Covariate-Adjusted Deep Causal Learning for Heterogeneous Panel Data Models

Guanhao Zhou, Yuefeng Han, and Xiufan Yu
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
{gzhou4, yuefeng.han, xiufan.yu}@nd.edu

Abstract

This paper studies the task of estimating heterogeneous treatment effects in causal panel data models with covariate effects. We propose a novel **Co**variate-Adjusted **DE**ep **CA**usal **L**earning (**CoDEAL**) framework that cohesively deals with the underlying heterogeneity and nonlinearity of both panel units and covariate effects. CoDEAL integrates nonlinear covariate effect components (parameterized by a feed-forward neural network) with nonlinear factor structures (modeled by a multi-output autoencoder) to form a heterogeneous causal panel model. The nonlinear covariate component flexibly captures complex covariate influences on outcomes, and the nonlinear factor decomposition enables CoDEAL to effectively capture both cross-sectional and temporal dependencies inherent in the data panel. This latent structural information is subsequently integrated into a customized matrix completion algorithm, thereby facilitating more accurate counterfactual imputation. Moreover, the use of a multi-output autoencoder explicitly accounts for heterogeneity across units and enhances the model interpretability of the latent factors. We establish theoretical guarantees on the convergence of the estimated counterfactuals and demonstrate the compelling performance of the proposed method using extensive simulation studies and real-world data applications.

1 Introduction

Causal inference in panel data settings has attracted growing interest in recent years, with broad applications across various fields such as economics [23], life science [35], political science [38], and social science [39]. Causal panel data analysis is inherently a missing data problem due to the *fundamental problem of causal inference*: for each unit and time period, we can only observe either the treated or the untreated outcome, but never both. The panel data structure brings unique challenges to causal analysis. Different units may get exposed to the treatment at various time periods, forming a block-wise non-random missing pattern [10]. The presence of both temporal and cross-sectional dependencies could bias the estimation if not properly addressed [62], and unit-specific heterogeneity can confound causal interpretation [46], especially if additional covariates are present.

To estimate causal effects, one natural approach is to first impute the missing counterfactual, and then examine the difference between the imputed counterfactuals and the actual observed outcomes. Athey et al. [12] formulated the imputation of missing counterfactuals as a matrix completion task with nuclear norm minimization. It is important to note that standard matrix completion approaches are not directly suitable for causal imputation [22]. Most conventional matrix completion methods operate under the assumption that missing entries are *missing at random* (MAR), which enables estimation of missing values without explicitly modeling the missingness mechanism. However, in causal panel data settings where different units receive treatment at different times, missing entries correspond to counterfactual outcomes that are systematically unobserved due to treatment assignments, invalidating

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the MAR assumption. As a result, applying standard matrix completion blindly without properly accounting for this structured, non-random missingness can lead to biased estimation [5].

In recent years, a growing body of work has focused on developing causal matrix completion methods that handle missing-not-at-random panel designs. Besides the optimization-based causal matrix completion [12, 22], another line of research leverages factor models [13, 14] to capture the latent low-rank structure of observed entries. Most factor-model-based causal matrix completion methods in the literature [5, 15, 68, 70, 72] posit a linear factor structure. Linear factor analysis helps capture cross-sectional and temporal dependencies to some extent [32–34, 75], but is limited in capturing complex nonlinear relationships, motivating the generalization to nonlinear factor models [60, 71].

Another key aspect in panel analysis is to account for unit-specific heterogeneity [46, 58]. Most relevant works in the literature still focus on the overall average treatment effects across units [9, 12], as opposed to unit-specific heterogeneous treatment effects. Pooled estimates may mask important variations across units. Units in a panel often differ in characteristics and respond to treatment differently [48, 65], which may bias the estimates of treatment effects, motivating a critical need to estimate unit-specific heterogeneous treatment effects and adjust for covariate effects if possible.

In this paper, we develop **Co**variate-Adjusted **DE**ep C**A**usal **L**earning (**CoDEAL**) for panel data models, a deep-learning-based causal learning method for estimating heterogeneous treatment effects. CoDEAL captures covariate effects with a deep neural network regression, and models the covariate-adjusted outcomes via a nonlinear factor structure. Our contributions can be summarized as follows.

- CoDEAL provides **a causal matrix completion method** that addresses the non-random missingness mechanism in causal panel data analysis. By leveraging nonlinear factor analysis, CoDEAL **exploits both cross-sectional and temporal dependencies**. This latent structural information is subsequently integrated into the matrix completion procedure, thereby facilitating more accurate imputation of missing counterfactual outcomes. Theoretically, we rigorously prove that CoDEAL provides consistent estimates of unit-specific heterogeneous treatment effects.
- CoDEAL employs flexible model structures and powerful neural network architectures to cohesively deal with the underlying **heterogeneity** and **nonlinearity** of **both panel units and covariate effects**. Algorithmically, we begin with a feed-forward deep neural network (DNN) to remove nonlinear covariate effects, followed by a multi-output autoencoder (AE) to recover nonlinear latent factors and explicitly account for unit-specific heterogeneity.
- CoDEAL offers **a unifying framework** for heterogeneous causal panel models in the presence of covariates. By combining nonlinear covariate adjustment with deep latent factor models, CoDEAL generalizes beyond traditional linear factor models and unifies a class of factor-model-based causal matrix completion approaches [e.g., 5, 15, 68, 72]. Moreover, with different choices of factor models and neural network architectures, the proposed framework can be further generalized to accommodate various data types such as spatial and tensor data.

Related Work. (a) An overview of causal learning methods in panel data models. One traditional approach to estimate treatment effects in causal panel models is *the difference-in-differences* (DiD) [10, 38, 66] that compares the average changes in outcomes over time between the treated and untreated units under a *parallel trend* assumption. Alternatively, one can estimate the causal effects by first imputing the missing counterfactual outcomes, and then examining the difference between the imputed counterfactuals and the actual observed outcomes. Following this strand, two approaches have emerged in recent causal panel literature: *the unconfoundedness-based approach* (also known as horizontal regression) [40, 53] and *the synthetic control method* (also known as vertical regression) [1–4, 27]. The close connections between the DiD, unconfoundedness-based methods, synthetic control approaches, and causal matrix completion estimators were revealed by Athey et al. [12]. Despite their distinct appearances, Athey et al. [12] showed that the linear versions of all four estimators can be characterized as solutions to the same optimization problem with the exact same objective function, but subject to different restrictions on parameters of that optimization. In parallel, a separate semiparametric literature develops doubly robust and machine-learning-based semiparametric estimators [20, 41, 43, 55, 63] under DiD-type identification assumptions. Our framework aligns conceptually with this push for flexibility; however, it departs from the semiparametric approach by adopting nonlinear, time-varying covariate functions and a nonlinear latent factor structure. We refer readers to [7] for a comprehensive review of past and recent progress on causal panel models.

(b) Causal matrix completion and deep matrix completion. Conceptually, framing causal estimation and inference as the recovery of systematically missing counterfactual information is widely adopted across heterogeneous data structures and missingness patterns [24–26, 37]. Here, we focus

specifically on matrix completion approaches that are most relevant to our study on causal panel data models. A surge of interest in the recent past has led to new research on causal matrix completion methods capable of addressing missing-not-at-random panel data, e.g., optimization-based methods [12, 22] and factor-model-based methods [5, 15, 68, 70, 72]. There has been a handful of deep-learning-based methods for matrix completion, e.g., [28, 29, 52, 60]. However, most of them assume MAR, and therefore, are not directly suitable for causal imputation in panel data models.

2 Methodology

2.1 Problem Setup

Heterogeneous Treatment Effects in Causal Panel Data Models. We begin by setting up the causal learning problem in panel data. Consider a panel data setting with N units over T periods. Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For each unit $i \in [N]$ and time point $t \in [T]$, there are two potential outcomes, $Y_{it}(1)$ and $Y_{it}(0)$, representing the outcomes the i-th unit would experience at time t under treatment and control, respectively. The treatment assignment is captured by a binary indicator variable $W_{it} \in \{0, 1\}$, where $W_{it} = 1$ indicates that the unit i receives the treatment at time t and $W_{it} = 0$ otherwise. Suppose we also have access to some unit-specific covariates $\mathbf{X}_i \in \mathbb{R}^P$ that could potentially affect the observations $\{Y_{it}, t \in [T]\}$. We denote the matrix forms of the observed outcomes, treatment indicators, and covariates by $\mathbf{Y} = (Y_{it}) \in \mathbb{R}^{N\times T}$, $\mathbf{\Omega} = (W_{it}) \in \mathbb{R}^{N\times T}$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^\top \in \mathbb{R}^{N\times P}$, respectively. Let $\mathbf{Y}(1) = (Y_{it}(1)) \in \mathbb{R}^{N\times T}$ and $\mathbf{Y}(0) = (Y_{it}(0)) \in \mathbb{R}^{N\times T}$ be the matrices of potential outcomes of the treated and untreated, respectively. Our aim in this work is to estimate the unit-specific (potentially heterogeneous) average treatment effects on the treated (ATT), $\tau_i^\star = \mathbb{E}_t[Y_{it}(1) - Y_{it}(0) \mid W_{it} = 1]$ for each treated unit i.

Staggered Adoption Design (SAD). This work focuses on SAD [10], where treatments can be rolled out at different times across various units and are irreversible once initiated. We assume there is at least one never-treated panel unit. Details of SAD are in Appendix A.1.

Causal Assumptions. To ensure valid interpretations of the estimators, we adopt the following assumptions throughout the analysis. (i) (Stable Unit Treatment Value Assumption (SUTVA).) The potential outcomes for a given unit at a particular time depend solely on the treatment administered to that unit at that time, and are unaffected by treatments assigned to other units or at other time points. (ii) (Static Treatment Effects.)[1] The treatment effects are non-dynamic, i.e., do not change over time.

Challenges and Limitations of Existing Methods. The *fundamental problem of causal inference* implies that $Y_{it}(0)$ and $Y_{it}(1)$ cannot be observed simultaneously. The panel data forms two partially observed data matrices $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, where the missing pattern is determined according to the treatment assignments. To estimate the treatment effects, one natural idea is to estimate the unobserved counterfactuals (impute the missing potential outcomes). Since our focus is the unit-specific ATT and $\{Y_{it}(1) : W_{it} = 1\}$ is observed, it suffices to estimate $\{Y_{it}(0) : W_{it} = 1\}$ to impute missing elements in $\mathbf{Y}(0)$. Let $\{\widehat{Y}_{it}(0) : W_{it} = 1\}$ denote imputed values. The unit-specific ATT can then be estimated by $\widehat{\tau}_i = \sum_{t:W_{it}=1}[Y_{it}(1) - \widehat{Y}_{it}(0)]/\sum_t W_{it}$ for each treated unit i.

There are three main limitations of existing methods that this work aims to overcome in the task of modeling unit-specific ATT and imputing missing potential outcomes. First, the vast majority of matrix completion methods rely on the MAR assumption, which is typically violated in SADs, resulting in biased estimates. Second, existing matrix-completion-based causal panel models mostly focus on the overall (non-unit-specific) ATT, without adequately accounting for unit-level heterogeneity. Third, many existing methods often lack the capacity to disentangle the causal impact of treatment from confounding covariate effects and to model the underlying complex nonlinear dependencies, limiting both the precision of counterfactual predictions and the interpretability of causal analysis.

2.2 Covariate-Adjusted Deep Causal Learning (CoDEAL) for Panel Data Models

To tackle the above challenges, we propose the CoDEAL method that (i) leverages the staggered adoption structure to fully exploit both cross-sectional and temporal dependencies for more accurate imputation of missing potential outcomes, (ii) models the unit-specific heterogeneity explicitly to ac-

[1]The static treatment assumption is not intrinsic to the proposed algorithm, but can be extended to accommodate time-varying contemporaneous treatment effects. We discuss this extension in Remark 3. For clarity of exposition, we maintain the static treatment assumption when presenting the methodology and theoretical results preceding Remark 3.

commodate variation in treatment effects across units, and (iii) involves flexible model structures and employs powerful neural network architectures to efficiently capture complex nonlinear relationships. For each unit $i \in [N]$ at time $t \in [T]$, we model the observed outcome by

$$Y_{it} = \phi_i^\star(\mathbf{F}_t^\star) + g_t^\star(\mathbf{X}_i) + \tau_i^\star \cdot \mathbb{1}\{W_{it} = 1\} + \varepsilon_{it}, \tag{1}$$

where $\mathbf{F}_t^\star \in \mathbb{R}^K$ is a K-dimensional vector of latent factors, $\phi_i^\star(\cdot) : \mathbb{R}^K \to \mathbb{R}$ is a potentially nonlinear factor loading function, $g_t^\star(\cdot) : \mathbb{R}^P \to \mathbb{R}$ is a potentially nonlinear function that captures covariate effects and $\mathbb{1}\{\cdot\}$ is an indicator function, and ε_{it} is the idiosyncratic noise. An illustrative DAG of Model (1) is in Appendix A.2. CoDEAL employs a nonlinear factor model [71] to characterize the underlying low-rank structure of the covariate-adjusted outcomes. By letting $\phi_i^\star(\cdot)$ take different forms, CoDEAL unifies a class of factor-model-based causal matrix completion approaches, e.g., [5, 15, 68, 70, 72]. As shown by (1), CoDEAL accommodates heterogeneous but non-dynamic treatment effects. The covariate effects can be heterogeneous across units and time-varying over periods. Here, we allow N and T to be large and diverging, while P and K are fixed and small.

For clarity, we first present the learning procedures using the simplest form of a 2×2 staggered adoption (also known as the four-block design) in Section 2.2.1, and generalize the algorithm to the more general SAD in Section 2.2.2. More details about the four-block and SAD as well as their graphical illustrations are in Appendix A.1. A graphical illustration of CoDEAL in the four-block design is provided in Figure 1. CoDEAL in general SAD is summarized in Algorithm 2.

2.2.1 CoDEAL in Four-Block Design

We begin with the four-block design, the simplest form of staggered adoption and often regarded as a foundational building block for more general SADs. The four-block design is a 2×2 SAD, where a subset of N_1 units never receive the treatment while the remaining $N_2 = N - N_1$ units are exposed to the irreversible treatment starting at time $t = T_1 + 1$. Without loss of generality, we rearrange the panel such that the first N_1 units ($i = 1, \ldots, N_1$) are untreated. We refer to the time periods $t \in [T_1]$ as pre-treatment periods and $t \in [T] \backslash [T_1]$ as post-treatment periods. Let $T_2 = T - T_1$ denote the duration of post-treatment periods. The design naturally decomposes the panel into four blocks:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{\mathcal{A}} & \mathbf{Y}_{\mathcal{B}} \\ \mathbf{Y}_{\mathcal{C}} & \mathbf{Y}_{\mathcal{D}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\mathcal{A}}(0) & \mathbf{Y}_{\mathcal{B}}(0) \\ \mathbf{Y}_{\mathcal{C}}(0) & \mathbf{Y}_{\mathcal{D}}(1) \end{bmatrix}, \text{ and } \mathbf{Y}(0) = \begin{bmatrix} \mathbf{Y}_{\mathcal{A}}(0) & \mathbf{Y}_{\mathcal{B}}(0) \\ \mathbf{Y}_{\mathcal{C}}(0) & ? \end{bmatrix} \tag{2}$$

where $\mathbf{Y}_{\mathcal{A}} \in \mathbb{R}^{N_1 \times T_1}$, $\mathbf{Y}_{\mathcal{B}} \in \mathbb{R}^{N_1 \times T_2}$, $\mathbf{Y}_{\mathcal{C}} \in \mathbb{R}^{N_2 \times T_1}$, and $\mathbf{Y}_{\mathcal{D}} \in \mathbb{R}^{N_2 \times T_2}$. By construction, $\mathbf{Y}_{\mathcal{A}}$, $\mathbf{Y}_{\mathcal{B}}$, $\mathbf{Y}_{\mathcal{C}}$ are observed outcomes of the untreated with $Y_{it} = Y_{it}(0)$ for $(i,t) \in \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$, while $\mathbf{Y}_{\mathcal{D}}$ are observed outcomes of the treated with $Y_{it} = Y_{it}(1)$ for $(i,t) \in \mathcal{D}$, leaving a missing block in $\mathbf{Y}(0)$. To estimate $\tau_i^\star$, it suffices to impute the missing values of $\mathbf{Y}_{\mathcal{D}}(0)$. Denote the imputed block as $\widehat{\mathbf{Y}}_{\mathcal{D}}(0)$. The unit-specific ATT can be then estimated by $\widehat{\tau}_i = \frac{1}{T_2} \sum_{t:(i,t)\in\mathcal{D}} [Y_{it}(1) - \widehat{Y}_{it}(0)]$ for $i \in [N] \backslash [N_1]$.

Modeling Covariate Effects. To account for the potentially nonlinear, heterogeneous, and time-varying effects brought by covariates, we consider a fully connected DNN for $g_t^\star(\cdot)$. We construct our model using a fully connected DNN with ReLU activation, due to its strong empirical performance. We refer to this architecture as a *deep ReLU network*. Let $L \in \mathbb{N}$ denote the network depth and let $\mathbf{d} = (d_1, \ldots, d_{L+1}) \in \mathbb{N}^{L+1}$ specify the layer widths. A deep ReLU network is a function mapping $\mathbb{R}^{d_0}$ to $\mathbb{R}^{d_{L+1}}$ and takes the form

$$h(x) = \mathcal{L}_{L+1} \circ \bar{\sigma}_L \circ \mathcal{L}_L \circ \bar{\sigma}_{L-1} \circ \cdots \circ \mathcal{L}_2 \circ \bar{\sigma}_1 \circ \mathcal{L}_1(x), \tag{3}$$

where each $\mathcal{L}_\ell(z) = M_\ell z + b_\ell$ is an affine transformation with weight matrix $M_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and bias vector $b_\ell \in \mathbb{R}^{d_\ell}$, and $\bar{\sigma}_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_\ell}$ applies the ReLU activation function entrywise. For simplicity, we refer to both M_ℓ and b_ℓ as the weights of the deep ReLU network.

Definition 1 (Deep ReLU network class)**.** *For any $L \in \mathbb{N}$, $\mathbf{d} \in \mathbb{N}^{L+1}$, and constants $B, C > 0$, we define the class of deep ReLU networks with depth L, width $\mathbf{d}$, and weights $\mathbf{M} = \{M_1, ..., M_{L+1}, b_1, ..., b_{L+1}\}$ bounded by C as*

$\mathcal{G}_{d_0}^{d_{L+1}}(L, \mathbf{d}, B, C) = \{\tilde{h}(x) = sgn(h(x))(|h(x)| \wedge B) : h \text{ of form } (3),\ \|M_\ell\|_\infty \le C,\ \|b_\ell\|_\infty \le C\}$.

The covariate effects can be estimated via

$$\widehat{g}_t(\cdot) = \underset{g_t \in \mathcal{G}_P^1(L,\mathbf{d},B,C)}{\arg\min} \sum_{i:W_{it}=0} (Y_{it} - g_t(\mathbf{X}_i))^2, \tag{4}$$

for a certain functional class $\mathcal{G}_P^1(L, \mathbf{d}, B, C)$ specified in Appendix E. We define the covariate-adjusted observed outcomes for unit $i \in [N]$ at time $t \in [T]$ as $\widetilde{Y}_{it} = Y_{it} - \widehat{g}_t(\mathbf{X}_i)$.

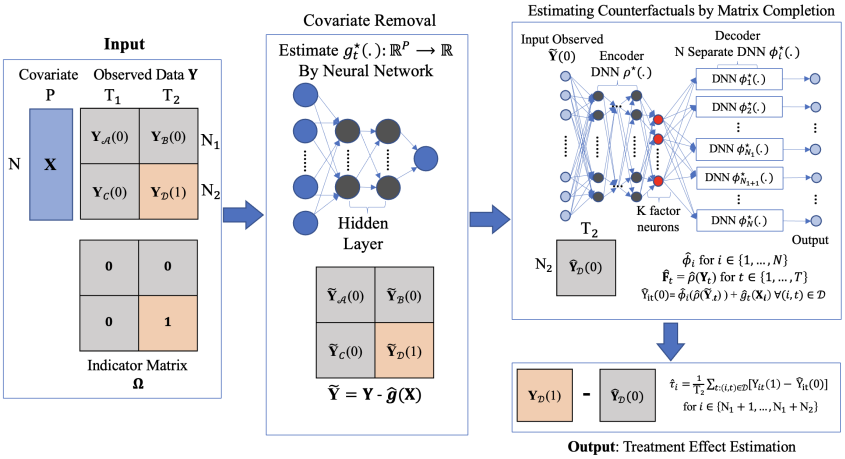

Figure 1: A graphical illustration of the proposed CoDEAL in a four-block design. Gray and orange refer to the untreated and treated blocks, respectively. Inputs are observed outcomes $\mathbf{Y} \in \mathbb{R}^{N\times T}$, covariates $\mathbf{X} \in \mathbb{R}^{N\times P}$, indicator matrix $\mathbf{\Omega} \in \mathbb{R}^{N\times T}$. Output is the estimated unit-specific ATT $\widehat{\tau}_i$.

Estimating Counterfactuals by Matrix Completion via a Multi-Output Autoencoder. With covariate-adjusted outcomes, it remains to model the nonlinear factor structure $\phi_i^\star(\mathbf{F}_t^\star)$ in (1). Inspired by [60], we consider tackling $\phi_i^\star(\mathbf{F}_t^\star)$ using a multi-output AE.

The architecture of the multi-output AE consists of a single DNN encoder and multiple independent DNN decoders. The encoder $\rho(\cdot) \in \mathcal{G}_N^{K_1}(L_1, \mathbf{d}_1, B, C)$ maps the input to a K_1-dimensional bottleneck representation. For each of the N output components, a distinct DNN decoder is used; specifically, the decoder for the i-th output is given by $\phi_i(\cdot) \in \mathcal{G}_{K_1}^1(L_2, \mathbf{d}_2, B, C)$ for $i \in [N]$. Consequently, the i-th output of the multi-output AE for an input x is given by $\phi_i \circ \rho(x)$. We formally define the function class of the multi-output AE as:

$$\mathcal{G}_{\text{AE}}^{K_1} := \Big\{(\phi_1, \ldots, \phi_N) \circ \rho : \rho \in \mathcal{G}_N^{K_1}(L_1, \mathbf{d}_1, B, C), \quad \phi_i \in \mathcal{G}_{K_1}^1(L_2, \mathbf{d}_2, B, C),\ i \in [N]\Big\} \tag{5}$$

where $K_1 \geq K$. More details on this function class are provided in Appendix E.

With the multi-output AE, the encoder $\rho^\star(\cdot)$ effectively approximates the latent factors, facilitating the factor-model-based matrix completion. The unit-specific decoders $\phi_i^\star(\cdot)$, $i \in [N]$, are better equipped to capture the heterogeneity across units, in contrast to a traditional single-output AE which limits $\phi_1^\star = \ldots = \phi_N^\star$. The DNNs $(\widehat{\phi}_1, \ldots, \widehat{\phi}_N) \circ \widehat{\rho}(\cdot)$ in the multi-output AE are estimated by

$$\underset{(\phi_1,\ldots,\phi_N)\circ\rho(\cdot)\in\mathcal{G}_{\text{AE}}^{K_1}}{\arg\min} \sum_{(i,t):W_{it}=0} \big(\phi_i(\rho(\widetilde{\mathbf{Y}}_{\cdot t})) - \widetilde{Y}_{it}\big)^2. \tag{6}$$

The counterfactuals $\mathbf{Y}_{\mathcal{D}}(0)$ can then be imputed by $\widehat{Y}_{it}(0) = \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_{\cdot t})) + \widehat{g}_t(\mathbf{X}_i)$ for any $(i,t) \in \mathcal{D}$. The complete steps of imputing counterfactuals in the four-block design are summarized as Algorithm 1 to facilitate further generalizations to the staggered design.

Remark 1 (Rationale of Sequential Estimation)**.** *We adopt a two-step estimation strategy over a joint estimation of covariate and factor effects for multiple reasons. It is not merely a procedural convenience, but rather a natural alignment with the data structure that improves stability and interpretability. We provide a detailed justification and supporting numerical evidence in Appendix A.3.*

Algorithm 1 Estimating Counterfactuals by Matrix Completion in the Four-Block Design

1: **Input:** Observed outcomes $\mathbf{Y}$, covariates $\mathbf{X}$, the estimated covariate effects $\widehat{g}_t(\cdot)$ from (4), treatment indicators $\mathbf{\Omega}$ with a four-block design in (2), latent factor dimension K
2: **Output:** Imputed outcome matrix $\widehat{\mathbf{Y}}_{\mathcal{D}}(0)$.
3: Obtain the covariate-adjusted observed outcomes by $\widetilde{Y}_{it} = Y_{it} - \widehat{g}_t(\mathbf{X}_i)$ for $i \in [N]$, $t \in [T]$.
4: Fit a multi-output AE using the entries in the untreated part $\{\widetilde{Y}_{it}(0) : W_{it} = 0\}$ and obtain $(\widehat{\phi}_1, \ldots, \widehat{\phi}_N, \widehat{\rho})$ from (6).
5: Estimate the missing elements by $\widehat{Y}_{it}(0) = \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_{\cdot t})) + \widehat{g}_t(\mathbf{X}_i)$ for any $(i,t) \in \mathcal{D}$.
6: **Return:** $\widehat{\mathbf{Y}}_{\mathcal{D}}(0)$.

Algorithm 2 CoDEAL: Staggered Adoption Design

Input: Observed outcomes $\mathbf{Y}$, covariates $\mathbf{X}$, treatment indicators $\mathbf{\Omega}$, latent factor dimension K.
Output: Estimated unit-specific ATT $\widehat{\tau}_i$ for each treated i.
Obtain the estimated covariate effect $\widehat{g}_t(\cdot)$ by (4)
Initialize $\widehat{\mathbf{Y}}(0) \in \mathbb{R}^{N\times T}$ by setting $\widehat{Y}_{it}(0) = Y_{it}$ for the untreated $\{(i,t) : W_{it} = 0\}$.
Extract block partitions $\{N_\xi\}_{1\le\xi\le r}$ and $\{T_\eta\}_{1\le\eta\le r}$ from $\mathbf{\Omega}$
for $\xi_0 = 2, \ldots, r$ **do**
  for $\eta_0 = r + 2 - \xi_0, \ldots, r$ **do**
    Construct the four-block submatrix $\mathbf{Y}^{(\xi_0,\eta_0)}$ and $\mathbf{\Omega}^{(\xi_0,\eta_0)}$ according to Equation (7).
    Call Algorithm 1 with inputs $\mathbf{Y}^{(\xi_0,\eta_0)}$, $\mathbf{X}^{(\xi_0)}$, $\widehat{g}_t(\cdot)$, $\mathbf{\Omega}^{(\xi_0,\eta_0)}$, and K to obtain the imputed block $\widehat{\mathbf{Y}}_{\mathcal{D}_{(\xi_0,\eta_0)}}(0)$, and extract the estimated block $\widehat{\mathbf{Y}}_{(\xi_0,\eta_0)}(0)$ from $\widehat{\mathbf{Y}}_{\mathcal{D}_{(\xi_0,\eta_0)}}(0)$.
    Update the matrix $\widehat{\mathbf{Y}}(0)$ with the imputed block $\widehat{\mathbf{Y}}_{(\xi_0,\eta_0)(0)}$.
  end for
end for
Compute unit-specific estimated ATT as $\widehat{\tau}_i = \sum_{t:W_{it}=1}[Y_{it}(1) - \widehat{Y}_{it}(0)]/\sum_t W_{it}$.
Return: Estimated unit-specific ATT $\widehat{\tau}_i$.

2.2.2 CoDEAL in Staggered Adoption Design

The extension to SAD is a natural generalization of the four-block design. Under the SAD, we rearrange the panels by sorting the N units based on the time at which they began receiving the treatment in a descending order (with never-treated units at the top, followed by the latest adopters, and then progressively earlier adopters towards the bottom). This sorting procedure yields a block structure in the treatment indicator matrix $\mathbf{\Omega}$. By grouping the units based on their adoption timing and segmenting the overall time periods into relevant intervals, we partition the indicator matrix $\mathbf{\Omega}$ and outcome $\mathbf{Y}$ into submatrices $\mathbf{\Omega}_{(\xi,\eta)} \in \mathbb{R}^{N_\xi\times T_\eta}$ and $\mathbf{Y}_{(\xi,\eta)} \in \mathbb{R}^{N_\xi\times T_\eta}$ for $\xi = 1, \ldots, r$ and $\eta = 1, \ldots, r$, where $\sum_{\xi=1}^r N_\xi = N$ and $\sum_{\eta=1}^r T_\eta = T$. Examples of block partitions in the staggered design are included in Appendix A.1. Let $\mathcal{I}_{tr} = \{(\xi,\eta) : \mathbf{\Omega}_{(\xi,\eta)} = 1\}$ denote the index set of the treated blocks. By the construction of these partitions, the indices in $\mathcal{I}_{tr}$ satisfy $\xi + \eta > r + 1$.

