MediEncoder: Nonlinear Representation Learning for High-Dimensional Causal Mediation Analysis

Shi Bo, Debarghya Mukherjee†, and AmirEmad Ghassami†
Department of Mathematics and Statistics, Boston University

First Version: June 5, 2026; Current Version: October 2, 2026

Abstract

Causal mediation analysis decomposes a treatment effect into indirect pathways through mediators and direct pathways not operating through them. Modern biomedical studies often involve high-dimensional covariates and mediators that are noisy proxies for lower-dimensional latent biological processes. Existing methods typically rely on sparsity, linear factor models, or ignore the connection among variables in the learned representations, which can be restrictive when measurements are nonlinear and covariate and mediator factors are structurally dependent. We propose `MediEncoder`, a representation-learning framework for nonlinear high-dimensional mediation analysis. MediEncoder jointly learns low-dimensional covariate and mediator representations using a coupled encoder–decoder architecture with a cross-factor network that links treatment and covariate representations to mediator representations. The learned features are then used in a cross-fitted efficient influence function-based estimator of natural direct and indirect effects. The resulting estimator is multiply robust and asymptotically normal under suitable regularity conditions. Simulations show that MediEncoder improves estimation accuracy over competing dimension-reduction approaches, and an application to Alzheimer's Disease Neuroimaging Initiative data illustrates its utility in high-dimensional biomedical causal mediation analysis.

Keywords— Auto-encoder; Causal Inference; Dimension Reduction; Factor Models; High-Dimensional Data

Contents

1 Introduction **3**
1.1 Related Work 4

2 Model Description **5**

3 Proposed Method: MediEncoder **6**
3.1 Coupled Representation Learning 7
3.2 Cross-Fitted Estimation 8

4 Theoretical Analysis **8**

5 Experiments **12**
5.1 Simulation Settings 12
5.2 Simulation and Ablation Study 13
5.3 Sensitivity Analyses 14

6 Real Data Application: Alzheimer's Disease Neuroimaging Initiative (ADNI) **16**
6.1 Data and Preprocessing 16
6.2 Results 17

7 Conclusion **18**

Supplementary Material **23**

S1 Nuisance Function Estimation **23**

S2 Wavelet Loading **23**

S3 A sufficient condition for factor recovery **24**

S4 Overview of Baseline Methods **25**

S5 Proofs **25**
S5.1 Proof of Theorem 4.1 25
S5.2 Proof of Lemma S5.1 36
S5.3 Proof of Theorem 4.2 36
S5.4 Proof of Proposition S3.1 43

S6 More on Simulation and Ablation Study **43**

S7 Comparisons **47**
S7.1 Comparison with Nath et al. (2023) 47
S7.2 Comparison with Jiang et al. (2023) 48
S7.3 Comparison with Existing Machine-Learning-Based Mediation Analysis 49

1 Introduction

Causal mediation analysis seeks to explain how a treatment affects an outcome by decomposing the total effect into pathways operating through intermediate variables, or mediators (Baron and Kenny, 1986; Pearl, 2001; Imai et al., 2010; VanderWeele and Vansteelandt, 2014). In the potential-outcomes framework (Rubin, 1974), this amounts to studying counterfactual means whose contrasts yield the natural indirect effect, natural direct effect, and total effect. Classical methods typically assume that the relevant covariates and mediators are directly observed and low-dimensional. Modern biomedical studies often violate both assumptions: the causal mechanisms of interest are latent, while the available measurements are noisy, high-dimensional proxies. This setting arises across genomics and multi-omics applications. High-dimensional molecular measurements, such as metabolomic or multi-omics profiles, are often indirect views of lower-dimensional biological states (Wörheide et al., 2021; Rahnenführer et al., 2023). For example, oxidative stress has been proposed as a pathway linking depression to Alzheimer's disease (Livingston et al., 2024; Buccellato et al., 2021), but reactive oxygen species are short-lived and difficult to measure directly (Katerji et al., 2019). Researchers therefore rely on downstream molecular signatures, including DNA damage and DNA methylation changes (Pandya et al., 2013; Black et al., 2015; Miranda et al., 2000). Such measurements may be nonlinear functions of latent biological processes (Johnson et al., 2017; Vershinina et al., 2021); moreover, upstream covariate factors, treatments, and mediator factors can interact in ways that complicate causal identification and estimation (VanderWeele, 2015). These features motivate a mediation framework that can handle latent structure, high dimensionality, nonlinear measurement maps, and treatment–covariate–mediator interactions simultaneously.

Existing high-dimensional mediation methods address only parts of this problem. Sparse and variable-selection approaches can screen many candidate mediators (Zhang, 2022; Zhao and Luo, 2021; Bo et al., 2024; Jones et al., 2025), but they usually treat the observed variables as the mediators and do not model latent structure. Dimension-reduction methods based on PCA or linear factor models provide low-dimensional summaries (Derkach et al., 2019; Fan et al., 2025), but linear measurement maps can be restrictive in complex biological systems (Angermueller et al., 2016). Autoencoders and related deep models can learn nonlinear representations (Hinton and Salakhutdinov, 2006; Goodfellow, 2016), yet applying them separately to covariates and mediators ignores the structural dependence between the two latent systems. Deep mediation estimators such as (Xu et al., 2022) improve robustness to nuisance model misspecification, but do not explicitly target nonlinear high-dimensional latent-factor settings. Variational autoencoder approaches for indirectly observed mediators (Jiang et al., 2023) model latent mediators, but do not jointly accommodate latent nonlinear covariate structure, structural dependence between covariate and mediator representations, and interaction effects; see Section 1.1 for further discussion. Overall, existing methods address individual aspects of the problem but do not simultaneously address all the aforementioned challenges. The key missing ingredient is a framework that learns nonlinear covariate and mediator representations jointly, aligns them with the latent mediation structure, and then uses these representations in a statistically valid causal estimator.

To bridge this gap, we propose `MediEncoder`, a nonlinear representation-learning framework for high-dimensional causal mediation analysis. The method jointly learns low-dimensional representations of covariates and mediators through a structured encoder–decoder architecture. A cross-factor network links the treatment and covariate representation to the mediator representation, encouraging the learned features to respect the causal structure of the latent mediation problem. Exact recovery of the original latent factors is not required: it suffices to learn representations that preserve the information needed for the mediation functional, and our theory formalizes invariance under left-invertible transformations of the latent factors.

Our contributions are fourfold. **(i)** First, we introduce a nonlinear latent-factor framework for mediation analysis with high-dimensional covariates and mediators, avoiding linearity and sparsity requirements. **(ii)** Second, we develop MediEncoder, a coupled representation-learning algorithm that uses a cross-factor network to capture structural dependence between covariate and mediator representations. **(iii)** Third, we develop a modular asymptotic theory for the resulting cross-fitted efficient influence-function estimator. The theory formalizes that exact factor recovery is unnecessary: it suffices for the learned representation to approximate a left-invertible transformation of the latent factors, and the resulting inference remains valid under high-level factor-recovery and nuisance-rate conditions. **(iv)** Fourth, we demonstrate the empirical performance of MediEncoder in extensive simulations and in a biomedical application using data from the Alzheimer's Disease

Neuroimaging Initiative (ADNI).

1.1 Related Work

High-Dimensional Mediation Analysis. Recent work extends causal mediation analysis to high-dimensional regimes where the number of mediators or covariates may exceed the sample size. Early approaches focus primarily on high-dimensional mediators, employing sparsity assumptions and variable selection techniques such as Lasso to estimate mediation effects (Zhao and Luo, 2016; Guo et al., 2022, 2023, 2024; Jones et al., 2025). Subsequent work incorporates interaction effects among treatment, mediators, and covariates, allowing for more flexible modeling of complex mechanisms (Rakshit and Guo, 2024). More recent developments consider settings where both mediators and covariates are high-dimensional and may interact, proposing estimation and inference procedures that account for their joint structure (Bo et al., 2024). Despite these advances, most existing methods rely on linear modeling assumptions and treat observed variables directly, which may be restrictive when the underlying mechanisms are nonlinear and only indirectly observed through noisy measurements.

Factor Models. Factor models provide a natural framework for representing high-dimensional data via a lower-dimensional set of latent variables. Classical linear factor models have been widely used in statistics and econometrics for dimension reduction (Bai and Ng, 2002, 2013; Fan et al., 2013, 2020; Fan and Liao, 2022). These ideas have also been incorporated into causal inference and mediation analysis to handle high-dimensional covariates or mediators by projecting them onto latent spaces (Derkach et al., 2019). However, existing approaches predominantly rely on linear factor structures, assuming observed variables are linear combinations of latent factors. In many scientific applications, particularly in biological and genomic settings, the mapping from latent processes to observed measurements is often nonlinear. Recent work has explored nonlinear extensions of factor models using neural networks to allow more flexible representations (Feng, 2023; Xiu and Shen, 2024; Fan et al., 2025), though their integration with mediation analysis remains limited. In parallel, factor-based methods have also been developed for high-dimensional average treatment effect (ATE) estimation under doubly robust frameworks. Fan et al. (2025) proposed a factor-informed AIPW estimator combining factor models and deep learning for high-dimensional causal inference. Their setting can be viewed as an ATE analogue of our problem, but still primarily relies on linear factor structures without considering nonlinear factor generation mechanisms.

Autoencoders and Mediation Analysis. Deep representation learning methods, especially autoencoders, provide a flexible approach for learning nonlinear low-dimensional representations of high-dimensional data (Hinton and Salakhutdinov, 2006; Goodfellow, 2016). They have been widely applied in large-scale settings for nonlinear dimension reduction and feature learning (LeCun et al., 2015; Khemakhem et al., 2020). Variational autoencoders further introduce probabilistic latent-variable models that enable structured and generative representations (Kingma and Welling, 2013). Related work on multi-view representation learning, including multi-view autoencoders and deep canonical correlation methods, aims to learn shared representations across multiple data modalities (Wang et al., 2015; Guerrero-López et al., 2022). Xu et al. (2022) develop a deep learning-based mediation estimator that is robust to model misspecification. However, their approach learns representations independently and does not explicitly model for high-dimensionality and latent factor structure. In particular, it does not consider a shared representation capturing the dependence between covariates and mediators, which is important in high-dimensional biological systems driven by associated latent factors. Jiang et al. (2023) propose a variational autoencoder-based mediation framework for indirectly observed mediators; however, their method does not accommodate latent and nonlinear structures in covariates, does not model the structural dependence between the latent representations of covariates and mediators, and does not allow treatment–covariate or treatment–mediator interactions. In addition, their estimation procedure may be sensitive to model misspecification.

2 Model Description

We consider a binary treatment (or exposure) $A \in \{0,1\}$ and a real-valued outcome $Y \in \mathbb{R}$. Let $\boldsymbol{f}_X \in \mathbb{R}^{\bar{p}}$ denote a vector of pre-treatment latent covariate factors and let $\boldsymbol{f}_M \in \mathbb{R}^{\bar{q}}$ denote a vector of latent mediator factors. The covariate factors may confound the treatment–mediator, treatment–outcome, and mediator–outcome relationships, whereas the mediator factors are the mechanisms through which part of the effect of A on Y may be transmitted. We use boldface notation for random latent vectors and write f_X and f_M for their generic values.

A compact structural representation of the latent mediation model is

$$\begin{aligned} \boldsymbol{f}_M &= (1-A)\{\mu_0^*(\boldsymbol{f}_X) + \boldsymbol{u}_{XM}\} + A\{\mu_1^*(\boldsymbol{f}_X) + \boldsymbol{u}'_{XM}\}, \\ Y &= (1-A)\{\mu_0(\boldsymbol{f}_X, \boldsymbol{f}_M) + \epsilon'\} + A\{\mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) + \epsilon\}, \end{aligned} \tag{2.1}$$

where $\mu_0^*, \mu_1^* : \mathbb{R}^{\bar{p}} \to \mathbb{R}^{\bar{q}}$ and $\mu_0, \mu_1 : \mathbb{R}^{\bar{p}+\bar{q}} \to \mathbb{R}$ are unknown structural functions and ϵ, ϵ', $\boldsymbol{u}_{XM} \in \mathbb{R}^{\bar{q}}$, and $\boldsymbol{u}'_{XM} \in \mathbb{R}^{\bar{q}}$ are centered noise terms.

