(t)^\top\sum_{k=1}^{t-1}\prod_{l=k+1}^{t-1}\Phi_{1,C_1}(l)\,\Phi_{0,C_1}(k)\,\bar{\tilde{X}}_N(k)-\alpha_{2,C_0}(t)^\top\sum_{k=1}^{t-1}\prod_{l=k+1}^{t-1}\Phi_{1,C_0}(l)\,\Phi_{0,C_0}(k)\,\bar{\tilde{X}}_N(k)\Bigg],$$
capturing how covariates influence the outcome indirectly through the state variables.

F Notation and Definitions

F.1 Formulation and Discussion of Straightforward Treatment Effect

$$\begin{aligned}\tau=&\sum_{t=1}^{m}(\alpha_{0,C_1}(t)-\alpha_{0,C_0}(t))+\sum_{t=1}^{m}\Big[\mathbb{E}(X_{C_1}(t))\alpha_{1,C_1}(t)-\mathbb{E}(X_{C_0}(t))\alpha_{1,C_0}(t)\Big]\\&+\sum_{t=1}^{m}\Big[\mathbb{E}(S_{C_1}(t))\alpha_{2,C_1}(t)-\mathbb{E}(S_{C_0}(t))\alpha_{2,C_0}(t)\Big].\end{aligned}$$

If the means of the predictive covariates are equal across groups, such that $\mathbb{E}(N^{-1}\sum_{i=1}^{N} X^i(t)) = \mathbb{E}(X_{C_1}(t)) = \mathbb{E}(X_{C_0}(t))$, and $\mathbb{E}(S(t, \mathbf{1}_{N,t})) = \mathbb{E}(S(t, \mathbf{0}_{N,t}))$, then τ coincides with the GATE. In practice, ensuring that the mean of the predictive covariates is balanced between the two groups is relatively straightforward using covariate balancing techniques such as matching or weighting. However, achieving equality in the expected potential outcomes of the state variables is more challenging, as the treatment history often directly influences their evolution over time.

F.2 Formulation of GATE Estimator

$$
\begin{aligned}
\widehat{\text{GATE}} = &\sum_{t=1}^{m}(\widetilde{\alpha}_{0,C_1}(t) - \widetilde{\alpha}_{0,C_0}(t)) + \sum_{t=1}^{m}\Big[n^{-1}\sum_{d=1}^{n} X_d(t)(\widetilde{\alpha}_{1,C_1}(t) - \widetilde{\alpha}_{1,C_0}(t))\Big] \\
&+ \sum_{t=1}^{m}\Big[\widetilde{\alpha}_{2,C_1}(t)^\top\Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}\widetilde{\Phi}_{1,C_1}(l)\Big[\widetilde{\gamma}_{0,C_1}(k) + \widetilde{\gamma}_{1,C_1}(k) + n^{-1}\widetilde{\Phi}_{0,C_1}(t)\sum_{d=1}^{n} H_{d,C_1}(t)\Big]\} + \prod_{l=1}^{t-1}\widetilde{\Phi}_{1,C_1}(l)\mathbb{E}(S(1))\Big) \\
&- \widetilde{\alpha}_{2,C_0}(t)^\top\Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}\widetilde{\Phi}_{1,C_0}(l)\Big[\widetilde{\gamma}_{0,C_0}(k) + n^{-1}\widetilde{\Phi}_{0,C_0}(t)\sum_{d=1}^{n} H_{d,C_0}(t)\Big]\} + \prod_{l=1}^{t-1}\widetilde{\Phi}_{1,C_0}(l)\mathbb{E}(S(1))\Big)\Big],
\end{aligned}
$$

where $X_d(t) := N^{-1}(N_0 X_{d,C_0}(t) + N_1 X_{d,C_1}(t))$.

G Assumptions for Theorem 5.1

In this section, we provide the regularity conditions for the bootstrap consistency theory.

Assumption G.1. The kernel function $K(\cdot)$ is a symmetric probability density function defined over the interval $[-1, 1]$. It is Lipschitz continuous and satisfies the condition $\int_{-1}^{1} |tK'(t)|dt < \infty$, signifying its finiteness under the integral of its absolute derivative.