To estimate the unit-specific ATT, it suffices to estimate the counterfactuals of $\mathbf{Y}_{(\xi,\eta)}$ for all $(\xi,\eta) \in \mathcal{I}_{tr}$. The imputation of $\mathbf{Y}_{(\xi,\eta)}(0)$ can be reduced to a four-block design problem [72], therefore the method proposed in the previous Section 2.2.1 can be directly applied. Specifically, we construct a four-block data matrix $\mathbf{Y}^{(\xi_0,\eta_0)}$ as follows to estimate $\mathbf{Y}_{(\xi_0,\eta_0)}(0)$ for some $(\xi_0,\eta_0) \in \mathcal{I}_{tr}$:

$$\mathbf{Y}^{(\xi_0,\eta_0)} = \begin{bmatrix} \mathbf{Y}_{\mathcal{A}_{(\xi_0,\eta_0)}} & \mathbf{Y}_{\mathcal{B}_{(\xi_0,\eta_0)}} \\ \mathbf{Y}_{\mathcal{C}_{(\xi_0,\eta_0)}} & \mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\mathcal{A}_{(\xi_0,\eta_0)}}(0) & \mathbf{Y}_{\mathcal{B}_{(\xi_0,\eta_0)}}(0) \\ \mathbf{Y}_{\mathcal{C}_{(\xi_0,\eta_0)}}(0) & \mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}}(1) \end{bmatrix} \tag{7}$$

where $\mathcal{A}_{(\xi_0,\eta_0)} = \{(\xi,\eta) : 1 \le \xi \le k_1, 1 \le \eta \le k_2\}$, $\mathcal{B}_{(\xi_0,\eta_0)} = \{(\xi,\eta) : 1 \le \xi \le k_1, k_2 + 1 \le \eta \le \eta_0\}$, $\mathcal{C}_{(\xi_0,\eta_0)} = \{(\xi,\eta) : k_1 + 1 \le \xi \le \xi_0, 1 \le j \le k_2\}$, and $\mathcal{D}_{(\xi_0,\eta_0)} = \{(\xi,\eta) : k_1 + 1 \le \xi \le \xi_0, k_2 + 1 \le j \le \eta_0\}$ with $k_1 = r + 1 - \eta_0$ and $k_2 = r + 1 - \xi_0$. The associated treatment indicator matrix $\mathbf{\Omega}^{(\xi_0,\eta_0)}$ is defined analogously and the associated covariate matrix $\mathbf{X}^{(\xi_0)}$ can be obtained by retaining the units included in $\mathbf{Y}^{(\xi_0,\eta_0)}$. The missing potential outcomes $\mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}}(0)$ can be directly estimated using Algorithm 1. By construction, $\mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}} \supseteq \mathbf{Y}_{(\xi_0,\eta_0)}$. In the case when $\mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}} \supset \mathbf{Y}_{(\xi_0,\eta_0)}$, we treat all the blocks of $\mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}}$ as missing and impute their values, but only extract $\widehat{\mathbf{Y}}_{(\xi_0,\eta_0)}(0)$ from the imputed $\widehat{\mathbf{Y}}_{\mathcal{D}_{(\xi_0,\eta_0)}}(0)$; see Figure 2 for an example.

	T_1	T_2	T_3	T_4	T_5
N_1	$\mathbf{Y}_{(1,1)}$	$\mathbf{Y}_{(1,2)}$	$\mathbf{Y}_{(1,3)}$	$\mathbf{Y}_{(1,4)}$	$\mathbf{Y}_{(1,5)}$
N_2	$\mathbf{Y}_{(2,1)}$	$\mathbf{Y}_{(2,2)}$	$\mathbf{Y}_{(2,3)}$	$\mathbf{Y}_{(2,4)}$	
N_3	$\mathbf{Y}_{(3,1)}$	$\mathbf{Y}_{(3,2)}$	$\mathbf{Y}_{(3,3)}$	?	
N_4	$\mathbf{Y}_{(4,1)}$	$\mathbf{Y}_{(4,2)}$			
N_5	$\mathbf{Y}_{(5,1)}$				

	T_1	T_2	T_3	T_4	T_5
N_1	$\mathbf{Y}_{(1,1)}$	$\mathbf{Y}_{(1,2)}$	$\mathbf{Y}_{(1,3)}$	$\mathbf{Y}_{(1,4)}$	$\mathbf{Y}_{(1,5)}$
N_2	$\mathbf{Y}_{(2,1)}$	$\mathbf{Y}_{(1,2)}$	$\mathbf{Y}_{(2,3)}$	$\mathbf{Y}_{(2,4)}$	
N_3	$\mathbf{Y}_{(3,1)}$	$\mathbf{Y}_{(3,2)}$	$\mathbf{Y}_{(3,3)}$		
N_4	$\mathbf{Y}_{(4,1)}$	$\mathbf{Y}_{(4,2)}$			
N_5	$\mathbf{Y}_{(5,1)}$		?		

Figure 2: Examples of constructing a four-block submatrix $\mathbf{Y}^{(\xi_0,\eta_0)}$ to estimate $\mathbf{Y}_{(\xi_0,\eta_0)}(0)$. Here, $r = 5$. The question mark denotes the block of interest, with (ξ_0,η_0) = (3,4) (Left) and (5,3) (Right). The associated four blocks are designed by green ($\mathbf{Y}_{\mathcal{A}_{(\xi_0,\eta_0)}}$), red ($\mathbf{Y}_{\mathcal{B}_{(\xi_0,\eta_0)}}$), blue ($\mathbf{Y}_{\mathcal{C}_{(\xi_0,\eta_0)}}$), and orange ($\mathbf{Y}_{\mathcal{D}_{(\xi_0,\eta_0)}}$) as in (7).

With the four-block construction in (7) and the CoDEAL method in four-block design introduced in Section 2.2.1, we obtain the imputed values of $\mathbf{Y}_{(\xi,\eta)}(0)$ for all $(\xi,\eta) \in \mathcal{I}_{tr}$. The unit-specific

ATT can then be estimated by $\widehat{\tau}_i = \sum_{t:W_{it}=1}[Y_{it}(1) - \widehat{Y}_{it}(0)]/\sum_t W_{it}$ for each treated unit i. The complete algorithm of CoDEAL in the SAD is summarized as Algorithm 2.

2.3 Theoretical Properties

This section establishes theoretical guarantees. We derive an explicit convergence rate for counterfactual recovery that characterizes the effects of nonparametric estimation, latent factor recovery, and finite-sample approximation. The results imply the estimation consistency of the CoDEAL estimator.

Assumption 1. *Suppose $\mathbf{X}_i$ is supported on $[0,1]^P$ and has a density function f satisfying $\sup_{x\in[0,1]^P} f(x) =: f_{\max} < \infty$, where $f_{\max}$ is independent of dimension P. We assume there exists a constant $B > 0$ such that $\mathbb{E}\mathbf{F}_t^\star = 0$ and $\|\mathbf{F}_t^\star\|_\infty \le B$ holds almost surely. We further assume that ε_{it} is sub-Gaussian with sub-Gaussian norm σ_ϵ^2, and that $\varepsilon_{it}, \mathbf{F}_t^\star, \mathbf{X}_i$ are mutually independent.*

Assumption 2. *The functions $\{g_t^\star, \phi_i^\star\}$ are assumed to be (β, C)-Hölder smooth, that is, each function is $\lfloor\beta\rfloor$ times continuously differentiable, and its $\lfloor\beta\rfloor$th derivative is Hölder continuous with exponent $\beta - \lfloor\beta\rfloor$ and constant C. Here, $\lfloor\beta\rfloor$ denotes the greatest integer less than β.*

Assumption 3 (Pervasiveness)**.** *There exist a matrix $M^\star \in \mathbb{R}^{K\times N}$ and a function $\rho^\star$, defined on the image of the mapping $M^\star\phi^\star$, denoted by $M^\star\phi^\star([-B,B]^K)$. Assume that $\rho^\star$ is also (β, C)-Hölder smooth, and $\|M^\star\|_\infty \lesssim n^{-1}$, $\|M^\star\|_0 \asymp n$ for some diverging integer $n > 0$. We further assume that*

$$\sup_{x\in[-B,B]^K} \|\rho^\star(M^\star\phi^\star(x)) - x\|^2 + \|M^\star\|^2 \lesssim n^{-1}. \tag{8}$$

Assumption 4 (Sequential exogeneity/no anticipation)**.** *Let $\mathcal{F}_t = \sigma(\mathbf{X}, \{\mathbf{F}_s^\star, \varepsilon_{\cdot s}\}_{s\le t})$. Assume (i) W_{it} is $\mathcal{F}_{t-1}$-measurable (adoption time is a stopping time); (ii) $\varepsilon_{\cdot t} \perp \mathcal{F}_{t-1}$ and $\mathbb{E}\varepsilon_{it} = 0$.*

Assumption 1 is a standard assumption in the literature of nonparametric regression. Note that assuming $\mathbf{X}_i$ has compact support is equivalent to assuming that the support is $[0,1]^P$ via centering and scaling. The smoothness condition is standard in the nonparametric regression literature (cf. [64], Chapter 1), as it governs the complexity of the underlying function class. A formal definition is provided in Appendix E. Assumption 3 effectively enables approximate recovery of the latent factors from the observed input: when x denotes the latent factors, condition (8) captures the tradeoff between approximation accuracy and model complexity, with n reflecting the number of encoder parameters. By construction, it is monotonically non-increasing in n, measuring how gains in encoder expressiveness translate into improved factor recovery. The n^{-1} rate could be generalized but is chosen for analytical simplicity. The constraint $\|M^\star\|_0 \asymp n$ regulates the necessary model complexity. To illustrate, consider the linear factor model in [13], where $\phi^\star(x) = Ax$ with $\|A\|_\infty \lesssim 1$ and $\|A\|_0 \asymp N$ indicating that each factor affects nearly all observed variables. Assumption 3 is satisfied by setting $M^\star = N^{-1}A^\top$ and $\rho^\star(x) = N(A^\top A)^{-1}x$ with $n = N$. For a nonlinear illustration, consider $\phi^\star(x) = \exp(Ax)$. Taking $M^\star = n^{-1}A^\top$, $M^\star\exp(Ax)$ has a positive definite Jacobian, ensuring local invertibility by the inverse function theorem. Hence there exists a smooth inverse $\rho^\star(M^\star\phi^\star(x)) = x$, yielding the same n^{-1} approximation error bound. Assumption 4 allows adoption to depend on covariates $\mathbf{X}_i$, the full latent factor path $\{\mathbf{F}_s^\star\}_{s<t}$ (already absorbed by the model), and lagged outcomes and shocks, covering design-based rollouts and policies responding to realized history. Excluded is dependence on contemporaneous or future shocks.

The following theorem establishes the convergence of the estimated counterfactuals under a four-block design. The result can be readily extended to general SADs.

Theorem 1. *Suppose Assumptions 1-3 hold, and let $K_1 \ge K$, with K_1 defined in* (5)*. Assume that $N_1 \asymp N$ and $T_1 \asymp T_2$. Then, under a suitable choice of parameters for the DNN function class (specified in Appendix E), we have*

$$\max_{N_1+1\le i\le N} \frac{1}{T_2}\sum\nolimits_{t=T_1+1}^{T} \mathbb{E}\left(\widehat{Y}_{it}(0) - \phi_i^\star(\mathbf{F}_t^\star) - g_t^\star(\mathbf{X}_i)\right)^2 = O_{\mathbb{P}}(\gamma),$$

$$\gamma = \left(\frac{(N+n)T^{-\frac{2\beta}{2\beta+K}}}{n} + \frac{n}{T} + \frac{nK_1+N}{n^2}\right)\log^4(NT) + N^{-\frac{2\beta}{2\beta+k}}\log^5 N.$$

Theorem 1 establishes the statistical consistency of the counterfactual estimator and guarantees the algorithm's approximation quality. Writing $r = \frac{2\beta}{2\beta+K}$, the four terms of γ are: (i) $\frac{N+n}{n}T^{-r}$, the Stone-optimal rate for regressing on the K-dimensional latent space with T time points (learning the N decoders), inflated by $\frac{N+n}{n}$ because factors are recovered through an n-parameter encoder; (ii) n/T, encoder variance (n effective parameters, T columns); (iii) $K_1/n + N/n^2$, factor-recovery bias from the n^{-1} approximation in Assumption 3 (equation (8)); (iv) the Stone rate $N^{-2\beta/(2\beta+P)}$

for $\widehat{g}_t$ on the P-dimensional covariate interactions with N cross-sectional observations. Terms (i),(iii) decrease and (ii) increases in n: a bias–variance tradeoff in encoder capacity. The curse of dimensionality enters only through the fixed, small K and P (never through N, T), which is precisely the benefit of the factor structure.

Remark 2 (Linear Pervasive Factors)**.** *Let $\phi^\star(x) = Ax$ with $\|A\|_\infty \lesssim 1$ and subset pervasiveness $\lambda_{\min}(A_S^\top A_S) \asymp |S|$ (e.g., i.i.d. loadings). Then Assumption 3 holds for* every *diverging $n \le N$ via $M^\star = n^{-1}A_S^\top$, $\rho^\star(z) = n(A_S^\top A_S)^{-1}z$, $|S| = n$ (exact inversion; $\|M^\star\|^2 \lesssim n^{-1}$). Optimizing $n \asymp (NT)^{1/3}$ gives $\gamma \asymp (N^{1/3}T^{-2/3} + N^{2/3}T^{-4/3})\log^4(NT) + N^{-\frac{2\beta}{2\beta+P}}\log^5 N$, which is consistent iff $N = o(T^2)$ (up to logs). For $N \asymp T$, we have $\gamma \asymp T^{-1/3}\log^4 T + T^{-\frac{2\beta}{2\beta+P}}\log^5 T$.*

Writing $\mathcal{D}_i = \{t : W_{it} = 1\}$, the unit-specific ATT estimation error satisfies

$$\widehat{\tau}_i - \tau_i^\star = \tfrac{1}{T_2}\textstyle\sum_{t\in\mathcal{D}_i}\varepsilon_{it} + \tfrac{1}{T_2}\sum_{t\in\mathcal{D}_i}\big[\phi_i^\star(\mathbf{F}_t^\star) + g_t^\star(\mathbf{X}_i) - \widehat{Y}_{it}(0)\big].$$

By Assumption 4, $\{W_{it}\varepsilon_{it}, \mathcal{F}_t\}$ is a martingale difference sequence, so $\mathbb{E}[\varepsilon_{it} \mid W_{it} = 1] = 0$ holds as a consequence. By Theorem 1 and the condition $\gamma = o(1)$, Azuma's inequality gives the uniform bound $\max_i |\widehat{\tau}_i - \tau_i^\star| = O_{\mathbb{P}}(\sqrt{\log N/T_2} + \sqrt{\gamma})$, which establishes consistency of the unit-specific ATT estimator. Proofs and more technical details are in Appendix E.

Remark 3 (Extension to Time-Varying Treatment Effects)**.** *In principle, CoDEAL can produce estimates of unit-time-specific treatment effect τ_{it} using $\widehat{\tau}_{it} = Y_{it}(1) - \widehat{Y}_{it}(0)$. The static-effect assumption enters only the interpretation of the averaged estimator $\widehat{\bar{\tau}}_i := T_2^{-1}\sum_{t:W_{it}=1}\widehat{\tau}_{it}$, not the imputation machinery. CoDEAL's counterfactual imputation $\widehat{Y}_{it}(0)$ is trained exclusively on untreated outcomes, so Theorem 1's guarantee on $\max_{N_1+1\le i\le N}\frac{1}{T_2}\sum_{t=T_1+1}^{T}\mathbb{E}(\widehat{Y}_{it}(0) - \phi_i^\star(\mathbf{F}_t^\star) - g_t^\star(\mathbf{X}_i))^2$ is entirely unaffected by whether the treatment effect is static or dynamic. When the treatment effect is static ($\tau_{it} \equiv \tau_i$), this result establishes the estimation consistency of $\widehat{\bar{\tau}}_i$ for τ_i. When the treatment effect τ_{it} varies over time, Theorem 1 implies that $\widehat{\bar{\tau}}_i$ consistently estimates the unit-wise average treatment effect $\bar{\tau}_i = T_2^{-1}\sum_{t:W_{it}=1}\tau_{it}$. Thus, CoDEAL permits the contemporaneous treatment effect to vary across time and units, while preserving consistency for the average effect. A pointwise guarantee of $\widehat{\tau}_{it}$ for τ_{it}, however, is not attainable by the current theory.*

3 Simulation Studies

We use simulation to demonstrate the finite-sample performance of CoDEAL[2]. Complete simulation studies are provided in Appendix B. We consider various benchmark methods, including a single-output AE (**Single AE**) (details in Appendix B.1), neuron-enhanced AE (**AEMC-NE**) [29], matrix completion with nuclear norm minimization (**MC-NNM**) [12], vertical regression (**Vert-Reg**) [27], difference-in-differences (**DiD**) [38], and kernel principal component analysis (**KernelPCA**) [57]. We also compare to three synthetic-control-style baselines, including the classic synthetic control [3], generalized synthetic control [70], and synthetic difference-in-differences [8], with results reported in the Appendix. We evaluate using mean squared error (MSE) and mean absolute error (MAE).

Data Generating Process and Simulation Configurations (with details in Appendix B.2). We carry out extensive simulations under various linear and nonlinear generating mechanisms (4 variants for generating factor effects and 6 variants for generating covariate effects), with different setups of unit numbers N, time periods T, latent factor dimension K, and treatment assignments. We consider configurations 1-4: **Four-Block Design**: (1) $(N, T, N_1, T_1, P, K) = (100, 200, 50, 100, 3, 4)$, (2) $(N, T, N_1, T_1, P, K) = (200, 120, 100, 60, 5, 3)$, **SAD**: (3) $(N, T, K) = (100, 200, 4)$, $r = \{5, 10\}$, and (4) $(N, T, K) = (200, 120, 3)$, $r = \{5, 10\}$. Hyperparameter settings and sensitivity analyses are presented in Appendix B.8. CoDEAL is robust to the learning rate and network depth over a wide range; the bottleneck size K_1 is the one choice that matters materially, and under-specifying it below the true factor dimension is far more damaging than over-specifying it. The hyperparameters may also be selected empirically via a masked-block validation strategy, which we detail in Remark S.1.

CoDEAL in the removal of covariate effects (with details in Appendix B.3). We compare the performance of CoDEAL using different methods, specifically, DNN and linear regression (LR), to remove the covariate effects, against an oracle benchmark when no covariate effects are present.

[2]The code for implementation is available at `https://github.com/chenezandie/CoDEAL-Code`.

Table 1: Comparisons of prediction accuracy (measured by MAE, averaged over 50 replications) in Config.1–4. Results of MSE are reported in Table S.6 of Appendix B. Numbers in parentheses are standard errors.

Factor	Method	Config. 1		Config. 2		Config. 3		Config. 4	
		Linear	Nonlinear	Linear	Nonlinear	$r = 5$	$r = 10$	$r = 5$	$r = 10$
Linear	CoDEAL	**0.646(0.009)**	**0.701(0.008)**	**0.609(0.004)**	**0.644(0.005)**	0.614(0.007)	0.625(0.005)	**0.532(0.003)**	**0.560(0.004)**
	Single AE	0.821(0.010)	0.851(0.008)	0.777(0.007)	0.795(0.007)	0.789(0.011)	0.795(0.008)	0.693(0.005)	0.714(0.006)
	AEMC-NE	0.838(0.009)	0.862(0.009)	0.759(0.007)	0.771(0.007)	0.731(0.010)	0.730(0.007)	0.647(0.005)	0.660(0.005)
	DiD	0.892(0.010)	0.916(0.009)	0.810(0.007)	0.828(0.008)	0.801(0.011)	0.802(0.008)	0.711(0.005)	0.726(0.006)
	Vert-Reg	0.757(0.006)	0.798(0.007)	0.761(0.004)	0.770(0.005)	0.940(0.005)	0.865(0.023)	0.691(0.003)	0.750(0.016)
	MC-NNM	0.699(0.006)	0.735(0.005)	0.650(0.005)	0.697(0.006)	**0.599(0.007)**	**0.616(0.006)**	0.537(0.003)	0.565(0.004)
	KernelPCA	0.811(0.007)	0.824(0.008)	0.835(0.009)	0.849(0.009)	0.798(0.007)	0.801(0.007)	0.707(0.005)	0.709(0.005)
Sine	CoDEAL	**1.078(0.012)**	**1.117(0.011)**	**0.894(0.007)**	**0.945(0.008)**	**0.973(0.010)**	**0.996(0.010)**	**0.850(0.007)**	**0.869(0.008)**
	Single AE	1.223(0.013)	1.250(0.012)	1.142(0.009)	1.185(0.009)	1.087(0.011)	1.122(0.011)	1.057(0.008)	1.061(0.008)
	AEMC-NE	1.209(0.013)	1.235(0.012)	1.128(0.009)	1.167(0.010)	1.055(0.012)	1.078(0.011)	1.011(0.008)	1.008(0.009)
	DiD	1.244(0.013)	1.272(0.012)	1.191(0.009)	1.234(0.009)	1.110(0.012)	1.146(0.012)	1.091(0.008)	1.088(0.009)
	Vert-Reg	1.169(0.012)	1.208(0.011)	1.012(0.007)	1.054(0.008)	1.549(0.090)	1.524(0.123)	0.974(0.009)	1.047(0.013)
	MC-NNM	1.201(0.013)	1.231(0.012)	1.105(0.009)	1.153(0.010)	1.066(0.011)	1.097(0.011)	0.998(0.008)	1.003(0.008)
	KernelPCA	1.195(0.009)	1.231(0.009)	1.170(0.013)	1.190(0.013)	1.109(0.012)	1.115(0.012)	1.075(0.009)	1.079(0.009)
Polynomial	CoDEAL	**0.908(0.013)**	**0.952(0.015)**	**0.772(0.010)**	**0.807(0.011)**	**0.785(0.011)**	**0.811(0.010)**	**0.662(0.007)**	**0.699(0.007)**
	Single AE	0.962(0.013)	0.997(0.016)	0.820(0.010)	0.854(0.011)	0.895(0.015)	0.926(0.016)	0.739(0.011)	0.785(0.010)
	AEMC-NE	0.997(0.014)	1.038(0.017)	0.856(0.011)	0.887(0.012)	0.860(0.014)	0.897(0.013)	0.720(0.009)	0.764(0.009)
	DiD	0.995(0.013)	1.025(0.016)	0.854(0.010)	0.886(0.011)	0.887(0.013)	0.917(0.014)	0.739(0.009)	0.773(0.008)
	Vert-Reg	0.953(0.013)	1.019(0.016)	0.855(0.007)	0.884(0.010)	1.155(0.035)	1.131(0.032)	0.788(0.005)	0.881(0.014)
	MC-NNM	0.951(0.013)	0.989(0.015)	0.819(0.009)	0.851(0.011)	0.839(0.012)	0.865(0.012)	0.703(0.008)	0.735(0.007)
	KernelPCA	0.927(0.013)	0.991(0.013)	1.029(0.015)	1.051(0.015)	0.935(0.011)	0.944(0.011)	0.776(0.008)	0.782(0.008)
ReLU MLP	CoDEAL	**0.797(0.009)**	**0.838(0.010)**	**0.698(0.005)**	**0.743(0.006)**	**0.632(0.004)**	**0.733(0.004)**	**0.633(0.004)**	**0.646(0.004)**
	Single AE	0.989(0.009)	1.026(0.011)	0.891(0.007)	0.930(0.008)	0.876(0.005)	0.913(0.006)	0.797(0.005)	0.809(0.005)
	AEMC-NE	1.045(0.010)	1.088(0.011)	0.959(0.007)	0.982(0.008)	0.934(0.006)	0.968(0.007)	0.860(0.006)	0.878(0.006)
	DiD	0.930(0.008)	0.963(0.010)	0.842(0.005)	0.870(0.007)	0.817(0.004)	0.842(0.004)	0.742(0.004)	0.748(0.005)
	Vert-Reg	0.888(0.010)	0.958(0.013)	0.847(0.005)	0.894(0.006)	1.045(0.031)	0.948(0.026)	0.755(0.003)	1.070(0.253)
	MC-NNM	0.843(0.007)	0.892(0.009)	0.767(0.005)	0.806(0.006)	0.721(0.003)	0.750(0.003)	0.656(0.004)	0.662(0.004)
	KernelPCA	1.105(0.008)	1.126(0.009)	1.156(0.011)	1.172(0.011)	1.111(0.008)	1.119(0.008)	1.005(0.006)	1.009(0.006)

Six scenarios of covariate effects were considered. The results show that LR is adequate for linear covariate effects but the DNN approach is essential in the presence of nonlinear covariate effects.

CoDEAL in causal matrix completion (with details in Appendix B.4). As shown in Table 1, CoDEAL consistently yields the lowest MAE across nearly all settings, and achieves substantial improvements relative to all competing methods. In SAD with linear factors, CoDEAL **ties MC-NNM within 3% while reducing MAE by 35–89% over other benchmarks**. In SAD with nonlinear factors, CoDEAL delivers evident advantages. Especially when under ReLU-based factors, CoDEAL delivers its largest margins, with **MAE reductions of 15–60% versus all competing approaches**. As r increases from 5 to 10, the missingness pattern grows more complex and extensive, so all methods show some drop in accuracy. CoDEAL's errors only suffer a slight increase, whereas simpler methods like Vert-Reg show a dramatic rise in MAE when missingness increases.

In summary, CoDEAL achieves performance comparable to MC-NNM in linear settings, while exhibiting substantial improvements over all considered methods in the presence of nonlinear effects. By integrating DNN-based covariate adjustments with deep multi-output-AE–based nonlinear factor analysis, CoDEAL is robustly effective across a wide range of linear and nonlinear scenarios.

4 Real Data Applications

We illustrate our methodology using data from the Oxford COVID-19 Government Response Tracker (OxCGRT) [30]. OxCGRT collects panel data on the timing and intensity of 24 different government interventions, such as school closures, travel restrictions, and vaccination mandates, across over 185 countries from January 2020 through 2022. We report two additional case studies in Appendix C.2, using the US Election Day Registration dataset [70] and the German reunification dataset [49].

We examine the impact of two specific policy interventions: (1) the **implementation of mandatory vaccination**, and (2) the **termination of internal travel restrictions** (results in Appendix C.1). Our analysis covers their impacts on COVID-19 confirmed cases and deaths across 64 states and territories in the United States and Canada from April 16, 2021, to December 11, 2021. This time frame aligns with the period following the detection of the Delta variant and precedes the rapid emergence of the Omicron variant in North America, after vaccines became widely available; restricting to this window limits major regime shifts in epidemic dynamics during the analysis. We use four OxCGRT indices (overall government response, containment and health, stringency, and economic support), averaged over the time frame, as covariates, resulting in data dimensions of $N = 64$, $T = 240$, and $P = 4$. The policy is adopted in different states at various times, forming a SAD (as visualized in Figure 3). Figure 3 plots the comparisons of the total confirmed cases and deaths across policy-in-effect states between observed results (with policy executed) and estimated counterfactuals (if no policy). Results

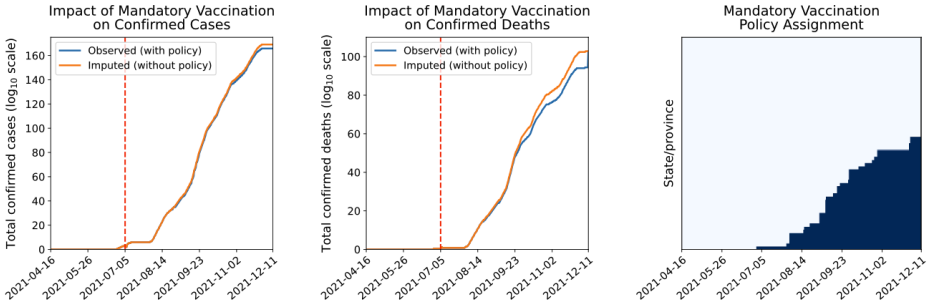

Figure 3: Left and Middle: Comparison of the total confirmed cases (Left) and deaths (Middle) across policy-in-effect states between observed results with policy executed and estimated counterfactuals if no policy. Values are reported on a $\log_{10}$ scale, and plots are visualized with 14-day moving averages. Right: Indicator matrix of observed policy assignment, with darker color referring to policy in effect. Enlarged figures are in Appendix C.1.

of other benchmarks besides CoDEAL are in the appendix. These results represent estimated counterfactual trajectories rather than directly verifiable causal effects, which cannot be directly verified due to the unobservability of counterfactual outcomes in real-world settings. The estimated trajectories allow for time-varying contemporaneous policy effects, while the estimated treatment effects are interpreted as averages over the post-adoption period, as discussed in Remark 3. As shown in the plot, if the mandatory vaccination policy were not implemented, there would be an increase in both confirmed cases and deaths, with a notably larger impact on mortality. This observation aligns with recent research highlighting a greater reduction in mortality than in infection rates following vaccination [36, 45, 67, 69]. Although the visual differences in the log-transformed graphs appear modest, converting these back to their original scales reveals substantial impacts: on December 11, 2021, the actual cumulative counts of confirmed cases and deaths for states with mandatory vaccination policies were $26.57\ million, 0.39\ million$, respectively. In contrast, our estimates suggest that without these policies, the counts would have risen to $29.81\ million(\uparrow 12.2\%), 0.52\ million(\uparrow 34.6\%)$.