In the applications motivating this work, $\boldsymbol{f}_X$ and $\boldsymbol{f}_M$ are not directly observed. Instead, we observe high-dimensional measurements $X \in \mathbb{R}^p$ and $M \in \mathbb{R}^q$, with $p \gg \bar{p}$ and $q \gg \bar{q}$, that serve as noisy proxies for the latent factors; see Figure 1. We assume the measurement model

$$X = \phi_X(\boldsymbol{f}_X) + \boldsymbol{u}_X, \qquad M = \phi_M(\boldsymbol{f}_M) + \boldsymbol{u}_M, \tag{2.2}$$

where $\phi_X : \mathbb{R}^{\bar{p}} \to \mathbb{R}^p$ and $\phi_M : \mathbb{R}^{\bar{q}} \to \mathbb{R}^q$ are unknown loading functions, and $\boldsymbol{u}_X$ and $\boldsymbol{u}_M$ are centered idiosyncratic errors. The classical linear factor model is a special case of (2.2), obtained by taking $\phi_X(f_X) = \Gamma_X f_X$ and $\phi_M(f_M) = \Gamma_M f_M$ for loading matrices $\Gamma_X \in \mathbb{R}^{p\times\bar{p}}$ and $\Gamma_M \in \mathbb{R}^{q\times\bar{q}}$.

We next define the causal estimands at the latent-factor level. Let $\boldsymbol{f}_M^{(a)}$ denote the mediator factors that would be observed if the treatment were set to a, and let $Y^{(a,f_M)}$ denote the potential outcome if the treatment were set to a and the mediator factors were set to the value f_M. We use the shorthand $Y^{(a)} = Y^{(a,\boldsymbol{f}_M^{(a)})}$. For treatment level 1 relative to 0, the total effect decomposes as

$$\underbrace{\mathbb{E}\{Y^{(1)} - Y^{(0)}\}}_{\text{Total effect}} = \underbrace{\mathbb{E}\left[Y^{(1,\boldsymbol{f}_M^{(1)})} - Y^{(1,\boldsymbol{f}_M^{(0)})}\right]}_{\text{Natural indirect effect}} + \underbrace{\mathbb{E}\left[Y^{(1,\boldsymbol{f}_M^{(0)})} - Y^{(0,\boldsymbol{f}_M^{(0)})}\right]}_{\text{Natural direct effect}}. \tag{2.3}$$

Identification of parameters of the form $\theta(a,a') = \mathbb{E}\{Y^{(a,\boldsymbol{f}_M^{(a')})}\}$ for any $a, a' \in \{0,1\}$ yields identifying the natural indirect effect, natural direct effect, and total effect through (2.3). Specifically, $\text{NIE} = \theta(1,1) - \theta(1,0)$, $\text{NDE} = \theta(1,0) - \theta(0,0)$, and $\text{TE} = \theta(1,1) - \theta(0,0)$. Hence, in what follows, we focus on the cross-world mean

$$\theta_0 := \theta(1,0) = \mathbb{E}\big[Y^{(1,\boldsymbol{f}_M^{(0)})}\big].$$

To identify θ_0, we impose the following standard causal assumptions.

Assumption 2.1 (Consistency)**.** *For all $a \in \{0,1\}$ and all mediator values f_M, if $A = a$ and $\boldsymbol{f}_M = f_M$, then $Y = Y^{(a,f_M)}$. In addition, if $A = a$, then $\boldsymbol{f}_M = \boldsymbol{f}_M^{(a)}$.*

Assumption 2.2 (Positivity)**.** *For every value f_X in the support of $\boldsymbol{f}_X$, $0 < \mathbb{P}(A = 1 \mid \boldsymbol{f}_X = f_X) < 1$. Moreover, the conditional density of the mediator is positive, i.e., $0 < p(f_M \mid A = a, \boldsymbol{f}_X = f_X)$ for all A, f_X, and f_M.*

Assumption 2.3 (Sequential exchangeability)**.** *Let $U \perp\!\!\!\perp V \mid W$ denote conditional independence of U and V given W. For all $a, a' \in \{0,1\}$ and all mediator values f_M,*

1. $\{Y^{(a,f_M)}, \boldsymbol{f}_M^{(a)}\} \perp\!\!\!\perp A \mid \boldsymbol{f}_X$; 2. $Y^{(a,f_M)} \perp\!\!\!\perp \boldsymbol{f}_M^{(a')} \mid \boldsymbol{f}_X$.

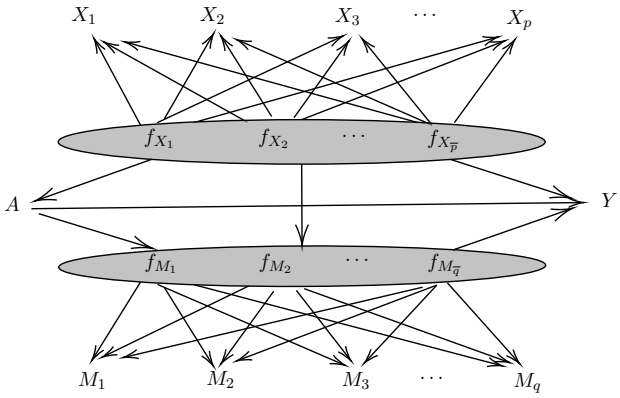


Figure 1: Causal relationship in the presence of latent factors $\mathbf{f}_X$ and $\mathbf{f}_M$

The first part rules out unmeasured confounding of the treatment–outcome and treatment–mediator relationships after conditioning on the latent covariate factors. The second part is the usual cross-world condition for natural direct and indirect effects; it rules out unmeasured mediator–outcome confounding not captured by $\boldsymbol{f}_X$ and excludes treatment-induced mediator–outcome confounders.

Under Assumptions 2.1–2.3, the mediation functional is identified by the mediation formula (Pearl, 2001; Imai et al., 2010). Then

$$\theta_0 = \mathbb{E}[\mathbb{E}\{\mathbb{E}\left(Y \mid A = 1, \boldsymbol{f}_X, \boldsymbol{f}_M\right) \mid A = 0, \boldsymbol{f}_X\}]\,. \tag{2.4}$$

A direct plug-in estimator based on (2.4) can be sensitive to nuisance-model misspecification. Following the semiparametric theory of Tchetgen Tchetgen and Shpitser (2012), we instead use the efficient influence function-based estimator of θ_0. Define the nuisance functions

$$\begin{aligned} \pi_1(f_X) &:= \mathbb{P}(A = 1 \mid \boldsymbol{f}_X = f_X), \quad & \mu_1(f_X, f_M) &:= \mathbb{E}\{Y \mid A = 1, \boldsymbol{f}_X = f_X, \boldsymbol{f}_M = f_M\}, \\ \pi_2(f_X, f_M) &:= \tfrac{p(f_M \mid A=0, f_X)}{p(f_M \mid A=1, f_X)}, & \mu_{10}(f_X) &:= \mathbb{E}\{\mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) \mid A = 0, \boldsymbol{f}_X = f_X\}. \end{aligned} \tag{2.5}$$

With these definitions, $\theta_0 = \mathbb{E}\{\mu_{10}(\boldsymbol{f}_X)\}$, and the efficient influence function is

$$\begin{aligned} \psi_0(O) = \mu_{10}(\boldsymbol{f}_X) - \theta_0 &+ \tfrac{A}{\pi_1(\boldsymbol{f}_X)}\pi_2(\boldsymbol{f}_X, \boldsymbol{f}_M)\{Y - \mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M)\} \\ &+ \tfrac{1-A}{1-\pi_1(\boldsymbol{f}_X)}\{\mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) - \mu_{10}(\boldsymbol{f}_X)\}\,, \end{aligned} \tag{2.6}$$

where $O = (Y, A, \boldsymbol{f}_X, \boldsymbol{f}_M)$. In the oracle setting where $(\boldsymbol{f}_X, \boldsymbol{f}_M)$ are observed, replacing the nuisance functions in (2.6) by estimates and solving the corresponding estimating equation gives

$$\begin{aligned} \hat{\theta}_0^{\mathrm{IF}} = \mathbb{P}_n \Big[&\hat{\mu}_{10}(\boldsymbol{f}_X) + \tfrac{A}{\hat{\pi}_1(\boldsymbol{f}_X)}\hat{\pi}_2(\boldsymbol{f}_X, \boldsymbol{f}_M)\{Y - \hat{\mu}_1(\boldsymbol{f}_X, \boldsymbol{f}_M)\} \\ &+ \tfrac{1-A}{1-\hat{\pi}_1(\boldsymbol{f}_X)}\{\hat{\mu}_1(\boldsymbol{f}_X, \boldsymbol{f}_M) - \hat{\mu}_{10}(\boldsymbol{f}_X)\}\Big], \end{aligned} \tag{2.7}$$

where $\mathbb{P}_n$ denotes the empirical average. However, in reality, we only observe (X, M), (ultra)-high-dimensional proxies of $(\boldsymbol{f}_X, \boldsymbol{f}_M)$, and the latent factors in this oracle estimator are replaced by learned representations, as elaborated in the next section.

3 Proposed Method: MediEncoder

We now describe MediEncoder, a cross-fitted procedure for estimating the mediation functional $\theta_0 = \theta(1, 0) = \mathbb{E}\{Y^{(1, \boldsymbol{f}_M^{(0)})}\}$. The method has two main components. First, it learns low-dimensional nonlinear representations

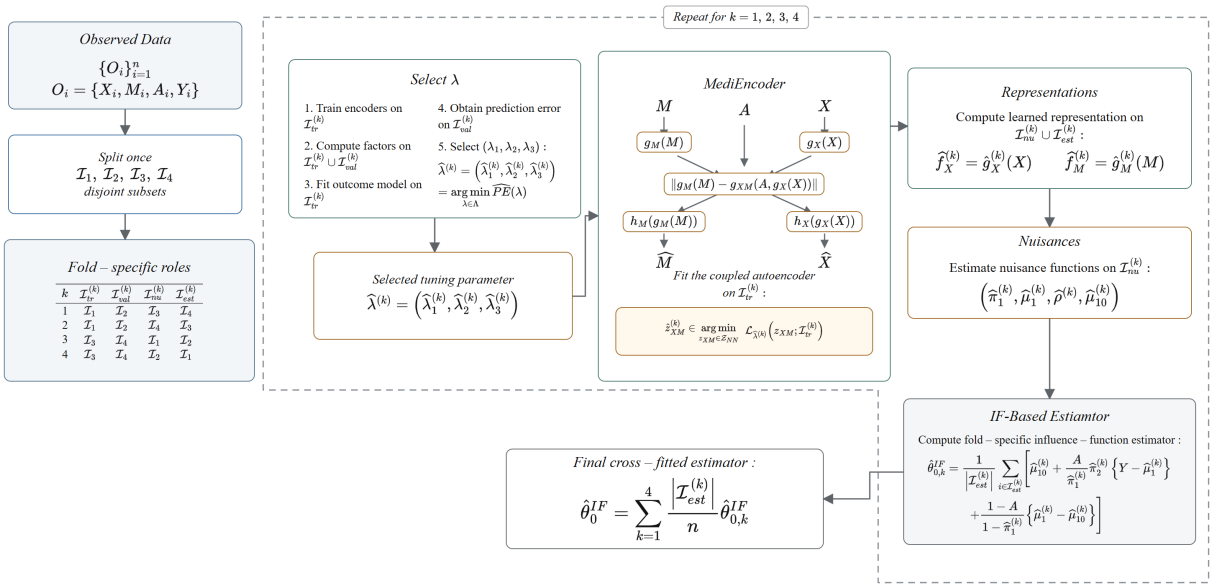


Figure 2: Schematic overview of MediEncoder. The observed data are split into disjoint subsets for representation learning, tuning, nuisance estimation, and final evaluation.

of the high-dimensional covariates and mediators through a coupled autoencoder. Second, it uses these learned representations to estimate the nuisance functions in the efficient-influence function estimator from Section 2.

Figure 2 summarizes the workflow.

3.1 Coupled Representation Learning

Let $z_{XM} = (g_X, g_M, g_{XM}, h_X, h_M) \in \mathcal{Z}_{\text{NN}}$, where $g_X : \mathbb{R}^p \to \mathbb{R}^{\tilde{p}}$ and $h_X : \mathbb{R}^{\tilde{p}} \to \mathbb{R}^p$ are the covariate encoder and decoder, $g_M : \mathbb{R}^q \to \mathbb{R}^{\tilde{q}}$ and $h_M : \mathbb{R}^{\tilde{q}} \to \mathbb{R}^q$ are the mediator encoder and decoder, and $g_{XM} : \{0,1\} \times \mathbb{R}^{\tilde{p}} \to \mathbb{R}^{\tilde{q}}$ maps the treatment and covariate representation to the mediator-representation space. For a training index set $\mathcal{I}$, define the empirical loss

$$\begin{aligned}\mathcal{L}_\lambda(z_{XM}; \mathcal{I}) = \tfrac{1}{|\mathcal{I}|} \textstyle\sum_{i \in \mathcal{I}} \Big[\lambda_1 \|X_i - h_X(g_X(X_i))\|_2^2 + \lambda_2 \|M_i - h_M(g_M(M_i))\|_2^2 \\ + \lambda_3 \|g_M(M_i) - g_{XM}(A_i, g_X(X_i))\|_2^2\Big],\end{aligned} \tag{3.1}$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ consists of nonnegative tuning parameters. The first two terms are reconstruction losses for X and M. The third term couples the two representations by encouraging the mediator representation to be predictable from the treatment and the covariate representation, matching the structural dependence of the latent mediator factors on $(A, \boldsymbol{f}_X)$. Unlike a standard multi-view autoencoder, the coupling is *directional and treatment-aware*: the mediator representation is encouraged to be predictable from $(A, g_X(X))$, obeying the latent structural equation for $\boldsymbol{f}_M$. After representation learning, only the encoders $\hat{g}_X$ and $\hat{g}_M$ are used in the influence function-based estimator; the decoders and g_{XM} act as regularizers during training.