Assumption G.2. The covariates used in outcome model and state model are independent and identically distributed across d and are sub-Gaussian (uniformly over t). Furthermore, for any integer t within the range $1 \leq t \leq m$, the smallest eigenvalue of each population second-moment matrix of the covariates is bounded away from zero.

Assumption G.3. All components of $\theta_{C_0}(ms)$, $\theta_{C_1}(ms)$, $\Theta_{C_0}(ms)$ and $\Theta_{C_1}(ms)$ possess second-order derivatives with respect to s.

Assumption G.4. There exist constants $0 < c_1 < 1$, $M_\Gamma, M_\beta > 0$, and $M_{min} > 0$ such that $\|\Phi_{1,C}(t)\|_\infty \leq c_1$, $\|\alpha_{2,C}(t)\|_\infty \leq M_\alpha$, and $\|\Phi_{0,C}(t)\|_\infty \leq M_\Phi$, for $\forall C \in \{C_0.C_1\}$.

Assumption G.1 is mild as the kernel $K(\cdot)$ is user-specified. Assumption G.2 has been commonly used in the literature on varying coefficient models (see e.g., Zhu et al. [2014]). Assumption G.3 serves as a routine smoothness condition that guarantees the local approximation accuracy of the kernel estimator. Assumption G.4 ensures that the time series is stationary, since $\Phi(t)$ is the autoregressive coefficient. It is commonly imposed in the literature on time series analysis [Shumway, 2006].

H Proofs of Propositions and Theorems

H.1 Proof of Proposition 3.5

Proof. The proof proceeds in two steps. First, we show that the function $R(\cdot)$ can be identified from the observed data using the conditional expectation of the observed outcomes. Second, we apply the law of iterated expectations to show that the GATE is identifiable.

Step 1: Identification of $R(\cdot)$

By the law of iterated expectations, we have the following relationship:

$$
\begin{aligned}
&\mathbb{E}\Big(Y^i(t)\,\Big|\,A_t^i = a, \{S(j), f(j)\}_{j\leq t}\Big) \qquad (16)\\
= \;&\mathbb{E}\Big[\mathbb{E}\Big[Y^i(t)\,\Big|\,A_t^i = a, \{S(j), f(j)\}_{j\leq t}, \{Y(j)\}_{j<t}\Big]\Big|A_t^i = a, \{S(j), f(j)\}_{j\leq t}\Big].
\end{aligned}
$$

By taking out the inner expectation in Equation (16), we can further derive:

$$
\begin{aligned}
&\mathbb{E}\Big[Y^i(t) \,\Big|\, A_t^i = a, \{S(j), f(j)\}_{j\le t}, \{Y(j)\}_{j<t}\Big] \\
\overset{(i)}{=}&\mathbb{E}\Big[Y^i(t) \,\Big|\, A_t^i = a, \bar{\boldsymbol{A}}_{N,t-1} = \bar{\boldsymbol{a}}_{N,t-1}, \{S(j), f(j)\}_{j\le t}, \{Y(j)\}_{j<t}\Big] \\
\overset{(ii)}{=}&\mathbb{E}\Big[Y^i(t, \bar{\boldsymbol{A}}_{N,t}) \,\Big|\, A_t^i = a, \bar{\boldsymbol{A}}_{N,t-1} = \bar{\boldsymbol{a}}_{N,t-1}, \{S(j, \bar{\boldsymbol{A}}_{N,j-1}), f(j)\}_{j\le t}, \{Y^i(j, \bar{\boldsymbol{A}}_{N,j})\}_{j<t}\Big] \\
\overset{(iii)}{=}&\mathbb{E}\Big[Y^i(t, \bar{\boldsymbol{a}}_{N,t}) \,\Big|\, A_t^i = a, \bar{\boldsymbol{A}}_{N,t-1} = \bar{\boldsymbol{a}}_{N,t-1}, \{S(j, \bar{\boldsymbol{A}}_{N,j-1}), f(j)\}_{j\le t}, \{Y^i(j, \bar{\boldsymbol{A}}_{N,j})\}_{j<t}\Big] \\
\overset{(iv)}{=}&\mathbb{E}\Big[Y^i(t, \bar{\boldsymbol{a}}_{N,t}) \,\Big|\, A_t^i = a, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}, \{Y^i(j, \bar{\boldsymbol{a}}_{N,j})\}_{j<t}\Big] \\
\overset{(v)}{=}&R\Big(a_t^i,\, \{S(t), f(j)\}_{j\le t}\Big).
\end{aligned} \tag{17}
$$