5 Discussion of Limitations and Extensions

We propose CoDEAL, a deep-learning-based causal matrix completion method and a unifying framework for estimating heterogeneous causal effects in panel data models. The proposed framework largely extends existing causal panel literature in the capacity of accommodating heterogeneity and nonlinearity. However, the extent to which these features are present in empirical applications is generally unknown. To guide practical implementation, Remark S.1 discusses a masked-block reconstruction strategy that enables data-driven comparison of model specifications and hyperparameter selection using observed untreated outcomes, given that the true counterfactual outcomes are inherently unobservable. We also clarify that several assumptions (e.g., time-invariant covariates, static treatment effects) are not intrinsic to the algorithm and can be relaxed to accommodate more general settings. Moreover, despite the additive model formulation, the proposed framework does not preclude interactions between covariates and latent factors, nor does it require them to be uncorrelated. The CoDEAL framework is amenable to further adaptation across data-driven contexts. Detailed discussions of these extensions are provided in Appendix D.

CoDEAL fundamentally differs from standard MAR-assumed matrix completion methods. Counterfactual missingness is systematically tied to treatment assignment. CoDEAL addresses the missing-not-at-random scenario explicitly through (1) a dedicated block-partition-and-learning algorithm that systematically leverages before-/after- treatment observations, and (2) a nonlinear factor model that captures the inherent dependence for imputation. The decomposition into $\phi_i^\star()$ and $\mathbf{F}_t^\star$ allows us to account for the underlying cross-sectional and temporal dependencies driving the outcome.

Methodologically, with different choices of factor models and advanced neural network architectures, the proposed CoDEAL framework can be further generalized to accommodate various data types with more complex data structures, such as tensors, images, and networks. For example, CoDEAL can be naturally extended to three-dimensional or higher-order tensors by involving tensor factorization methodologies [17, 18, 31, 76], providing a more general approach compared to matrix factor analysis [19, 44, 73, 74]. That being said, to extend to multiple treatments, how to handle the increased level of missingness brought by the treatment assignments in a higher-dimensional tensor remains challenging and requires careful modeling strategies. We leave the multiple treatment settings as future work.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for helpful suggestions. The work of Yuefeng Han is supported in part by National Science Foundation Grants DMS-2412578. The work of Guanhao Zhou and Xiufan Yu is supported in part by National Institutes of Health grant R01GM152812.

References

[1] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

[2] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003.

[3] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

[4] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

[5] Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3821–3826. PMLR, 2023.

[6] Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions: Extending synthetic controls to multiple treatments. *Operations Research*, 74(2):840–859, 2026.

[7] Dmitry Arkhangelsky and Guido Imbens. Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 27(3):C1–C61, 2024.

[8] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

[9] Susan Athey and Guido Imbens. Identification of average treatment effects in nonparametric panel models. *arXiv preprint arXiv:2503.19873*, 2025.

[10] Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

[11] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51, 2019.

[12] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

[13] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

[14] Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

[15] Jushan Bai and Serena Ng. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021.

[16] Sohom Bhattacharya, Jianqing Fan, and Debarghya Mukherjee. Deep neural networks for nonparametric interaction models with diverging dimension. *The Annals of Statistics*, 52(6): 2738–2766, 2024.

[17] Bin Chen, Yuefeng Han, and Qiyang Yu. Diffusion index forecasting with tensor data. *Journal of Econometrics*, 254:106204, 2026.

[18] Bin Chen, Yuefeng Han, and Qiyang Yu. Estimation and inference for CP tensor factor models. *Journal of Econometrics*, 253:106167, 2026.

[19] Elynn Chen, Yuefeng Han, Jiayu Li, and Ke Xu. Modewise additive factor model for matrix time series. *arXiv preprint arXiv:2512.25025*, 2025.

[20] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[21] Matteo Chinazzi, Jessica T. Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, Cécile Viboud, Xinyue Xiong, Hongjie Yu, M. Elizabeth Halloran, Ira M. Longini, Jr., and Alessandro Vespignani. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489):395–400, 2020.

[22] Jungjun Choi and Ming Yuan. Matrix completion when missing is not at random and its applications in causal panel data models. *Journal of the American Statistical Association*, pages 1–15, 2024.

[23] Damian Clarke, Daniel Pailañir, Susan Athey, and Guido Imbens. On synthetic difference-in-differences and related estimation methods in Stata. *The Stata Journal*, 24(4):557–598, 2024.

[24] Johan de Aguas, Leonard Henckel, Johan Pensar, and Guido Biele. Causal inference amid missingness-specific independences and mechanism shifts. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 31–44. PMLR, 2025.

[25] Johan de Aguas, Johan Pensar, Tomás Varnet Pérez, and Guido Biele. Recovery and inference of causal effects with sequential adjustment for confounding and attrition. *Journal of Causal Inference*, 13(1):20240009, 2025.

[26] Peng Ding and Fan Li. Causal inference: A missing data perspective. *Statistical Science*, 33(2): 214–237, 2018.

[27] Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research, 2016.

[28] Jicong Fan and Jieyu Cheng. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34–41, 2018.

[29] Jicong Fan, Rui Chen, Zhao Zhang, and Chris Ding. Neuron-enhanced AutoEncoder matrix completion and collaborative filtering: Theory and practice. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 5(4):529–538, 2021.

[31] Yuefeng Han and Cun-Hui Zhang. Tensor principal component analysis in high dimensional CP models. *IEEE Transactions on Information Theory*, 69(2):1147–1167, 2023.

[32] Yuefeng Han, Rong Chen, and Cun-Hui Zhang. Rank determination in tensor factor model. *Electronic Journal of Statistics*, 16(1):1726–1803, 2022.

[33] Yuefeng Han, Rong Chen, Dan Yang, and Cun-Hui Zhang. Tensor factor model estimation by iterative projection. *The Annals of Statistics*, 52(6):2641–2667, 2024.

[34] Yuefeng Han, Dan Yang, Cun-Hui Zhang, and Rong Chen. CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1383–1413, 2024.

[35] Jouni Helske and Santtu Tikka. Estimating causal effects from panel data with dynamic multivariate panel models. *Advances in Life Course Research*, 60:100617, 2024.

[36] Porfirio Felipe Hernández Bautista, Concepción Grajales Muñiz, David Alejandro Cabrera Gaytán, Teresita Rojas Mendoza, Alfonso Vallejos Parás, Clara Esperanza Santacruz Tinoco, Julio Elias Alvarado Yaah, Yu Mei Anguiano Hernández, Nancy Sandoval Gutiérrez, and Leticia Jaimes Betancourt. Impact of vaccination on infection or death from COVID-19 in individuals with laboratory-confirmed cases: Case-control study. *PLOS ONE*, 18(8):e0265698, 2023.

[37] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[38] Kosuke Imai and In Song Kim. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415, 2021.

[39] Guido W Imbens. Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11:123–152, 2024.

[40] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[41] Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. In Eric Laber, Bibhas Chakraborty, Erica E. M. Moodie, Tianxi Cai, and Mark van der Laan, editors, *Handbook of Statistical Methods for Precision Medicine*, chapter 10, pages 207–236. Chapman and Hall/CRC, 2024.

[42] Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

[43] Hui Lan, Haoge Chang, Eleanor Wiske Dillon, and Vasilis Syrgkanis. A meta-learner for heterogeneous effects in difference-in-differences. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 32420–32451. PMLR, 2025.

[44] Wei Luo, Lingzhou Xue, Jiawei Yao, and Xiufan Yu. Inverse moment methods for sufficient forecasting using high-dimensional predictors. *Biometrika*, 109(2):473–487, 2022.

[45] Margaux M I Meslé, Jeremy Brown, Piers Mook, Mark A Katz, José Hagan, Roberta Pastore, Bernhard Benka, Monika Redlberger-Fritz, Nathalie Bossuyt, Veerle Stouten, Catharina Vernemmen, Elisabet Constantinou, Marek Maly, Jan Kynčl, Ondrej Sanca, Tyra Grove Krause, Lasse Skafte Vestergaard, Tuija Leino, Eero Poukka, Kassiani Gkolfinopoulou, Kassiani Mellou, Maria Tsintziloni, Zsuzsanna Molnár, Gudrun Aspelund, Marianna Thordardottir, Lisa Domegan, Eva Kelly, Joan O'Donell, Alberto-Mateo Urdiales, Flavia Riccardo, Chiara Sacco, Viktoras Bumšteinas, Rasa Liausediene, Joël Mossong, Anne Vergison, Maria-Louise Borg, Tanya Melillo, Dragan Kocinski, Enkela Pollozhani, Hinta Meijerink, Diana Costa, João Paulo Gomes, Pedro Pinto Leite, Alina Druc, Veaceslav Gutu, Valentin Mita, Mihaela Lazar, Rodica Popescu, Odette Popovici, Monika Musilová, Maja Mrzel, Maja Socan, Veronika Učakar, Aurora Limia, Clara Mazagatos, Carmen Olmedo, Gavin Dabrera, Meaghan Kall, Mary Sinnathamby, Graham McGowan, Jim McMenamin, Kirsty Morrison, Dorit Nitzan, Marc-Alain Widdowson, Catherine Smallwood, and Richard Pebody. Estimated number of lives directly saved by COVID-19 vaccination programmes in the WHO European Region from December, 2020, to March, 2023: a retrospective surveillance study. *The Lancet Respiratory Medicine*, 12 (9):714–727, 2024. On behalf of the WHO European Respiratory Surveillance Network.

[46] Daniel L. Millimet and Marc F. Bellemare. On the (Mis) Use of the fixed effects estimator. *Oxford Bulletin of Economics and Statistics*, 88(3):484–505, 2026.

[47] Yayoi Murano, Ryo Ueno, Shoi Shi, Takayuki Kawashima, Yuta Tanoue, Shiori Tanaka, Shuhei Nomura, Hiromichi Shoji, Toshiaki Shimizu, Huy Nguyen, Hiroaki Miyata, Stuart Gilmour, and Daisuke Yoneoka. Impact of domestic travel restrictions on transmission of COVID-19 infection using public transportation network approach. *Scientific Reports*, 11(1):3109, 2021.

[48] Preetam Nandy, Xiufan Yu, Wanjun Liu, Ye Tu, Kinjal Basu, and Shaunak Chatterjee. Generalized causal tree for uplift modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 788–798. IEEE, 2023.

[49] Xun Pang, Licheng Liu, and Yiqing Xu. A Bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288, 2022.

[50] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[51] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009.

[52] Adityanarayanan Radhakrishnan, George Stefanakis, Mikhail Belkin, and Caroline Uhler. Simple, fast, and flexible framework for matrix completion with infinite width neural networks. *Proceedings of the National Academy of Sciences*, 119(16):e2115064119, 2022.

[53] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[54] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[55] Pedro H. C. Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

[56] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

[57] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural Networks — ICANN '97*, volume 1327 of *Lecture Notes in Computer Science*, pages 583–588. Springer, 1997.

[58] Vira Semenova, Matt Goldman, Victor Chernozhukov, and Matt Taddy. Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510, 2023.

[59] Azeem M Shaikh and Panos Toulis. Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848, 2021.

[60] Zhouyu Shen and Dacheng Xiu. Deep autoencoders for nonlinear factor models: Theory and applications. Research Paper 25-14, University of Chicago Booth School of Business, 2024. Manuscript dated December 28, 2024, revised June 2026.

[61] Chandler Squires, Dennis Shen, Anish Agarwal, Devavrat Shah, and Caroline Uhler. Causal imputation via synthetic interventions. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 688–711. PMLR, 2022.

[62] Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

[63] Lorenzo Testa, Edward H Kennedy, and Matthew Reimherr. Efficient difference-in-differences estimation when outcomes are missing at random. *arXiv preprint arXiv:2509.25009*, 2025.

[64] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

[65] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[66] Coady Wing, Madeline Yozwiak, Alex Hollingsworth, Seth Freedman, and Kosali Simon. Designing difference-in-difference studies with staggered treatment adoption: Key concepts and practical guidelines. *Annual Review of Public Health*, 45:485–505, 2024.

[67] Nana Wu, Keven Joyal-Desmarais, Paula A. B. Ribeiro, Ariany Marques Vieira, Jovana Stojanovic, Comfort Sanuade, Doro Yip, and Simon L Bacon. Long-term effectiveness of COVID-19 vaccines against infections, hospitalisations, and mortality in adults: findings from a rapid living systematic evidence synthesis and meta-analysis up to December, 2022. *The Lancet Respiratory Medicine*, 11(5):439–452, 2023.

[68] Ruoxuan Xiong and Markus Pelger. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301, 2023.

[69] Kun Xu, Zihan Wang, Maorong Qin, Yangyu Gao, Na Luo, Wanting Xie, Yihan Zou, Jie Wang, and Xingming Ma. A systematic review and meta-analysis of the effectiveness and safety of COVID-19 vaccination in older adults. *Frontiers in Immunology*, 14:1113156, 2023.

[70] Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

[71] Ilker Yalcin and Yasuo Amemiya. Nonlinear factor analysis as a statistical method. *Statistical Science*, 16(3):275–294, 2001.

[72] Yuling Yan and Martin J Wainwright. Entrywise inference for missing panel data: A simple and instance-optimal approach. *arXiv preprint arXiv:2401.13665*, 2024.

[73] Ruofan Yu, Rong Chen, Han Xiao, and Yuefeng Han. Dynamic matrix factor models for high dimensional time series. *arXiv preprint arXiv:2407.05624*, 2024.

[74] Xiufan Yu, Jiawei Yao, and Lingzhou Xue. Nonparametric estimation and conformal inference of the sufficient forecasting with a diverging number of factors. *Journal of Business & Economic Statistics*, 40(1):342–354, 2022.

[75] Xiufan Yu, Jiawei Yao, and Lingzhou Xue. Power enhancement for testing multi-factor asset pricing models via Fisher's method. *Journal of Econometrics*, 239(2):105458, 2024.

[76] Guanhao Zhou, Yuefeng Han, and Xiufan Yu. Factor augmented tensor-on-tensor neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22928–22936, 2025.

Appendices

The Appendices of this paper contain five parts. Appendix A includes additional technical notes for clearer exposition on the methodological framework, including an introduction to staggered adoption design along with detailed illustrative examples, an illustrative DAG of the proposed model, as well as discussions on the rationale of adopting a sequential estimation regime. Appendix B presents a complete report, along with full details, of the simulation studies. Appendix C provides additional results of the real data analysis using the Oxford COVID-19 Government Response Tracker (OxCGRT), and two additional empirical studies. Appendix D provides further discussions on broader impacts and directions of extensions. Appendix E provides additional theoretical properties and complete proofs of the theoretical guarantees of the convergence of CoDEAL.

A Technical Notes on Methodological Framework

A.1 Staggered Adoption Design

In this section, we introduce the concept of staggered adoption design [10], which is a commonly encountered framework in causal panel data analysis [12, 22, 59, 66, 72].

A.1.1 Four-Block Design and Staggered Adoption Design

Throughout this section, we assume the treatment adoption is irreversible. That is, once the treatment is initiated for a unit, it remains exposed to the treatment afterward.

We begin with the simplest design of treatment assignment in causal panel data analysis, when there is one control group and one treatment group. All units in the control group are never exposed to the treatment, and all units in the treatment group begin receiving the treatment at the same time. This design is considered the building block of more complex designs in causal panel models, often called the Four-Block design or Differences-in-Differences design in the literature. A graphical representation of the four-block design is plotted in the left panel of Figure S.1.

Staggered adoption design accommodates more complex treatment assignments by allowing different units to get exposed to treatment at varying points in time. The four-block design can be regarded as a special case of 2×2 staggered adoption. A graphical representation of the general staggered design is plotted in the right panel of Figure S.1. Without loss of generality, we can arrange the panel units by sorting the units based on the time at which they began receiving the treatment in a descending or ascending order, forming a block-wise missing pattern in the matrices of potential outcomes; see Section A.1.2 for an illustrative example.

Figure S.1 visually illustrates the distinction between the four-block and staggered adoption designs. Each row in these panels represents an individual observational unit, while each column indicates a distinct time period. Orange blocks indicate periods during which units have adopted treatment, thus resulting in missing counterfactual control outcomes. Gray blocks represent observable periods under control conditions. The left panel presents the four-block design, in which one group of units simultaneously transitions from untreated (gray cells) to treated status (orange cells) at a single shared time point. The right panel depicts the staggered adoption scenario, characterized by multiple adoption times and varying durations of observed control outcomes.

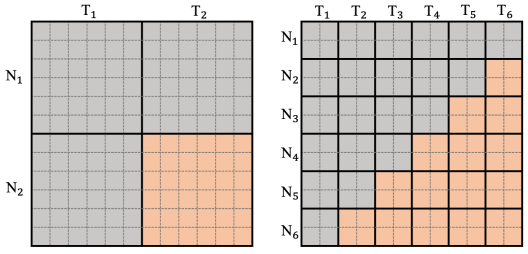

Figure S.1: Examples that illustrate the four-block design (Left) and staggered adoption design with $r = 6$ (Right). Rows stand for units and columns stand for time. Orange blocks stand for treated.

A.1.2 Illustrative Example of Staggered Adoption Designs

We use an explicit example to clearly illustrate the block-wise missing patterns of staggered adoption structure. In a staggered adoption design, various units receive irreversible treatment at different times. Without loss of generality, we arrange the panel by sorting the units based on the time at which they began receiving the treatment in a descending order (with never-treated units at the top, followed by the latest adopters, and then progressively earlier adopters towards the bottom).

We consider the following example with $N = 6$ units over $T = 5$ time periods. For treatment adoption designs, we suppose

- **Unit 1 (Never adopter):** never receives treatment during the study periods, so all control outcomes are observed throughout the study.
- **Unit 2 (Late adopter):** adopts treatment towards the end of the observation period, resulting in observed control outcomes at earlier time points.
- **Unit 3 and Unit 4 (Medium adopters):** adopt treatment midway through the study, leading to missing control outcomes for periods following the adoption.
- **Unit 5 and Unit 6 (Early adopters):** adopt treatment early in the observation period, causing most subsequent outcomes to become unobserved control outcomes.

The matrix of control outcome $\mathbf{Y}(0)$ is shown in Equation (S.1), where question marks denote the entries whose control outcomes are missing as this unit is under treatment at the time. The corresponding matrix of treatment indicators and its block-wise representation are presented in Equation (S.3), with 1 for treated and 0 otherwise.

As revealed by Equations (S.1) and (S.3), there is naturally a block structure in the treatment indicator matrix $\mathbf{\Omega}$, or equivalently speaking, a block-wise missing pattern in the matrix of control outcome $\mathbf{Y}(0)$. Though it is possible to impute the missing elements one by one, it is generally computationally expensive, especially if either T or N is large. For efficient imputation, as illustrated in Section 2.2.2, we partition the indicator matrix (as well as the control outcome matrix) into submatrices $\mathbf{\Omega}_{(\xi,\eta)} \in \mathbb{R}^{N_\xi \times T_\eta}$ and $\mathbf{Y}_{(\xi,\eta)} \in \mathbb{R}^{N_\xi \times T_\eta}$ for $\xi = 1, \ldots, r$ and $\eta = 1, \ldots, r$, where $\sum_{\xi=1}^{r} N_\xi = N$ and $\sum_{\eta=1}^{r} T_\eta = T$. The matrix imputation can subsequently be conducted using the block structured submatrices, following the procedures introduced in Algorithm 2 of Section 2.2.2. In this example, details of the partitions are provided in Equation (S.2), resulting in r taking the value of 4.

$$\mathbf{Y}(0) = \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} & \text{never adopter} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & ? & \text{late adopter} \\ Y_{31} & Y_{32} & ? & ? & ? & \text{medium adopter} \\ Y_{41} & Y_{42} & ? & ? & ? & \text{medium adopter} \\ Y_{51} & ? & ? & ? & ? & \text{early adopter} \\ Y_{61} & ? & ? & ? & ? & \text{early adopter} \end{bmatrix} \tag{S.1}$$

$$= \left[\begin{array}{c|c|ccc} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ \hline Y_{21} & Y_{22} & Y_{23} & Y_{24} & ? \\ \hline Y_{31} & Y_{32} & ? & ? & ? \\ Y_{41} & Y_{42} & ? & ? & ? \\ \hline Y_{51} & ? & ? & ? & ? \\ Y_{61} & ? & ? & ? & ? \end{array}\right] := \left[\begin{array}{c|c|cc} \mathbf{Y}_{(1,1)} & \mathbf{Y}_{(1,2)} & \mathbf{Y}_{(1,3)} & \mathbf{Y}_{(1,4)} \\ \hline \mathbf{Y}_{(2,1)} & \mathbf{Y}_{(2,2)} & \mathbf{Y}_{(2,3)} & \mathbf{?} \\ \hline \mathbf{Y}_{(3,1)} & \mathbf{Y}_{(3,2)} & \mathbf{?} & \mathbf{?} \\ \hline \mathbf{Y}_{(4,1)} & \mathbf{?} & \mathbf{?} & \mathbf{?} \end{array}\right] \tag{S.2}$$

$$\text{and } \mathbf{\Omega} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} = \left[\begin{array}{c|c|cc} 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ \hline 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right] = \left[\begin{array}{c|c|cc} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ \hline \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{array}\right] \tag{S.3}$$

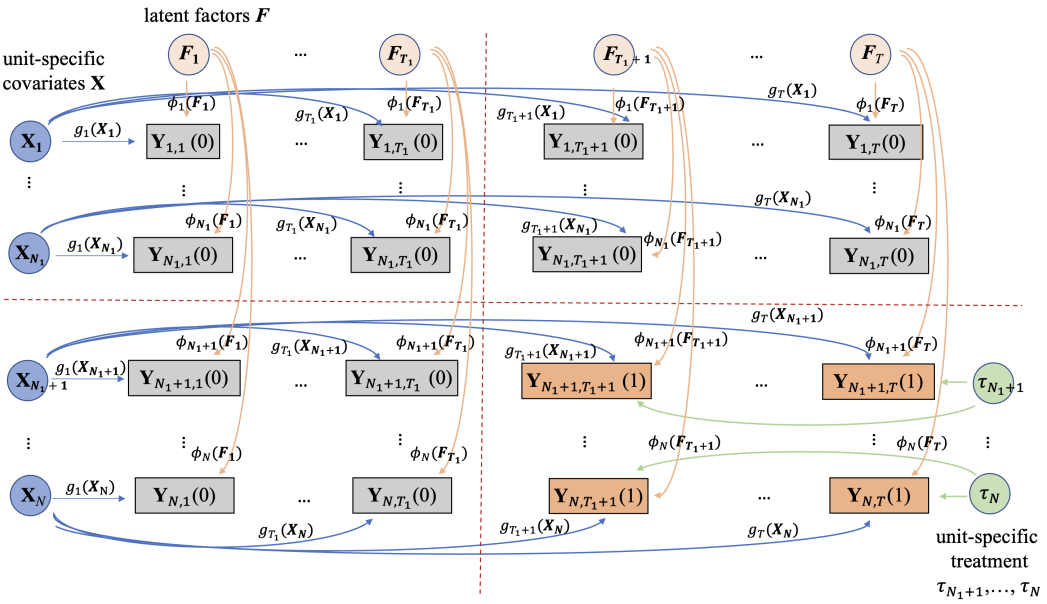

Figure S.2: Illustrative diagram of the CoDEAL model. Observed covariates $\mathbf{X}_i$ (left) affect outcomes through $g_t(\mathbf{X}_i)$, while latent time-varying factors $\mathbf{F}_t$ influence outcomes through $\phi_i(\mathbf{F}_t)$. The pre-treatment period (left of the vertical dashed line) contains only control outcomes, while the post-treatment period (right) includes treated units with unit-specific effects τ_i.

A.2 An Illustrative DAG of Model (1)

Figure S.2 plots an illustrative DAG of model (1):

$$Y_{it} = \phi_i^\star(\mathbf{F}_t^\star) + g_t^\star(\mathbf{X}_i) + \tau_i^\star \cdot \mathbb{1}\{W_{it} = 1\} + \varepsilon_{it}.$$

For clarity, the illustrative DAG emphasizes the key features of our model: (i) $\mathbf{F}_t^\star$ acts as a time-varying confounder, (ii) treatment has a contemporaneous effect on outcomes, (iii) time-varying covariate effects can be accommodated in the proposed framework, and (iv) temporal dependence operates through $\mathbf{F}_t^\star$ rather than through explicit lagged outcomes or treatments.

While an illustrative DAG offers an intuitive visualization, a fully rigorous causal DAG would be nontrivial. Despite the high-level mathematical equivalence between Rubin's potential outcome framework [40, 54] and Pearl's causal DAG framework [50, 51], mapping the causal panel data model onto a faithful graphical structure presents significant challenges. In our setting, temporal dependence is governed through time-varying latent factors (i.e., $\mathbf{F}_t^\star$), which act as unobserved confounders; representing these in a causal DAG requires introducing unobserved nodes with specific structural assumptions about their evolution and connections across time. Notably, in the relevant literature on causal panel models (e.g., [11, 15, 22], including the recent literature review by [7]), there has been no formal statement framing causal panel models within a causal graphical framework in rigorous terms. Establishing a complete and rigorous equivalence between the two formulations in the causal panel setting is beyond the scope of this paper.

A.3 Rationale of Sequential Estimation and Comparisons with Joint Estimation

For CoDEAL, we employ a two-step estimation strategy: first, we model covariate effects to obtain adjusted observed outcomes; second, we estimate counterfactuals using a multi-output autoencoder (AE). We adopt this sequential approach over joint estimation that estimates covariate and factor effects simultaneously for several strategic reasons.

First, sequential estimation improves stability and interpretability. Joint estimation risks entangling the effects of latent factor $\mathbf{F}_t^\star$ and observed covariates $\mathbf{X}_i^\star$. If optimized jointly, the two neural networks may compete to explain the same signal, making it difficult to disentangle observed covariate effects from latent structure, and may potentially cause identification issues. Sequential estimation enforces

a clean separation: covariate effects are estimated first from observed information, and only residual variation is attributed to the latent structure. This prevents the autoencoder (AE) from "absorbing" covariate effects in an uninterpretable way.

Moreover, sequential design is not merely a procedural convenience, but is a natural alignment with the data structure. $\mathbf{X}_i$ is observed and time-invariant, whereas $\mathbf{F}_t^\star$ is unobserved and time-varying. This asymmetry motivates a sequential ordering. It also aligns with the spirit of double/debiased machine learning, where nuisance functions are partialled out in a first stage so that the second stage operates on a less confounded signal.

To provide empirical grounding for this choice, we perform simulation experiments to evaluate the relative performance of sequential (SEQ) versus joint (JOINT) estimation strategies. Following the data-generating process detailed in Section 3, we generate data with

$$g_t^\star(X_i) = 1 + \left[\tanh\left(|X_i^\top w_t|\right)\right]^{1/2} + b_t,$$

and use ReLU-MLP factors for $\phi_i^\star(F_t^\star)$, following the setup in Table S.1. Across 20 replications, the average ground-truth component values are $g^\star = 1.7968, \phi^\star = 0.1515$, and $g^\star + \phi^\star = 1.9483$. Table S.1 reports the summary statistics of the estimated covariate and latent-factor components.

Table S.1: Comparison between sequential estimation (SEQ) and joint estimation (JOINT) over 20 replications.

	SEQ			JOINT		
	$\hat{g}$	$\hat{\phi}$	$\hat{g}+\hat{\phi}$	$\hat{g}$	$\hat{\phi}$	$\hat{g}+\hat{\phi}$
Mean	1.9010	0.0409	1.9417	0.2814	1.6465	1.9279
SD	1.0508	0.1043	1.0934	0.1433	1.0283	1.1063
MSE	0.9832	0.9573	0.0033	2.3166	2.2593	0.0069

The results show that SEQ yields smaller MSEs than JOINT for both component-wise estimation and aggregate estimation. At the aggregate level, both methods estimate $g^\star + \phi^\star$ reasonably well. However, the component-wise decomposition differs substantially. JOINT assigns most of the signal to the latent factor network, whereas SEQ provides a decomposition closer to the intended model structure. In particular, JOINT has much larger component-wise MSE for $\hat{g}$ and $\hat{\phi}$, despite producing a comparable estimate of their sum. This suggests that joint optimization can obscure the separation between observed covariate effects and latent factor effects.

To further reveal the entanglement phenomenon between $\widehat{g}_t(X_i)$ and $\widehat{\phi}_i(\widehat{F}_t)$ in JOINT, Table S.2 reports the per-replication decomposition. In JOINT, the constant component in $g_t^\star(X_i)$ is frequently absorbed by the latent factor network $\hat{\phi}$. This is likely due to the flexibility of the autoencoder, which can absorb constant or shared signals into the latent representation. In contrast, SEQ removes the constant component during the covariate-removal step, leaving the remaining residual signal for factor extraction.