The working dimensions $(\tilde{p}, \tilde{q})$ need not equal the true latent dimensions $(\bar{p}, \bar{q})$. The theoretical results in Section 4 require the learned representation to approximate a sufficiently informative, left-invertible transformation of the latent factors. Thus, exact recovery of the original factors is not required, but the chosen representation dimensions must be large enough to retain the information relevant for the mediation functional.

3.2 Cross-Fitted Estimation

Algorithm 1 gives the full procedure for estimating θ_0. The data are split into four disjoint subsets. In each fold, these subsets play four distinct roles: representation training, tuning, nuisance estimation, and final evaluation. This separation ensures that the summands used to evaluate the influence function-based estimator are not used to train either the representation or the nuisance functions. The particular rotations below are one convenient choice; any rotations with the same disjointness property can be used.

The estimator in (3.4) is the empirical analogue of the efficient influence function representation in Section 2, with the latent factors replaced by fold-specific learned representations. The auxiliary propensity score $\rho(f_X, f_M) = \mathbb{P}(A = 1 \mid \boldsymbol{f}_X = f_X, \boldsymbol{f}_M = f_M)$ is used only to estimate the mediator density ratio π_2. Indeed,

$$\frac{p(f_M \mid A=0, f_X)}{p(f_M \mid A=1, f_X)} = \frac{\mathbb{P}(A=0 \mid f_X, f_M)}{\mathbb{P}(A=1 \mid f_X, f_M)} \frac{\mathbb{P}(A=1 \mid f_X)}{\mathbb{P}(A=0 \mid f_X)}, \tag{3.2}$$

which yields (S1.2) after substituting ρ and π_1.

The same construction estimates any mediation functional $\theta(a, a') = \mathbb{E}\{Y^{(a, \boldsymbol{f}_M^{(a')})}\}$. Consequently,

$$\widehat{\mathrm{NIE}} = \hat{\theta}^{\mathrm{IF}}(1,1) - \hat{\theta}^{\mathrm{IF}}(1,0), \qquad \widehat{\mathrm{NDE}} = \hat{\theta}^{\mathrm{IF}}(1,0) - \hat{\theta}^{\mathrm{IF}}(0,0), \qquad \widehat{\mathrm{TE}} = \hat{\theta}^{\mathrm{IF}}(1,1) - \hat{\theta}^{\mathrm{IF}}(0,0).$$

Section 4 establishes conditions under which $\hat{\theta}_0^{\mathrm{IF}}$ is $\sqrt{n}$-consistent and asymptotically normal.

Remark 3.1 (Capacity of the cross-factor network g_{XM})**.** *The cross-factor network g_{XM} should have moderate capacity. If it is overly expressive, it can approximate an arbitrary mapping from $(A, g_X(X))$ to $g_M(M)$, driving the alignment loss in (3.1) close to zero regardless of the quality of g_X and g_M. In that case, g_{XM} absorbs discrepancies between the two representations, and the alignment term no longer enforces a meaningful structural relationship. Conversely, if g_{XM} is too restrictive, it may fail to capture the dependence of the mediator factors on the treatment and covariate factors. In practice, we use a lower- or moderate-capacity architecture for g_{XM}, relative to the encoders and decoders, so that the alignment term couples the representations without allowing the cross-factor network to overfit. The sensitivity analysis in Section 5.3, particularly Table 5, examines how the capacity of g_{XM} affects finite-sample estimation and inference.*

The auxiliary regression $\widehat{m}_1^{(\lambda)}$ is used only to select the representation learning tuning parameter λ and is not used in the final influence function-based estimator. After $\hat{\lambda}$ is selected, Algorithm 1 refits the coupled autoencoder on the fold-specific representation-training set and estimates the nuisance functions on a separate nuisance-estimation set.

4 Theoretical Analysis

In this section, we present theoretical results establishing the $\sqrt{n}$-consistency and asymptotic normality of the proposed estimator $\hat{\theta}_0^{\mathrm{IF}}$. As is clear from Algorithm 1, the estimation procedure relies on two key ingredients: (i) constructing $(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M)$, a surrogate of the latent factors $(\boldsymbol{f}_X, \boldsymbol{f}_M)$, from the high-dimensional observations (X, M); and (ii) estimating the nuisance functions that enter the influence function representation by regressing either Y or A on these learned representations. Moreover, as discussed in the Section 1, we do not assume that the true factor dimensions are known. Instead, we only require an upper bound on them, and our theoretical framework is flexible enough to accommodate this setting. We impose the following regularity condition on the data-generating process, in addition to the standard causal identification assumptions in Assumptions 2.1–2.3.

Assumption 4.1. *There exists a constant $C < \infty$ such that $\mathbb{E}[(Y - \mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M))^2 \mid A = 1, \boldsymbol{f}_X, \boldsymbol{f}_M] \le C$ and $\mathbb{E}[(\mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) - \mu_{10}(\boldsymbol{f}_X))^2 \mid A = 0, \boldsymbol{f}_X] \le C$.*

Assumption 4.1 is mild. It requires the outcome noise and the relevant conditional mean contrast to have finite conditional second moments, which is a basic regularity requirement for establishing a central limit theorem. We further have the following requirement on the estimated factor surrogates.

Algorithm 1 `MediEncoder`: Estimating θ_0 with cross-fitting

Require: Data $\{(X_i, A_i, M_i, Y_i)\}_{i=1}^n$; representation dimensions $(\tilde{p}, \tilde{q})$; tuning grid Λ

Ensure: Cross-fitted estimator $\hat{\theta}_0^{\mathrm{IF}}$

1: Split $\{1, \ldots, n\}$ into four disjoint subsets $\mathcal{I}_1, \ldots, \mathcal{I}_4$, each of size approximately $n/4$. Use the following fold assignments, where the four columns denote the representation-training, validation, nuisance-estimation, and final-evaluation sets:

k	$\mathcal{I}_{\mathrm{tr}}^{(k)}$	$\mathcal{I}_{\mathrm{val}}^{(k)}$	$\mathcal{I}_{\mathrm{nu}}^{(k)}$	$\mathcal{I}_{\mathrm{est}}^{(k)}$
1	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	$\mathcal{I}_4$
2	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_4$	$\mathcal{I}_3$
3	$\mathcal{I}_3$	$\mathcal{I}_4$	$\mathcal{I}_1$	$\mathcal{I}_2$
4	$\mathcal{I}_3$	$\mathcal{I}_4$	$\mathcal{I}_2$	$\mathcal{I}_1$

2: **for** $k = 1, \ldots, 4$ **do**

3: Set $\mathcal{I}_{\mathrm{tr}}^{(k)}, \mathcal{I}_{\mathrm{val}}^{(k)}, \mathcal{I}_{\mathrm{nu}}^{(k)}, \mathcal{I}_{\mathrm{est}}^{(k)}$ according to the k-th row above.

4: Select $\hat{\lambda}^{(k)} = (\hat{\lambda}_1^{(k)}, \hat{\lambda}_2^{(k)}, \hat{\lambda}_3^{(k)}) \in \Lambda$ using the training and validation sets $(\mathcal{I}_{\mathrm{tr}}^{(k)}, \mathcal{I}_{\mathrm{val}}^{(k)})$ via Algorithm 2.

5: Fit the coupled autoencoder on $\mathcal{I}_{\mathrm{tr}}^{(k)}$:

$$\widehat{z}_{XM}^{(k)} = (\hat{g}_X^{(k)}, \hat{g}_M^{(k)}, \hat{g}_{XM}^{(k)}, \hat{h}_X^{(k)}, \hat{h}_M^{(k)}) \in \arg\min_{z_{XM} \in \mathcal{Z}_{\mathrm{NN}}} \mathcal{L}_{\hat{\lambda}^{(k)}}(z_{XM}; \mathcal{I}_{\mathrm{tr}}^{(k)}). \tag{3.3}$$

6: Compute learned representations for all $i \in \mathcal{I}_{\mathrm{nu}}^{(k)} \cup \mathcal{I}_{\mathrm{est}}^{(k)}$:

$$\widehat{\boldsymbol{f}}_{X,i}^{(k)} = \hat{g}_X^{(k)}(X_i), \qquad \widehat{\boldsymbol{f}}_{M,i}^{(k)} = \hat{g}_M^{(k)}(M_i).$$

7: Estimate nuisance functions via $\hat{\pi}_1^{(k)}$, $\hat{\mu}_1^{(k)}$, $\hat{\mu}_{10}^{(k)}$, and $\hat{\rho}^{(k)}$ on $\mathcal{I}_{\mathrm{nu}}^{(k)}$ using the learners provided in Section S1 of the Supplementary Material. Set $\hat{\pi}_2^{(k)}$ as per Equation (3.2).

8: Compute the fold-specific influence-function estimator:

$$\begin{aligned}\hat{\theta}_{0,k}^{\mathrm{IF}} = \tfrac{1}{|\mathcal{I}_{\mathrm{est}}^{(k)}|} \textstyle\sum_{i \in \mathcal{I}_{\mathrm{est}}^{(k)}} \Bigg[& \hat{\mu}_{10}^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)}) + \tfrac{A_i}{\hat{\pi}_1^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)})} \hat{\pi}_2^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)}) \Big\{ Y_i - \hat{\mu}_1^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)}) \Big\} \\ & + \tfrac{1-A_i}{1-\hat{\pi}_1^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)})} \Big\{ \hat{\mu}_1^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)}) - \hat{\mu}_{10}^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)}) \Big\} \Bigg].\end{aligned} \tag{3.4}$$

9: **end for**

10: Return the weighted cross-fitted estimator

$$\hat{\theta}_0^{\mathrm{IF}} = \sum_{k=1}^{4} \frac{|\mathcal{I}_{\mathrm{est}}^{(k)}|}{n} \hat{\theta}_{0,k}^{\mathrm{IF}}.$$

Assumption 4.2. *The estimated factors $\hat{\boldsymbol{f}} = (\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M)$ satisfy:*

$$\|\hat{\boldsymbol{f}} - \nu(\boldsymbol{f})\|_{L_2(P)} = O_p(\delta_{n,f}), \qquad \delta_{n,f} \downarrow 0 \text{ as } n \uparrow \infty, \tag{4.1}$$

for some $\nu : \mathbb{R}^{\bar{p}+\bar{q}} \mapsto \mathbb{R}^{\tilde{p}+\tilde{q}}$, where $\nu = (\nu_X, \nu_M)$ admits a Lipschitz left inverse $\nu^{-1} : \mathbb{R}^{\tilde{p}+\tilde{q}} \mapsto \mathbb{R}^{\bar{p}+\bar{q}}$.

In Section S3 of the Supplementary Material (Proposition S3.1), we give a sufficient condition showing that Equation (4.1) follows from three interpretable ingredients: (i) the existence of oracle encoders that recover a left-invertible transformation of the latent factors up to denoising error; (ii) a local stability, or quadratic-growth, condition for the population MediEncoder objective around the oracle representation; and (iii) a small excess-risk bound for the fitted MediEncoder. This result shows that the factor condition is not tied to a particular algorithm, but can be verified whenever the representation learning objective is stable, and

Algorithm 2 Tuning $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ for `MediEncoder`

Require: Training indices $\mathcal{I}_{\text{tr}}$, validation indices $\mathcal{I}_{\text{val}}$, tuning grid $\Lambda \subset [0,\infty)^3 \setminus \{(0,0,0)\}$, representation class $\mathcal{Z}_{\text{NN}}$, auxiliary outcome-regression class $\mathcal{F}_\mu^{\text{tun}}$

Ensure: Selected tuning parameter $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)$

1: Define the treated training and validation subsets

$$\mathcal{I}_{\text{tr},1} := \{i \in \mathcal{I}_{\text{tr}} : A_i = 1\}, \qquad \mathcal{I}_{\text{val},1} := \{i \in \mathcal{I}_{\text{val}} : A_i = 1\}.$$

2: Assume $|\mathcal{I}_{\text{tr},1}| > 0$ and $|\mathcal{I}_{\text{val},1}| > 0$; otherwise, redraw the sample split.