Here, equality (i) follows from Assumption 3.1, Assumption 3.3 and 3.4; equalities (ii)-(iv) follow from the consistency Assumption 3.2.

According to the definition of $R(\cdot)$ and the independence of $\{Y^i(j, \bar{\boldsymbol{a}}_{N,j})\}_{j<t}$, substituting the result of Equation (17) into Equation (16), and using the result of Assumption 3.3, we obtain immediately that:

$$
R\Big(a_t^i,\, \{S(t), f(j)\}_{j\le t}\Big) = \mathbb{E}\Big(Y^i(t) \,\Big|\, A_t^i = a, \{S(j), f(j)\}_{j\le t}\Big).
$$

Consequently, $R(\cdot)$ can be estimated by the sample average over units receiving treatment a:

$$
R\Big(a_t^i,\, \{S(j), f(j)\}_{j\le t}\Big) = N_a^{-1}\mathbb{E}\Big(\sum_{\{i:a_t^i=a\}} Y^i(t) \,\Big|\, A_t^i = a_t^i, \{S(j), f(j)\}_{j<t}\Big),
$$

which completes the derivation of Equation (4).

Step 2: Identification of GATE

We next show Equation (5). By the law of iterated expectations, we have

$$
\begin{aligned}
&\mathbb{E}\Big[R\big(a, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}\big)\Big] \\
=\;&\mathbb{E}\Big[\mathbb{E}\Big\{R\big(a, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}\big)\Big|\bar{A}_{N,1} = \bar{\boldsymbol{a}}_{N,1}, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}, \{Y(j, \bar{\boldsymbol{a}}_{N,t})\}_{j<t}\Big\}\Big].
\end{aligned}
$$

Under Assumption 3.2, we can replace $Y(1, \bar{\boldsymbol{a}}_{N,1})$ and $S(2, \bar{\boldsymbol{a}}_{N,1})$ with $Y(1)$ and $S(2)$, respectively. Under Assumption 3.3 and 3.4, the event $(A_2^1, \ldots, A_2^N) = (a_2^1, \ldots, a_2^N)$ can be included in the conditioning set. This yields that

$$
\begin{aligned}
\mathbb{E}\Big[\big(a, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}\big)\Big] = \mathbb{E}\Big[\mathbb{E}\Big\{&R\big(a, \{S(t, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t},\big)\Big| \\
&\bar{A}_{N,2} = \bar{\boldsymbol{a}}_{N,2}, \{S(j, \bar{\boldsymbol{a}}_{N,j-1}), f(j)\}_{j\le t}, \{Y(j, \bar{\boldsymbol{a}}_{N,t})\}_{j<t}, S(1), S(2), Y(1)\Big\}\Big].
\end{aligned}
$$

Iteratively applying this argument allows us to repeatedly replace the counterfactual variables with the observed ones. At the end, all the potential outcomes/states will be replaced with the observed versions conditional on the actions. The proof is hence completed. □

H.2 Proof of Proposition 4.1

Proof. Let $\zeta(t) := N^{-1}\mathbb{E}(\sum_{i=1}^N Y^i(t, \mathbf{1}_{N,t})) - N^{-1}\mathbb{E}(\sum_{i=1}^N Y^i(t, \mathbf{0}_{N,t}))$, which means that GATE $= \sum_{t=1}^m \zeta(t)$. Thus, we only need to prove that