This experiment provides empirical evidence for the entanglement issue in joint estimation: JOINT risks making the decomposition into latent-factor and observed-covariate contributions less stable and less interpretable. While JOINT can still produce a reasonable estimate of the aggregate signal $\hat{g} + \hat{\phi}$, it may yield a misleading decomposition between observed covariate effects and latent factor effects. This can compromise the interpretability of the learned factors and deviate from the intended model design, where $g_t(\cdot)$ captures observed covariate effects and $\phi_i(\cdot)$ captures latent effects.

More generally, when two flexible neural networks jointly model an additive structure of the form $g(\cdot) + \phi(\cdot)$, the decomposition is not identifiable without additional restrictions. For any overlapping component $h(\cdot)$, the same aggregate function can be represented as

$$g(\cdot) + \phi(\cdot) = \{g(\cdot) - h(\cdot)\} + \{\phi(\cdot) + h(\cdot)\},$$

which leaves the prediction unchanged but reallocates signal between the two components. We therefore interpret the above results not as a formal identification failure, but as evidence that joint optimization can absorb shared signals into different channels in a non-robust way. This supports the use of sequential estimation in CoDEAL.

Table S.2: Per-replication comparison between SEQ and JOINT.

Sim	True value			SEQ			JOINT		
	$g^\star$	$\phi^\star$	$g^\star+\phi^\star$	$\hat{g}$	$\hat{\phi}$	$\hat{g}+\hat{\phi}$	$\hat{g}$	$\hat{\phi}$	$\hat{g}+\hat{\phi}$
1	1.6327	-1.1332	0.4995	0.4476	-0.0081	0.4395	0.0671	0.2076	0.2748
2	1.8617	1.2469	3.1086	2.8624	0.2424	3.1048	0.1713	2.9729	3.1442
3	1.8485	-0.5985	1.2500	1.3981	-0.1415	1.2567	0.1582	1.2026	1.3609
4	1.8246	0.0237	1.8483	1.8629	0.0875	1.9504	0.3391	1.5944	1.9335
5	1.8295	0.0629	1.8924	1.8260	-0.0135	1.8125	0.4222	1.4780	1.9002
6	2.0058	1.1126	3.1184	3.0681	0.1774	3.2456	0.3539	2.6920	3.0459
7	1.7880	-0.7071	1.0809	1.1180	-0.0462	1.0718	0.0906	1.0832	1.1738
8	1.5881	-1.8680	-0.2799	-0.2799	-0.0166	-0.2965	0.0122	-0.4029	-0.3907
9	1.8263	0.5298	2.3561	2.2064	0.1415	2.3480	0.3307	1.9936	2.3242
10	1.8187	0.1177	1.9363	1.8625	0.1091	1.9717	0.1847	1.7044	1.8891
11	1.8074	0.4165	2.2238	2.1950	0.0999	2.2949	0.2813	1.8516	2.1329
12	1.8108	0.0605	1.8713	1.8906	-0.0706	1.8199	0.3713	1.5437	1.9150
13	1.8646	-1.4704	0.3942	0.3745	-0.0495	0.3251	0.3569	-0.0833	0.2736
14	1.8628	-0.1993	1.6635	1.5455	0.0408	1.5862	0.1689	1.4396	1.6085
15	1.7500	1.9254	3.6753	3.5302	0.1655	3.6956	0.3782	3.2931	3.6713
16	1.7984	1.3016	3.1000	3.0414	-0.0008	3.0405	0.2314	2.8129	3.0444
17	1.8398	0.6253	2.4651	2.3634	0.1366	2.5000	0.3930	2.1074	2.5005
18	1.7183	-0.8995	0.8188	0.7065	0.0899	0.7963	0.2918	0.4728	0.7647
19	1.5954	1.0235	2.6189	2.6049	-0.0050	2.5999	0.4415	2.2092	2.6506
20	1.8643	1.4608	3.3251	3.3918	-0.1201	3.2717	0.5840	2.7569	3.3409

B Supplement to Simulation Studies in Section 3

In this section, we use extensive simulation studies to demonstrate the finite-sample performance of the proposed CoDEAL. The context in this appendix supplements Section 3 of the main context.

B.1 Benchmark Methods

We consider a variety of benchmark methods, including a single-output AE (**Single AE**) (detailed below), neuron-enhanced AE (**AEMC-NE**) [29], matrix completion with nuclear norm minimization (**MC-NNM**) [12], vertical regression (**Vert-Reg**) [27], difference-in-differences (**DiD**) [38], and kernel principal component analysis (**KernelPCA**) [57]. In the assessment of performance in causal matrix completion, we further compare to three synthetic-control-style baselines, including the classic synthetic control (**SCM**)[3], generalized synthetic control (**GSC**) [70], and synthetic difference-in-differences (**SDID**) [8].

Additionally, we evaluate the algorithm's performance in terms of covariate effect removal. To this end, we compare the performance of CoDEAL using different methods to remove the covariate effects, specifically, DNN-based removal (as illustrated in Section 2.2), and linear-regression(LR)-based removal (detailed below).

We evaluate the algorithm's performance using two evaluation metrics: the mean absolute error $MAE = (\sum_{(i,t):W_{it}=1} |Y_{it}(1) - \widehat{Y}_{it}(0) - \tau_i^\star|)/(\sum_{(i,t)} W_{it})$ and the mean squared error $MSE = (\sum_{(i,t):W_{it}=1} (Y_{it}(1) - \widehat{Y}_{it}(0) - \tau_i^\star)^2)/(\sum_{(i,t)} W_{it})$.

Matrix Completion by Single-Output AE (as a benchmark method). Our proposed CoDEAL utilizes a multi-output autoencoder (as illustrated in Figure 1) to recover nonlinear latent factors while explicitly modeling the unit-specific heterogeneity through multiple decoders. The multi-output autoencoder uses a shared encoder network $\rho(\cdot)$ to map each column $\widetilde{\mathbf{Y}}_{\cdot t}$ into a K-dimensional latent space, followed by N separate decoder networks $\{\phi_1, \ldots, \phi_N\}$ to reconstruct each unit's value. To demonstrate CoDEAL's ability to account for the unit-specific heterogeneity, we include an additional benchmark that replaces the multi-output autoencoder with a single-output autoencoder in the CoDEAL algorithm. This benchmark is referred to as the Single AE in later discussions. A single-output autoencoder uses the same encoder $\rho(\cdot)$ but only one decoder $\phi(\cdot)$ (i.e. $\phi = \phi_1 = \cdots = \phi_N$) for all units. By design, the common decoder makes the single-output AE inherently unable to model the unit-specific heterogeneity effectively. For implementation, DNNs in the single-output AE are

estimated by

$$\widehat{\phi} \circ \widehat{\rho}(\cdot) = \underset{\phi \circ \rho(\cdot) \in \mathcal{G}_{\text{AE}}^{K_1}}{\arg\min} \sum_{(i,t):W_{it}=0} \Big(\phi(\rho(\widetilde{\mathbf{Y}}_{\cdot t})) - \widetilde{Y}_{it}\Big)^2.$$

Covariate Effect Removal by Linear Regression (as a benchmark method). Our proposed CoDEAL uses a DNN to remove the potentially nonlinear, heterogeneous, and time-varying effects brought by covariates. To evaluate the algorithm's performance in terms of the covariate effect removal, we include a linear-regression(LR)-based removal as a benchmark. The LR-based removal assumes that covariates have a linear impact on the outcome. With covariates $\mathbf{X}_i \in \mathbb{R}^P$, $i \in [N]$, we denote the effect of covariates on the outcome by a coefficient matrix $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_T] \in \mathbb{R}^{P \times T}$, and we estimate $\mathbf{H}_t \in \mathbb{R}^P$ separately for each period. For each pre-treatment time $t \in \{1, \ldots, T_1\}$, we estimate the covariate effect vector $\widehat{\mathbf{H}}_t \in \mathbb{R}^P$ using OLS on all units: $\widehat{\mathbf{H}}_t = \arg\min_{\mathbf{H}_t} \sum_{i=1}^N \left(Y_{it} - \mathbf{X}_i^\top \mathbf{H}_t\right)^2$. For each post-treatment period $t \in \{T_1 + 1, \ldots, T\}$, we use only control units to estimate the covariate effect: $\widehat{\mathbf{H}}_t = \arg\min_{\mathbf{H}_t} \sum_{i:W_{it}=0} \left(Y_{it} - \mathbf{X}_i^\top \mathbf{H}_t\right)^2$. With the estimated $\widehat{\mathbf{H}}_t, t \in [T]$, we obtain the LR-based covariate-adjusted observed outcomes by $\widetilde{Y}_{it} = Y_{it} - \mathbf{X}_i^\top \widehat{\mathbf{H}}_t$ for each $i \in [N]$ and $t \in [T]$.

B.2 Data Generating Process and Simulation Configurations

We perform extensive simulation studies under various linear and nonlinear generating mechanisms, with different setups of unit numbers N, time periods T, matrix ranks K, and treatment assignments. Specifically, we consider the following configurations 1-4:

1. Four-Block Design: $(N, T, N_1, T_1, P, K) = (100, 200, 50, 100, 3, 4)$
2. Four-Block Design: $(N, T, N_1, T_1, P, K) = (200, 120, 100, 60, 5, 3)$
3. Staggered Adoption Design: $(N, T, K) = (100, 200, 4)$, $r = \{5, 10\}$
4. Staggered Adoption Design: $(N, T, K) = (200, 120, 3)$, $r = \{5, 10\}$.

For the generation of factor effects and covariate effects, we first generate covariates $\mathbf{X} = (X_{it}) \in \mathbb{R}^{N \times P}$, factors $\mathbf{F} = (F_{it}) \in \mathbb{R}^{T \times K}$, factor loadings $\mathbf{\Lambda} = (\Lambda_{it}) \in \mathbb{R}^{N \times K}$ with $X_{it}, F_{it}, \Lambda_{it} \overset{\text{i.i.d.}}{\sim} N(0, 1)$. With these building blocks, we consider various linear and nonlinear generating mechanisms to generate factor effects $\phi_i^\star(\mathbf{F}_t^\star)$ and covariate effects $g_t^\star(\mathbf{X}_i)$. Specifically, we experiment with 4 variants for generating factor effects and 6 variants for generating covariate effects, as listed in Table S.3. The unit-specific causal effect $\tau_i^\star \overset{i.i.d}{\sim} N(12, 5)$ is designed to be added to treated units in treated periods. The observed outcome matrix $\mathbf{Y} = (Y_{it}) \in \mathbb{R}^{N \times T}$ is then obtained by $Y_{it} = Y_{it}^\star + \tau_i^\star \cdot \mathcal{I}\{W_{it} = 1\} + \varepsilon_{it}$ where W_{it} are determined based on the treatment assignments according to the simulation configuration, and the idiosyncratic error ε_{it} follows $\varepsilon_{it} \overset{i.i.d}{\sim} \mathcal{N}(0, 0.5^2)$.

Table S.3: Data Generation Processes for Latent Factor Effects and Covariate Effects

Factor Effect Generation ($\phi_i^\star(\mathbf{F}_t^\star)$)		**Covariate Effect Generation** ($g_t^\star(\mathbf{X}_i)$)	
Linear	$0.5C_1 \mathbf{\Lambda}_i^\top \mathbf{F}_t, C_1 \overset{\text{i.i.d.}}{\sim} N(0,1)$	**Matrix Linear**	$\mathbf{X}_i^\top \mathbf{U}_t, \mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_t) \overset{\text{i.i.d.}}{\sim} N_{P \times T}(1,1)$
Sine	$2C_1 \sin(\mathbf{\Lambda}_i^\top \mathbf{F}_t)$	**Vector Linear**	$\mathbf{X}_i^\top \mathbf{U}, \mathbf{U} \overset{\text{i.i.d.}}{\sim} N_P(1,1)$
Polynomial	$0.2C_1 \mathbf{\Lambda}_i^\top \mathbf{F}_t + 0.2C_2 \mathbf{F}_t^\top \mathbf{F}_t, C_2 \overset{\text{i.i.d.}}{\sim} N(0,1)$	**tanh**	$[\tanh(\lvert\mathbf{X}_i^\top \boldsymbol{w}_t\rvert)]^{1/2} + b_t, \boldsymbol{w}_t \sim N_P(0,1)$
ReLU MLP	$\mathbf{R}_i^{(2)} \text{ReLU}(\mathbf{R}_i^{(1)} \mathbf{F}_t + \mathbf{b}_i^{(1)}) + b_i^{(2)}$	**poly**	$\lvert\mathbf{X}_i^\top \boldsymbol{w}_t\rvert^{1/2} + b_t$
	$b_i^{(2)} \overset{\text{i.i.d.}}{\sim} N(0, 0.5^2), \mathbf{R}_i^{(1)} \overset{\text{i.i.d.}}{\sim} N_{h_1 \times r}(0, 0.5^2)$	**log**	$\log(\lvert\mathbf{X}_i^\top \boldsymbol{w}_t + b_t\rvert), b_t \overset{\text{i.i.d.}}{\sim} N(0,1)$
	$\mathbf{b}_i^{(1)}, \mathbf{R}_i^{(2)} \overset{\text{i.i.d.}}{\sim} N_{h_1}(0, 0.5^2), h_1 = 10$	**ReLU**	$\text{ReLU}(\mathbf{X}_i \mathbf{R}_1^{(t)} + \mathbf{b}_1^{(t)}) \mathbf{R}_2^{(t)} + b_2^{(t)}, b_2^{(t)} \overset{\text{i.i.d.}}{\sim} N(0,1)$
			$\mathbf{R}_1^{(t)} \overset{\text{i.i.d.}}{\sim} N_{P \times h_2}(0, \frac{2}{P}), \mathbf{b}_1^{(t)} \overset{\text{i.i.d.}}{\sim} N_{h_2}(0,1)$
			$\mathbf{R}_2^{(t)} \overset{\text{i.i.d.}}{\sim} N_{h_2}(0, \frac{2}{h_2}), h_2 = 32$

Note: The vector linear covariate effect generation setting mimics a situation in which covariate effects are heterogeneous across panel units but remain time-invariant over time. The matrix linear covariate effect generation setting, however, allows for heterogeneous and time-varying covariate effects.

Table S.4: Comparisons of prediction accuracy (measured by MAE and MSE) with covariate effects removed by **LR** and **DNN** under various covariate generations in Config.1. Results are averaged over 50 replications. Numbers in parentheses are standard errors.

Covariate	Method	**LR**, Linear Factor		**NN**, Linear Factor		**LR**, Sine Factor		**NN**, Sine Factor	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
No covariate	CoDEAL	**0.560(0.009)**	0.578(0.028)	**0.560(0.009)**	0.578(0.028)	**0.933(0.015)**	**1.722(0.066)**	**0.933(0.015)**	**1.722(0.066)**
	Single AE	0.724(0.009)	1.081(0.034)	0.724(0.009)	1.081(0.034)	1.071(0.013)	2.131(0.056)	1.071(0.013)	2.131(0.056)
	AEMC-NE	0.734(0.009)	1.142(0.039)	0.734(0.009)	1.142(0.039)	1.090(0.013)	2.172(0.056)	1.090(0.013)	2.172(0.056)
	DiD	0.804(0.010)	1.364(0.043)	0.804(0.010)	1.364(0.043)	1.141(0.013)	2.343(0.059)	1.141(0.013)	2.343(0.059)
	Vert-Reg	0.601(0.002)	**0.574(0.004)**	0.601(0.002)	**0.574(0.004)**	1.024(0.012)	2.228(0.069)	1.024(0.012)	2.228(0.069)
	MC-NNM	0.580(0.010)	0.593(0.030)	0.580(0.010)	0.593(0.030)	1.093(0.013)	2.147(0.054)	1.093(0.013)	2.147(0.054)
	KernelPCA	0.773(0.009)	1.252(0.037)	0.773(0.009)	1.252(0.037)	1.099(0.009)	2.184(0.056)	1.099(0.009)	2.184(0.056)
Matrix Linear	CoDEAL	**0.585(0.013)**	**0.615(0.024)**	**0.646(0.009)**	**0.722(0.023)**	**0.983(0.011)**	**1.889(0.046)**	**1.079(0.012)**	**2.135(0.054)**
	Single AE	0.751(0.008)	1.131(0.032)	0.821(0.010)	1.320(0.039)	1.123(0.014)	2.283(0.060)	1.226(0.013)	2.626(0.062)
	AEMC-NE	0.779(0.010)	1.244(0.041)	0.838(0.009)	1.394(0.041)	1.145(0.013)	2.336(0.057)	1.209(0.013)	2.564(0.060)
	DiD	0.837(0.010)	1.444(0.045)	0.892(0.010)	1.587(0.045)	1.187(0.013)	2.490(0.060)	1.244(0.013)	2.696(0.063)
	Vert-Reg	0.686(0.006)	0.766(0.015)	0.757(0.006)	0.957(0.019)	1.103(0.012)	2.498(0.066)	1.169(0.012)	2.647(0.064)
	MC-NNM	0.604(0.009)	0.664(0.025)	0.699(0.006)	0.895(0.020)	1.140(0.013)	2.293(0.055)	1.201(0.013)	2.510(0.059)
	KernelPCA	0.830(0.008)	1.413(0.037)	0.811(0.007)	1.260(0.030)	1.179(0 012)	2.433(0.049)	1.195(0.009)	2.496(0.039)
Vector Linear	CoDEAL	**0.577(0.010)**	**0.592(0.026)**	**0.652(0.006)**	**0.715(0.017)**	**0.964(0.016)**	**1.775(0.060)**	**1.053(0.011)**	**2.035(0.045)**
	Single AE	0.746(0.007)	1.106(0.026)	0.841(0.008)	1.373(0.030)	1.106(0.012)	2.226(0.054)	1.205(0.012)	2.549(0.053)
	AEMC-NE	0.778(0.010)	1.205(0.032)	0.823(0.008)	1.317(0.029)	1.135(0.013)	2.298(0.053)	1.191(0.012)	2.497(0.053)
	DiD	0.825(0.008)	1.373(0.031)	0.872(0.009)	1.489(0.033)	1.173(0.012)	2.437(0.053)	1.224(0.012)	2.620(0.056)
	Vert-Reg	0.669(0.005)	0.727(0.012)	0.733(0.006)	0.887(0.017)	1.085(0.011)	2.413(0.057)	1.145(0.011)	2.520(0.053)
	MC-NNM	0.591(0.008)	0.617(0.020)	0.678(0.007)	0.826(0.020)	1.106(0.017)	2.219(0.099)	1.175(0.012)	2.412(0.053)
	KernelPCA	0.819(0.009)	1.402(0.035)	0.806(0.008)	1.208(0.029)	1.156(0 011)	2.398(0.047)	1.189(0.008)	2.437(0.038)
tanh	CoDEAL	0.778(0.008)	1.168(0.029)	**0.701(0.008)**	**0.871(0.031)**	**1.142(0.013)**	2.366(0.059)	**1.080(0.013)**	**2.175(0.059)**
	Single AE	0.889(0.008)	1.554(0.035)	0.851(0.008)	1.383(0.033)	1.299(0.014)	2.910(0.067)	1.257(0.014)	2.748(0.066)
	AEMC-NE	0.988(0.008)	1.822(0.036)	0.862(0.009)	1.445(0.036)	1.305(0.013)	2.925(0.064)	1.239(0.014)	2.677(0.065)
	DiD	0.855(0.008)	1.473(0.035)	0.916(0.009)	1.642(0.037)	1.199(0.013)	2.536(0.061)	1.274(0.014)	2.817(0.067)
	Vert-Reg	**0.721(0.005)**	**0.844(0.015)**	0.798(0.007)	1.075(0.024)	1.146(0.012)	2.660(0.071)	1.212(0.013)	2.787(0.072)
	MC-NNM	0.837(0.007)	1.131(0.021)	0.735(0.005)	1.007(0.017)	1.153(0.013)	**2.339(0.056)**	1.234(0.013)	2.636(0.063)
	KernelPCA	0.856(0.008)	1.471(0.036)	0.824(0.008)	1.292(0.033)	1.196(0.011)	2.496(0.049)	1.231(0.009)	2.626(0.041)
log	CoDEAL	**1.020(0.006)**	**1.873(0.026)**	**0.905(0.010)**	**1.517(0.030**	**1.345(0.012)**	**3.233(0.065)**	**1.253(0.013)**	**2.849(0.064)**
	Single AE	1.119(0.006)	2.266(0.031)	1.013(0.008)	1.943(0.033)	1.427(0.012)	3.502(0.065)	1.386(0.013)	3.332(0.068)
	AEMC-NE	1.137(0.007)	2.386(0.036)	1.024(0.008)	2.007(0.036)	1.421(0.012)	3.483(0.064)	1.372(0.014)	3.281(0.069)
	DiD	1.127(0.007)	2.371(0.037)	1.069(0.008)	2.182(0.038)	1.408(0.013)	3.437(0.066)	1.401(0.014)	3.402(0.071)
	Vert-Reg	1.100(0.006)	2.087(0.024)	1.055(0.007)	1.940(0.028)	1.485(0.012)	4.054(0.077)	1.439(0.012)	3.737(0.074)
	MC-NNM	1.044(0.006)	1.998(0.028)	0.939(0.005)	1.649(0.020)	1.373(0.012)	3.266(0.061)	1.366(0.013)	3.233(0.067)
	KernelPCA	1.136(0.008)	2.426(0.040)	1.064(0.009)	2.174(0.041)	1.417(0.010)	3.455(0.050)	1.391(0.011)	3.326(0.052)
poly	CoDEAL	0.833(0.007)	1.290(0.028)	**0.715(0.007)**	**0.913(0.035)**	**1.182(0.013)**	**2.507(0.059)**	**1.024(0.013)**	**2.086(0.062)**
	Single AE	0.946(0.008)	1.704(0.036)	0.854(0.008)	1.395(0.033)	1.345(0.014)	3.088(0.068)	1.261(0.014)	2.760(0.066)
	AEMC-NE	1.060(0.008)	2.038(0.037)	0.868(0.008)	1.461(0.036)	1.348(0.012)	3.088(0.063)	1.245(0.014)	2.701(0.065)
	DiD	0.906(0.008)	1.596(0.036)	0.921(0.009)	1.653(0.037)	1.235(0.013)	2.659(0.062)	1.279(0.014)	2.832(0.067)
	Vert-Reg	**0.756(0.006)**	**0.934(0.017)**	0.800(0.007)	1.078(0.023)	1.281(0.012)	2.983(0.072)	1.216(0.013	2.790(0.070)
	MC-NNM	0.847(0.006)	1.305(0.032)	0.747(0.005)	1.014(0.017)	1.290(0.013)	2.763(0.057)	1.238(0.013)	2.652(0.063)
	KernelPCA	0.924(0.008)	1.637(0.037)	0.920(0.009)	1.563(0.038)	1.274(0.012)	2.785(0.053)	1.243(0.011)	2.659(0.049)
ReLU	CoDEAL	0.946(0.009)	1.702(0.042)	**0.717(0.008)**	**0.916(0.025)**	1.205(0.011)	2.645(0.056)	**1.121(0.012)**	**2.269(0.054)**
	Single AE	1.020(0.010)	1.971(0.046)	0.858(0.010)	1.407(0.039)	1.317(0.011)	2.976(0.050)	1.222(0.014)	2.617(0.062)
	AEMC-NE	1.288(0.011)	2.982(0.053)	0.868(0.010)	1.471(0.043)	1.544(0.012)	4.003(0.062)	1.235(0.013)	2.660(0.062)
	DiD	0.960(0.009)	1.777(0.046)	0.922(0.010)	1.670(0.047)	1.274(0.011)	2.802(0.050)	1.269(0.014)	2.791(0.064)
	Vert-Reg	**0.779(0.006)**	**1.013(0.020)**	0.782(0.006)	1.027(0.021)	**1.195(0.010)**	2.786(0.059)	1.199(0.012)	2.753(0.067)
	MC-NNM	0.967(0.009)	1.893(0.035)	0.764(0.009)	1.073(0.024)	1.233(0.011)	**2.620(0.048)**	1.227(0.013)	2.608(0.060)
	KernelPCA	0.955(0.008)	1.746(0.038)	0.919(0.009)	1.645(0.039)	1.270(0.011)	2.773(0.049)	1.264(0.012)	2.747(0.052)

B.3 Evaluation on the Performance of CoDEAL in the Removal of Covariate Effects

To evaluate the algorithm's performance in terms of the covariate effect removal, we compare the performance of CoDEAL using different methods, specifically, DNN (as illustrated in Section 2.2) and LR (detailed in Section B.1), to remove the covariate effects. Six scenarios of covariate effects are considered as detailed in Table S.3. We also include an oracle benchmark, when no covariate effects are present and no covariate-adjusted operations are needed.

Table S.4 and Table S.5 report the comparisons of MAE and MSE using the two covariate effect removal methods in Config.1 and Config.2, respectively. Under each configuration, we consider both linear factor effects and nonlinear Sine factor effects, with all six types of covariate effects.

As shown in Tables S.4 and S.5, LR-based removal and DNN-based removal behave very differently depending on the underlying true types of covariate effects. In cases where the covariate effect is strictly linear, LR often outperforms NN by a small margin, indicating that a simple linear adjustment suffices for purely linear effects. In contrast, whenever the covariate effects enter nonlinearly (e.g. tanh, log, polynomial, or ReLU), the proposed DNN covariate removal achieves substantially lower errors, with MAE reduced by 10–30 % and MSE by similar amounts relative to LR. These results confirm that LR is adequate for linear covariate effects but that the DNN approach is essential to capture complex, nonlinear covariate–outcome relationships.

Table S.5: Comparisons of prediction accuracy (measured by MAE and MSE) with covariate effects removed by **LR** and **DNN** under various covariate generations in Config.2. Results are averaged over 50 replications. Numbers in parentheses are standard errors.