3: **for** each $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in \Lambda$ **do**

4: Fit the coupled autoencoder on $\mathcal{I}_{\text{tr}}$:

$$\begin{aligned}\widehat{z}_{XM}^{(\lambda)} &= \left(\hat{g}_X^{(\lambda)}, \hat{g}_M^{(\lambda)}, \hat{g}_{XM}^{(\lambda)}, \hat{h}_X^{(\lambda)}, \hat{h}_M^{(\lambda)}\right) \\ &\in \arg \min_{z_{XM}=(g_X,g_M,g_{XM},h_X,h_M)\in\mathcal{Z}_{\text{NN}}} \frac{1}{|\mathcal{I}_{\text{tr}}|} \sum_{i\in\mathcal{I}_{\text{tr}}} \Big[\lambda_1 \|X_i - h_X(g_X(X_i))\|_2^2 \\ &\qquad + \lambda_2 \|M_i - h_M(g_M(M_i))\|_2^2 \\ &\qquad + \lambda_3 \|g_M(M_i) - g_{XM}(A_i, g_X(X_i))\|_2^2\Big].\end{aligned}$$

5: Compute learned representations for all $i \in \mathcal{I}_{\text{tr}} \cup \mathcal{I}_{\text{val}}$:

$$\widehat{\boldsymbol{f}}_{X,i}^{(\lambda)} = \hat{g}_X^{(\lambda)}(X_i), \qquad \widehat{\boldsymbol{f}}_{M,i}^{(\lambda)} = \hat{g}_M^{(\lambda)}(M_i).$$

6: Fit an auxiliary treated-outcome regression on the training subset:

$$\widehat{m}_1^{(\lambda)} \in \arg \min_{g\in\mathcal{F}_\mu^{\text{tun}}} \frac{1}{|\mathcal{I}_{\text{tr},1}|} \sum_{i\in\mathcal{I}_{\text{tr},1}} \left[Y_i - g\left(\widehat{\boldsymbol{f}}_{X,i}^{(\lambda)}, \widehat{\boldsymbol{f}}_{M,i}^{(\lambda)}\right)\right]^2.$$

7: Compute the held-out validation prediction error:

$$\widehat{\text{PE}}(\lambda) = \frac{1}{|\mathcal{I}_{\text{val},1}|} \sum_{i\in\mathcal{I}_{\text{val},1}} \left[Y_i - \widehat{m}_1^{(\lambda)}\left(\widehat{\boldsymbol{f}}_{X,i}^{(\lambda)}, \widehat{\boldsymbol{f}}_{M,i}^{(\lambda)}\right)\right]^2.$$

8: **end for**

9: Select any minimizer

$$\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3) \in \arg\min_{\lambda\in\Lambda} \widehat{\text{PE}}(\lambda).$$

10: **return** $\hat{\lambda}$

the learned encoder has small population excess risk. In linear factor models with pervasive factors, diversified projection provides one such example Fan et al. (2013); Fan and Gu (2023); Fan and Liao (2022). For nonlinear factor models, recent autoencoder theory Xiu and Shen (2024) provides related recovery guarantees up to functional transformations. Translating these guarantees into (4.1) requires additional stability conditions on the learned bottleneck map; we discuss this connection in Section S3 of the Supplementary Material.

We now present our first main result which shows that if the estimated factor surrogates and the nuisance estimators converge at appropriate rates, then the final estimator $\hat{\theta}^{\text{IF}}$ is $\sqrt{n}$-consistent and asymptotically normal.

Theorem 4.1. *Suppose Assumptions 2.1-4.2 hold and the nuisance estimators from Line 7 of Algorithm 1 are Lipschitz continuous, and $(\pi_1, \hat{\pi}_1, \rho, \hat{\rho})$ are uniformly bounded away from 0 and 1, and*

$$\begin{aligned}\|\hat{\pi}_1 - \pi_1 \circ \nu_X^{-1}\|_{L_2(P)} &= O_p(\delta_{n,\pi_1}), & \|\hat{\mu}_1 - \mu_1 \circ \nu^{-1}\|_{L_2(P)} &= O_p(\delta_{n,\mu_1}), \\ \|\hat{\mu}_{10} - \mu_{10} \circ \nu_X^{-1}\|_{L_2(P)} &= O_p(\delta_{n,\mu_{10}}), & \|\hat{\rho} - \rho \circ \nu^{-1}\|_{L_2(P)} &= O_p(\delta_{n,\rho}).\end{aligned} \tag{4.2}$$

for some rates $\{\delta_{n,\eta} : \eta \in \Xi\} \downarrow 0$ *as* $n \uparrow \infty$*, that satisfy:*

$$\begin{aligned}\sqrt{n}\,(\delta_{n,\pi_1} + \delta_{n,\rho} + \delta_{n,f} + n^{-1/2})(\delta_{n,\mu_1} + \delta_{n,f} + n^{-1/2}) = o(1),\\ \sqrt{n}\,(\delta_{n,\pi_1} + \delta_{n,f} + n^{-1/2})(\delta_{n,\mu_{10}} + \delta_{n,f} + n^{-1/2}) = o(1).\end{aligned} \tag{4.3}$$

Then, the estimator $\hat{\theta}_0^{\mathrm{IF}}$ *satisfies* $\sqrt{n}\big(\hat{\theta}_0^{\mathrm{IF}} - \theta_0\big) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}(0, \sigma_{\mathrm{eff}}^2)$*, where* σ_{eff}^2 *is the semiparametrically efficient variance of the oracle latent-factor model. A sufficient condition for Equation (4.3) is* $\max\{\delta_{n,\pi_1}, \delta_{n,\rho}, \delta_{n,\mu_1}, \delta_{n,\mu_{10}}, \delta_{n,f}\} = o(n^{-1/4})$.

Remark 4.1. *It is important to note that the transformation* ν *is an unknown population-level map and is not estimated by our procedure. It is used only to formalize the class of latent representations under which the mediation functional remains estimable at* $\sqrt{n}$ *rate. Thus, our theory does not require recovering the true factors themselves; it only requires the learned representation to be close to some left-invertible transformation of them.*

Theorem 4.1 only requires weak rate requirements on the learned representation and nuisance functions irrespective of what algorithm is being used. We now verify these rates under a compositional smoothness assumption for DNN-based estimators as used in Algorithm 1. However, the nuisances are estimated using learned representation $\widehat{\boldsymbol{f}}$, but their population targets are functions of $\boldsymbol{f}$ or equivalently $\nu(\boldsymbol{f})$ as ν is left-invertible; hence, the factor learning error $\delta_{n,f}$ contributes to the nuisance parameter estimation. We start with a definition of the compositional function class, namely Hierarchical Compositional Model (HCM):

Definition 4.1 (Hierarchical Composition Model)**.** *Let* $\mathcal{P} \subset [1, \infty) \times \mathbb{N}$*. For* $\ell = 1$*, define a collection* $\mathcal{H}(d, 1, \mathcal{P})$ *consisting of functions* $h(x) = g(x_{\pi(1)}, \ldots, x_{\pi(t)})$*, where* $g : \mathbb{R}^t \to \mathbb{R}$ *is* (β, C) *Holder-smooth for some* $(\beta, t) \in \mathcal{P}$*. For* $\ell > 1$*,* $\mathcal{H}(d, \ell, \mathcal{P})$ *is defined recursively by* $h(x) = g(h_1(x), \ldots, h_t(x))$*, where* $g : \mathbb{R}^t \to \mathbb{R}$ *is* (β, C) *Hölder-smooth for some* $(\beta, t) \in \mathcal{P}$*, and each* $h_j \in \mathcal{H}(d, \ell - 1, \mathcal{P})$*. The dimensional-adjusted smoothness (DAS) is defined as* $\gamma^\star := \inf_{(\beta,t)\in\mathcal{P}}(\beta/t)$.

HCM includes ordinary Hölder classes as well as additive, single-index, and other low-dimensional compositional models. For instance, if $h(x) = \sum_{j=1}^d h_j(x_j)$ or $h(x) = g(x^\top \beta_0)$ with β-smooth univariate components or link functions, then the effective index is $\gamma^\star = \beta$ up to smooth linear and aggregation maps, whereas a generic d-variate Hölder function has $\gamma^\star = \beta/d$. Thus $\gamma^\star$ can be much larger than the ambient smoothness index, and DNN-based estimators adaptively achieve faster rates than those dictated by the ambient dimension. To that end, we require the following condition.

Assumption 4.3. *The mean functions* μ_1 *and* μ_{10} *are uniformly bounded, and* $(\epsilon, \epsilon', \mathbf{u}_X, \mathbf{u}_M)$ *in Equation (2.1) are independent of* $(\boldsymbol{f}_X, \boldsymbol{f}_M)$ *and centered sub-Gaussian random variables.*

Theorem 4.2. *Suppose Assumptions 2.1-2.3, 4.2, and 4.3 hold. Define the transformed nuisances as* $\eta_{\mu_1} := \mu_1 \circ \nu^{-1}, \eta_{10} := \mu_{10} \circ \nu_X^{-1}, \eta_{\pi_1} := \pi_1 \circ \nu_X^{-1}$ *and* $\eta_\rho := \rho \circ \nu^{-1}$*. Assume that each of these* η_j *admits a bounded Lipschitz HCM extension with DAS* $\gamma_j^\star$ *to a compact neighborhood containing both the transformed* $(\nu(\boldsymbol{f}))$ *and learned representations* $(\hat{\boldsymbol{f}})$*. Assume also the overlap conditions that* π_1 *and* ρ *are bounded away from* 0 *and* 1*. Then, under a proper choice of width and depth (see Supplementary Material, Section S5.3 for details) of deep ReLU networks, the estimators obtained in Line 7 of Algorithm 1 satisfy:*

$$\|\hat{\eta}_j(\hat{\boldsymbol{f}}) - \eta_j(\nu(\boldsymbol{f}))\|_{L_2(P)}^2 = O_p(\delta_{n,j}^2), \quad \text{with } \delta_{n,j}^2 = n^{-2\gamma_j^*/(2\gamma_j^*+1)}(\log n)^{a_j} + \delta_{n,f}^2,$$

for $j \in \{\pi_1, \rho, \mu_1\}$*, and* $a_j > 0$ *fixed constant. Further* $\hat{\eta}_{10}$ *satisfies:*

$$\|\hat{\eta}_{10}(\hat{\boldsymbol{f}}_X) - \eta_{10}(\nu_X(\boldsymbol{f}_X))\|_{L_2(P)}^2 = O_p(\delta_{n,10}^2), \quad \delta_{n,10}^2 = n^{-2\gamma_{10}^\star/(2\gamma_{10}^\star+1)}(\log n)^{a_{10}} + \delta_{n,f}^2 + \delta_{n,\mu_1}^2\,.$$

Theorem 4.2 implies that, if $\gamma_j^\star > 1/2$ and $\delta_{n,f} = o(n^{-1/4})$, then all nuisance functions are estimable at rates faster than $n^{-1/4}$. Consequently, Equation (4.3) in Theorem 4.1 holds, and $\widehat{\theta}^{\mathrm{IF}}$ is $\sqrt{n}$-CAN. The rate for μ_{10} additionally depends on the rate for μ_1, because μ_{10} is estimated by regressing the estimated responses $\widehat{\mu}_1(\widehat{\boldsymbol{f}}_X, \widehat{\boldsymbol{f}}_M)$ on $\widehat{\boldsymbol{f}}_X$. Our theory extends the results of Fan and Gu (2024) to nonlinear factor estimation and nuisance estimation under logistic-link losses.

5 Experiments

5.1 Simulation Settings

In this section, we present simulation studies to assess the performance of our proposed methodology. We evaluate our cross-fitted estimator ($\hat{\theta}_0^{\mathrm{IF}}$ in Algorithm 1).The data generating process is as follows:

1. We generate latent covariates $\boldsymbol{f}_X \in \mathbb{R}^{\bar{p}}$ from $\boldsymbol{f}_X \sim \mathcal{U}(-1,1)^{\bar{p}}$. The treatment variable A is generated according to a logistic model based on $\boldsymbol{f}_X$:

$$\mathbb{P}(A=1 \mid \boldsymbol{f}_X) = \frac{\mathsf{exp}(\boldsymbol{f}_X^\top \alpha)}{1+\mathsf{exp}(\boldsymbol{f}_X^\top \alpha)}, \qquad A \sim \text{Bernoulli}\left(\frac{\mathsf{exp}(\boldsymbol{f}_X^\top \alpha)}{1+\mathsf{exp}(\boldsymbol{f}_X^\top \alpha)}\right),$$

   where $\alpha \in \mathbb{R}^{\bar{p}}$ is drawn once from $\mathcal{U}(0,2)^{\bar{p}}$.