$$
\begin{aligned}
\zeta(t) = &(\alpha_{0,C_1}(t) - \alpha_{0,C_0}(t)) + \Big[(\alpha_{1,C_1}(t) - \alpha_{1,C_0}(t))^\top \mathbb{E}(N^{-1}\sum_{i=1}^N X^i(t))\Big] \\
&+ \Big[\alpha_{2,C_1}(t)^\top\Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}\Phi_{1,C_1}(l)[\gamma_{0,C_1}(k) + \gamma_{1,C_1}(k) + \Phi_{0,C_1}(k)\mathbb{E}(N^{-1}\sum_{i=1}^N H_{C_1}^i(k))]\} + \prod_{l=1}^{t-1}\Phi_{1,C_1}(l)\mathbb{E}(S(1))\Big) \\
&- \alpha_{2,C_0}(t)^\top\Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}\Phi_{1,C_0}(l)[\gamma_{0,C_0}(k) + \Phi_{0,C_0}(k)\mathbb{E}(N^{-1}\sum_{i=1}^N H_{C_0}^i(k))]\} + \prod_{l=1}^{t-1}\Phi_{1,C_0}(l)\mathbb{E}(S(1))\Big)\Big].
\end{aligned}
$$

Since potential outcomes admits the form of models in (6)-(7), we have

$$
\begin{aligned}
&\zeta(t) = (\alpha_{0,C_1}(t) - \alpha_{0,C_0}(t)) + (\alpha_{1,C_1}(t) - \alpha_{1,C_0}(t))^\top \mathbb{E}(N^{-1}\sum_{i=1}^{N} X^i(t)) \\
&+ \alpha_{2,C_1}(t)^\top \mathbb{E}(S_{C_1}(t, \mathbf{1}_{N,t-1})) - \alpha_{2,C_0}(t)^\top \mathbb{E}(S_{C_0}(t, \mathbf{0}_{N,t-1})) \\
=&(\alpha_{0,C_1}(t) - \alpha_{0,C_0}(t)) + (\alpha_{1,C_1}(t) - \alpha_{1,C_0}(t))^\top \mathbb{E}(N^{-1}\sum_{i=1}^{N} X^i(t)) \\
&+ \alpha_{2,C_1}(t)^\top [\gamma_{0,C_1}(t-1) + \gamma_{1,C_1}(t-1) + \Phi_{0,C_1}(t-1)\mathbb{E}(S_{C_1}(t-1, \mathbf{1}_{N,t-2})) + \Phi_{1,C_1}(t-1)\mathbb{E}(N^{-1}\sum_{i=1}^{N} H^i(t-1))] \\
&- \alpha_{2,C_0}(t)^\top [\gamma_{0,C_0}(t-1) + \Phi_{0,C_0}(t-1)\mathbb{E}(S_{C_0}(t-1, \mathbf{1}_{N,t-2})) + \Phi_{1,C_0}(t-1)\mathbb{E}(N^{-1}\sum_{i=1}^{N} H^i(t-1))] \\
=&\cdots \\
=&(\alpha_{0,C_1}(t) - \alpha_{0,C_0}(t)) + \Big[(\alpha_{1,C_1}(t) - \alpha_{1,C_0}(t))^\top \mathbb{E}(N^{-1}\sum_{i=1}^{N} X^i(t))\Big] \\
&+ \Big[\alpha_{2,C_1}(t)^\top \Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1} \Phi_{1,C_1}(l)[\gamma_{0,C_1}(k) + \gamma_{1,C_1}(k) + \Phi_{0,C_1}(k)\mathbb{E}(N^{-1}\sum_{i=1}^{N} H^i_{C_1}(k))]\} + \prod_{l=1}^{t-1}\Phi_{1,C_1}(l)\mathbb{E}(S(1))\Big) \\
&- \alpha_{2,C_0}(t)^\top \Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1} \Phi_{1,C_0}(l)[\gamma_{0,C_0}(k) + \Phi_{0,C_0}(k)\mathbb{E}(N^{-1}\sum_{i=1}^{N} H^i_{C_0}(k))]\} + \prod_{l=1}^{t-1}\Phi_{1,C_0}(l)\mathbb{E}(S(1))\Big)\Big].
\end{aligned}
$$