Covariate	Method	**LR**, Linear Factor		**NN**, Linear Factor		**LR**, Sine Factor		**NN**, Sine Factor	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
No covariate	CoDEAL	**0.507(0.004)**	**0.425(0.010)**	**0.507(0.004)**	**0.425(0.010)**	**0.779(0.007)**	**1.173(0.027)**	**0.779(0.007)**	**1.173(0.027)**
	Single AE	0.622(0.006)	0.761(0.021)	0.622(0.006)	0.761(0.021)	0.963(0.010)	1.772(0.038)	0.963(0.010)	1.772(0.038)
	AEMC-NE	0.647(0.006)	0.850(0.023)	0.647(0.006)	0.850(0.023)	1.004(0.009)	1.884(0.036)	1.004(0.009)	1.884(0.036)
	DiD	0.713(0.007)	1.043(0.028)	0.713(0.007)	1.043(0.028)	1.085(0.010)	2.163(0.041)	1.085(0.010)	2.163(0.041)
	Vert-Reg	0.651(0.002)	0.671(0.004)	0.651(0.002)	0.671(0.004)	0.865(0.007)	1.422(0.034)	0.865(0.007)	1.422(0.034)
	MC-NNM	0.535(0.005)	0.495(0.013)	0.535(0.005)	0.495(0.013)	1.020(0.009)	1.856(0.034)	1.020(0.009)	1.856(0.034)
	KernelPCA	0.703(0.007)	1.012(0.027)	0.703(0.007)	1.012(0.027)	1.081(0.009)	2.128(0.035)	1.081(0.009)	2.128(0.035)
Matrix Linear	CoDEAL	**0.510(0.005)**	**0.436(0.013)**	**0.609(0.004)**	**0.622(0.010)**	**0.781(0.007)**	**1.177(0.025)**	**0.896(0.008)**	**1.456(0.029)**
	Single AE	0.645(0.007)	0.807(0.022)	0.777(0.007)	1.147(0.024)	1.007(0.009)	1.885(0.037)	1.146(0.010)	2.330(0.042)
	AEMC-NE	0.682(0.007)	0.919(0.024)	0.759(0.007)	1.093(0.025)	1.051(0.009)	2.018(0.037)	1.127(0.009)	2.259(0.040)
	DiD	0.736(0.008)	1.093(0.029)	0.810(0.007)	1.262(0.029)	1.122(0.010)	2.275(0.041)	1.191(0.009)	2.503(0.042)
	Vert-Reg	0.680(0.003)	0.733(0.006)	0.761(0.004)	0.936(0.011)	0.917(0.007)	1.550(0.031)	1.110(0.007)	2.024(0.027)
	MC-NNM	0.537(0.005)	0.498(0.013)	0.650(0.005)	0.736(0.015)	1.025(0.009)	1.889(0.033)	1.106(0.009)	2.147(0.037)
	KernelPCA	0.729(0.007)	1.067(0.030)	0.835(0.009)	1.399(0.038)	1.119(0 010)	2.241(0.037)	1.170(0.013)	2.415(0.056)
Vector Linear	CoDEAL	**0.508(0.005)**	**0.428(0.010)**	**0.589(0.004)**	**0.574(0.010)**	**0.784(0.007)**	**1.184(0.026)**	**0.883(0.007)**	**1.426(0.025)**
	Single AE	0.659(0.006)	0.843(0.018)	0.772(0.007)	1.135(0.024)	1.010(0.008)	1.890(0.031)	1.147(0.009)	2.333(0.036)
	AEMC-NE	0.700(0.009)	0.952(0.024)	0.761(0.008)	1.099(0.026)	1.071(0.010)	2.073(0.036)	1.130(0.009)	2.265(0.034)
	DiD	0.737(0.006)	1.093(0.023)	0.800(0.007)	1.234(0.025)	1.125(0.009)	2.279(0.036)	1.186(0.009)	2.478(0.036)
	Vert-Reg	0.675(0.003)	0.724(0.007)	0.742(0.004)	0.884(0.010)	0.915(0.007)	1.555(0.031)	0.995(0.007)	1.768(0.030)
	MC-NNM	0.536(0.005)	0.499(0.012)	0.640(0.006)	0.714(0.017)	1.019(0.010)	1.870(0.036)	1.099(0.008)	2.124(0.030)
	KernelPCA	0.720(0.008)	1.006(0.031)	0.829(0.008)	1.309(0.038)	1.112(0 009)	2.186(0.040)	1.154(0.014)	2.308(0.052)
tanh	CoDEAL	0.700(0.007)	0.920(0.024)	**0.644(0.005)**	**0.704(0.015)**	1.012(0.008)	1.913(0.034)	**0.904(0.008)**	**1.492(0.030)**
	Single AE	0.880(0.009)	1.442(0.032)	0.795(0.007)	1.190(0.027)	1.227(0.011)	2.638(0.047)	1.180(0.010)	2.446(0.043)
	AEMC-NE	0.902(0.008)	1.489(0.028)	0.771(0.007)	1.120(0.026)	1.243(0.009)	2.681(0.041)	1.160(0.010)	2.365(0.042)
	DiD	0.758(0.007)	1.139(0.028)	0.828(0.008)	1.309(0.030)	1.135(0.010)	2.313(0.041)	1.224(0.010)	2.613(0.044)
	Vert-Reg	0.731(0.003)	0.849(0.008)	0.770(0.005)	0.968(0.015)	**0.968(0.007)**	**1.698(0.032)**	1.048(0.007)	1.921(0.029)
	MC-NNM	**0.570(0.005)**	**0.555(0.013)**	0.697(0.006)	0.859(0.019)	1.136(0.009)	2.317(0.036)	1.145(0.009)	2.281(0.040)
	KernelPCA	0.811(0.008)	1.259(0.031)	0.849(0.009)	1.429(0.038)	1.194(0.009)	2.494(0.040)	1.190(0.013)	2.476(0.057)
log	CoDEAL	1.030(0.005)	1.867(0.017)	**0.878(0.004)**	**1.385(0.013)**	**1.236(0.008)**	**2.683(0.038)**	**1.107(0.007)**	**2.264(0.030)**
	Single AE	1.085(0.006)	2.125(0.026)	0.985(0.006)	1.808(0.026)	1.367(0.008)	3.241(0.040)	1.338(0.009)	3.116(0.042)
	AEMC-NE	1.078(0.006)	2.115(0.028)	0.968(0.007)	1.751(0.027)	1.360(0.008)	3.210(0.040)	1.319(0.009)	3.034(0.042)
	DiD	1.079(0.007)	2.135(0.030)	1.013(0.007)	1.919(0.030)	1.382(0.009)	3.312(0.043)	1.374(0.009)	3.268(0.045)
	Vert-Reg	1.275(0.005)	2.722(0.020)	1.108(0.005)	2.079(0.017)	1.516(0.008)	3.956(0.049)	1.388(0.008)	3.301(0.039)
	MC-NNM	**0.999(0.005)**	**1.795(0.021)**	0.907(0.005)	1.503(0.019)	1.306(0.008	2.957(0.038)	1.305(0.009)	2.951(0.040)
	KernelPCA	1.087(0.007)	2.168(0.032)	1.010(0.007)	1.907(0.033)	1.389(0.008)	3.329(0.039)	1.376(0.009)	3.266(0.042)
poly	CoDEAL	0.798(0.006)	1.128(0.022)	**0.656(0.005)**	**0.725(0.013)**	**1.082(0.009)**	**2.105(0.038)**	**0.962(0.008)**	**1.640(0.030)**
	Single AE	0.965(0.007)	1.672(0.029)	0.808(0.007)	1.223(0.027)	1.298(0.009)	2.895(0.044)	1.187(0.009)	2.463(0.041)
	AEMC-NE	0.999(0.007)	1.756(0.027)	0.789(0.007)	1.162(0.025)	1.313(0.010	2.934(0.044)	1.173(0.009)	2.412(0.041)
	DiD	0.848(0.007)	1.337(0.029)	0.841(0.008)	1.339(0.030)	1.197(0.009)	2.510(0.041)	1.234(0.010)	2.650(0.044)
	Vert-Reg	0.809(0.004)	1.045(0.010)	0.791(0.005)	1.017(0.015)	1.159(0.007)	2.288(0.033)	1.072(0.008)	2.001(0.032)
	MC-NNM	**0.710(0.005)**	**0.853(0.014)**	0.714(0.006)	0.896(0.018)	1.204(0.009)	2.321(0.036)	1.158(0.009)	2.324(0.039)
	KernelPCA	0.858(0.007)	1.349(0.030)	0.840(0.008)	1.333(0.033)	1.206(0.008)	2.517(0.037)	1.239(0.009)	2.658(0.041)
ReLU	CoDEAL	0.771(0.007)	1.205(0.027)	**0.612(0.005)**	**0.629(0.012)**	1.187(0.010)	2.498(0.044)	**0.902(0.008)**	**1.478(0.029)**
	Single AE	1.078(0.011)	2.110(0.043)	0.764(0.006)	1.069(0.020)	1.387(0.013)	3.302(0.064)	1.121(0.010)	2.232(0.040)
	AEMC-NE	1.234(0.011)	2.692(0.048)	0.784(0.007)	1.148(0.024)	1.513(0.011)	3.870(0.055)	1.156(0.009)	2.355(0.040)
	DiD	0.868(0.007)	1.407(0.029)	0.840(0.007)	1.334(0.029)	1.217(0.009)	2.580(0.038)	1.220(0.010)	2.603(0.042)
	Vert-Reg	0.808(0.004)	1.051(0.011)	0.783(0.004)	0.992(0.011)	**1.052(0.009)**	**1.981(0.035)**	1.049(0.007)	1.938(0.030)
	MC-NNM	**0.766(0.006)**	**1.034(0.019)**	0.651(0.005)	0.737(0.016)	1.123(0.009)	2.188(0.036)	1.107(0.009)	2.149(0.036)
	KernelPCA	0.860(0.007)	1.373(0.030)	0.835(0.008)	1.315(0.031)	1.211(0.009)	2.548(0.037)	1.224(0.009)	2.602(0.039)

B.4 Evaluation of the Performance of CoDEAL in Causal Matrix Completion

Table 1 (in the main text) reports the comparisons of prediction accuracy (MSE) between the proposed CoDEAL and six benchmark methods (detailed in Section B.1) in Config.1-4. Under each configuration, we consider all four types of factor effects (listed in Table S.3), with linear covariate effects and nonlinear tanh covariate effects.

CoDEAL in Four-Block Design. As shown in Config.1 and Config.2 in Table S.6, CoDEAL consistently **achieves the lowest MAE and MSE across nearly all settings** of factor effects and covariate effects, and achieves substantial improvements in MAE and MSE relative to all competing methods. Under the simplest linear factor model, CoDEAL reduces MAE by approximately 5–10% and MSE by about 10–20% compared to MC-NNM, the second-best performing method, and significantly outperforms other baseline methods by even larger margins. For scenarios involving nonlinear factor models, CoDEAL continues to show superior performance, typically surpassing MC-NNM by approximately 5–20% reduction in MAE and 10–20% reduction in MSE. Overall, these findings clearly demonstrate that CoDEAL's integration of DNN-based covariate adjustment and deep autoencoder-based latent factor extraction is robust and highly effective across diverse linear and nonlinear data-generating settings.

Table S.6: Comparisons of prediction accuracy (measured by MAE and MSE) under various factor and covariate generations in Config.1-4. Results are averaged over 50 replications. Numbers in parentheses are standard errors.

Factor	Method	Config. 1, linear covariate		Config. 1, nonlinear covariate		Config. 2, linear covariate		Config. 2, nonlinear covariate	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Linear	CoDEAL	**0.646(0.009)**	**0.722(0.023)**	**0.701(0.008)**	**0.871(0.031)**	**0.609(0.004)**	**0.622(0.010)**	**0.644(0.005)**	**0.704(0.015)**
	Single AE	0.821(0.010)	1.320(0.039)	0.851(0.008)	1.383(0.033)	0.777(0.007)	1.147(0.024)	0.795(0.007)	1.190(0.027)
	AEMC-NE	0.838(0.009)	1.394(0.041)	0.862(0.009)	1.445(0.036)	0.759(0.007)	1.093(0.025)	0.771(0.007)	1.120(0.026)
	DiD	0.892(0.010)	1.587(0.045)	0.916(0.009)	1.642(0.037)	0.810(0.007)	1.262(0.029)	0.828(0.008)	1.309(0.030)
	Vert-Reg	0.757(0.006)	0.957(0.019)	0.798(0.007)	1.075(0.024)	0.761(0.004)	0.936(0.011)	0.770(0.005)	0.968(0.015)
	MC-NNM	0.699(0.006)	0.895(0.020)	0.735(0.005)	1.007(0.017)	0.650(0.005)	0.736(0.015)	0.697(0.006)	0.859(0.019)
	KernelPCA	0.811(0.007)	1.260(0.030)	0.824(0.008)	1.292(0.033)	0.835(0.009)	1.399(0.038)	0.849(0.009)	1.429(0.038)
Sine	CoDEAL	**1.078(0.012)**	**2.136(0.055)**	**1.117(0.011)**	**2.238(0.050)**	**0.894(0.007)**	**1.454(0.028)**	**0.945(0.008)**	**1.593(0.030)**
	Single AE	1.223(0.013)	2.618(0.061)	1.250(0.012)	2.705(0.053)	1.142(0.009)	2.312(0.039)	1.185(0.009)	2.458(0.039)
	AEMC-NE	1.209(0.013)	2.566(0.061)	1.235(0.012)	2.645(0.053)	1.128(0.009)	2.256(0.039)	1.167(0.010)	2.389(0.041)
	DiD	1.244(0.013)	2.697(0.063)	1.272(0.012)	2.785(0.055)	1.191(0.009)	2.498(0.042)	1.234(0.009)	2.650(0.041)
	Vert-Reg	1.169(0.012)	2.633(0.063)	1.208(0.011)	2.736(0.059)	1.012(0.007)	1.829(0.031)	1.054(0.008)	1.935(0.034)
	MC-NNM	1.201(0.013)	2.511(0.059)	1.231(0.012)	2.610(0.053)	1.105(0.009)	2.142(0.036)	1.153(0.010)	2.305(0.040)
	KernelPCA	1.195(0.009)	2.496(0.039)	1.231(0.009)	2.626(0.041)	1.170(0.013)	2.415(0.056)	1.190(0.013)	2.476(0.057)
Polynomial	CoDEAL	**0.908(0.013)**	2.250(0.100)	**0.952(0.015)**	2.477(0.126)	**0.772(0.010)**	1.399(0.076)	**0.807(0.011)**	1.539(0.070)
	Single AE	0.962(0.013)	2.727(0.116)	0.997(0.016)	3.014(0.164)	0.820(0.010)	1.759(0.087)	0.854(0.011)	1.964(0.084)
	AEMC-NE	0.997(0.014)	3.198(0.142)	1.038(0.017)	3.500(0.192)	0.856(0.011)	2.132(0.110)	0.887(0.012)	2.339(0.111)
	DiD	0.995(0.013)	2.708(0.111)	1.025(0.016)	2.939(0.144)	0.854(0.010)	1.860(0.083)	0.886(0.011)	2.065(0.092)
	Vert-Reg	0.953(0.013)	**1.906(0.076)**	1.019(0.016)	**2.327(0.113)**	0.855(0.007)	**1.289(0.032)**	0.884(0.010)	**1.456(0.042)**
	MC-NNM	0.951(0.013)	2.419(0.101)	0.989(0.015)	2.627(0.130)	0.819(0.009)	1.654(0.07)	0.851(0.011)	1.839(0.078)
	KernelPCA	0.927(0.013)	2.582(0.121)	0.991(0.013)	2.662(0.122)	1.029(0.015)	3.676(0.205)	1.051(0.015)	3.752(0.206)
ReLU MLP	CoDEAL	**0.797(0.009)**	**1.097(0.027)**	**0.838(0.010)**	**1.217(0.032)**	**0.698(0.005)**	**0.820(0.014)**	**0.743(0.006)**	**0.929(0.015)**
	Single AE	0.989(0.009)	1.712(0.034)	1.026(0.011)	1.845(0.044)	0.891(0.007)	1.375(0.022)	0.930(0.008)	1.502(0.028)
	AEMC-NE	1.045(0.010)	1.919(0.038)	1.088(0.011)	2.084(0.048)	0.959(0.007)	1.605(0.026)	0.982(0.008)	1.679(0.029)
	DiD	0.930(0.008)	1.499(0.028)	0.963(0.010)	1.616(0.036)	0.842(0.005)	1.218(0.017)	0.870(0.007)	1.303(0.021)
	Vert-Reg	0.888(0.010)	1.315(0.035)	0.958(0.013)	1.544(0.046)	0.847(0.005)	1.166(0.014)	0.894(0.006)	1.302(0.020)
	MC-NNM	0.843(0.007)	1.227(0.025)	0.892(0.009)	1.453(0.031)	0.767(0.005)	1.001(0.013)	0.806(0.006)	1.110(0.019)
	KernelPCA	1.105(0.008)	2.082(0.036)	1.126(0.009)	2.156(0.036)	1.156(0.011)	2.315(0.045)	1.172(0.011)	2.372(0.045)
Factor	**Method**	Config. 3, $r = 5$		Config. 3, $r = 10$		Config. 4, $r = 5$		Config. 4, $r = 10$	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Linear	CoDEAL	0.614(0.007)	0.721(0.023)	0.625(0.005)	0.754(0.018)	**0.532(0.003)**	**0.506(0.009)**	**0.560(0.004)**	**0.591(0.012)**
	Single AE	0.789(0.011)	1.263(0.044)	0.795(0.008)	1.270(0.033)	0.693(0.005)	0.948(0.017)	0.714(0.006)	1.009(0.023)
	AEMC-NE	0.731(0.010)	1.125(0.042)	0.730(0.007)	1.110(0.029)	0.647(0.005)	0.846(0.018)	0.660(0.005)	0.886(0.021)
	DiD	0.801(0.011)	1.341(0.047)	0.802(0.008)	1.326(0.033)	0.711(0.005)	1.031(0.021)	0.726(0.006)	1.081(0.025)
	Vert-Reg	0.940(0.005)	6.880(2.378)	0.865(0.023)	8.238(2.625)	0.691(0.003)	0.888(0.012)	0.750(0.016)	3.071(0.914)
	MC-NNM	**0.599(0.007)**	**0.685(0.023)**	**0.616(0.006)**	**0.733(0.019)**	0.537(0.003)	0.520(0.009)	0.565(0.004)	0.596(0.014)
	KernelPCA	0.798(0.007)	1.332(0.031)	0.801(0.007)	1.334(0.028)	0.707(0.005)	1.032(0.020)	0.709(0.005)	1.038(0.020)
Sine	CoDEAL	**0.973(0.010)**	**1.798(0.041)**	**0.996(0.010)**	**1.907(0.039)**	**0.850(0.007)**	**1.389(0.024)**	**0.869(0.008)**	**1.463(0.028)**
	Single AE	1.087(0.011)	2.138(0.047)	1.122(0.011)	2.288(0.047)	1.057(0.008)	2.050(0.030)	1.061(0.008)	2.082(0.035)
	AEMC-NE	1.055(0.012)	2.029(0.047)	1.078(0.011)	2.128(0.044)	1.011(0.008)	1.890(0.029)	1.008(0.009)	1.897(0.034)
	DiD	1.110(0.012)	2.209(0.049)	1.146(0.012)	2.361(0.048)	1.091(0.008)	2.172(0.031)	1.088(0.009)	2.189(0.037)
	Vert-Reg	1.549(0.090)	8.954(2.462)	1.524(0.123)	15.689(4.057)	0.974(0.009)	2.153(0.059)	1.047(0.013)	4.384(0.544)
	MC-NNM	1.066(0.011)	2.033(0.046)	1.097(0.011)	2.164(0.043)	0.998(0.008)	1.819(0.031)	1.003(0.008)	1.858(0.032)
	KernelPCA	1.109(0.012)	2.205(0.051)	1.115(0.012)	2.221(0.049)	1.075(0.009)	2.084(0.033)	1.079(0.009)	2.096(0.031)
Polynomial	CoDEAL	**0.785(0.011)**	2.146(0.112)	**0.811(0.010)**	2.379(0.161)	**0.662(0.007)**	1.311(0.073)	**0.699(0.007)**	**1.530(0.064)**
	Single AE	0.895(0.015)	2.423(0.117)	0.926(0.016)	2.710(0.172)	0.739(0.011)	1.582(0.086)	0.785(0.010)	1.798(0.071)
	AEMC-NE	0.860(0.014)	2.921(0.160)	0.897(0.013)	3.273(0.220)	0.720(0.009)	1.889(0.108)	0.764(0.009)	2.166(0.093)
	DiD	0.887(0.013)	2.422(0.116)	0.917(0.014)	2.693(0.165)	0.739(0.009)	1.613(0.082)	0.773(0.008)	1.801(0.068)
	Vert-Reg	1.155(0.035)	6.771(1.168)	1.131(0.032)	5.530(0.666)	0.788(0.005)	**1.246(0.028)**	0.881(0.014)	3.254(0.526)
	MC-NNM	0.839(0.012)	**2.117(0.104)**	0.865(0.012)	**2.308(0.129)**	0.703(0.008)	1.397(0.071)	0.735(0.007)	1.569(0.058)
	KernelPCA	0.935(0.011)	3.302(0.148)	0.944(0.011)	3.330(0.139)	0.776(0.008)	2.255(0.101)	0.782(0.008)	2.260(0.097)
ReLU MLP	CoDEAL	**0.632(0.004)**	**0.886(0.012)**	**0.733(0.004)**	**0.953(0.013)**	**0.633(0.004)**	**0.698(0.011)**	**0.646(0.004)**	**0.730(0.011)**
	Single AE	0.876(0.005)	1.338(0.017)	0.913(0.006)	1.448(0.019)	0.797(0.005)	1.108(0.015)	0.809(0.005)	1.142(0.015)
	AEMC-NE	0.934(0.006)	1.574(0.021)	0.968(0.007)	1.677(0.024)	0.860(0.006)	1.330(0.023)	0.878(0.006)	1.381(0.022)
	DiD	0.817(0.004)	1.163(0.013)	0.842(0.004)	1.238(0.014)	0.742(0.004)	0.962(0.013)	0.748(0.005)	0.976(0.014)
	Vert-Reg	1.045(0.031)	5.157(0.846)	0.948(0.026)	10.387(6.292)	0.755(0.003)	1.060(0.015)	1.070(0.253)	8.399(2.377)
	MC-NNM	0.721(0.003)	0.897(0.009)	0.750(0.003)	0.976(0.010)	0.656(0.004)	0.741(0.011)	0.662(0.004)	0.756(0.011)
	KernelPCA	1.111(0.008)	2.142(0.032)	1.119(0.008)	2.164(0.030)	1.005(0.006)	1.758(0.024)	1.009(0.006)	1.770(0.024)

CoDEAL in Staggered Adoption Design. As Config.3 and Config.4 of Table S.6 demonstrate, CoDEAL achieves substantial percentage gains in MAE and MSE relative to all competing methods across staggered adoption designs. Under *Linear* factors, CoDEAL **ties MC-NNM within 3% while reducing MAE by 16–35% over other benchmarks**. In staggered design with nonlinear factors, CoDEAL delivers evident advantages. Especially when under *ReLU*-based factors, CoDEAL delivers its largest margins, with **MAE reductions of 12–40% and MSE reductions of 15–41% versus all competing approaches**. Finally, as r increases from 5 to 10, the missingness pattern grows more complex and extensive, so all methods show some drop in accuracy. However, CoDEAL's errors only suffer a slight increase, whereas simpler methods like Vert-Reg show a dramatic rise in both MAE and MSE when missingness increases. The above results reemphasize CoDEAL's consistent superiority and robustness in staggered-treatment settings.

Comparisons with Synthetic-Control-Type Methods. We further compare to the classical synthetic control [3], fitted with the standard donor construction and simplex weighting, the generalized synthetic control (**GSC**)[70], and synthetic difference-in-differences (**SDID**) [8]. We note that GSC

corresponds to the special case of Model (1) in which both $g_t^\star(\cdot)$ and $\phi_i^\star(\cdot)$ are restricted to be linear. Table S.7 reports the results under four factor generating mechanisms and three covariate generating mechanisms. The results demonstrate CoDEAL's comparative advantage in the presence of nonlinearity, while GSC attains the best performance in every linear covariate setting. Since GSC corresponds to the special case of CoDEAL in which both $g_t^\star(\cdot)$ and $\phi_i^\star(\cdot)$ are restricted to be linear, these findings highlight the trade-off associated with the additional modeling flexibility.

In summary, CoDEAL achieves performance comparable to linearity-assumed methods, such as MC-NNM, in linear settings, while exhibiting substantial improvements in the presence of nonlinear effects. By integrating DNN-based covariate adjustments with deep multi-output-AE–based nonlinear factor analysis, CoDEAL is robustly effective across a wide range of linear and nonlinear scenarios.

Table S.7: Comparisons with synthetic-control-type methods under various factor and covariate generations. Results are averaged over 50 replications. Numbers in parentheses are standard errors.

Factor	**Method**	Config. 1, linear covariate		Config. 1, tanh covariate		Config. 1, poly covariate	
		MAE	MSE	MAE	MSE	MAE	MSE
Linear	CoDEAL	0.709(0.007)	0.914(0.022)	**0.746(0.007)**	1.009(0.025)	**0.751(0.007)**	**1.018(0.024)**
	SCM	0.804(0.011)	1.306(0.045)	0.909(0.010)	1.240(0.027)	1.126(0.012)	1.785(0.034)
	GSC	**0.484(0.006)**	**0.370(0.009)**	0.847(0.011)	**0.997(0.022)**	1.088(0.013)	1.590(0.032)
	SDID	1.584(0.013)	4.422(0.075)	1.089(0.010)	1.995(0.038)	1.282(0.011)	2.559(0.042)
Sine	CoDEAL	1.116(0.011)	2.232(0.050)	**1.156(0.011)**	2.362(0.052)	**1.155(0.011)**	**2.356(0.048)**
	SCM	1.146(0.012)	2.435(0.054)	1.168(0.010)	**2.269(0.042)**	1.345(0.011)	2.840(0.045)
	GSC	**0.999(0.011)**	**1.897(0.047)**	1.223(0.012)	2.522(0.052)	1.396(0.012)	3.097(0.055)
	SDID	1.791(0.013)	5.423(0.078)	1.344(0.013)	2.982(0.056)	1.493(0.013)	3.534(0.060)
Polynomial	CoDEAL	0.921(0.015)	2.564(0.154)	**0.953(0.016)**	2.698(0.159)	**0.960(0.015)**	**2.713(0.152)**
	SCM	0.960(0.016)	2.553(0.138)	1.052(0.018)	2.428(0.121)	1.253(0.019)	2.972(0.124)
	GSC	**0.826(0.014)**	**1.752(0.086)**	1.094(0.020)	**2.360(0.097)**	1.292(0.021)	2.937(0.104)
	SDID	1.660(0.016)	5.521(0.151)	1.172(0.022)	3.117(0.139)	1.358(0.024)	3.671(0.145)
ReLU MLP	CoDEAL	0.823(0.009)	1.172(0.030)	**0.872(0.011)**	**1.321(0.036)**	**0.877(0.011)**	**1.332(0.036)**
	SCM	0.892(0.011)	1.447(0.041)	0.940(0.014)	1.354(0.036)	1.143(0.016)	1.896(0.046)
	GSC	**0.664(0.007)**	**0.739(0.016)**	0.939(0.015)	1.335(0.037)	1.148(0.017)	1.907(0.048)
	SDID	1.573(0.013)	4.260(0.075)	1.068(0.015)	1.809(0.043)	1.246(0.016)	2.352(0.053)
		Config. 2, linear covariate		Config. 2, tanh covariate		Config. 2, poly covariate	
		MAE	MSE	MAE	MSE	MAE	MSE
Linear	CoDEAL	0.648(0.006)	0.730(0.016)	**0.673(0.006)**	**0.788(0.019)**	**0.693(0.006)**	**0.832(0.018)**
	SCM	0.838(0.008)	1.372(0.028)	0.905(0.011)	1.147(0.024)	1.234(0.013)	2.018(0.035)
	GSC	**0.465(0.007)**	**0.341(0.011)**	0.882(0.012)	1.052(0.024)	1.226(0.014)	1.976(0.037)
	SDID	1.884(0.012)	6.024(0.075)	1.050(0.011)	1.752(0.035)	1.347(0.012)	2.642(0.043)
Sine	CoDEAL	0.964(0.007)	1.662(0.027)	**1.017(0.009)**	1.821(0.033)	**1.032(0.008)**	**1.867(0.033)**
	SCM	1.113(0.008)	2.344(0.035)	1.048(0.010)	**1.727(0.028)**	1.346(0.011)	2.639(0.037)
	GSC	**0.800(0.008)**	**1.252(0.027)**	1.108(0.010)	1.976(0.034)	1.398(0.012)	2.901(0.043)
	SDID	2.069(0.011)	7.121(0.073)	1.324(0.010)	2.870(0.041)	1.562(0.012)	3.763(0.050)
Polynomial	CoDEAL	0.793(0.010)	1.608(0.071)	**0.822(0.011)**	1.720(0.076)	**0.836(0.011)**	**1.761(0.078)**
	SCM	0.921(0.010)	1.982(0.064)	0.991(0.015)	**1.675(0.050)**	1.314(0.017)	2.588(0.062)
	GSC	**0.678(0.010)**	**1.054(0.045)**	1.031(0.016)	1.788(0.053)	1.348(0.018)	2.728(0.065)
	SDID	1.915(0.015)	6.578(0.120)	1.094(0.016)	2.366(0.082)	1.397(0.017)	3.279(0.088)
ReLU MLP	CoDEAL	0.760(0.006)	0.986(0.018)	**0.804(0.007)**	**1.106(0.019)**	**0.818(0.007)**	**1.135(0.020)**
	SCM	0.911(0.007)	1.527(0.026)	0.947(0.013)	1.297(0.029)	1.267(0.015)	2.190(0.043)
	GSC	**0.622(0.007)**	**0.647(0.014)**	0.963(0.013)	1.353(0.031)	1.281(0.016)	2.275(0.046)
	SDID	1.864(0.013)	5.879(0.081)	1.045(0.012)	1.668(0.032)	1.335(0.014)	2.563(0.046)

B.5 Sensitivity to the Dependence between Covariates and Latent Heterogeneity

Although Model (1) admits an additive formulation in the covariate effect and the latent factor effect, as discussed in Appendix D.2.3, this does not preclude dependence between the observed covariates and the unobserved unit-level heterogeneity. To examine this empirically, we generate the factor loadings as

$$\boldsymbol{\Lambda}_i = \rho\,\mathbf{B}\mathbf{X}_i + \sqrt{1-\rho^2}\,\mathbf{U}_i, \qquad \rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\},$$

where $\mathbf{B}$ is fixed within a replication and $\mathbf{U}_i$ is independent noise, so that ρ controls the strength of the dependence between $\mathbf{X}_i$ and the latent heterogeneity. We additionally introduce an explicit

non-additive interaction term $\delta\, X_{i1} \tanh(F_{t1})$ with $\delta \in \{0, 0.5, 1.0\}$, which lies outside the additive specification of Model (1) altogether.

Table S.8 reports the MSE under Configuration 2 with sine factors. CoDEAL is stable across the full range of ρ and δ: its error varies by less than 5% over the grid, and it is in fact smallest at $\rho = 1$, where the loadings are a deterministic function of the covariates. Single AE, which uses a shared decoder and therefore cannot represent covariate-dependent loadings, is uniformly worse and deteriorates as the dependence strengthens. We emphasise that this experiment concerns the accuracy of the total counterfactual prediction; as noted in Appendix A.3, the decomposition of that prediction into a covariate component and a latent component is identified only up to reallocation between the two, and good predictive accuracy should not be read as evidence that the two components are separately recovered.

Table S.8: Prediction accuracy (MSE, averaged over 50 replications) under Configuration 2 with sine factors, as a function of the covariate–loading dependence ρ and the non-additive interaction strength δ.