2. The latent mediators $\boldsymbol{f}_M \in \mathbb{R}^{\bar{q}}$ are generated through a nonlinear transformation of $\boldsymbol{f}_X$, the treatment indicator A, and additive noise:

$$\boldsymbol{f}_M = (1-A)\big(\boldsymbol{\delta}_0 \boldsymbol{f}_X^{\circ 2} + \mathbf{u}_{XM}\big) + A\big(\boldsymbol{\delta}_1 \boldsymbol{f}_X^{\circ 2} + \mathbf{u}'_{XM}\big), \quad \mathbf{u}_{XM}, \mathbf{u}'_{XM} \sim \mathcal{N}(0, \Sigma_U),$$

   where $\boldsymbol{f}_X^{\circ 2}$ denotes the element-wise square of $\boldsymbol{f}_X$. The control-arm loading $\boldsymbol{\delta}_0 \in \mathbb{R}^{\bar{q}\times\bar{p}}$ has entries drawn once from $\mathcal{U}(0.5,1.5)$, and the treated-arm loading is $\boldsymbol{\delta}_1 = \mathcal{U}(0.5,1.5) + V$, where $V \in \mathbb{R}^{\bar{q}\times\bar{p}}$ has i.i.d. $\mathcal{N}(0,1)$ entries. The residual covariance is $\Sigma_U = R_U \Lambda_U R_U^\top$, where $R_U \in \mathbb{R}^{\bar{q}\times\bar{q}}$ is an orthonormal matrix and $\Lambda_U = \sigma_U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_{\bar{q}})$ with $\lambda_j \sim \mathcal{U}(1,2)$ and $\sigma_U > 0$ a residual-scale parameter.

3. The outcome variable Y is generated via a nonlinear equation:

$$\begin{aligned} Y = (1-A)\Big(&\sin(5\boldsymbol{f}_X)^\top \beta_0 + \log\big(1+\mathsf{exp}(\boldsymbol{f}_X^\top \kappa_0)\big) + (\boldsymbol{f}_M \circ \boldsymbol{f}_X^{\circ 2})^\top \gamma_0 + \epsilon'_Y\Big) \\ &+ A\Big(\sin(5\boldsymbol{f}_X)^\top \beta_1 + \log\big(1+\mathsf{exp}(\boldsymbol{f}_X^\top \kappa_1)\big) + (\boldsymbol{f}_M \circ \boldsymbol{f}_X^{\circ 2})^\top \gamma_1 + \epsilon_Y\Big), \end{aligned}$$

   where $\beta_0, \beta_1, \kappa_0, \kappa_1 \in \mathbb{R}^{\bar{p}}$, $\gamma_0, \gamma_1 \in \mathbb{R}^{\bar{q}}$, each drawn once from $\mathcal{U}(0.5,1.5)$, and $\epsilon_Y, \epsilon'_Y \sim \mathcal{N}(0, \sigma_Y^2)$.

4. We generate the observed high-dimensional covariates $X \in \mathbb{R}^p$ and mediators $M \in \mathbb{R}^q$ through nonlinear factor loadings with additive noise. Specifically, for each observed coordinate $j = 1, \ldots, p$ and $k = 1, \ldots, q$, we generate

$$X_j = \phi_{X,j}(\boldsymbol{f}_X) + \boldsymbol{u}_{X,j}, \qquad M_k = \phi_{M,k}(\boldsymbol{f}_M) + \boldsymbol{u}_{M,k},$$

   where $\boldsymbol{f}_X \in \mathbb{R}^{\bar{p}}$ and $\boldsymbol{f}_M \in \mathbb{R}^{\bar{q}}$ denote the latent factors.

   Nonlinear loading families. We consider nonlinear factor loadings for $g_{X,j}(\cdot)$ and $g_{M,k}(\cdot)$ in wavelet (see Section S2 of the Supplementary Material for more details) based on Haar atoms, and we use an additive structure over latent coordinates:

$$g_{X,j}(\boldsymbol{f}_X) = \sum_{i=1}^{\bar{p}} h_{X,ji}(\boldsymbol{f}_{X_i}), \qquad g_{M,k}(\boldsymbol{f}_M) = \sum_{i=1}^{\bar{q}} h_{M,ki}(\boldsymbol{f}_{M_i}),$$

   where the univariate functions $h_{X,jm}(\cdot)$ and $h_{M,km}(\cdot)$ are instantiated using the wavelet bases.

5. The idiosyncratic components satisfy

$$\mathbf{u}_X \sim \mathcal{N}(0, \sigma_X^2 I_p), \qquad \mathbf{u}_M \sim \mathcal{N}(0, \sigma_M^2 I_q),$$

   independently of the latent factors.

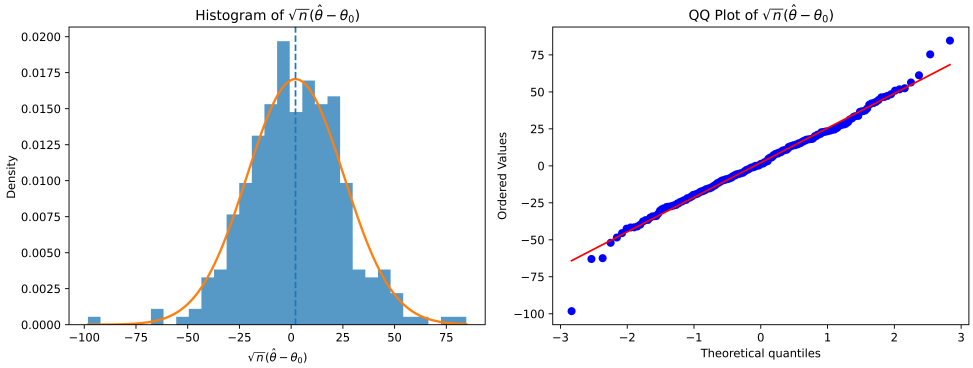


Figure 3: Histogram and Q–Q plot of $\sqrt{n}(\hat{\theta}_0^{\mathrm{IF}} - \theta_0)$ for $n = 2000$ and $p + q = 1000$ under the nonlinear wavelet DGP.

5.2 Simulation and Ablation Study

We evaluate `MediEncoder` in nonlinear high-dimensional settings generated from Section 2 (Equation (2.1)). The target is $\theta_0 = \theta(1, 0)$, estimated by the cross-fitted efficient influence function-based estimator in (2.7). We compare with the baseline representation methods described in Section S4 of the Supplementary Material. Across $B = 200$ Monte Carlo replications, we report the empirical standard deviation (SD), root mean squared error (RMSE), average length of the nominal 95% confidence interval, and empirical coverage. The full data-generating process is given in Section 5.1.

Simulation settings. For each replication, we generate $n \in \{100, 300, 800, 1200, 2000, 3000\}$ observations and consider $p+q \in \{1000, 3000, 5000, 10000\}$. The latent dimensions are fixed at $\bar{p} = \bar{q} = 5$, while the learned representation dimensions are set to $\tilde{p} = \tilde{q} = 10$. The mediator and outcome noise standard deviations are $\sigma_M = \sigma_Y = 1$; the covariate noise level is set to $\sigma_X = 2$ for the first two dimension settings and $\sigma_X = 1.5$ for the larger settings. Observed covariates and mediators are generated from nonlinear additive Haar-wavelet loading functions with independent Gaussian idiosyncratic noise.

All representation and nuisance models are multilayer perceptrons with ReLU activations. The encoders g_X and g_M use two hidden layers of widths 300 and 200, and the decoders h_X and h_M use the reverse architecture. The cross-factor network g_{XM} uses one hidden layer of width 50, giving it lower capacity than the encoders and decoders. For nuisance estimation, we fit π_1, ρ, μ_1, and μ_{10} using fully connected networks; the density ratio π_2 is then computed from $\hat{\pi}_1$ and $\hat{\rho}$ using Bayes' rule as in Algorithm 1. Networks are trained with Adam using initial learning rate 10^{-3}, weight decay 10^{-3}, a learning-rate decay factor of 0.5 every 30 epochs, and at most 300 epochs. We use mean squared error for regression losses and binary cross-entropy for classification losses, with early stopping based on validation loss.

Results. Table 1 reports the full results for $p + q = 3000$ across all six sample sizes; additional results are given in Supplementary Material, Section S6. In this setting, `MediEncoder` has the smallest SD, RMSE, and average confidence-interval length for all displayed sample sizes, while maintaining coverage close to the nominal 95% level. The improvement becomes clearer as n increases: for example, at $n = 3000$, `MediEncoder` reduces RMSE from 0.494 for VAE to 0.470. The supplementary results show the same trend for moderate and large sample sizes, although very small samples can favor simpler representation learners because deep representations are harder to train stably. Figure 3 shows that the scaled errors are approximately centered and have nearly linear Q–Q behavior, consistent with the asymptotic normal approximation developed in Section 4.

Ablation study on λ_3. To assess the role of the alignment loss, we set $\lambda_3 = 0$ and compare this ablated version with the fully tuned `MediEncoder`. The ablation results in Table 2 and Supplementary Material, Section S6 show that removing the alignment term generally increases SD, RMSE, and confidence-interval length, with the largest gap in smaller samples. This supports the use of the cross-factor alignment component

for stabilizing the learned mediator representation and improving estimation efficiency, while the advantage naturally becomes less pronounced as the sample size grows.

Tables 1 and 2 provide the full simulation and ablation results for $p = 2000$ and $q = 1000$.

n	**Estimator**	**SD**	**RMSE**	**CI Length**	**Coverage**
100	Projection	31.166	31.089	122.168	0.985
	Autoencoder	17.740	18.024	69.538	0.955
	VAE	47.131	47.079	184.752	0.995
	MediEncoder	11.739	11.720	46.018	0.955
300	Projection	9.143	9.154	35.839	0.995
	Autoencoder	3.862	3.859	15.137	0.965
	VAE	2.267	2.275	8.885	0.940
	MediEncoder	1.890	1.931	7.407	0.955
800	Projection	1.098	1.096	4.305	0.950
	Autoencoder	1.051	1.053	4.121	0.965
	VAE	0.852	0.850	3.340	0.965
	MediEncoder	0.768	0.766	3.011	0.955
1200	Projection	0.753	0.755	2.952	0.945
	Autoencoder	0.764	0.764	2.996	0.950
	VAE	0.716	0.717	2.806	0.965
	MediEncoder	0.661	0.661	2.593	0.950
2000	Projection	0.559	0.558	2.190	0.960
	Autoencoder	0.621	0.623	2.433	0.955
	VAE	0.528	0.528	2.069	0.970
	MediEncoder	0.514	0.513	2.014	0.965
3000	Projection	0.561	0.569	2.201	0.955
	Autoencoder	0.922	0.926	3.614	0.990
	VAE	0.494	0.494	1.937	0.945
	MediEncoder	0.471	0.470	1.846	0.955

Table 1: Estimator performance comparison by sample size under the nonlinear wavelet DGP with $p = 2000$, $q = 1000$, $\sigma_X = 2$, $\sigma_M = 1$, $\sigma_Y = 1$, $\bar{p} = \bar{q} = 5$, $\tilde{p} = \tilde{q} = 10$, and $B = 200$ replications.

5.3 Sensitivity Analyses

In this section, we provide three additional sensitivity analyses to examine the robustness of MediEncoder to practical choices in representation learning. Specifically, we study: (i) sensitivity to the working representation dimensions $(\tilde{p}, \tilde{q})$; (ii) whether replacing the deterministic autoencoders with a variational formulation materially changes estimation and inference; and (iii) sensitivity to the capacity of the cross-factor network g_{XM}. These experiments complement the ablation study in Supplementary Material, Section S6 and examine whether the performance of MediEncoder depends on a narrowly specified representation architecture.

Sensitivity to the working representation dimensions. Our framework does not require the working representation dimensions $(\tilde{p}, \tilde{q})$ to equal the true latent dimensions $(\bar{p}, \bar{q})$; rather, the learned representations need to retain sufficient information about the latent factors. To examine the finite-sample

	SD		RMSE		CI Length	
	$\lambda_3 = 0$	Tuning	$\lambda_3 = 0$	Tuning	$\lambda_3 = 0$	Tuning
100	111.633	11.739	111.493	11.720	437.591	46.018
300	2.455	1.890	2.451	1.931	9.623	7.407
800	0.977	0.768	0.976	0.766	3.830	3.011
1200	0.729	0.661	0.736	0.661	2.856	2.593
2000	0.571	0.514	0.569	0.513	2.237	2.014
3000	0.602	0.471	0.601	0.470	2.360	1.846

Table 2: Ablation study on the alignment term (λ_3) in `MediEncoder` with $p = 2000$, $q = 1000$, $\sigma_X = 2$, $\sigma_M = 1$, $\sigma_Y = 1$, $\bar{p} = \bar{q} = 5$, and $\tilde{p} = \tilde{q} = 10$.

effect of dimension over-specification, Table 3 varies $\tilde{p} = \tilde{q} \in \{5, 10, 15, 20\}$ while the true latent dimensions are fixed at $\bar{p} = \bar{q} = 3$.