The proof is hence completed. □

H.3 Proof of Theorem 5.1

Proof. We focus on providing an upper error bound for

$$\rho^*(z) = \left|\mathbb{P}\left(\frac{1}{m}T - \frac{1}{m}\text{GATE} \le z\right) - \mathbb{P}\left(\frac{1}{m}T^b - \frac{1}{m}T \le z \middle| Data\right)\right|.$$

We begin with some notations. Note that, in outcome model, $\widetilde{\theta}_C(t), C \in \{C_0, C_1\}$ can be expressed as

$$\widetilde{\theta}_C(t) = \theta_{s,C}(t) + \frac{1}{n}\sum_{d=1}^{n}\left(\sum_{j=1}^{m} B_{d,j,C}(t) e_{d,C}(j)\right),$$

where

$$B_{d,j,C}(t) = \omega_{j,h}(t)\left(\frac{1}{n}\sum_{d'=1}^{n} Z_{d',C}(j)^\top Z_{d',C}(j)\right)^{-1} Z_{d,C}(j),$$

are independent of the random part $e_{d,C}(j)$, and $\theta_{s,C}(t) = \sum_{j=1}^{m}\omega_{j,h}(t)\theta_C(j)$. Let $e^\theta_{d,C}(t) = \sum_{j=1}^{m} B_{d,j,C}(t)e_{d,C}(j) = \left\{e^{\alpha_{0,C}}_{d,C}(t), \left(e^{\alpha_{1,C}}_{d,C}(t)\right)^\top, \left(e^{\alpha_{2,C}}_{d,C}(t)\right)^\top\right\}^\top$ and $e^\theta_C(t) = n^{-1/2}\sum_{d=1}^{n} e^\theta_{d,C}(t)$. Similarly, we can represent $\widetilde{\Theta}(t)$ as

$$\widetilde{\Theta}(t) = \widetilde{\Theta}_{s,C}(t) + \frac{1}{n}\sum_{d=1}^{n}\left(\sum_{j=1}^{m-1} \widetilde{B}_{d,j,C}(t) E_{d,C}(t)\right),$$

where

$$\widetilde{B}_{d,j,C}(t) = \omega_{j,h}(t)\left(\frac{1}{n}\sum_{d'=1}^{D} \widetilde{Z}_{d',C}(j)^\top \widetilde{Z}_{d',C}(j)\right)^{-1} \widetilde{Z}_{d,C}(j),$$

are independent of the random part $E_{d,C}(j)$, and $\Theta_{s,C}(t) = \sum_{j=1}^{m} \omega_{j,h}(t)\Theta_C(j)$. Let $E^{\Theta}_{d,C}(t) = \sum_{j=1}^{m} \widetilde{B}_{d,j,C}(t)E_{d,C}(j) = \left\{E^{\gamma_{0,C}}_{d,C}(t), \left(E^{\Phi_{0,C}}_{d,C}(t)\right)^\top, \left(E^{\Phi_{1,C}}_{d,C}(t)\right)^\top\right\}^\top$ and $E^{\Theta}_{C}(t) = n^{-1/2}\sum_{d=1}^{n} E^{\Theta}_{d,C}(t)$.

The OLS estimation corresponds to the special case $h = 0$. We remark that $E^{\Theta}_{C}(t)$ is asymptotically normal when $h = 0$ and degenerates to a point distribution when $mh \to \infty$. To make the following analysis hold for the OLS-based test statistic, we view $E^{\Theta}_{C}(t)$ as a random variable in the discussion below.

For simplicity, let vec$(\cdot)$ be the operator that reshapes a matrix into a vector by stacking its columns on top of one another. Denote

$$x_{d,C}(t) = \left[e^{\alpha_{0,C}}_{d,C}(t), \left(e^{\alpha_{1,C}}_{d,C}(t)\right)^\top, \left(e^{\alpha_{2,C}}_{d,C}(t)\right)^\top, \left(E^{\gamma_{0,C}}_{d,C}(t)\right)^\top, \left\{\text{vec}\left(E^{\Phi_{0,C}}_{d,C}(t)\right)\right\}^\top, \left\{\text{vec}\left(E^{\Phi_{1,C}}_{d,C}(t)\right)\right\}^\top\right]^\top,$$
$$x_d(t) = \left[x_{d,C_0}(t)^\top, x_{d,C_1}(t)^\top\right]^\top \in \mathbb{R}^{2+p+2q(q+p_H+2)},$$
$$x_d = \left(x_d(2)^\top, x_d(3)^\top, \ldots, x_d(m)^\top\right)^\top \in \mathbb{R}^{p_x}, \quad p_x = (m-1)[2+p+2q(q+p_H+2)],$$

where p_H is the dimension of $\widetilde{X}_{d,C}$ and $\widetilde{X}_{d,S}$.