ρ	$\delta = 0.0$		$\delta = 0.5$		$\delta = 1.0$	
	CoDEAL	Single AE	CoDEAL	Single AE	CoDEAL	Single AE
0.0	**1.476**	2.287	**1.476**	2.303	**1.509**	2.379
0.2	**1.488**	2.262	**1.473**	2.271	**1.504**	2.304
0.4	**1.452**	2.261	**1.496**	2.287	**1.501**	2.325
0.6	**1.446**	2.246	**1.451**	2.250	**1.475**	2.312
0.8	**1.418**	2.181	**1.417**	2.189	**1.443**	2.273
1.0	**1.252**	2.061	**1.268**	2.065	**1.294**	2.134

B.6 Evaluation in the Presence of Homogeneous Treatment Effects

We conduct an additional study to evaluate the empirical differences between Single AE and Multi-output AE in CoDEAL. Recall that the major difference between Single AE and CoDEAL is the structure of decoder: Single AE estimates a common mapping $\phi(\cdot)$ shared across all units, whereas CoDEAL estimates unit-specific mappings $\phi_i(\cdot)$. In Table 1 where the treatment effects are heterogeneous, CoDEAL consistently outperforms Single AE. Since Single AE and CoDEAL use the same covariate-adjustment procedure, the performance gap is not driven by differences in covariate removal. Rather, it reflects their different abilities to model unit heterogeneity. Since the data-generating process in Table S.6 is heterogeneous in $\phi_i(\cdot)$, Single AE is at a structural disadvantage relative to CoDEAL.

To further illustrate this point, we additionally consider settings with homogeneous $\phi(\cdot)$ and homogeneous treatment effects $\tau^\star$. In this case, Single AE becomes comparable to, and in some settings slightly better than, CoDEAL, except under the ReLU-MLP factor design, where CoDEAL still retains a clear advantage.

Table S.9: MAE and MSE comparison between Single AE and CoDEAL under homogeneous $\phi(\cdot)$ and homogeneous $\tau^\star$.

Setting	Metric	Method	Linear	Sine	Polynomial	ReLU-MLP
Config 1, no covariates	MSE	Single AE	0.263	0.262	0.314	0.857
		CoDEAL	0.285	0.277	0.365	0.451
	MAE	Single AE	0.409	0.409	0.442	0.702
		CoDEAL	0.424	0.428	0.464	0.525
Config 1, with covariates	MSE	Single AE	0.656	0.796	0.927	1.299
		CoDEAL	0.623	0.709	0.973	0.963
	MAE	Single AE	0.570	0.676	0.656	0.869
		CoDEAL	0.595	0.729	0.667	0.752

The results confirm that the advantage of CoDEAL is closely tied to its ability to model heterogeneous latent structure. When the latent factor component is homogeneous, the performance gap between Single AE and CoDEAL becomes much smaller, and Single AE can even be slightly preferable in simpler settings. By contrast, under the more nonlinear ReLU-MLP factor design, CoDEAL

continues to outperform Single AE, suggesting that its flexibility remains beneficial even beyond heterogeneity alone.

B.7 Evaluation of the Impact of the Proportion of Treated Periods and Units

We also study the algorithm's performance with respect to different proportions of treated periods and units. To illustrate, we examine different missing proportions with $N = 200$ and $T = 50$ with sine factor and no covariates. The following tables report MAE and MSE with different numbers of never-treated units and treatment start timestamps. From Table S.10 and S.11, we could observe that performance improves with more never-treated units as they provide essential baseline information for latent factor estimation. We also observe that while larger never-treated cohorts steadily improve accuracy, CoDEAL remains robust even when control units are relatively scarce.

Table S.10: CoDEAL: comparison error under varying numbers of never-treated units with treatment start at $T_1 = 25$.

# Never-treated units	MAE	MSE
40	0.864	1.470
60	0.858	1.450
80	0.855	1.448
100	0.846	1.436
120	0.843	1.439
140	0.833	1.405

Table S.11: CoDEAL: comparison error under varying treatment start timestamps with $N_1 = 100$ never-treated units.

Treatment start time	MAE	MSE
15	0.953	1.869
20	0.887	1.584
25	0.846	1.436
30	0.809	1.310
35	0.780	1.208

B.8 Hyperparameter Setting and Sensitivity Analysis

The simulation experiments in Section 3 and Appendix B.3 were repeated 50 times for each setup. For DNN-based covariate effect removal, we used a single hidden-layer neural network of width 32, trained with the Adam optimizer (learning rate $\eta = 0.001$), Xavier initialization, for 400 epochs. Both single-output AE and multi-output AE for matrix completion employed three hidden layers in their encoders and decoders; these were trained with Adam (learning rate $\eta = 0.002$) for 50 epochs. The AEMC-NE model featured layer-wise hidden sizes $[30, 30]$ and element-wise sizes $[20, 20]$, was trained with AdamW (learning rate $\eta = 0.001$) for 30 epochs.

We also conduct a sensitivity analysis for CoDEAL under Configuration 2 with sine factors, where the true latent factor dimension is $K = 3$. We vary the learning rate η, the number of DNN hidden layers L, and the AE bottleneck size K_1. The results are reported in Table S.12.

There are three main findings from Table S.12. First, CoDEAL exhibits stable performance across a broad range of learning rates, suggesting that its accuracy is relatively insensitive to this choice, provided the learning rate is not excessively small. Second, moderate network depth generally yields better performance, indicating that additional layers do not necessarily improve accuracy. Third, and most consequential in practice, the bottleneck dimension K_1 is the one choice that materially affects the performance. In particular, under-specifying the bottleneck dimension leads to substantial accuracy deterioration, whereas over-specification does not adversely affect performance and may even yield modest improvements. This suggests that K_1 should be chosen to be at least as large as the anticipated number of latent factors. Overall, these findings demonstrate that CoDEAL is relatively

Table S.12: Sensitivity analysis of CoDEAL under Configuration 2 with sine factors ($K = 3$).

Learning rate η	0.0005	0.001	0.002	0.003	0.004	0.005
MSE	1.624	1.284	1.173	1.078	1.098	1.122
Hidden layers L	1	2	3	4	5	
MSE	1.268	1.189	1.173	1.269	1.413	
Bottleneck K_1	2	3	4	5	6	
MSE	1.464	1.173	1.108	1.063	1.023	

robust to the choice of learning rate and network depth, with the bottleneck dimension being the most important hyperparameter to consider in practice.

Beyond the sensitivity analysis presented above, hyperparameters can also be selected in a data-driven manner by minimizing the out-of-sample reconstruction error on held-out blocks of the observed untreated panel. We further discuss this approach in Remark S.1 as a stand-alone empirical evaluation strategy, particularly useful for hyperparameter tuning and assessing performance in real-world applications where ground-truth counterfactual outcomes are inherently unobservable.

Remark S.1 (An empirical strategy for out-of-sample evaluation and hyperparameter selection)**.** *One may construct pseudo-treatment blocks within the observed untreated portion of the panel, mask these blocks, fit competing methods, and evaluate their reconstruction errors on the held-out observations. Since the true untreated outcomes are observed for these masked entries, the resulting prediction error provides an out-of-sample measure of counterfactual reconstruction accuracy. This masked-block evaluation offers a practical approach to assessing counterfactual reconstruction performance in empirical applications, where the true counterfactual outcomes are inherently unobservable. The same strategy can also be used for data-driven hyperparameter selection.*

B.9 Compute Resources and Computation Time

All simulations were executed on a 16-core CPU. Over 50 replications, Configurations 1–2 (four-block design) took roughly 1 hour and 500 MB max memory for every combination of factor and covariate generation, while Configurations 3–4 (staggered-adoption design) required about 3 hours running time and max memory used 1GB for each factor generation type.

We report the per-replication runtime comparisons in Table S.13. The results show that although the inclusion of separate decoders in CoDEAL increases computational cost, the overall overhead remains modest, even when N is large.

Table S.13: Runtime (in seconds) comparison in Config. 1 and 2 across methods with linear/sine factors and nonlinear (tanh) covariates

	Config. 1		Config. 2	
	Linear	Sine	Linear	Sine
CoDEAL	1.295	1.286	2.429	2.495
Single AE	0.025	0.026	0.025	0.027
AEMC-NE	0.115	0.118	0.125	0.129
DiD	0.000	0.001	0.002	0.001
Vert-Reg	0.034	0.037	0.083	0.083
MC-NNM	1.421	1.408	3.876	3.820
KernelPCA	0.006	0.004	0.004	0.003

C Supplement to Real Data Analysis in Section 4

C.1 Additional Results for OxCGRT dataset

This section reports supplementary results of the OxCGRT analysis. Figure S.3 and Figure S.5 are enlarged plots of Figure 3 of the main text for clarity, showing the estimated impact of the mandatory vaccination policy and an indicator matrix of the observed policy assignment.

We also examine the impact of the termination of internal travel restrictions, with an indicator matrix of the observed restriction-lifting assignments shown in Figure S.6. Figure S.4 plots the comparisons of the total confirmed cases and deaths across policy-lifted states between observed results (with travel restrictions lifted) and estimated counterfactuals (had the restrictions not been lifted). Results in Figure S.4 show that imposing internal travel restrictions led to similar proportional declines in both COVID-19 cases and deaths. This pattern is supported by recent studies [21, 47]: by limiting movement, restrictions reduce virus transmission (and thus new cases), which in turn lowers the number of deaths, even though restricting trips does not directly affect individual mortality risk.

In Table S.14, we also report the comparisons of the estimated counterfactual confirmed cases and deaths (in millions) for other benchmark methods besides CoDEAL. The results are similar in terms of the magnitude, but differ in the exact numbers. The vertical regression estimates for deaths are not reported because the staggered adoption design yields very short pre-treatment windows and a small size of control units, which makes the regression unstable in this case. This leads to highly sensitive coefficient estimates and severe overextrapolation, producing implausibly large counterfactual results.

However, because of the fundamental problem of causal inference, we are unable to observe the ground-truth counterfactuals in real-world datasets. Therefore, while we can report and compare the predicted values by each method, it is not possible to objectively determine which method performs best in terms of prediction accuracy in this dataset.

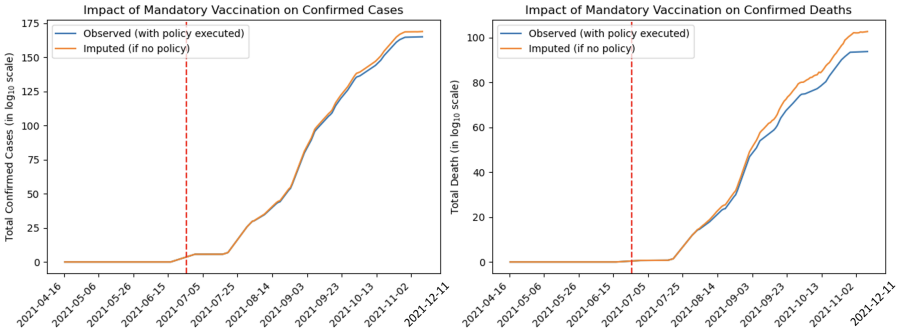

Figure S.3: Comparison of the total confirmed cases (Left) and deaths (Right) across policy-in-effect states between observed results with the mandatory vaccination policy executed and estimated counterfactuals if no policy. Values are reported on a $\log_{10}$ scale, and plots are visualized with 14-day moving averages.

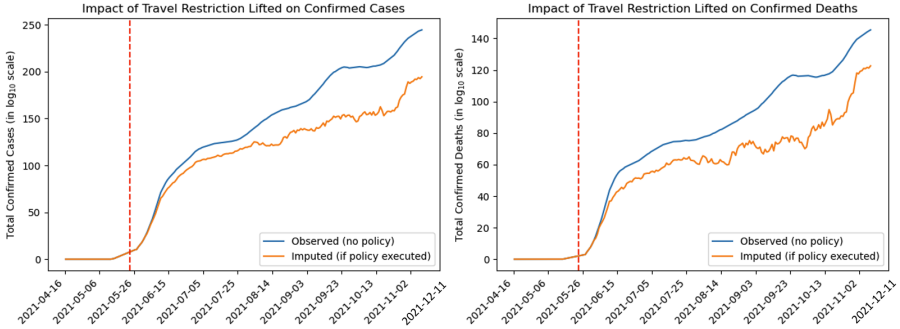

Figure S.4: Comparison of the total confirmed cases (Left) and deaths (Right) across restriction-lifted states between observed results with the travel restrictions lifted and estimated counterfactuals if not. Values are reported in $\log_{10}$ scale, and plots are visualized with 14-day moving averages.

Table S.14: Comparisons of the estimated counterfactual confirmed cases and deaths (in millions) on the last day, if no mandatory vaccination policy were implemented

Method	Confirmed Cases	Deaths
Ground Truth	26.6	0.39
CoDEAL	29.8	0.52
Single AE	34.1	0.61
AEMC-NE	25.3	0.35
DiD	31.5	0.45
Vert-Reg	28.3	–
MC-NNM	27.5	0.41
KernelPCA	15.9	0.33

C.2 Additional Real-World Case Studies

We further evaluate CoDEAL on two additional real-world datasets that are commonly used in the causal panel data literature. Across both examples, CoDEAL produces estimates that are broadly consistent with existing findings.

US Election Day Registration (EDR). We use the dataset in Xu [70] to estimate the effect of Election Day Registration (EDR) laws on voter turnout. CoDEAL estimates that EDR increases turnout by 6.84%. For comparison, the estimated effects are $+1.70\%$ for Single AE, $+9.23\%$ for AEMC-NE, $+3.60\%$ for MC-NNM, $+1.26\%$ for DiD, $+4.30\%$ for Vert-Reg, and $+12.39\%$ for KernelPCA. Overall, CoDEAL yields a positive and moderate estimate, lying within the range of competing methods and supporting the conclusion that EDR has a favorable effect on voter participation.

German Reunification and GDP. We use the dataset in Pang et al. [49] to estimate the effect of German reunification on log-GDP per capita in Western Germany. CoDEAL estimates a decrease of 20.14%. The corresponding estimates from the benchmarks are -21.71% for Single AE, -24.81% for AEMC-NE, -25.94% for MC-NNM, -19.34% for DiD, -21.86% for Vert-Reg, and -22.12% for KernelPCA. In this example, CoDEAL again produces an estimate aligned with the broader literature and close to the competing methods, suggesting a substantial negative impact of reunification on Western Germany's log-GDP per capita in the post-treatment period.

These two additional case studies provide further evidence that CoDEAL yields empirically plausible estimates on real panel datasets and remains stable across applications with different treatment mechanisms and substantive interpretations.

D Further Discussions

D.1 Discussions on Broader Impacts

In this paper, we propose the CoDEAL framework and demonstrate its effectiveness for estimating heterogeneous treatment effects in causal panel data models. Though we presented our methods within the context of panel data, the proposed CoDEAL does not have to be confined to panel data but can accommodate a variety of data types. The main idea of CoDEAL is to integrate nonlinear covariate effect components with nonlinear factor structures to effectively capture the inherent data dependencies and model the data heterogeneity. In panel data models (in the matrix form) with unit-specific covariates (in the vector form), we employ a feed-forward deep neural network to model the covariate effects and a multi-output autoencoder to characterize the nonlinear factor model. Methodologically, with different choices of factor models and advanced neural network architectures, the proposed CoDEAL framework can be further generalized to accommodate various data types with more complex data structures, such as tensors, images, and networks. The broad potential for generalization of CoDEAL to diverse data types brings a broader impact across both research and practice, unlocking innovations well beyond traditional panel data settings. Such generalization can offer new insights and opportunities to the broader research community beyond the field of panel data modeling. In addition, the generalization to more complex data structures can also equip practitioners with a powerful new toolkit for tackling an expanded array of real-world problems.

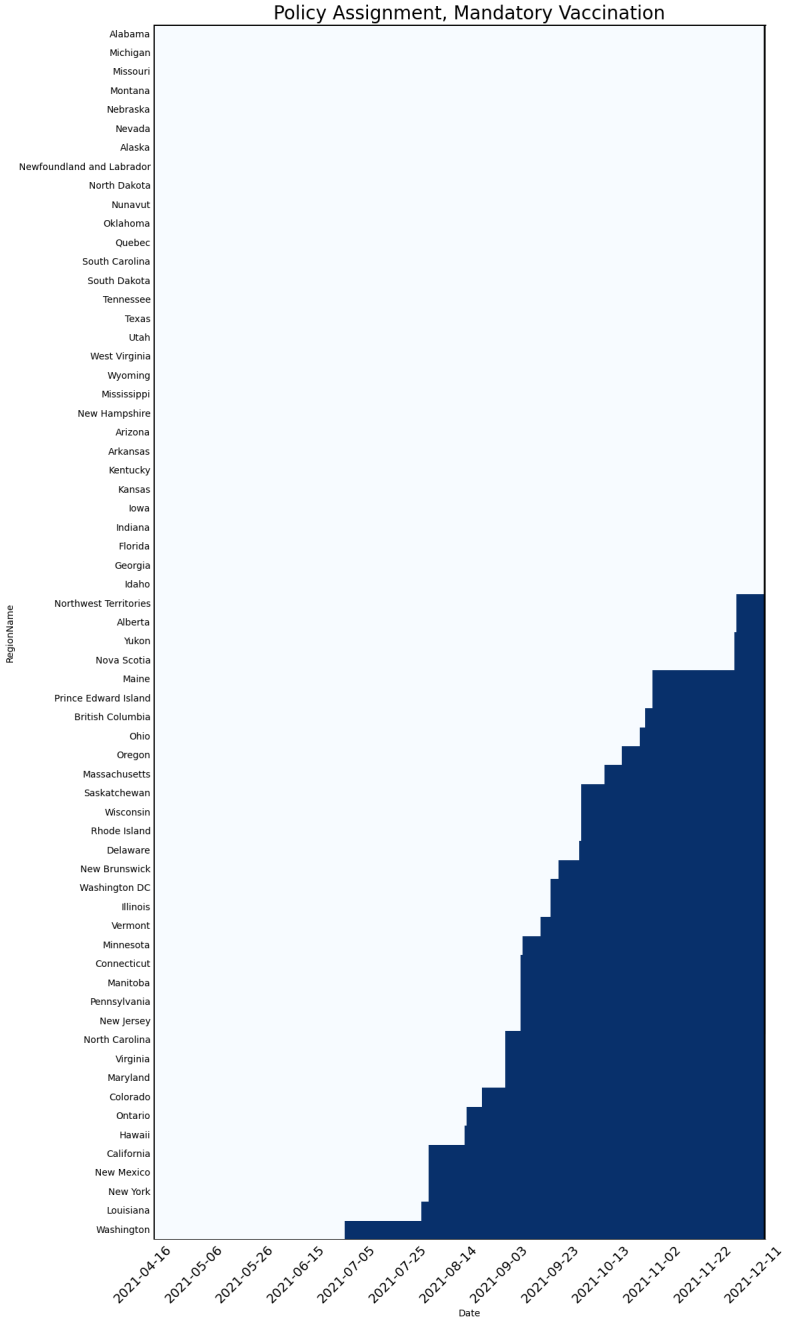

Figure S.5: Indicator matrix of the implementation of the mandatory vaccination policy. Blocks with darker color refer to policy executed.

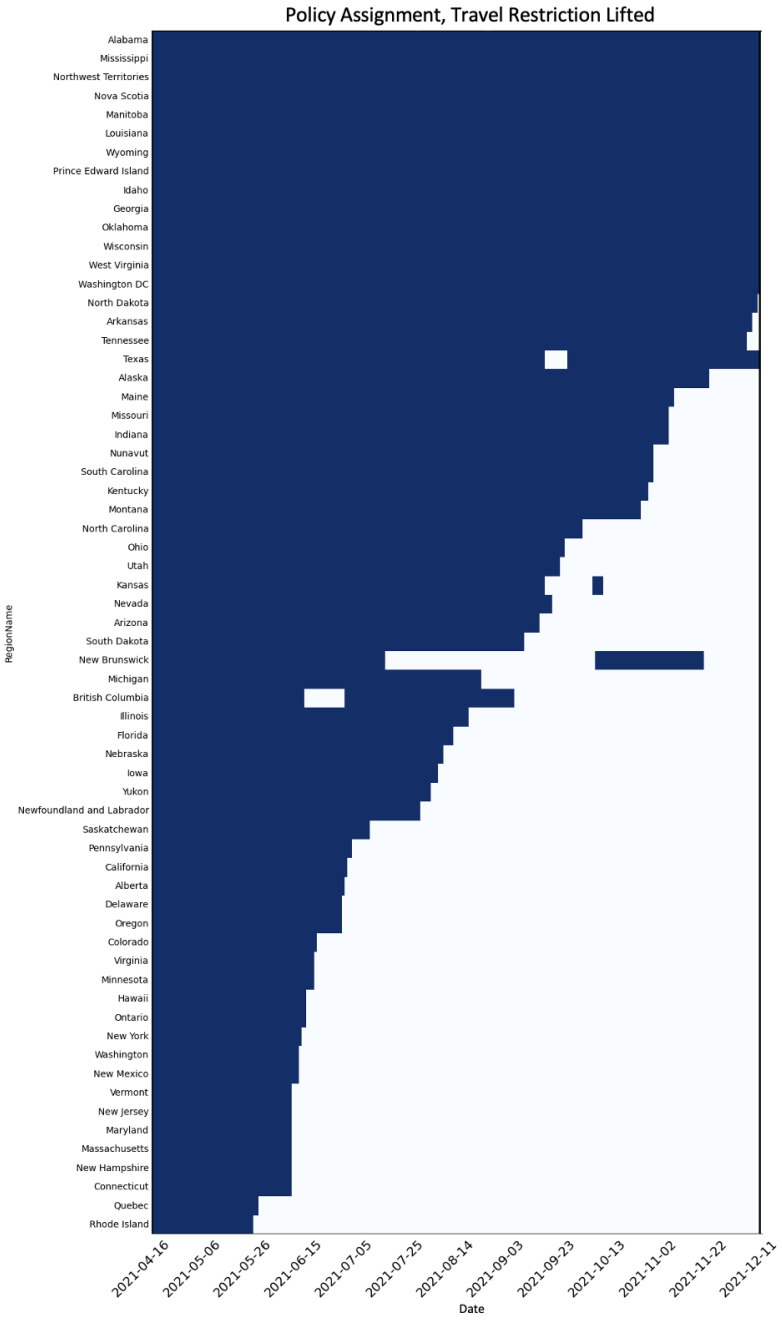

Figure S.6: Indicator matrix of the termination of interval travel restrictions. Blocks with darker color refer to travel restriction lifted.

D.2 Discussions on Further Extensions

The proposed framework largely extends existing causal panel literature in the capacity of accommodating heterogeneity and nonlinearity. We also clarify that several assumptions (e.g., time-invariant covariates, static treatment effects) are not intrinsic to the algorithm and can be relaxed to accommodate more general settings. Moreover, despite the additive model formulation, the proposed framework does not preclude interactions between covariates and latent factors, nor does it require them to be uncorrelated. The CoDEAL framework is amenable to further adaptation across data-driven contexts.

D.2.1 Dynamic Treatment Effects

As discussed in Remark 3, the static treatment assumption is not intrinsic to the proposed algorithm, but can be extended to accommodate time-varying contemporaneous treatment effects. In principle, CoDEAL can produce estimates of unit-time-specific treatment effect τ_{it} using $\widehat{\tau}_{it} = Y_{it}(1) - \widehat{Y}_{it}(0)$. The static-effect assumption enters only the interpretation of the averaged estimator $\widehat{\bar{\tau}}_i := T_2^{-1} \sum_{t:W_{it}=1} \widehat{\tau}_{it}$, not the imputation machinery. CoDEAL's counterfactual imputation $\widehat{Y}_{it}(0)$ is trained exclusively on untreated outcomes, so Theorem 1's guarantee on $\max_{N_1+1\le i\le N} \frac{1}{T_2} \sum_{t=T_1+1}^{T} \mathbb{E}(\widehat{Y}_{it}(0) - \phi_i^\star(\mathbf{F}_t^\star) - g_t^\star(\mathbf{X}_i))^2$ is entirely unaffected by whether the treatment effect is static or dynamic. When the treatment effect is static ($\tau_{it} \equiv \tau_i$), this result establishes the estimation consistency of $\widehat{\bar{\tau}}_i$ for τ_i. When the treatment effect τ_{it} varies over time, Theorem 1 implies that $\widehat{\bar{\tau}}_i$ consistently estimates the unit-wise average treatment effect $\bar{\tau}_i = T_2^{-1} \sum_{t:W_{it}=1} \tau_{it}$. Thus, CoDEAL permits the contemporaneous treatment effect to vary across time and units, while preserving consistency for the average effect.

A pointwise guarantee of $\widehat{\tau}_{it}$ for τ_{it}, however, is not attainable by the current theory. Estimating dynamic effect trajectories with pointwise guarantees would require additional assumptions and substantially more sophisticated techniques, which we leave for future research. We also clarify that the current theoretical framework for CoDEAL does not cover general lagged or fading effects. Modeling such dynamic treatment responses would require potential outcomes indexed by event time or treatment history, together with additional identifying assumptions and theoretical analysis, and we leave these important extensions for future work.

D.2.2 Time-Covariate Interaction and Time-Varying Covariates

While we present the methodology with unit-specific time-invariant covariates $\mathbf{X}_i$, $i \in [N]$, the time-covariate interaction is not absent from the modeling. The covariate effect is allowed to vary over time through $g_t^\star(\mathbf{X}_i)$. The framework also extends straightforwardly to time-varying covariates by replacing $g_t^\star(\mathbf{X}_i)$ with $g_t^\star(\mathbf{X}_{it})$.

D.2.3 Dependence between the Observed Covariates and Latent Factors

Although Model (1) admits an additive formulation in the covariate effect and the latent factor effect, the proposed framework does not preclude the dependence between the observed covariates and the unobserved unit-level heterogeneity. Formally, if the truth is $Y_{it} = \psi(\mathbf{X}_i, \mathbf{U}_i, \mathbf{F}_t^\star) + g_t^\star(\mathbf{X}_i) + \tau_i^\star \mathbf{1}\{W_{it} = 1\} + \varepsilon_{it}$ for arbitrary unobserved heterogeneity $\mathbf{U}_i$ (precisely the case where heterogeneity depends on covariates) then setting $\phi_i^\star(\cdot) := \psi(\mathbf{X}_i, \mathbf{U}_i, \cdot)$ recovers Model (1) exactly. The theory applies verbatim. The multi-output decoder is exactly the estimator of this structure, as a single-output AE (i.e., SingleAE) cannot absorb X-dependent loadings. We also conduct experiments in which $\mathbf{X}_i$ and $\mathbf{F}_t$ are dependent, with details presented in Appendix B.5. The results show that SingleAE shows deteriorating performance as dependence gets stronger, while CoDEAL is robust to dependence and heterogeneity.

What additivity does restrict is when interactions may vary over time only (i) through the fully unrestricted $g_t(\mathbf{X}_i)$ (arbitrary covariate-time interaction), or (ii) through the K-dimensional $\mathbf{F}_t^\star$ via ϕ_i. Excluded are components $\Psi(\mathbf{X}_i, \mathbf{U}_i, t)$ full-rank in t outside the factor span, essentially the premise of any factor-based method, not an additional cost of additivity. Moreover, the g/ϕ split is identified only up to the reallocation discussed in Appendix A.3, and both $\widehat{Y}_{it}(0)$ and Theorem 1 depend on the sum alone; misallocated interaction affects component interpretation but not the ATT.

D.2.4 Potential Extensions to Multiple Treatment Settings

One limitation of CoDEAL is that it is only designed for binary treatment settings. Causal panel data with multiple treatments is an open research area with limited existing work in the literature [1], yet this topic is gradually gaining attention due to its growing relevance in practical applications [6, 61]. To extend CoDEAL to multiple treatment settings, one possible direction is to adopt a similar strategy to [6] in which the treatment type is incorporated as an additional data dimension. Together with the original panel data structures, the inclusion of treatment dimension forms a three-dimensional tensor.

E Supplementary Theoretical Properties and Proofs of Main Theorem

E.1 Preliminary

The set of natural numbers is denoted by $\mathbb{N}$, and we define $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we define the supremum norm as $\|f\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^d} |f(\mathbf{x})|$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, we define the entrywise max norm as $\|A\|_\infty = \max_{i,j} |a_{ij}|$. Let $\mathbb{P}_\mathbf{X}$ denote the distribution of the random vector X. Then, the $\mathcal{L}_2(\mathbb{P}_\mathbf{X})$ norm of a function f is defined as its $\mathcal{L}_2$ norm with respect to the probability measure $\mathbb{P}_\mathbf{X}$, i.e.,

$$\|f\|_{\mathcal{L}_2(\mathbb{P}_\mathbf{X})} := \left(\int |f(\mathbf{y})|^2 \, d\mathbb{P}_\mathbf{X}(\mathbf{y}) \right)^{1/2}.$$

For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ (resp. $a_n \asymp b_n$) if there exists a constant C such that $|a_n| \le C|b_n|$ (resp. $1/C \le a_n/b_n \le C$) for all sufficiently large n, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. Write $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) if there exists a constant C such that $a_n \le Cb_n$ (resp. $a_n \ge Cb_n$). Write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. We use $C, C_1, c, c_1, ...$ to denote generic constants, whose actual values may vary from line to line.