Across all three sample sizes, the estimator remains stable over a wide range of working dimensions. In particular, empirical coverage remains close to the nominal 95% level, ranging from 0.935 to 0.975, and there is no systematic deterioration in either SD or RMSE as the representation dimension increases. The best-performing dimension varies somewhat across sample sizes, indicating that there is no single dimension that is uniformly optimal in finite samples. Overall, these results suggest that MediEncoder is not particularly sensitive to moderate or even substantial over-specification of the latent representation dimension.

n	Dimension	SD	RMSE	CI Length	Coverage
800	$\tilde{p} = \tilde{q} = 5$	0.474	0.474	1.857	0.965
	$\tilde{p} = \tilde{q} = 10$	0.410	0.410	1.605	0.955
	$\tilde{p} = \tilde{q} = 15$	0.373	0.373	1.461	0.970
	$\tilde{p} = \tilde{q} = 20$	0.398	0.398	1.559	0.975
1200	$\tilde{p} = \tilde{q} = 5$	0.337	0.339	1.322	0.935
	$\tilde{p} = \tilde{q} = 10$	0.329	0.329	1.288	0.950
	$\tilde{p} = \tilde{q} = 15$	0.385	0.385	1.508	0.950
	$\tilde{p} = \tilde{q} = 20$	0.352	0.352	1.380	0.945
2000	$\tilde{p} = \tilde{q} = 5$	0.318	0.319	1.246	0.950
	$\tilde{p} = \tilde{q} = 10$	0.347	0.347	1.361	0.940
	$\tilde{p} = \tilde{q} = 15$	0.325	0.325	1.274	0.960
	$\tilde{p} = \tilde{q} = 20$	0.341	0.341	1.335	0.960

Table 3: Sensitivity to the encoding dimensions. MediEncoder with $\tilde{p} = \tilde{q} \in \{5, 10, 15, 20\}$ against the true $\bar{p} = \bar{q} = 3$, at $p = 800$, $q = 200$ and $B = 200$ replicates.

Variational regularization of the learned representations. We next investigate whether the deterministic autoencoder architecture used by MediEncoder is essential for its performance. Table 4 compares three representation learners: a standard VAE applied to the two high-dimensional measurements, the proposed deterministic MediEncoder, and a variational version of MediEncoder, which we denote by MediVAE. MediVAE replaces the deterministic autoencoder components with variational autoencoders while retaining the treatment-aware coupling through g_{XM}.

All three methods maintain coverage close to the nominal level across the considered sample sizes. MediVAE

has a lower RMSE than MediEncoder at $n = 800$, while the two coupled methods perform nearly identically at $n = 1200$ and $n = 2000$. The standard VAE is also competitive in the largest sample. Thus, no representation architecture uniformly dominates across all settings. These results indicate that variational regularization can be incorporated into our framework without compromising inferential validity, we will consider this as an interesting future direction.

n	Method	SD	RMSE	CI Length	Coverage
800	VAE	0.462	0.463	1.810	0.950
	MediVAE	0.329	0.336	1.291	0.965
	MediEncoder	0.410	0.410	1.605	0.955
1200	VAE	0.371	0.372	1.455	0.945
	MediVAE	0.320	0.330	1.214	0.950
	MediEncoder	0.320	**0.328**	1.288	0.950
2000	VAE	0.326	0.326	1.279	0.945
	MediVAE	0.302	**0.347**	1.185	0.955
	MediEncoder	0.347	**0.347**	1.361	0.950

Table 4: Sensitivity to variational regularization with $\bar{p} = \bar{q} = 3$, $\tilde{p} = \tilde{q} = 10$, and $B = 200$ Monte Carlo replications.

Sensitivity to the capacity of the cross-factor network. Finally, Table 5 studies the capacity of the auxiliary cross-factor network g_{XM} while keeping the remaining components of MediEncoder fixed. This experiment directly examines the tradeoff discussed in Remark 3.1: a g_{XM} that is too restrictive may fail to capture the dependence of the mediator representation on $(A, g_X(X))$, whereas an excessively expressive g_{XM} may absorb discrepancies between the two learned representation spaces and thereby weaken the structural role of the alignment loss.

The results are consistent with this intuition. At $n = 800$ and $n = 1200$, the moderate-capacity $(50, 50)$ architecture substantially outperforms both the very small (2) architecture and the larger $(200, 200)$ and $(500, 500)$ architectures in terms of RMSE. At $n = 2000$, the differences become substantially smaller, with the $(50, 50)$ and $(200, 200)$ architectures performing similarly. Thus, the precise capacity of g_{XM} becomes less consequential as the sample size increases, but a moderate-capacity network provides the most stable finite-sample performance overall. This supports our use of a relatively low- to moderate-capacity g_{XM} rather than an unrestricted auxiliary network.

6 Real Data Application: Alzheimer's Disease Neuroimaging Initiative (ADNI)

6.1 Data and Preprocessing

We apply our proposed method to investigate how depressive symptoms influence cognitive decline in Alzheimer's disease, and whether this relationship is mediated by DNA methylation. Understanding this pathway is important because prior studies have established associations between depression and cognitive impairment, but the biological mechanisms—particularly epigenetic mediation—remain less clear. Here, we present the detailed data preprocessing procedure.

Dataset. We conduct our analysis using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal, multi-center observational study designed to validate biomarkers for Alzheimer's disease (AD)

n	g_{XM}	SD	RMSE	CI Length	Coverage
800	(2)	0.614	0.615	2.408	0.950
	(50, 50)	0.410	**0.409**	1.605	0.955
	(200, 200)	0.615	0.615	2.409	0.970
	(500, 500)	0.772	0.773	3.028	0.975
1200	(2)	0.464	0.464	1.819	0.945
	(50, 50)	0.329	**0.328**	1.288	0.950
	(200, 200)	0.357	0.357	1.398	0.940
	(500, 500)	0.385	0.385	1.508	0.935
2000	(2)	0.350	0.359	1.372	0.950
	(50, 50)	0.347	**0.346**	1.361	0.950
	(200, 200)	0.321	0.347	1.260	0.960
	(500, 500)	0.359	0.361	1.407	0.960

Table 5: Sensitivity to the capacity of the cross-factor network g_{XM}. All rows use MediEncoder with $\tilde{p} = \tilde{q} = 10$, $\bar{p} = \bar{q} = 3$, and $B = 200$ Monte Carlo replications; only the architecture of g_{XM} is varied.

clinical trials (Mueller et al., 2005). The dataset is accessible via https://adni.loni.usc.edu/. Our treatment variable A is the Geriatric Depression Scale (GDS), a self-reported measure ranging from 0 to 15, with higher values indicating more severe depressive symptoms. We binarize GDS using the clinically standard threshold (e.g., GDS > 5 indicates depression). The outcome Y is the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog), ranging from 0 to 85, where higher scores reflect greater cognitive impairment (Cano et al., 2010; Raghavan et al., 2013).

Preprocessing. Covariates X comprise demographic and health-related variables, with missing values imputed by `mice` (Van Buuren and Groothuis-Oudshoorn, 2011); adding polynomial and interaction terms gives $\dim(X) = 171$. The mediator M is DNA methylation from the Illumina HumanMethylationEPIC array, preprocessed with `minfi` (Aryee et al., 2014), restricted to CpG sites reported as significantly associated with Alzheimer's disease in large-scale EWAS (Zhang et al., 2020; Lunnon et al., 2014; Battram et al., 2022) and to sites with no missing values, which yields $\dim(M) = 3{,}206$. This leaves $n = 649$ complete cases. Both X and M are standardized column-wise to zero mean and unit variance; the outcome is left on its original scale. The exposure is the binarized depression indicator $A = \mathbf{1}\{\text{GDS} > 5\}$, positive for 132 of the 649 subjects. We report $(\tilde{p}, \tilde{q}) = (18, 2)$ which accounts for approximate 40% of the variance of X and M.

Nuisance estimation. The nuisances entering the efficient influence function are estimated parametrically on the learned representations: ridge regression with the penalty chosen by cross-validation for the continuous nuisances, and ℓ_2-penalized logistic regression with the penalty chosen by cross-validated log-loss for the two propensities. We use a truncation of the two fitted propensities to $[\varepsilon, 1 - \varepsilon]$ with $\varepsilon = 1/\sqrt{n}$ rule. Finally, the reported estimate averages 10 independent four-fold splits. Within each sampling, the $(\lambda_1, \lambda_2, \lambda_3)$ is re-selected by Algorithm 2 of every bootstrap replicate.

6.2 Results

We apply `MediEncoder` to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) to study whether the effect of depressive symptoms on cognitive decline is mediated by DNA methylation. The treatment A is the binarized Geriatric Depression Scale (GDS), with $A = 1$ indicating

Effect	Estimate	Bootstrap Mean	Quantile CI (95%)
Direct	2.8983	3.1340	[1.0097, 5.6699]
Indirect	0.0639	0.0091	[−0.3456, 0.3815]
Total	2.9621	3.1432	[0.9767, 5.5275]

Table 6: Estimated natural direct, natural indirect, and total effects in the ADNI analysis using the binarized threshold GDS > 5.

GDS > 5, a commonly used threshold for elevated depressive symptoms (Greenberg, 2012). The outcome Y is the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog), where higher scores indicate worse cognitive function (Cano et al., 2010; Raghavan et al., 2013). The covariates X include demographic and clinical variables, and the mediators M are high-dimensional DNA methylation features.

After preprocessing (Section 6.1), the analysis includes $n = 649$ subjects, with $\dim(X) = 171$ and $\dim(M) =$ 3,206. Table 6 reports the estimated natural direct effect (NDE), natural indirect effect (NIE), and total effect (TE). The total effect estimate is positive, suggesting that elevated depressive symptoms are associated, under the causal assumptions in Section 2, with worse cognitive outcomes. The point estimates indicate that most of the effect operates through the direct pathway, while the estimated DNA-methylation-mediated indirect effect is positive but smaller. The bootstrap intervals for the direct and total effects exclude zero, whereas the interval for the indirect effect includes zero. Thus, the data provide evidence for an overall adverse effect of elevated depressive symptoms on cognitive outcomes, primarily through the direct pathway, while the DNA-methylation-mediated indirect effect remains statistically uncertain.

7 Conclusion

We introduced `MediEncoder`, a coupled representation-learning framework for high-dimensional causal mediation analysis with nonlinear latent covariate and mediator structures. By combining a structured autoencoder with a cross-fitted efficient influence function-based estimator, the proposed method targets natural direct and indirect effects while accommodating complex high-dimensional measurements. Across simulations, `MediEncoder` achieved lower RMSE and more stable inference than competing methods, and the ablation study showed that the alignment term improves estimation accuracy and efficiency. In the ADNI application, we found a positive total effect of depressive symptoms on cognitive decline; the point estimates suggested that most of this effect operated through the direct pathway, while the estimated DNA-methylation-mediated indirect effect was smaller and more uncertain.

Limitations and future work. The proposed method relies on standard causal assumptions for natural mediation analysis, including no unmeasured mediator–outcome confounding. It is also currently developed for binary treatments, requires tuning representation dimensions and network hyperparameters, and the learned mediator representations may be difficult to interpret biologically. These limitations motivate future work on sensitivity analysis for unmeasured mediator–outcome confounding, extensions to continuous or multi-valued treatments, and methods for improving interpretability of the learned mediator representations.

AI use statement

Generative AI tools were used to polish the language and improve the clarity of the manuscript. The authors take full responsibility for the final content.

Supplementary Material

The supplementary material contains nuisance-function estimation details, wavelet loading specifications, a sufficient condition for factor recovery, descriptions of the baseline methods, proofs of the theoretical results,

additional simulation and ablation results, and detailed comparisons with existing mediation methods.

References

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, 12(7):MSB156651.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of econometrics*, 176(1):18–29.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17.

Battram, T., Yousefi, P., Crawford, G., Prince, C., Babaei, M. S., Sharp, G., Hatcher, C., Vega-Salas, M. J., Khodabakhsh, S., Whitehurst, O., et al. (2022). The ewas catalog: a database of epigenome-wide association studies. *Wellcome open research*, 7:41.

Black, C. N., Bot, M., Scheffer, P. G., Cuijpers, P., and Penninx, B. W. (2015). Is depression associated with increased oxidative stress? a systematic review and meta-analysis. *Psychoneuroendocrinology*, 51:164–175.