Let $\{y_d\}_d$ be independent mean zero Gaussian vectors with $\mathbb{E}y_d y_d^\top = \mathbb{E}x_d x_d^\top$. We similarly represent y_d as

$$y_{d,C}(t) = \left[\bar{e}^{\alpha_{0,C}}_{d,C}(t), \left(\bar{e}^{\alpha_{1,C}}_{d,C}(t)\right)^\top, \left(\bar{e}^{\alpha_{2,C}}_{d,C}(t)\right)^\top, \left(\bar{E}^{\gamma_{0,C}}_{d,C}(t)\right)^\top, \left\{\text{vec}\left(\bar{E}^{\Phi_{0,C}}_{d,C}(t)\right)\right\}^\top, \left\{\text{vec}\left(\bar{E}^{\Phi_{1,C}}_{d,C}(t)\right)\right\}^\top\right]^\top,$$
$$y_d(t) = \left[y_{d,C_0}(t)^\top, y_{d,C_1}(t)^\top\right]^\top \in \mathbb{R}^{2+p+2q(q+p_H+2)},$$
$$y_d = \left(y_d(2)^\top, y_d(3)^\top, \ldots, y_d(m)^\top\right)^\top \in \mathbb{R}^{p_x}, \quad p_x = (m-1)[2+p+2q(q+p_H+2)],$$

Let $\{e^b_{d,C_0}(j), e^b_{d,C_1}(j), E^b_{d,C_0}(j), E^b_{d,C_1}(j)\}$ be the empirical Gaussian analogs of $\{e_{d,C_0}(j), e_{d,C_1}(j), E_{d,C_0}(j), E_{d,C_1}(j)\}$. In other words, for $d = 1, \ldots, n$, $j = 1, \ldots, m$, let

$$e^b_{d,C_0}(j) = \hat{e}_{d,C_0}(j)\xi_d,\ e^b_{d,C_1}(j) = \hat{e}_{d,C_1}(j)\xi_d,\ E^b_{d,C_0}(j) = \hat{E}_{d,C_0}(j)\xi_d,\ E^b_{d,C_1}(j) = \hat{E}_{d,C_1}(j)\xi_d,$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. standard normal random variables. We next define

$$w_{d,C}(t) = \left[\bar{e}^{\alpha_{0,C},b}_{d,C}(t), \left(\bar{e}^{\alpha_{1,C},b}_{d,C}(t)\right)^\top, \left(\bar{e}^{\alpha_{2,C},b}_{d,C}(t)\right)^\top, \left(\bar{E}^{\gamma_{0,C},b}_{d,C}(t)\right)^\top, \left\{\text{vec}\left(\bar{E}^{\Phi_{0,C},b}_{d,C}(t)\right)\right\}^\top, \left\{\text{vec}\left(\bar{E}^{\Phi_{1,C},b}_{d,C}(t)\right)\right\}^\top\right]^\top,$$
$$w_d(t) = \left[w_{d,C_0}(t)^\top, w_{d,C_1}(t)^\top\right]^\top \in \mathbb{R}^{2+p+2q(q+p_H+2)},$$
$$w_d = \left(w_d(2)^\top, w_d(3)^\top, \ldots, w_d(m)^\top\right)^\top \in \mathbb{R}^{p_x}, \quad p_x = (m-1)[2+p+2q(q+p_H+2)],$$