Definition 2. *Let $\beta = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \le 1$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (β, C)-Hölder smooth, if for every $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$, the partial derivative*

$$\frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

exists and satisfies

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(\mathbf{x}) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(\mathbf{z}) \right| \le C \cdot \|\mathbf{x} - \mathbf{z}\|^s$$

for all $x, z \in \mathbb{R}^d$, where $\|\cdot\|$ denotes the Euclidean norm.

E.2 Covariate Adjustment

We begin by analyzing the estimation accuracy of the covariate function $g_t^\star$. For technical convenience, we assume that the covariate function $g_t^\star$ in model (1) admits the following hierarchical decomposition

$$g_t^\star(\mathbf{X}_i) = \sum_{j=1}^P h_{0t,j}(X_{i,j}) + \sum_{j_1<j_2} h_{0t,j_1j_2}(X_{i,j_1}, X_{i,j_2}) + \cdots + \sum_{j_1<\cdots<j_k} h_{0t,j_1j_2\ldots j_k}(X_{i,j_1}, ..., X_{i,j_k}), \tag{S.4}$$

where $k \le P$, each $h_{0t,j}$ is a univariate function and each $h_{0t,j_1\cdots j_k}$ is a k-variate interaction function. Under this decomposition, the smoothness condition in Assumption 2 for $g_t^\star$ is equivalent to the same smoothness condition imposed on all component functions $h_{0t,j_1\cdots j_k}$. To ensure identifiability and enable estimation of these nonparametric components, we impose additional structural constraints to avoid non-identifiable additive shifts. To that end, we make the following assumption.

Assumption 5 (Identifiability and boundedness)**.** *We assume the following conditions for identifiability purposes and boundedness:*

1. *All the k-variate interaction functions in* (S.4) *have 0 marginals, that is,*

$$\int_0^1 h_{0t,j_1\cdots j_k}(x_1, ..., x_k)\, dx_1 = \cdots = \int_0^1 h_{0t,j_1\cdots j_k}(x_1, ..., x_k)\, dx_k = 0.$$

2. $\|g_t^\star\|_\infty \le B$ *for some* $B > 0$.

Note that the above assumption implies that the total integral of $h_{0t,j_1\cdots j_k}$ is also 0. Assumption 5 ensures identifiability of all the functions involved in (S.4).

We now briefly describe our estimation procedure for covariate adjustment. Consider the model in equation (S.4). For each time point t, we estimate the component functions using neural networks $(\psi_{t1}, ..., \psi_{tk})$ by minimizing the squared-error loss:

$$(\hat{\psi}_{t1}, ..., \hat{\psi}_{tk}) = \arg \min_{\psi_{t1}\in\mathcal{F}^1_{NN},...,\psi_{tk}\in\mathcal{F}^k_{NN}} \frac{1}{N}\sum_{i=1}^{N}(1-W_{it})(Y_{it}-\psi_{t1}(\mathbf{X}_i)-\cdots-\psi_{tk}(\mathbf{X}_i))^2. \quad \text{(S.5)}$$

We then define the aggregated estimate as $\hat{g}_{t,\text{raw}} := \hat{\psi}_{t1} + \cdots + \hat{\psi}_{tk}$, and obtain the final estimator $\widehat{g}_t$ by truncating $\hat{g}_{t,\text{raw}}$ at level B:

$$\widehat{g}_t := \text{sgn}(\hat{g}_{t,\text{raw}})(|\hat{g}_{t,\text{raw}}| \wedge B)$$

where B is chosen such that $\|g_t^\star\|_\infty \le B$, as specified in point 2 of Assumption 5. The function classes $\mathcal{F}^k_{NN}, 1 \le k \le p$, are defined as

$$\mathcal{F}^k_{NN} = \mathcal{F}_{NN}(p, c_1 p^k M_k \log M_k, c_2 L_k \log L_k, c_3 p^k (M_k \log M_k)^2 L_k \log L_k, 1) \quad \text{(S.6)}$$

for some constants c_1, c_2, c_3. These are not fully connected networks over the entire covariate vector; rather, we use fully connected neural networks to estimate each k-variate component function separately, with width $M_k \log M_k$ and depth $L_k \log L_k$, and then sum the outputs. With slight abuse of notation, we continue to denote these specialized architectures by $\mathcal{F}^k_{NN}$. The parameters M_k and L_k typically depend on the sample size N and will be specified later. As with any nonparametric estimator, the *generalization error* of $\widehat{g}_t$ can be decomposed into two components:

$$\|\widehat{g}_t - g_t^\star\|_{\mathcal{L}_2(\mathbb{P}_\mathbf{X})} \le \underbrace{\|\widehat{g}_t - g_{0t}\|_{\mathcal{L}_2(\mathbb{P}_\mathbf{X})}}_{\text{statistical error}} + \underbrace{\|g_{0t} - g_t^\star\|_{\mathcal{L}_2(\mathbb{P}_\mathbf{X})}}_{\text{approximation error}}, \quad \text{(S.7)}$$

where g_{0t} denotes the best approximation to $g_t^\star$ in $\mathcal{L}_2(\mathbb{P}_\mathbf{X})$ norm within the specified class of neural networks. The following lemma presents a bound on the overall generalization error, combining both statistical and approximation errors.

Lemma 1. *Under Assumptions 1, 2, 5, the estimator* $\widehat{g}_t$ *of* $g_t^\star$ *defined in equation* (S.5) *satisfies*

$$\mathbb{E}[(\widehat{g}_t(\mathbf{X}_i) - g_t^\star(\mathbf{X}_i))^2 \mid \mathbf{X}] = O_\mathbb{P}\left(P^k N^{-\frac{2\beta}{2\beta+k}} \log^4 N\right).$$

The proof of this lemma follows the same argument as Theorem 2.11 in [16] and is therefore omitted. A careful examination of that proof also yields the following proposition in a four-block design.

Proposition 1. *Under Assumptions 1, 2, 5, the estimator* $\widehat{g}_t$ *of* $g_t^\star$ *defined in equation* (S.5) *satisfies*

$$\max_{N_1+1\le i\le N} \frac{1}{T_2}\sum_{t=T_1+1}^{T_1+T_2} \mathbb{E}[(\widehat{g}_t(\mathbf{X}_i) - g_t^\star(\mathbf{X}_i))^2 \mid \mathbf{X}] = O_\mathbb{P}\left(P^k N^{-\frac{2\beta}{2\beta+k}} \log^5 N\right).$$

When P is fixed and a fully connected neural network is used to estimate $g_t^\star$, this is equivalent to setting $k = P$ in Lemma 1 and Proposition 1 above.

E.3 Proof of Theorem 1

Consider the covariate-adjusted observed outcomes as

$$\widetilde{Y}_{it}(0) = Y_{it}(0) - \widehat{g}_t(\mathbf{X}_i) = \phi_i^\star(\mathbf{F}_t^\star) + g_t^\star(\mathbf{X}_i) - \widehat{g}_t(\mathbf{X}_i) + \varepsilon_{it} = \phi_i^\star(\mathbf{F}_t^\star) + \varepsilon_{it} + o_\mathbb{P}(1).$$

In a four-block design, only within blocks $\mathcal{A}, \mathcal{B}, \mathcal{C}$, we have untreated outcomes $\widetilde{\mathbf{Y}}(0)$. Let $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})^\top$. Slightly abusing notation, we first consider the case without covariates, namely $\widetilde{Y}_{it}(0) = \phi_i^\star(\mathbf{F}_t^\star) + \varepsilon_{it}$. Let $\widetilde{\mathbf{Y}}_t(0) = \boldsymbol{\phi}^\star(\mathbf{F}_t^\star) + \varepsilon_t$, where $\boldsymbol{\phi}^\star = (\phi_1, ...\phi_N)^\top$. We then can apply the triangle inequality to incorporate the additional estimation error arising from covariate adjustment.

Define the function class of the multi-output autoencoder is

$$\mathcal{G}_{\mathrm{AE}}^{K_1} := \Big\{(\phi_1,\ldots,\phi_N)\circ\rho : \rho\in\mathcal{G}_N^{K_1}(L_1,\mathbf{d}_1,B,T^{5\beta+5}),\ \phi_i\in\mathcal{G}_{K_1}^1(L_2,\mathbf{d}_2,B,T^{5\beta+5}),\ i\in[N]\Big\},$$

where $K_1\geq K$. The proof of Theorem 1 requires the following additional conditions

$$\begin{aligned}
&K\leq K_1\leq \min\{\max_{1\leq\ell\leq L_1} d_{1,\ell},\max_{1\leq\ell\leq L_2} d_{2,\ell}\},\ L_1\asymp L_2\asymp\log(T),\\
&\max_{1\leq\ell\leq L_1} d_{1,\ell}\asymp\max_{1\leq\ell\leq L_2} d_{2,\ell}\asymp T^{\frac{K}{2(2\beta+K)}},\\
&\text{total number of weight parameters in the encoder is } O(n+T^{\frac{K}{2\beta+K}}\log T).
\end{aligned}\tag{S.8}$$

Let $d_1=\max_{1\leq\ell\leq L_1} d_{1,\ell}, d_2=\max_{1\leq\ell\leq L_2} d_{2,\ell}$.

Proof. Using the facts that

$$\frac{1}{T_1}\sum_{t=1}^{T_1}\Big(\phi_i^\star(\mathbf{F}_t^\star)-\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\Big)^2\leq 4B^2,\exp(-cT)=o(T^{-1}),\exp(-cn)=o(n^{-1})$$

and Proposition 2, we have

$$\begin{aligned}
&\frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}\Big(\phi_i^\star(\mathbf{F}_t^\star)-\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\Big)^2\\
\lesssim&\Big(n^{-1}K_1+Nn^{-1}T^{-\frac{2\beta}{2\beta+K}}+Nn^{-2}\Big)\log^4(T)+4B^2\cdot(\exp(-cT)+\exp(-cn))\\
\lesssim&\Big(n^{-1}K_1+Nn^{-1}T^{-\frac{2\beta}{2\beta+K}}+Nn^{-2}\Big)\log^4(T)+4B^2\cdot o(T^{-1}+n^{-1})\\
\lesssim&\Big(n^{-1}K_1+Nn^{-1}T^{-\frac{2\beta}{2\beta+K}}+Nn^{-2}\Big)\log^4(T)+T^{-1}.
\end{aligned}\tag{S.9}$$

Assuming that $\{(\widetilde{\mathbf{F}}_t^\star,\widetilde{\varepsilon}_t)_{t=1}^T,\check{\mathbf{Y}}(0)\}$ is an independent copy of $\{(\mathbf{F}_t^\star,\varepsilon_t)_{t=1}^T,\widetilde{\mathbf{Y}}(0)\}$, we have, for any $T_1+1\leq s\leq T$,

$$\begin{aligned}
&\left|\mathbb{E}\Big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0)))-\phi_i^\star(\mathbf{F}_s^\star)\Big)^2-\frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}\Big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))-\phi_i^\star(\mathbf{F}_t^\star)\Big)^2\right|\\
=&\left|\frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}\Big(\widehat{\phi}_i(\widehat{\rho}(\check{\mathbf{Y}}_t(0)))-\phi_i^\star(\widetilde{\mathbf{F}}_t^\star)\Big)^2-\frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}\Big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))-\phi_i^\star(\mathbf{F}_t^\star)\Big)^2\right|.
\end{aligned}\tag{S.10}$$

Observe that $\phi_i\circ\rho\in\mathcal{G}_N^1(L_1+L_2,\mathbf{d}_1+\mathbf{d}_2,B,T^{5\beta+5})$ and its total number of non-zero weights are asymptotically bounded by $n+T^{\frac{K}{2\beta+K}}\log T$. Therefore, by Lemma 7, there exists

$$\mathcal{G}_{N,\delta}^1(L_1+L_2,\mathbf{d}_1+\mathbf{d}_2,B,T^{5\beta+5})$$

with $\delta=T^{-1}$ satisfying $\log|\mathcal{G}_{N,\delta}^1|\lesssim(n+T^{\frac{K}{2\beta+K}})\log^4(NT)$. In addition, for any $\phi_i\circ\rho$ satisfying the conditions in (S.8) and Theorem 1, there exists a function $f(\cdot)\in\mathcal{G}_{N,\delta}^1$ such that $\|\phi_i\circ\rho(x)-f(x)\|\leq T^{-1}$ whenever $\|x\|_\infty\leq 2NT$. Without loss of generality, we can assume all the functions in $\mathcal{G}_{N,\delta}^1$ are bounded by B. Write $\mathcal{G}_{N,\delta}^1=\{f_j,j\in[|\mathcal{G}_{N,\delta}^1|]\}$ and assume $f_j\in\mathcal{G}_{N,\delta}^1$ such that

$$\|\widehat{\phi}_i\circ\widehat{\rho}(x)-f_{j^\star}(x)\|_\infty\leq\delta=T^{-1},\tag{S.11}$$

for $\|x\|_\infty\leq 2NT$. Note that by Lemma 3 and the fact that $\|\phi_i^\star(\mathbf{F}_t^\star)\|_\infty\lesssim 1,\|\phi_i^\star(\widetilde{\mathbf{F}}_t^\star)\|_\infty\lesssim 1$, it is easy to show that with probability at least $1-C\exp(-cT)$, $\|\widetilde{\mathbf{Y}}_{\mathcal{A}}(0)\|_\infty\leq 2NT,\|\widetilde{\mathbf{Y}}_{\mathcal{B}}(0)\|_\infty\leq 2NT,\|\widetilde{\mathbf{Y}}_{\mathcal{C}}(0)\|_\infty\leq 2NT,\|\check{\mathbf{Y}}(0)\|_\infty\leq 2NT$. Therefore, under the same event, (S.11) holds for $\widetilde{\mathbf{Y}}_1(0),...,\widetilde{\mathbf{Y}}_T(0),\check{\mathbf{Y}}_1(0),...,\check{\mathbf{Y}}_T(0)$. Let us define

$$r_{ij}:=\max\left(\sqrt{T^{-1}\log|\mathcal{G}_{N,\delta}^1|},\sqrt{\mathbb{E}\left[\Big(f_j(\widetilde{\mathbf{Y}}_s(0))-\phi_i^\star(\mathbf{F}_s^\star)\Big)^2\right]}\right),$$

$$V_i := \max_{j\in[|\mathcal{G}^1_{N,\delta}|]} \left| \frac{1}{r_{ij}B} \sum_{t=1}^{T_1} \left[(f_j(\check{\mathbf{Y}}_t(0)) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star))^2 - (f_j(\widetilde{\mathbf{Y}}_t(0)) - \phi_i^\star(\mathbf{F}_t^\star))^2 \right] \right|,$$

for some $s > T$. Then by (S.11), we have

$$\begin{aligned} r_{ij^\star} &= \max\left(\sqrt{T^{-1}\log|\mathcal{G}^1_{N,\delta}|}, \mathbb{E}^{1/2}\left[\left(f_{j^\star}(\widetilde{\mathbf{Y}}_s(0)) - \phi_i^\star(\mathbf{F}_s^\star) \right)^2 \big| \widetilde{\mathbf{Y}}(0) \right] \right) \\ &\le \sqrt{T^{-1}\log|\mathcal{G}^1_{N,\delta}|} + \mathbb{E}^{1/2}\left[\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0))) - \phi_i^\star(\mathbf{F}_s^\star) \right)^2 \big| \widetilde{\mathbf{Y}}(0) \right] + T^{-1}. \end{aligned} \tag{S.12}$$

Moreover, by (S.11), with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned} &\left| \frac{1}{T_1}\sum_{t=1}^{T_1} \left(\widehat{\phi}_i(\widehat{\rho}(\check{\mathbf{Y}}_t(0))) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star) \right)^2 - \frac{1}{T_1}\sum_{t=1}^{T_1} \left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star) \right)^2 \right| \\ \le &\left| \frac{1}{T_1}\sum_{t=1}^{T_1} \left(f_{j^\star}(\check{\mathbf{Y}}_t(0)) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star) \right)^2 - \frac{1}{T_1}\sum_{t=1}^{T_1} \left(f_{j^\star}(\widetilde{\mathbf{Y}}_t(0)) - \phi_i^\star(\mathbf{F}_t^\star) \right)^2 \right| + 9T^{-1}B. \end{aligned}$$

Using the fact that $\exp(-cT) = o(T^{-1})$, it follows that

$$\begin{aligned} &\left| \frac{1}{T_1}\sum_{t=1}^{T_1} \mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\check{\mathbf{Y}}_t(0))) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star) \right)^2 - \frac{1}{T_1}\sum_{t=1}^{T_1} \mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star) \right)^2 \right| \\ \le &\frac{1}{T_1}\sum_{t=1}^{T_1} \mathbb{E}\left(f_{j^\star}(\check{\mathbf{Y}}_t(0)) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star) \right)^2 - \frac{1}{T_1}\sum_{t=1}^{T_1} \mathbb{E}\left(f_{j^\star}(\widetilde{\mathbf{Y}}_t(0)) - \phi_i^\star(\mathbf{F}_t^\star) \right)^2 + 10T^{-1}B \\ \le &\frac{B}{T_1}\mathbb{E}(V_i r_{ij^\star}) + 10T^{-1}B. \end{aligned} \tag{S.13}$$

By Cauchy-Schwarz inequality and (S.12), we have

$$\begin{aligned} \mathbb{E}(V_i r_{ij^\star}) \le &\left(\mathbb{E}\left(\mathbb{E}\left[\left(\widehat{\phi}_i \circ \widehat{\rho}(\widetilde{\mathbf{Y}}_s(0)) - \phi_i^\star(\mathbf{F}_s^\star) \right)^2 \big| \widetilde{\mathbf{Y}}(0) \right] \right)^2 \right)^{1/2} (\mathbb{E}V_i^2)^{1/2} \\ &+ \sqrt{T^{-1}\log|\mathcal{G}^1_{N,\delta}|}\,\mathbb{E}V_i + T^{-1}\mathbb{E}V_i \\ = &\left(\mathbb{E}\left(\widehat{\phi}_i \circ \widehat{\rho}(\widetilde{\mathbf{Y}}_s(0)) - \phi_i^\star(\mathbf{F}_s^\star) \right)^2 \right)^{1/2} (\mathbb{E}V_i^2)^{1/2} + \sqrt{T^{-1}\log|\mathcal{G}^1_{N,\delta}|}\,\mathbb{E}V_i + T^{-1}\mathbb{E}V_i. \end{aligned} \tag{S.14}$$

Then we need to bound V_i. For any fixed $j \in [|\mathcal{G}^1_{N,\delta}|]$, define

$$Z_{tj} := [f_j(\check{\mathbf{Y}}_t(0)) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star)]^2 - [f_j(\widetilde{\mathbf{Y}}_t(0)) - \phi_i^\star(\mathbf{F}_t^\star)]^2.$$

Then it is straightforward to see that $\{Z_{tj}\}_{t=1}^T$ is a series of i.i.d. random variables satisfying $\mathbb{E}Z_{tj} = 0$, $|Z_{tj}| \le 4B^2$, and

$$\mathrm{Var}(Z_{tj}) = 2\,\mathrm{Var}\left(\left(f_j(\check{\mathbf{Y}}_t(0)) - \phi_i^\star(\widetilde{\mathbf{F}}_t^\star) \right)^2 \right) \le 8B^2 r_{ij}^2.$$

Applying Bernstein's inequality, we will obtain

$$\mathbb{P}\left(V_i \ge x\right) = \mathbb{P}\left(\max_{j\in[|\mathcal{G}^1_{N,\delta}|]} \left| \frac{1}{r_{ij}B}\sum_{t=1}^{T_1} Z_{tj} \right| \ge x \right) \le 2|\mathcal{G}^1_{N,\delta}| \max_{j\in[|\mathcal{G}^1_{N,\delta}|]} \exp\left(-\frac{x^2}{8x/(3r_{ij}) + 16T} \right).$$

Since $r_{ij} \ge \sqrt{T^{-1}\log|\mathcal{G}^1_{N,\delta}|}$, setting $x \ge 6\sqrt{T\log|\mathcal{G}^1_{N,\delta}|}$, it holds that

$$\mathbb{P}(V_i \ge x) \le 2\exp\left(-3x\sqrt{\log|\mathcal{G}^1_{N,\delta}|}/(16\sqrt{T}) \right).$$

Thus, for sufficiently large T,

$$\mathbb{E}V_i = \int_0^\infty \mathbb{P}(V_i \geq x)\, dx \leq 6\sqrt{T\log|\mathcal{G}^1_{N,\delta}|} + \int_{6\sqrt{T\log|\mathcal{G}^1_{N,\delta}|}}^\infty 2|\mathcal{G}^1_{N,\delta}| \exp\left(\frac{-3x\sqrt{\log|\mathcal{G}^1_{N,\delta}|}}{16\sqrt{T}}\right) dx$$
$$\leq 6\sqrt{T\log|\mathcal{G}^1_{N,\delta}|} + \frac{32}{3}\cdot\sqrt{\frac{T}{\log|\mathcal{G}^1_{N,\delta}|}} < 7\sqrt{T\log|\mathcal{G}^1_{N,\delta}|}.$$

Similarly, we can show $\mathbb{E}V_i^2 \leq 40T\log|\mathcal{G}^1_{N,\delta}|$. By (S.14), since $\log|\mathcal{G}^1_{N,\delta}| \lesssim (T^{\frac{K}{2\beta+K}} + n)\log^4(NT)$, it follows that

$$\begin{aligned}
&\frac{B}{T_1}\mathbb{E}(V_i r_{ij^\star}) \\
\lesssim&\frac{B}{T_1}\left(\sqrt{\mathbb{E}\left(\widehat{\phi}_i\circ\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0)) - \phi_i^\star(\mathbf{F}_s^\star)\right)^2}\sqrt{T\log|\mathcal{G}^1_{N,\delta}|} + \sqrt{T^{-1}\log|\mathcal{G}^1_{N,\delta}|} + \log|\mathcal{G}^1_{N,\delta}|\right) \\
\lesssim&\left(T^{-\frac{\beta}{2\beta+K}} + T^{-1/2}n^{1/2}\right)\log^2(NT)\sqrt{\mathbb{E}\left(\widehat{\phi}_i\circ\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0)) - \phi_i^\star(\mathbf{F}_s^\star)\right)^2} + \left(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}n\right)\log^4(NT).
\end{aligned}$$

Employing (S.10), (S.13), we have, for $s > T, i > N_1$,

$$\begin{aligned}
&\left|\mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0))) - \phi_i^\star(\mathbf{F}_s^\star)\right)^2 - \frac{1}{T_1}\sum_{t=1}^{T_1}\mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star)\right)^2\right| \\
\lesssim&\left(T^{-\frac{\beta}{2\beta+K}} + T^{-1/2}n^{1/2}\right)\log^2(NT)\sqrt{\mathbb{E}\left(\widehat{\phi}_i\circ\widehat{\rho}(\widetilde{\mathbf{Y}}_{\mathcal{B},s}(0)) - \phi_i^\star(\mathbf{F}_s^\star)\right)^2} + \left(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}n\right)\log^4(NT).
\end{aligned} \tag{S.15}$$

By basic inequality and (S.9), we have with probability at least $1 - C\exp(-cT)$, for $s > T, i > N_1$,

$$\begin{aligned}
&\left|\mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0))) - \phi_i^\star(\mathbf{F}_s^\star)\right)^2\right| \\
\lesssim&\left(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}n\right)\log^4(NT) + (n^{-1}K_1 + Nn^{-1}T^{-\frac{2\beta}{2\beta+K}} + Nn^{-2})\log^4(T) + T^{-1}.
\end{aligned}$$

As $\exp(-cT) = o(T^{-1})$ and $N_1 \asymp N, T_1 \asymp T$, we further conclude with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\frac{1}{T_2}\sum_{s=T_1+1}^{T}\mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_s(0))) - \phi_i^\star(\mathbf{F}_s^\star)\right)^2 \\
\lesssim&\left((N+n)n^{-1}T^{-\frac{2\beta}{2\beta+K}} + T^{-1}n + n^{-1}K_1 + Nn^{-2}\right)\log^4(NT).
\end{aligned} \tag{S.16}$$

Therefore, by the triangle inequality, Proposition 1 and (S.16), we have

$$\begin{aligned}
&\frac{1}{T_2}\sum_{t=T_1+1}^{T}\mathbb{E}\left(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) + \widehat{g}_t(\mathbf{X}_i) - \phi_i^\star(\mathbf{F}_t^\star) - g_t^\star(\mathbf{X}_i)\right)^2 \\
=&O_{\mathbb{P}}\left(\left((N+n)n^{-1}T^{-\frac{2\beta}{2\beta+K}} + T^{-1}n + n^{-1}K_1 + Nn^{-2}\right)\log^4(NT) + N^{-\frac{2\beta}{2\beta+k}}\log^5 N\right).
\end{aligned}$$

□

E.4 Technical Lemmas

Proposition 2. *Under the condition of Theorem 1, with probability at least* $1 - C\exp(-cT) - C\exp(-cn)$*, for* N, T *sufficiently large, we have*

$$\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\left(\phi_i^\star(\mathbf{F}_t^\star) - \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\right)^2 \lesssim \left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + n^{-1}\right)\log^4(T),$$

and

$$\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\phi_i^\star(\mathbf{F}_t^\star)-\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big\|^2\lesssim\Big(n^{-1}K_1+Nn^{-1}T^{-\frac{2\beta}{2\beta+K}}+Nn^{-2}\Big)\log^4(T).$$

Proof. For simplicity, denote $\mathcal{G}_1=\mathcal{G}_N^{K_1}(L_1,\mathbf{d}_1,B,T^{5\beta+5})$, $\mathcal{G}_2=\mathcal{G}_{K_1}^1(L_2,\mathbf{d}_2,B,T^{5\beta+5})$. By the definition of $(\widehat{\phi}_1,\ldots,\widehat{\phi}_N)\circ\widehat{\rho}$, for any $(\phi_1,\ldots,\phi_N)\circ\rho\in\mathcal{G}_{AE}^{K_1}$, it holds that

$$\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\widetilde{Y}_{it}(0)-\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big)^2\le\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\widetilde{Y}_{it}(0)-\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)^2.$$

As a consequence, we obtain

$$\begin{aligned}
&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star)-\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big)^2\\
=&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\Big[\big(\widetilde{Y}_{it}(0)-\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big)^2-\varepsilon_{it}^2+2\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\varepsilon_{it}-2\phi_i^\star(\mathbf{F}_t^\star)\varepsilon_{it}\Big]\\
\le&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\Big[\big(\widetilde{Y}_{it}(0)-\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)^2-\varepsilon_{it}^2+2\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\varepsilon_{it}-2\phi_i^\star(\mathbf{F}_t^\star)\varepsilon_{it}\Big]\\
=&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\Big[\big(\phi_i^\star(\mathbf{F}_t^\star)-\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)^2+2\varepsilon_{it}\Big(\big(\phi_i^\star(\mathbf{F}_t^\star)-\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)+\big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))-\phi_i^\star(\mathbf{F}_t^\star)\big)\Big)\Big].
\end{aligned}\tag{S.17}$$

Below, we establish three inequalities, denoted (I) to (III).

(I): Assume $(\phi_1,\ldots,\phi_N)\circ\rho\in\mathcal{G}_{AE}^{K_1}$. As long as $\phi_1,\ldots,\phi_N$ are deterministic, with probability at least $1-C\exp(-cT)$,

$$\begin{aligned}
&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star)-\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)\\
\lesssim&T^{1/2}K_1^{1/2}\log(T)\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star)-\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)^2\right)^{1/2}+K_1^{1/2}T^{1/2}N^{1/2}\log(T)+N.
\end{aligned}$$

(II): With probability at least $1-C\exp(-cT)$, we have

$$\begin{aligned}
&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))-\phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\\
\lesssim&\Big(NT^{\frac{K}{2\beta+K}}+TK_1\Big)^{1/2}\log^2(T)\left(\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))-\phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2}+N^{1/2}\right)+N.
\end{aligned}$$

(III): There exists $(\phi_1^\dagger,\ldots,\phi_N^\dagger)\circ\rho^\dagger\in\mathcal{G}_{AE}^{K_1}$ such that $\phi_1^\dagger,\ldots,\phi_N^\dagger$ are deterministic and

$$\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star)-\phi_i^\dagger(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0)))\big)^2\lesssim T^{-\frac{2\beta}{2\beta+K}}+T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\big\|\rho(\widetilde{\mathbf{Y}}_t(0))-\mathbf{F}_t^\star\big\|^2.$$

Step 1: Proof of (I).