Bo, S., Ghassami, A., and Mukherjee, D. (2024). A debiased estimator for the mediation functional in ultra-high-dimensional setting in the presence of interaction effects. *arXiv preprint arXiv:2412.08827*.

Buccellato, F. R., D'Anca, M., Fenoglio, C., Scarpini, E., and Galimberti, D. (2021). Role of oxidative damage in alzheimer's disease and neurodegeneration: From pathogenic mechanisms to biomarker discovery. *Antioxidants*, 10(9):1353.

Cano, S. J., Posner, H. B., Moline, M. L., Hurt, S. W., Swartz, J., Hsu, T., and Hobart, J. C. (2010). The adas-cog in alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(12):1363–1368.

Derkach, A., Pfeiffer, R. M., Chen, T.-H., and Sampson, J. N. (2019). High dimensional mediation analysis with latent variables. *Biometrics*, 75(3):745–756.

Fan, J. and Gu, Y. (2023). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of the American Statistical Association*, pages 1–15.

Fan, J. and Gu, Y. (2024). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of the American Statistical Association*, 119(548):2680–2694.

Fan, J., Jana, S., Kulkarni, S., and Yin, Q. (2025). Factor informed double deep learning for average treatment effect estimation. *arXiv preprint arXiv:2508.17136*.

Fan, J., Ke, Y., and Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of econometrics*, 216(1):71–85.

Fan, J. and Liao, Y. (2022). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association*, 117(538):909–924.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680.

Feng, Y. (2023). Optimal estimation of large-dimensional nonlinear factor models. *arXiv preprint arXiv:2311.07243*.

Goodfellow, I. (2016). Deep learning.

Greenberg, S. A. (2012). The geriatric depression scale (gds). *Best Practices in Nursing Care to Older Adults*, 4(1):1–2.

Guerrero-López, A., Sevilla-Salcedo, C., Gómez-Verdejo, V., and Olmos, P. M. (2022). Multimodal hierarchical variational autoencoders with factor analysis latent space. *arXiv preprint arXiv:2207.09185*.

Guo, X., Li, R., Liu, J., and Zeng, M. (2022). High-dimensional mediation analysis for selecting dna methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association*, 117(539):1110–1121.

Guo, X., Li, R., Liu, J., and Zeng, M. (2023). Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to covid-19 pandemic. *Journal of Econometrics*, 235(1):166–179.

Guo, X., Li, R., Liu, J., and Zeng, M. (2024). Estimations and tests for generalized mediation models with high-dimensional potential mediators. *Journal of Business & Economic Statistics*, 42(1):243–256.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*.

Jiang, Z., Liu, Y., Klein, M. H., Aloui, A., Ren, Y., Li, K., Tarokh, V., and Carlson, D. (2023). Causal mediation analysis with multi-dimensional and indirectly observed mediators. *arXiv preprint arXiv:2306.07918*.

Johnson, N. D., Wiener, H. W., Smith, A. K., Nishitani, S., Absher, D. M., Arnett, D. K., Aslibekyan, S., and Conneely, K. N. (2017). Non-linear patterns in age-related dna methylation may reflect cd4+ t cell differentiation. *Epigenetics*, 12(6):492–503.

Jones, J., Ertefaie, A., and Strawderman, R. L. (2025). Causal mediation analysis: selection with asymptotically valid inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):678–700.

Katerji, M., Filippova, M., and Duerksen-Hughes, P. (2019). Approaches and methods to measure oxidative stress in clinical samples: research applications in the cancer field. *Oxidative medicine and cellular longevity*, 2019(1):1279250.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Livingston, G., Huntley, J., Liu, K. Y., Costafreda, S. G., Selbæk, G., Alladi, S., Ames, D., Banerjee, S., Burns, A., Brayne, C., et al. (2024). Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. *The lancet*, 404(10452):572–628.

Lunnon, K., Smith, R., Hannon, E., De Jager, P. L., Srivastava, G., Volta, M., Troakes, C., Al-Sarraj, S., Burrage, J., Macdonald, R., et al. (2014). Methylomic profiling implicates cortical deregulation of ank1 in alzheimer's disease. *Nature neuroscience*, 17(9):1164–1170.

Miranda, S., Opazo, C., Larrondo, L. F., Muñoz, F. J., Ruiz, F., Leighton, F., and Inestrosa, N. C. (2000). The role of oxidative stress in the toxicity induced by amyloid β-peptide in alzheimer's disease. *Progress in neurobiology*, 62(6):633–648.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877.

Nath, T., Caffo, B., Wager, T., and Lindquist, M. A. (2023). A machine learning based approach towards high-dimensional mediation analysis. *NeuroImage*, 268:119843.

Pandya, C. D., Howell, K. R., and Pillai, A. (2013). Antioxidants as potential therapeutics for neuropsychiatric disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 46:214–223.

Pearl, J. (2001). Direct and indirect effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 373.

Raghavan, N., Samtani, M. N., Farnum, M., Yang, E., Novak, G., Grundman, M., Narayan, V., DiBernardo, A., Initiative, A. D. N., et al. (2013). The adas-cog revisited: novel composite scales based on adas-cog to improve efficiency in mci and early ad trials. *Alzheimer's & Dementia*, 9(1):S21–S31.

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A.-L., Migliavacca, E., Binder, H., Michiels, S., Sauerbrei, W., et al. (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. *BMC medicine*, 21(1):182.

Rakshit, P. and Guo, Z. (2024). Statistical inference in high-dimensional poisson regression with applications to mediation analysis. *arXiv preprint arXiv:2410.20671*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

Vershinina, O., Bacalini, M., Zaikin, A., Franceschi, C., and Ivanchenko, M. (2021). Disentangling age-dependent dna methylation: deterministic, stochastic, and nonlinear. *Scientific reports*, 11(1):9201.

Wang, S. and Huang, Y. (2024). Dp2lm: leveraging deep learning approach for estimation and hypothesis testing on mediation effects with high-dimensional mediators and complex confounders. *Biostatistics*, 25(3):818–832.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2015). On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR.

Wörheide, M. A., Krumsiek, J., Kastenmüller, G., and Arnold, M. (2021). Multi-omics integration in biomedical research–a metabolomics-centric review. *Analytica chimica acta*, 1141:144–162.

Xiu, D. and Shen, Z. (2024). Deep autoencoders for nonlinear factor models: Theory and applications. *Chicago Booth Research Paper*, (25-14).

Xu, S., Liu, L., and Liu, Z. (2022). Deepmed: Semiparametric causal mediation analysis with debiased deep learning. *Advances in Neural Information Processing Systems*, 35:28238–28251.

Zhang, L., Silva, T. C., Young, J. I., Gomez, L., Schmidt, M. A., Hamilton-Nelson, K. L., Kunkle, B. W., Chen, X., Martin, E. R., and Wang, L. (2020). Epigenome-wide meta-analysis of dna methylation differences in prefrontal cortex implicates the immune processes in alzheimer's disease. *Nature communications*, 11(1):6114.

Zhang, Q. (2022). High-dimensional mediation analysis with applications to causal gene identification. *Statistics in biosciences*, 14(3):432–451.

Zhao, Y. and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749*.

Zhao, Y. and Luo, X. (2021). Pathway lasso: pathway estimation and selection with high-dimensional mediators. *Statistics and its Interface*, 15(1):39.

Supplementary Material for

"MediEncoder: Nonlinear Representation Learning for High-Dimensional Causal Mediation Analysis"

This supplementary material provides nuisance-function estimation details (Section S1), wavelet loading specifications (Section S2), a sufficient condition for factor recovery (Section S3), and descriptions of the baseline methods (Section S4). It also contains proofs of the theoretical results (Section S5), additional simulation and ablation results (Section S6), and methodological and numerical comparisons with existing mediation methods (Section S7).

S1 Nuisance Function Estimation

In this section, we give the nuisance-function fitting objectives and density-ratio construction used in Line 7 of Algorithm 1. The nuisance functions are fitted on the designated nuisance-estimation fold using the learned representations.

- Estimate nuisance functions on $\mathcal{I}_{\mathrm{nu}}^{(k)}$:

$$\begin{aligned}
\hat{\pi}_1^{(k)} &\in \arg\min_{g\in\mathcal{F}_\pi} \sum_{i\in\mathcal{I}_{\mathrm{nu}}^{(k)}} \Big[-A_i \log g(\widehat{\boldsymbol{f}}_{X,i}^{(k)}) - (1-A_i)\log\{1-g(\widehat{\boldsymbol{f}}_{X,i}^{(k)})\}\Big], \\
\hat{\mu}_1^{(k)} &\in \arg\min_{g\in\mathcal{F}_\mu} \sum_{\substack{i\in\mathcal{I}_{\mathrm{nu}}^{(k)} \\ A_i=1}} \Big[Y_i - g(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)})\Big]^2, \\
\hat{\rho}^{(k)} &\in \arg\min_{g\in\mathcal{F}_\rho} \sum_{i\in\mathcal{I}_{\mathrm{nu}}^{(k)}} \Big[-A_i \log g(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)}) - (1-A_i)\log\{1-g(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)})\}\Big], \\
\hat{\mu}_{10}^{(k)} &\in \arg\min_{g\in\mathcal{F}_{\mu_{10}}} \sum_{\substack{i\in\mathcal{I}_{\mathrm{nu}}^{(k)} \\ A_i=0}} \Big[\hat{\mu}_1^{(k)}(\widehat{\boldsymbol{f}}_{X,i}^{(k)}, \widehat{\boldsymbol{f}}_{M,i}^{(k)}) - g(\widehat{\boldsymbol{f}}_{X,i}^{(k)})\Big]^2.
\end{aligned} \tag{S1.1}$$

- Estimate the density ratio $\pi_2(f_X, f_M) = p(f_M \mid A=0, f_X)/p(f_M \mid A=1, f_X)$ by Bayes' rule:

$$\hat{\pi}_2^{(k)}(u,v) = \frac{1-\hat{\rho}^{(k)}(u,v)}{\hat{\rho}^{(k)}(u,v)} \frac{\hat{\pi}_1^{(k)}(u)}{1-\hat{\pi}_1^{(k)}(u)}. \tag{S1.2}$$

S2 Wavelet Loading

In this section, we specify the Haar-wavelet loading functions used in the simulation design of Section 5.1. We describe the basis functions, their scales and locations, and the coefficients in the additive loadings over the latent coordinates.

Haar mother wavelet. We adopt the standard Haar mother wavelet $\psi : \mathbb{R} \to \mathbb{R}$, defined by

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1, \\ -1, & 1 \le t < 2, \\ 0, & \text{otherwise.} \end{cases}$$

This function satisfies

$$\int_{\mathbb{R}} \psi(t)\, dt = 0, \qquad \|\psi\|_{L^2} = \sqrt{2}.$$

Wavelet atoms. For each wavelet atom ϕ_ℓ, we independently sample an integer-valued scale parameter $r_\ell \sim \mathrm{Unif}\{r_{\min}, \dots, r_{\max}\}$. To ensure that the wavelet atoms overlap with the latent support, we sample a location $t_\ell \sim \mathrm{Unif}[x_{\min}, x_{\max}]$, where $[x_{\min}, x_{\max}]$ denotes the range of the latent variables. The corresponding translation parameter is then defined as $s_\ell = \lfloor 2^{r_\ell} t_\ell \rfloor$. Given the sampled pair (r_ℓ, s_ℓ), the wavelet atom is $\psi_{r_\ell, s_\ell}(t) = 2^{r_\ell/2}\, \psi(2^{r_\ell} t - s_\ell)$.

Additive nonlinear loading. For each latent coordinate i, we select a collection of L wavelet atoms $\{\phi_\ell(\cdot)\}_{\ell=1}^{L} \subset \{\psi_{r,s}(\cdot)\}$. The loading functions are defined as

$$g_{X,j}(f_X) = \sum_{i=1}^{\bar{p}} \sum_{\ell=1}^{L} \lambda_{X,ji\ell}^{\text{wavelet}}\, \phi_\ell(f_{X_i}), \qquad g_{M,k}(f_M) = \sum_{i=1}^{\bar{q}} \sum_{\ell=1}^{L} \lambda_{M,ki\ell}^{\text{wavelet}}\, \phi_\ell(f_{M_i}),$$

where $\lambda_{X,jm\ell}^{\text{wavelet}} \sim \mathcal{N}\left(0, \frac{1}{L}\right), \lambda_{M,km\ell}^{\text{wavelet}} \sim \mathcal{N}\left(0, \frac{1}{L}\right)$, drawn once at initialization.