Let

$$X = (X_2^\top, X_3^\top, \ldots, X_m^\top) = \frac{1}{\sqrt{n}}\sum_{d=1}^{n} x_d,$$
$$Y = (Y_2^\top, Y_3^\top, \ldots, Y_m^\top) = \frac{1}{\sqrt{n}}\sum_{d=1}^{n} y_d,$$
$$W = (W_2^\top, W_3^\top, \ldots, W_m^\top) = \frac{1}{\sqrt{n}}\sum_{d=1}^{n} w_d.$$

Define the following function

$$
\begin{aligned}
& F_{\text{GATE}}(X;\theta_{C_0},\theta_{C_1},\Theta_{C_0},\Theta_{C_1}) \\
\equiv\ & \frac{1}{m}\sum_{t=2}^{m}(\alpha_{0,C_1}(t)-\alpha_{0,C_0}(t)+\frac{e_{i,C}^{\alpha_{0,C_1}}(t)}{\sqrt{n}}-\frac{e_{i,C}^{\alpha_{0,C_0}}(t)}{\sqrt{n}}) \\
& +\frac{1}{m}\sum_{t=1}^{m}\Big[\mathbb{E}(n^{-1}\sum_{i=1}^{n}X^i(t))(\alpha_{1,C_1}(t)-\alpha_{1,C_0}(t)+\frac{e_{i,C}^{\alpha_{1,C_1}}(t)}{\sqrt{n}}-\frac{e_{i,C}^{\alpha_{1,C_0}}(t)}{\sqrt{n}})\Big] \\
& +\frac{1}{m}\sum_{t=1}^{m}\Big[\Big(\alpha_{2,C_1}(t)+\frac{e_{i,C}^{\alpha_{2,C_1}}(t)}{\sqrt{n}}\Big)^{\top}\Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}(\Phi_{2,C_1}(l)+\frac{E_{i,C}^{\Phi_{2,C_1}}(l)}{\sqrt{n}})(\Phi_{0,C_1}(k)+\frac{E_{i,C}^{\Phi_{0,C_1}}(k)}{\sqrt{n}} \\
& +(\Phi_{1,C_1}(t)+\frac{E_{i,C}^{\Phi_{1,C_1}}(t)}{\sqrt{n}})\mathbb{E}(n^{-1}\sum_{i=1}^{n}H_{C_1}^i(t)))\}+\prod_{l=1}^{t-1}(\Phi_{2,C_1}(l)+\frac{E_{i,C}^{\Phi_{2,C_1}}(l)}{\sqrt{n}})\mathbb{E}(S(1))\Big) \\
& -\Big(\alpha_{2,C_0}(t)+\frac{e_{i,C}^{\alpha_{2,C_0}}(t)}{\sqrt{n}}\Big)^{\top}\Big(\sum_{k=1}^{t-1}\{\prod_{l=k+1}^{t-1}(\Phi_{2,C_0}(l)+\frac{E_{i,C}^{\Phi_{2,C_0}}(l)}{\sqrt{n}})(\Phi_{0,C_0}(k)+\frac{E_{i,C}^{\Phi_{0,C_0}}(k)}{\sqrt{D}} \\
& +(\Phi_{1,C_0}(t)+\frac{E_{i,C}^{\Phi_{1,C_0}}(t)}{\sqrt{n}})\mathbb{E}(n^{-1}\sum_{i=1}^{n}H_{C_0}^i(t)))\}+\prod_{l=1}^{t-1}(\Phi_{2,C_0}(l)+\frac{E_{i,C}^{\Phi_{2,C_0}}(l)}{\sqrt{D}})\mathbb{E}(S(1))\Big)\Big].
\end{aligned}
$$