Write

$$\epsilon_T=K_1^{-1}T^{-1-(5\beta+5)(L_2+1)}d_2^{-L_2}$$

and define the set

$$S=\big\{x\in\mathbb{R}^{K_1}:\ \forall\,1\le i\le K_1,\ \exists\,1\le j(\in\mathbb{N})\le 2B\epsilon_T^{-1},\ \text{s.t. } x_i=-B+j\epsilon_T\big\}.\tag{S.18}$$

By definition, we see that for any $F \in [-B, B]^{K_1}$, there exists a $\bar{F} \in S$ such that $\|F - \bar{F}\|_\infty \leq \epsilon_T$. Write the set as $\{\bar{F}_k,\, 1 \leq k \leq |S|\}$. It is not hard to verify that

$$\log |S| \leq K_1 \log\Big(2BK_1 T^{1+(5\beta+5)(L_2+1)} d_2^{L_2}\Big) \lesssim K_1 \log^2(T). \tag{S.19}$$

For $1 \leq t \leq T$, assume $\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \bar{F}_{k_t^\star}\|_\infty \leq \epsilon_T$. As a consequence, by Lemma 6,

$$\big|\phi_i(\rho(\widetilde{\mathbf{Y}}_t(0))) - \phi_i(\bar{F}_{k_t^\star})\big| \leq T^{-1}. \tag{S.20}$$

Together with Lemma 3, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big) \\
&\lesssim \sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t^\star})\big) + N \\
&\leq \max_{\substack{k_t\in[|S|]\\ 1\leq t\leq T}} \frac{\Big|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} \left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t^\star})\big)^2\right)^{1/2} + N \\
&\lesssim \max_{\substack{k_t\in[|S|]\\ 1\leq t\leq T}} \frac{\Big|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} \left(\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\rho(\widetilde{\mathbf{Y}}_t(0)))\big)^2\right)^{1/2} + N^{1/2}\right) + N,
\end{aligned} \tag{S.21}$$

where we use (S.20) and Lemma 3 in the first inequality and (S.20) in the last inequality. For any $u \geq 0$, by Lemma 2 and the property of subGaussian distribution, as well as the fact that $\phi_1, \ldots, \phi_N$, $\bar{F}_{k_1}, \ldots, \bar{F}_{k_T}$, and $\{\phi_i^\star(\mathbf{F}_t^\star)\}_{i,t}$ are deterministic given $\{\mathbf{F}_t^\star\}_{t=1}^T$,

$$\mathbb{P}\left(\frac{\Big|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} > u \,\middle|\, \{\mathbf{F}_t^\star\}_{t=1}^T\right) \leq 2\exp\left(-\frac{u^2}{2\sigma_\epsilon^2}\right).$$

Therefore, by union bound inequality and the arbitrariness of $\{\mathbf{F}_t^\star\}_{t=1}^T$ in the above inequality,

$$\mathbb{P}\left(\max_{\substack{k_t\in[|S|]\\ 1\leq t\leq T}} \frac{\Big|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \varepsilon_{it}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} > u\right) \leq 2|S|^T \exp\left(-\frac{u^2}{2\sigma_\epsilon^2}\right).$$

By (S.19), (S.21), and setting $u = 2\sigma_\epsilon\,(T\log|S|)^{1/2}$, we complete the proof for (I).

Step 2: Proof of (II).

To prove this assertion, we rely on Lemma 7. The lemma states that for $\delta = T^{-1}$, there exists a function class $\mathcal{G}_{2,\delta} \subset \mathcal{G}_2$ with

$$\log|\mathcal{G}_{2,\delta}| \lesssim \big(d_2^2 L_2 + K_1 d_2\big)\,\log\Big(8\delta^{-1}BT^{(5\beta+5)(L_2+1)}(1+d_2)^{L_2}K_1L_2\Big) \lesssim T^{\frac{K}{2\beta+K}}\log^4(T) \tag{S.22}$$

such that for any $\phi \in \mathcal{G}_2$, there exists $\bar{\phi} \in \mathcal{G}_{2,\delta}$ meeting the requirement $|\bar{\phi}(x) - \phi(x)| \leq T^{-1}$, $\forall \|x\|_\infty \leq B$. Since $\|\phi\|_\infty \leq B$, we have $|\bar{\phi}(x)| \leq T^{-1} + B$. Hence we can assume all the functions in $\mathcal{G}_{2,\delta}$ are bounded by $2B$ for simplicity, i.e., $\mathcal{G}_{2,\delta} \in \mathcal{G}_{K_1}^1(L_2, \mathbf{d}_2, 2B, T^{5\beta+5})$. Write $\mathcal{G}_{2,\delta} = \{\bar{\phi}_j,\, 1 \leq j \leq |\mathcal{G}_{2,\delta}|\}$. By definition, there exists $\ell_i^\star \in [|\mathcal{G}_{2,\delta}|]$ such that $|\widehat{\phi}_i(x) - \bar{\phi}_{\ell_i^\star}(x)| \leq T^{-1}, \forall \|x\|_\infty \leq B$. Additionally, recall the definition of S in (S.18), there exists $k_t^\star \in [|S|]$ such that $\|\bar{F}_{k_t^\star} - \widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))\|_\infty \leq \epsilon_T$. Since $\widehat{\phi}_i \in \mathcal{G}_{K_1}^1(L_2, \mathbf{d}_2, B, T^{5\beta+5})$, by Lemma 6, we have

$$\big|\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \bar{\phi}_{\ell_i^\star}(\bar{F}_{k_t^\star})\big| \leq \big|\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \widehat{\phi}_i(\bar{F}_{k_t^\star})\big| + \big|\widehat{\phi}_i(\bar{F}_{k_t^\star}) - \bar{\phi}_{\ell_i^\star}(\bar{F}_{k_t^\star})\big| \leq 2T^{-1}.$$

Together with Lemma 2, with probability at least $1 - C\exp(-cT)$ we have

$$
\begin{aligned}
&\left|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\right| \\
&\lesssim \left|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i^\star}(\bar{F}_{k_t^\star}) - \phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\right| + N \\
&\leq \max_{\substack{k_t\in[|S|],\,1\le t\le T\\ \ell_i\in[|\mathcal{G}_{2,\delta}|],\,1\le i\le N}} \frac{\left|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i}(\bar{F}_{k_t}) - \phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\right|}{\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i}(\bar{F}_{k_t}) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2}}\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i^\star}(\bar{F}_{k_t^\star}) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2} + N \\
&\lesssim \max_{\substack{k_t\in[|S|],\,1\le t\le T\\ \ell_i\in[|\mathcal{G}_{2,\delta}|],\,1\le i\le N}} \frac{\left|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i}(\bar{F}_{k_t}) - \phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\right|}{\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i}(\bar{F}_{k_t}) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2}}\left(\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2} + N^{1/2}\right) + N
\end{aligned}
\tag{S.23}
$$

By Lemma 2, the property of sub-Gaussian distribution, as well as the union bound inequality and the arbitrariness of $\{\mathbf{F}_t^\star\}_{t=1}^T$, we obtain

$$
\mathbb{P}\left(\max_{\substack{k_t\in[|S|],\,1\le t\le T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],\,1\le i\le N}} \frac{\left|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i}(\bar{F}_{k_t}) - \phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\right|}{\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\bar{\phi}_{\ell_i}(\bar{F}_{k_t}) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2}} > u\right) \leq 2|S|^T\cdot|\mathcal{G}_{2,\delta}|^N\exp\left(-\frac{u^2}{2\sigma_\epsilon^2}\right).
$$

The conclusion follows by (S.19), (S.22) and (S.23) with $u = 2\sigma_\epsilon\big(T\log|S| + N\log|\mathcal{G}_{2,\delta}|\big)^{1/2}$.

Step 3: Proof of (III).

It is sufficient to construct $(\phi_1^\dagger,\ldots,\phi_N^\dagger)\circ\rho^\dagger \in \mathcal{G}_{AE}^K$ satisfying the inequality. If $K_1 > K$, we can introduce useless neurons into the bottleneck with zero weights and biases, ensuring that the output of the network remains unchanged.

Assume

$$
\rho^\dagger(\cdot) := \arg\min_{\rho\in\mathcal{G}_N^K}\sum_{t=1}^T\big\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \mathbf{F}_t^\star\big\|^2.
$$

According to Proposition 3, there exist deterministic $\phi_i^\dagger \in \mathcal{G}_1^K(L_2,\mathbf{d}_2,B,T^{5\beta+5})$ such that:

$$
\big|\phi_i^\dagger(x) - \phi_i^\star(x)\big| \lesssim T^{-\frac{\beta}{2\beta+K}}, \qquad \forall x\in[-B,B]^K \text{ and } i=1,\ldots,N. \tag{S.24}
$$

By the triangle inequality, the following inequality holds:

$$
\begin{aligned}
&\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i^\dagger(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0)))\big)^2 \\
&\leq \frac{2}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \phi_i^\star(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0)))\big)^2 \\
&\quad + \frac{2}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\dagger(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0)))\big)^2 =: A_1 + A_2.
\end{aligned}
$$

By Assumptions 2, we confirm:

$$
\big|\phi_i^\star(x) - \phi_i^\star(y)\big| \lesssim \|x-y\|_2, \qquad \forall i = 1,\ldots,N. \tag{S.25}
$$

Therefore,

$$
A_1 \lesssim \frac{1}{T}\sum_{t=1}^T\big\|\mathbf{F}_t^\star - \rho^\dagger(\widetilde{\mathbf{Y}}_t(0))\big\|^2 = \inf_{\rho\in\mathcal{G}_N^K}\frac{1}{T}\sum_{t=1}^T\big\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \mathbf{F}_t^\star\big\|^2.
$$

For A_2, by (S.24), we have $A_2 \lesssim T^{-\frac{2\beta}{2\beta+K}}$, which completes the proof.

Step 4: Total in Sample Error.

Note that by (I) and (III), it can be proven that, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\left|\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \big(\phi_i^\dagger(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star)\big)\varepsilon_{it}\right| \\
&\lesssim T^{1/2}K_1^{1/2}\log T\left(\left(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \big(\phi_i^\dagger(\rho^\dagger(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2} + N^{1/2}\right) + N \\
&\lesssim T^{1/2}N^{1/2}K_1^{1/2}\log T\left(T^{1/2}\left(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{G}_N^K}\sum_{t=1}^T\big\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \mathbf{F}_t^\star\big\|^2\right)^{1/2} + 1\right) + N \\
&\lesssim NT\left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{G}_N^K}\sum_{t=1}^T\big\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \mathbf{F}_t^\star\big\|^2\right)\log^2(T).
\end{aligned}$$

Together with (II), (III) and (S.17) with $(\phi_1,\ldots,\phi_N)\circ\rho$ on the right hand side replaced by $(\phi_1^\dagger,\ldots,\phi_N^\dagger)\circ\rho^\dagger$, we conclude that with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \big(\phi_i^\star(\mathbf{F}_t^\star) - \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big)^2 \\
&\lesssim \left(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1\right)^{1/2}\log^2(T)\left(\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \big(\widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \phi_i^\star(\mathbf{F}_t^\star)\big)^2\right)^{1/2} \\
&\quad + \left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{G}_N^K}\sum_{t=1}^T\big\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \mathbf{F}_t^\star\big\|^2\right)\log^2(T).
\end{aligned}$$

By basic inequality $x \le a\sqrt{x} + b \Rightarrow x \le a^2 + 4b$, we have

$$\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} \big(\phi_i^\star(\mathbf{F}_t^\star) - \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big)^2 \lesssim \left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{G}_N^K}\sum_{t=1}^T\big\|\rho(\widetilde{\mathbf{Y}}_t(0)) - \mathbf{F}_t^\star\big\|^2\right)\log^4(T). \tag{S.26}$$

Notably, since $\|\phi^\star\|_\infty \le B$ and $\|M^\star\|_\infty \lesssim n^{-1}$, $\|M^\star\|_0 \asymp n$ by Assumption 3, we observe that the domain, $M^\star\phi^\star([-B,B]^K)$, a compact set, is contained within a bounded set $[-\tilde{C},\tilde{C}]^K$. Therefore, by the Whitney extension theorem, $\rho^\star$ can be extended to (β, C)-Hölder smooth on $[-\tilde{C}-1,\tilde{C}+1]^K$ for some fixed constant C.

According to Proposition 3, there exist $\tilde{\rho}^\dagger \in \mathcal{G}_K^K(L_1-1, \mathbf{d}_1, B, T^{5\beta+5})$ such that:

$$\big\|\tilde{\rho}^\dagger(x+1+\tilde{C}) - \rho^\star(x)\big\|_\infty \lesssim T^{-\frac{\beta}{2\beta+K}}, \qquad \forall x\in[-\tilde{C}-1,\tilde{C}+1]^K. \tag{S.27}$$

Let $v^\dagger \in \mathbb{R}^K$ be a vector with all elements equal to $-1-\tilde{C}$ and define $\rho^\dagger(x) := \tilde{\rho}^\dagger(\bar{\sigma}(M^\star x - v^\dagger)) \in \mathcal{G}_K^K(L_1, \mathbf{d}_1, B, T^{5\beta+5})$. By Lemma 4, with probability at least $1-\exp(-cn)$, $\|M^\star\widetilde{\mathbf{Y}}_t(0)\|_\infty \le 1+\tilde{C}$. As a consequence, by Lemma 5, (S.27) and the fact that $\rho^\star$ is Lipschitz continuous, with probability

at least $1 - C\exp(-cT) - C\exp(-cn)$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\big\|\mathbf{F}_t^\star - \rho^\dagger(\widetilde{\mathbf{Y}}_t(0))\big\|^2 \\
\lesssim&\frac{1}{T}\sum_{t=1}^{T}\Big(\big\|\mathbf{F}_t^\star - \rho^\star(M^\star\boldsymbol{\phi}^\star(\mathbf{F}_t^\star))\big\|^2 + \big\|\rho^\star(M^\star\boldsymbol{\phi}^\star(\mathbf{F}_t^\star)) - \rho^\star(M^\star\widetilde{\mathbf{Y}}_t(0))\big\|^2 + \big\|\rho^\star(M^\star\widetilde{\mathbf{Y}}_t(0)) - \tilde{\rho}^\dagger(\bar{\sigma}(M^\star\widetilde{\mathbf{Y}}_t(0) - v^\dagger))\big\|^2\Big) \\
\lesssim&\frac{1}{T}\sum_{t=1}^{T}\Big(\|M^\star\varepsilon_t\|^2 + \big\|\rho^\star(M^\star\widetilde{\mathbf{Y}}_t(0)) - \tilde{\rho}^\dagger(M^\star\widetilde{\mathbf{Y}}_t(0) + 1 + \tilde{C})\big\|^2\Big) \\
\lesssim&n^{-1} + T^{-\frac{2\beta}{2\beta+K}}.
\end{aligned}$$

Substituting into (S.26), we have

$$\frac{1}{NT}\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}\big(\phi_i^\star(\mathbf{F}_t^\star) - \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big)^2 \lesssim \Big(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + n^{-1}\Big)\log^4(T). \tag{S.28}$$

Step 5: Unit Level in Sample Error.

By Assumption 3, (S.25), $\|M^\star\|^2 \le \mathrm{Tr}((M^\star)^\top M^\star) \lesssim n^{-1}$, and (S.28), we have with probability at least $1 - C\exp(-cT) - C\exp(-cn)$,

$$\begin{aligned}
&\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\rho^\star(M^\star\widehat{\boldsymbol{\phi}}(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))) - \mathbf{F}_t^\star\big\|^2 \\
\lesssim&\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\rho^\star(M^\star\widehat{\boldsymbol{\phi}}(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))) - \rho^\star\big(M^\star\boldsymbol{\phi}^\star(\mathbf{F}_t^\star)\big)\big\|^2 + \frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\rho^\star\big(M^\star\boldsymbol{\phi}^\star(\mathbf{F}_t^\star)\big) - \mathbf{F}_t^\star\big\|^2 \\
\lesssim&\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|M^\star\widehat{\boldsymbol{\phi}}(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - M^\star\boldsymbol{\phi}^\star(\mathbf{F}_t^\star)\big\|^2 + n^{-1} \lesssim \frac{1}{T_1}\sum_{t=1}^{T_1}\|M^\star\|^2\big\|\widehat{\boldsymbol{\phi}}(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \boldsymbol{\phi}^\star(\mathbf{F}_t^\star)\big\|^2 + n^{-1} \\
\lesssim&\frac{1}{T_1 n}\sum_{t=1}^{T_1}\big\|\widehat{\boldsymbol{\phi}}(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))) - \boldsymbol{\phi}^\star(\mathbf{F}_t^\star)\big\|^2 + n^{-1} \lesssim \Big(n^{-1}K_1 + Nn^{-1}T^{-\frac{2\beta}{2\beta+K}} + Nn^{-2}\Big)\log^4(T).
\end{aligned}$$

Let $\widehat{\mathbf{F}}_t = \widehat{\rho}(\widetilde{\mathbf{Y}}_t(0))$. By Assumption 3, (S.25), and the fact that $\rho^\star$ is Lipschitz continuous, we have with probability at least $1 - C\exp(-cT) - C\exp(-cn)$,

$$\begin{aligned}
&\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\phi_i^\star(\mathbf{F}_t^\star) - \widehat{\phi}_i(\widehat{\rho}(\widetilde{\mathbf{Y}}_t(0)))\big\|^2 \\
\lesssim&\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\phi_i^\star(\rho^\star(M^\star\widehat{\phi}(\widehat{\mathbf{F}}_t))) - \phi_i^\star(\mathbf{F}_t^\star)\big\|^2 + \frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\phi_i^\star(\rho^\star(M^\star\widehat{\phi}(\widehat{\mathbf{F}}_t))) - \widehat{\phi}_i(\widehat{\mathbf{F}}_t)\big\|^2 \\
\lesssim&\frac{1}{T_1}\sum_{t=1}^{T_1}\big\|\rho^\star(M^\star\widehat{\phi}(\widehat{\mathbf{F}}_t)) - \mathbf{F}_t^\star\big\|^2 + n^{-1} \\
\lesssim&\Big(n^{-1}K_1 + Nn^{-1}T^{-\frac{2\beta}{2\beta+K}} + Nn^{-2}\Big)\log^4(T).
\end{aligned}$$

□

Proposition 3. *Assume f is (p, C)-Hölder smooth on $[-a, a]^r$ for some $p = q + s$, $q \in \mathbb{N}_0$ and $s \in (0, 1]$, and $C > 0$. Then for $a \ge 1$ and ζ sufficiently large (only depends on fixed constants including p, a, C and $\|f\|_{C^q}$), when*

$$L \asymp \big\lceil \log_4\big(\zeta^{2p}\big)\big\rceil \cdot \Big(\big\lceil \log_2\big(\max\{q, r\} + 1\big)\big\rceil + 1\Big) + 1$$

and

$$d = \max_{1\le\ell\le L} d_\ell \asymp 2^r \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r,$$

there exists a neural network $\hat{f}_{\text{wide}} \in \mathcal{G}_r^1\big(L, \mathbf{d}, B, \zeta^{5p+5}\big)$ *such that*

$$\big\|f - \hat{f}_{\text{wide}}\big\|_\infty \le C\,\zeta^{-2p},$$

for some constant C *only depends on fixed parameters.*

Proposition 3 is proved in Proposition C1 of Shen and Xiu [60], and follows the same argument as in Kohler and Langer [42].

Lemma 2. *Under Assumption 1, given* $\{\mathbf{F}_t^\star\}_{t=1}^T$*, for any constants* $\{C_{it}\}_{i=1,\ldots,N,\ t=1,\ldots,T}$*,*

$$\frac{\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}\varepsilon_{it}}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}^2\Big)^{1/2}}$$

is subGaussian with its subGaussian norm bounded by some fixed constant σ_ϵ^2*.*

Proof. Recall that the subGaussian norm of ε_{it} is bounded by σ_ϵ^2. Let $C = \{C_{it}\}_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}, \varepsilon = \{\varepsilon_{it}\}_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}}$, we have

$$\frac{\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}\varepsilon_{it}}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}^2\Big)^{1/2}} = \frac{\mathrm{vec}(C)^\top \mathrm{vec}(\varepsilon)}{\Big(\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}^2\Big)^{1/2}}.$$

Since $\mathrm{vec}(\varepsilon)$ is composed of independent subGaussian random variables with bounded subGaussian norm, the subGaussian norm of the right hand side is bounded by

$$\frac{\sigma_\epsilon^2\,\mathrm{vec}(C)^\top \mathrm{vec}(C)}{\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}^2} \le \frac{\sigma_\epsilon^2 \|\mathrm{vec}(C)\|^2}{\sum_{(i,t)\in\mathcal{A},\mathcal{B},\mathcal{C}} C_{it}^2} = \sigma_\epsilon^2.$$

□

Lemma 3. *Under Assumption 1, with probability at least* $1 - C\exp(-cT)$*,*

$$\sum_{t=1}^T \|\varepsilon_t\|_1 \lesssim NT.$$

Proof. Recall that the subGaussian norm of ε_{it} is bounded by σ_ϵ^2. Observe that

$$\begin{aligned}\sum_{t=1}^T \|\varepsilon_t\|_1 &\le \sum_{t=1}^T N^{1/2}\|\varepsilon_t\| \le N^{1/2}T^{1/2}\left(\sum_{i=1}^N\sum_{t=1}^T \varepsilon_{it}^2\right)^{1/2} \\ &= N^{1/2}T^{1/2}\big(\mathrm{vec}(\varepsilon)^\top\mathrm{vec}(\varepsilon)\big)^{1/2}.\end{aligned}$$

Therefore, we only need to prove $N^{-1}T^{-1}\mathrm{vec}(\varepsilon)^\top\mathrm{vec}(\varepsilon)$ is bounded by some constant. By Hanson–Wright inequality,

$$\mathbb{P}\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\big(\varepsilon_{it}^2 - \mathbb{E}\varepsilon_{it}^2\big) \ge \sigma_\epsilon^2\right) \le 2\exp(-cNT).$$

By the property of sub-Gaussian variable, we have $\mathbb{E}\varepsilon_{it}^2 \le 2\sigma_\epsilon^2$. As a consequence, it holds that

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \varepsilon_{it}^2 \le 3\sigma_\epsilon^2$$

with probability at least $1 - C\exp(-cT)$. □

Lemma 4. *Under Assumptions 1 and 3, if* $\log T = o(n)$*, with probability at least* $1 - C\exp(-cn)$*,*

$$\|M^\star\varepsilon_t\|_\infty \le 1 \quad \textit{for all } 1 \le t \le T.$$

Proof. For $i = 1, \ldots, K$, write $M_i^\star$ as the i-th row of $M^\star$. Then we have

$$\|M^\star \varepsilon_t\|_\infty = \max_{i=1,\ldots,K} \big|M_i^\star \varepsilon_t\big|.$$

Observe that $M_i^\star \varepsilon_t$ is a subGaussian random variable with its subGaussian norm bounded by

$$\sigma_\epsilon^2 \, M_i^\star M_i^{\star\top} \lesssim \|M_i^\star\|^2 \lesssim n^{-1}.$$

By the property of subGaussian variable,

$$\mathbb{P}\big(\big|M_i^\star \varepsilon_t\big| \geq 1\big) \leq C \exp(-cn).$$

Together with union bound inequality and the fact that $\log(T) = o(L)$,

$$\mathbb{P}\bigg(\max_{1\leq t\leq T} \max_{1\leq i\leq K} \big|M_i^\star \varepsilon_t\big| \geq 1\bigg) \leq CTK \exp(-cn) \lesssim C \exp(-cn/2).$$

□

Lemma 5. *With probability at least* $1 - C\exp(-cT)$*,*

$$T^{-1} \sum_{t=1}^{T} \|M^\star \varepsilon_t\|^2 \lesssim n^{-1}.$$

Proof. Let $\varepsilon = (\varepsilon_1, ..., \varepsilon_T) \in \mathbb{R}^{N\times T}$ and $Q = I_T \otimes M^{\star\top} M^\star$. Observe that

$$T^{-1} \sum_{t=1}^{T} \|M^\star \varepsilon_t\|^2 = T^{-1} \mathrm{vec}(\varepsilon)^\top Q \mathrm{vec}(\varepsilon).$$

Note that $\|Q\| \leq \|M^{\star\top} M^\star\| \leq \mathrm{Tr}((M^\star)^\top M^\star) \lesssim n^{-1}$. Additionally, observe that $\mathrm{rank}(Q) \leq T \,\mathrm{rank}(M^\star) = TK$. As a consequence, we have

$$T^{-1}\mathrm{Tr}(Q) \leq \|T^{-1}Q\| \,\mathrm{rank}(Q) \lesssim n^{-1},$$

$$\big\|T^{-1}Q\big\|_F^2 \lesssim \big\|T^{-1}Q\big\|^2 \,\mathrm{rank}(Q) \lesssim T^{-1}n^{-2}.$$

By applying Hanson–Wright inequality, we conclude that

$$\mathbb{P}\big(\big|(T^{-1}\mathrm{vec}(\varepsilon)^\top Q \mathrm{vec}(\varepsilon) - \mathbb{E}T^{-1}\mathrm{vec}(\varepsilon)^\top Q \mathrm{vec}(\varepsilon)\big| \geq n^{-1}\big) \leq C\exp(-cT).$$

Together with the fact that

$$\mathbb{E}\, T^{-1}\mathrm{vec}(\varepsilon)^\top Q \mathrm{vec}(\varepsilon) \asymp T^{-1}\mathrm{Tr}(Q) \lesssim n^{-1},$$

we complete the proof. □

Lemma 6. *For any* $\phi \in \mathcal{G}_{d_0}^{d_{L+1}}(L, \mathbf{d}, T^{5\beta+5})$*, it holds that*

$$|\phi(x) - \phi(y)| \leq d_0 T^{(5\beta+5)(L+1)} \, d^L \, \|x - y\|_\infty.$$

Proof. Write $\phi(x) = M_{L+1}\bar{\sigma}_{v_L} M_L \bar{\sigma}_{v_{L-1}} \cdots M_2 \bar{\sigma}_{v_1} M_1 x$ where $M_i \in \mathbb{R}^{d_i \times d_{i-1}}$ and $\bar{\sigma}_{v_i} M_i x = M_i x - v_i$. It is not hard to verify that $\|M_i x - M_i y\|_\infty \leq d_{i-1} T^{5\beta+5} \|x-y\|_\infty$ and $\|\bar{\sigma}_{v_i} x - \bar{\sigma}_{v_i} y\|_\infty \leq \|x-y\|_\infty$. Note that for two Lipschitz functions f_1 and f_2 satisfying $\|f_1(x) - f_1(y)\|_\infty \leq c_1 \|x-y\|_\infty$ and $\|f_2(x) - f_2(y)\|_\infty \leq c_2 \|x-y\|_\infty$, it holds that

$$\|f_2 \circ f_1(x) - f_2 \circ f_1(y)\|_\infty \leq c_1 c_2 \|x - y\|_\infty.$$

By repeatedly applying this fact to $M_{L+1}\bar{\sigma}_{v_L} M_L \bar{\sigma}_{v_{L-1}} \cdots M_2 \bar{\sigma}_{v_1} M_1 x$, we have

$$\|\phi(x) - \phi(y)\| \leq T^{(5\beta+5)(L+1)} \left(\prod_{i=0}^{L} d_i\right) \|x - y\|_\infty \leq d_0 T^{(5\beta+5)(L+1)} \, d^L \, \|x - y\|_\infty. \qquad \square$$

□

Lemma 7. *For the function class $\mathcal{G}_{d_0}^{d_{L+1}}(L, \mathbf{d}, B, T^{5\beta+5})$ with width vector $\mathbf{d} = (d_1, \ldots, d_{L+1})$, assume the total number of non-zero weights are bounded by S. There exists a function class $\mathcal{G}_{d_0,\delta}^{d_{L+1}} \subset \mathcal{G}_{d_0}^{d_{L+1}}(L, \mathbf{d}, B, T^{5\beta+5})$ such that*

(a)

$$\left|\mathcal{G}_{d_0,\delta}^{d_{L+1}}\right| \leq \left(8\delta^{-1}CT^{(5\beta+5)(L+1)}(1+d)^L d_0 d_{L+1} L\right)^{2S};$$

(b) For any $\phi \in \mathcal{G}_{d_0}^{d_{L+1}}(L, \mathbf{d}, B, T^{5\beta+5})$, there exists $\bar{\phi} \in \mathcal{G}_{d_0,\delta}^{d_{L+1}}$ satisfying

$$\|\phi(x) - \bar{\phi}(x)\|_\infty \leq \delta \qquad \textit{whenever } \|x\|_\infty \leq C.$$

Lemma 7 is proved in Lemma 7 of Shen and Xiu [60], and follows the same idea as Lemma 5 of Schmidt-Hieber [56].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work are discussed in Appendix D.2.4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results, assumptions, and proofs are presented in Section 2.3 and Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Methodological details of the proposed methods are provided in Section 2.2, and experimental details necessary to reproduce the results are provided in Sections 3-4 and Appendices B-C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our code is provided in the supplement. We use open-access datasets, and we cite original sources of this dataset in Section 4.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental settings and details are provided in Sections 3 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard errors of MAE and MSE from multiple replications are reported along with the experiment results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources needed to reproduce the experiments are reported in Appendix B.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussions of broader impacts are included in Appendix D.1.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper uses open-access datasets, and we cite original sources of this dataset in Section 4.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code for implementing the proposed CoDEAL method is available in the supplementary material, along with the proper documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.