We next represent the proposed test statistic and the bootstrap samples based on F_{GATE}. Recall that $\Theta_{s,C}(t)=\sum_{j=1}^{m}\omega_{j,h}(t)\Theta_C(j)$ and $\theta_{s,C}(t)=\sum_{j=1}^{m}\omega_{j,h}(t)\theta_C(j)$ are the smoothed parameters, and $\widetilde{\theta}_C,\widetilde{\Theta}_C$ correspond to the estimates. The difference between the proposed test statistic and the oracle indirect effect $m^{-1}(\hat{T}-\text{GATE})$ can be represented as $T_0^*=F_{\text{GATE}}(X;\theta_{s,C_0},\theta_{s,C_1},\Theta_{s,C_0},\Theta_{s,C_1})-F_{\text{GATE}}(0;\theta_{C_0},\theta_{C_1},\Theta_{C_0},\Theta_{C_1})$. Similarly, we can represent $T^{-1}(\widehat{\text{GATE}}^{b}-\widehat{\text{GATE}})$ by $W_0^*=F_{\text{GATE}}(W;\widetilde{\theta}_{C_0},\widetilde{\theta}_{C_1},\widetilde{\Theta}_{C_0},\widetilde{\Theta}_{C_1})-F_{\text{GATE}}(0;\widetilde{\theta}_{C_0},\widetilde{\theta}_{C_1},\widetilde{\Theta}_{C_0},\widetilde{\Theta}_{C_1})$. By definition, we have

$$\rho^*(z)=\Big|P\{T_0^*\le z\}-P\{W_0^*\le z\}\Big|.$$

We also define the oracle statistics: $T_0=F_{\text{GATE}}(X;\theta_{C_0},\theta_{C_1},\Theta_{C_0},\Theta_{C_1})-F_{\text{GATE}}(0;\theta_{C_0},\theta_{C_1},\Theta_{C_0},\Theta_{C_1})=F_{\text{GATE}}(X)-F_{\text{GATE}}(0)$, $W_0=F_{\text{GATE}}(W;\theta_{C_0},\theta_{C_1},\Theta_{C_0},\Theta_{C_1})-F_{\text{GATE}}(0;\theta_{C_0},\theta_{C_1},\Theta_{C_0},\Theta_{C_1})=F_{\text{GATE}}(W)-F_{\text{GATE}}(0)$ by replacing $\theta_{s,C}$, $\widetilde{\theta}_C$, $\Theta_{s,C}$ and $\widetilde{\Theta}_C$, $C\in\{C_0,C_1\}$, with the oracle values. This yields an upper bound for

$$\rho(z)=\Big|P\{T_0\le z\}-P\{W_0\le z\}\Big|.$$

Following the derivation of Lemma 4 in Luo et al. [2024], we know $\sup_z\rho(z)\le\widetilde{C}n^{-1/8}$, for some constant $\widetilde{C}>0$.

Notice that

$$
\begin{aligned}
\rho^*(z) &= \Big|P\{T_0^*\le z\}-P\{W_0^*\le z\}\Big| \\
&\le \Big|P\{T_0^*\le z\}-P\{T_0\le z\}\Big|+\Big|P\{W_0^*\le z\}-P\{W_0\le z\}\Big|+\Big|P\{T_0\le z\}-P\{W_0\le z\}\Big| \\
&\le I_1+I_2+\widetilde{C}n^{-1/8},
\end{aligned}
$$

where I_1 and I_2 denote the above first two components, respectively.

Define $T_{01}^*:=F_{\text{GATE}}(X;\theta_{s,C_0},\theta_{s,C_1},\Theta_{s,C_0},\Theta_{s,C_1})-F_{\text{GATE}}(0;\theta_{s,C_0},\theta_{s,C_1},\Theta_{s,C_0},\Theta_{s,C_1})$. Then I_1 can be further divided into two part:

$$
\begin{aligned}
I_1 &= \Big|P\{T_0^*\le z\}-P\{T_0\le z\}\Big| \\
&\le \Big|P\{T_0^*\le z\}-P\{T_{01}^*\le z\}\Big|+\Big|P\{T_{01}^*\le z\}-P\{T_0\le z\}\Big| \\
&= I_{11}+I_{12},
\end{aligned}
$$

where I_{11} and I_{12} denote the above two components, respectively.

Similar to the proof of Theorem 2 in Luo et al. [2024], we can obtain that

$$I_2 \leq \widetilde{C} n^{-1/8},\ I_{12} \leq \widetilde{C} n^{-1/8}.$$

Also, according to the proof of Theorem 2 in Luo et al. [2024], we know that

$$I_{11} = O(n^{1/2} h^2 + n^{1/2} m^{-1}),$$

holds with probability 1 as $n \to \infty$. The proof is hence completed.

□