S3 A sufficient condition for factor recovery

In this section, we relate the representation-recovery requirement in Section 4 to the population objective of MediEncoder. Proposition S3.1 bounds factor-recovery error in terms of excess risk and encoder approximation error under a quotient-stability condition; its proof is given in Section S5.4.

Proposition S3.1 (Sufficient condition for factor recovery)**.** *Let*

$$\mathcal{L}_\lambda(z_{XM}) = \mathbb{E}\left[\lambda_1 \|X - h_X(g_X(X))\|_2^2 + \lambda_2 \|M - h_M(g_M(M))\|_2^2 + \lambda_3 \|g_M(M) - g_{XM}(A, g_X(X))\|_2^2\right]$$

denote the population MediEncoder objective, where $z = (g_X, g_M, g_{XM}, h_X, h_M) \in \mathcal{Z}$. Let

$$\mathcal{L}_\lambda^\star := \inf_{z \in \mathcal{Z}} \mathcal{L}_\lambda(z).$$

Let $\mathcal{V}$ be a class of maps $\nu = (\nu_X, \nu_M) : \mathbb{R}^{\bar{p}+\bar{q}} \to \mathbb{R}^{\tilde{p}+\tilde{q}}$ such that each $\nu \in \mathcal{V}$ admits a Lipschitz left inverse on the support of $(\boldsymbol{f}_X, \boldsymbol{f}_M)$. Assume that the representation class contains oracle encoders $g_X^\star$ and $g_M^\star$, together with a transformation $\nu^\star = (\nu_X^\star, \nu_M^\star) \in \mathcal{V}$, such that

$$\|g_X^\star(X) - \nu_X^\star(\boldsymbol{f}_X)\|_{L_2(P)} \le \delta_{X,\text{app}}, \qquad \|g_M^\star(M) - \nu_M^\star(\boldsymbol{f}_M)\|_{L_2(P)} \le \delta_{M,\text{app}}.$$

Assume further that the population objective satisfies the quotient-stability condition

$$d_{\mathcal{V}}^2(z) \le C\{\mathcal{L}_\lambda(z) - \mathcal{L}_\lambda^\star\} + C(\delta_{X,\text{app}}^2 + \delta_{M,\text{app}}^2), \tag{S3.1}$$

where, for $z \in \mathcal{Z}$,

$$d_{\mathcal{V}}^2(z) := \inf_{\nu \in \mathcal{V}} \left\{ \|g_X(X) - \nu_X(\boldsymbol{f}_X)\|_{L_2(P)}^2 + \|g_M(M) - \nu_M(\boldsymbol{f}_M)\|_{L_2(P)}^2 \right\}.$$

Suppose the fitted MediEncoder $\hat{z} = (\hat{g}_X, \hat{g}_M, \hat{g}_{XM}, \hat{h}_X, \hat{h}_M)$ satisfies the excess-risk bound

$$\mathcal{L}_\lambda(\hat{z}) - \mathcal{L}_\lambda^\star = O_p(r_n^2). \tag{S3.2}$$

Then there exists a possibly n-dependent transformation $\nu_n = (\nu_{X,n}, \nu_{M,n}) \in \mathcal{V}$ such that

$$\|\hat{g}_X(X) - \nu_{X,n}(\boldsymbol{f}_X)\|_{L_2(P)} + \|\hat{g}_M(M) - \nu_{M,n}(\boldsymbol{f}_M)\|_{L_2(P)} = O_p(r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}}).$$

In particular, if $r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}} = o(n^{-1/4})$, then the factor-recovery condition

$$\|\hat{\boldsymbol{f}} - \nu_n(\boldsymbol{f})\|_{L_2(P)} = o_p(n^{-1/4})$$

holds, where

$$\hat{\boldsymbol{f}} := (\hat{g}_X(X), \hat{g}_M(M)), \qquad \nu_n(\boldsymbol{f}) := (\nu_{X,n}(\boldsymbol{f}_X), \nu_{M,n}(\boldsymbol{f}_M)).$$

Remark S3.1. *Condition (S3.1) is the natural analog of a curvature condition for autoencoder-based representations. A fixed encoder is generally not identifiable from reconstruction loss, since invertible reparametrizations of the bottleneck coordinates can be absorbed by the decoder. The distance $d_{\mathcal{V}}$ therefore measures recovery only up to the equivalence class of Lipschitz left-invertible transformations, which is precisely the level of recovery required by Theorem 4.1. The terms $\delta_{X,\mathrm{app}}$ and $\delta_{M,\mathrm{app}}$ represent the approximation errors, which depend on the expressability of the underlying function class used for the encoders.*

S4 Overview of Baseline Methods

In this section, we summarize the diversified-projection, autoencoder, and variational-autoencoder baselines considered in Section 5. These methods supply low-dimensional representations of the covariates X and mediators M for downstream mediation effect estimation.

Diversified Projection. Diversified projection method (Fan and Liao, 2022; Fan and Gu, 2023) constructs a linear surrogate of latent factors using a projection matrix $W \in \mathbb{R}^{p\times\tilde{r}}$ that satisfies boundedness and significance conditions. The estimated factors are obtained as

$$\hat{\boldsymbol{f}} = p^{-1}W^{\top}x,$$

which can be viewed as an affine approximation of the true latent factor f:

$$\hat{\boldsymbol{f}} = H\boldsymbol{f} + \xi, \quad \text{with } H = p^{-1}W^{\top}B.$$

Under mild conditions, the signal term Hf dominates the noise ξ, allowing $\tilde{f}$ to serve as a consistent proxy for f. This approach avoids explicit estimation of the factor loading matrix and does not require precise knowledge of the factor dimension.

Autoencoder. Autoencoders (AE) learn nonlinear low-dimensional representations by jointly training an encoder $g(\cdot)$ and a decoder $h(\cdot)$. Given input x, the latent representation is $z = g(x)$ and the reconstruction is $\hat{x} = h(z)$. The model is trained by minimizing the reconstruction loss:

$$\min_{g,h} \sum_{i=1}^{n} \|x_i - h(g(x_i))\|_2^2.$$

In our setting, autoencoders are applied separately to X and M to obtain latent representations, which may fail to capture shared structure across modalities.

Variational Autoencoder. Variational autoencoders (VAE) extend autoencoders by introducing a probabilistic latent-variable model. The encoder defines a variational posterior $q_\phi(z|x)$, while the decoder specifies a likelihood $p_\theta(x|z)$. The model is trained by maximizing the evidence lower bound (ELBO), equivalently minimizing:

$$\mathcal{L}_{\mathrm{VAE}} = \mathbb{E}_{q_\phi(z|x)}\big[\|x - \hat{x}\|_2^2\big] + \mathrm{KL}\big(q_\phi(z|x)\,\|\,p(z)\big),$$

where $p(z)$ is a prior distribution. Similar to AE, VAEs are typically applied independently to X and M, which may not capture their joint dependence structure.

S5 Proofs

S5.1 Proof of Theorem 4.1

For notational simplicity, we present the analysis for a single fold (e.g., $k = 1$, so $|\mathcal{I}_4|$ is used for evaluation of $\hat{\theta}_0^{\mathrm{IF}}$ and write $n = |\mathcal{I}_4|$ for simplicity). The arguments for the remaining folds follow analogously by permuting

the roles of $\mathcal{I}_1, \ldots, \mathcal{I}_4$ (see Algorithm 1).

Decomposition of $\hat{\theta}_0^{\mathrm{IF}} - \theta_0$. Our goal in this step is to decompose $\hat{\theta}_0^{\mathrm{IF}} - \theta_0$ as the leading influence-function term and the remainder terms. Although the Algorithm 1 estimates the auxiliary propensity $\rho(\boldsymbol{f}_X, \boldsymbol{f}_M) := \mathbb{P}(A = 1 \mid \boldsymbol{f}_X, \boldsymbol{f}_M)$ rather than estimating the mediator density ratio directly, we keep the notation π_2 for readability in this part of the proof. Throughout the proof, π_2 is understood as the Bayes-rule representation $\pi_2(\boldsymbol{f}_X, \boldsymbol{f}_M) = \frac{1-\rho(\boldsymbol{f}_X, \boldsymbol{f}_M)}{\rho(\boldsymbol{f}_X, \boldsymbol{f}_M)} \cdot \frac{\pi_1(\boldsymbol{f}_X)}{1-\pi_1(\boldsymbol{f}_X)}$.

We first claim that the estimator admits the decomposition

$$\hat{\theta}_0^{\mathrm{IF}} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \Psi_i + R_0 + R_1 + R_2 + R_3 + R_4,$$

where the oracle influence-function score

$$\begin{aligned}\Psi_i := \mu_{10}(\boldsymbol{f}_{X,i}) &+ \frac{A_i}{\pi_1(\boldsymbol{f}_{X,i})}\pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \\ &+ \frac{1 - A_i}{1 - \pi_1(\boldsymbol{f}_{X,i})}\big(\mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big) - \theta_0.\end{aligned}$$

The remainder terms collect the differences between the estimated nuisance quantities and their population counterparts. Specifically,

$$\begin{aligned}
R_0 &= \frac{1}{n}\sum_{i=1}^{n} A_i \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})}\right)\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big), \\
R_1 &= \frac{1}{n}\sum_{i=1}^{n} A_i \frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big), \\
R_2 &= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1 - A_i}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - A_i \frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\right]\big(\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big), \\
R_3 &= \frac{1}{n}\sum_{i=1}^{n} (1 - A_i)\big(\mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big)\left(\frac{1}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{1 - \pi_1(\boldsymbol{f}_{X,i})}\right), \\
R_4 &= \frac{1}{n}\sum_{i=1}^{n} \frac{A_i - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\big(\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big).
\end{aligned}$$

For readibility, we suppress the arguments of all functions throughout the following derivation. For example, we write $\hat{\pi}_1 = \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}), \pi_1 = \pi_1(\boldsymbol{f}_{X,i}), \hat{\pi}_2 = \hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}), \pi_2 = \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})$, and similarly for $\hat{\mu}_1$, μ_1 and $\hat{\mu}_{10}$, μ_{10}.

We now prove this decomposition step by step. By definition of $\hat{\theta}_0^{\mathrm{IF}}$, we have

$$\hat{\theta}_0^{\mathrm{IF}} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_{10} + \frac{A_i}{\hat{\pi}_1}\hat{\pi}_2\big(Y_i - \hat{\mu}_1\big) + \frac{1 - A_i}{1 - \hat{\pi}_1}\big(\hat{\mu}_1 - \hat{\mu}_{10}\big)\right] - \theta_0. \tag{S5.1}$$

Introduce

$$\begin{aligned}
Z_{1,i} &:= \frac{A_i}{\hat{\pi}_1}\hat{\pi}_2\big(Y_i - \hat{\mu}_1\big) - \frac{A_i}{\pi_1}\pi_2\big(Y_i - \mu_1\big), \\
Z_{0,i} &:= \frac{1 - A_i}{1 - \hat{\pi}_1}\big(\hat{\mu}_1 - \hat{\mu}_{10}\big) - \frac{1 - A_i}{1 - \pi_1}\big(\mu_1 - \mu_{10}\big).
\end{aligned}$$

Then (S5.1) can be rewritten as

$$\hat{\theta}_0^{\mathrm{IF}} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\Psi_i + \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\mu}_{10} - \mu_{10}\big) + \frac{1}{n}\sum_{i=1}^{n} Z_{1,i} + \frac{1}{n}\sum_{i=1}^{n} Z_{0,i}, \tag{S5.2}$$

where (S5.2) is obtained by adding and subtracting Ψ_i and regrouping terms. We next expand $Z_{1,i}$ in (S5.2). By direct algebra,

$$
\begin{aligned}
Z_{1,i} &= \frac{A_i}{\hat{\pi}_1}\hat{\pi}_2\big(Y_i - \hat{\mu}_1\big) - \frac{A_i}{\pi_1}\pi_2\big(Y_i - \mu_1\big) \\
&= A_i\left[\frac{\hat{\pi}_2}{\hat{\pi}_1}\big(Y_i - \hat{\mu}_1\big) - \frac{\pi_2}{\pi_1}\big(Y_i - \mu_1\big)\right] \\
&= A_i\left[\frac{\hat{\pi}_2}{\hat{\pi}_1}\big(Y_i - \hat{\mu}_1\big) - \frac{\hat{\pi}_2}{\hat{\pi}_1}\big(Y_i - \mu_1\big) + \frac{\hat{\pi}_2}{\hat{\pi}_1}\big(Y_i - \mu_1\big) - \frac{\pi_2}{\pi_1}\big(Y_i - \mu_1\big)\right] \\
&= A_i\left[\frac{\hat{\pi}_2}{\hat{\pi}_1}\big(\mu_1 - \hat{\mu}_1\big) + \left(\frac{\hat{\pi}_2}{\hat{\pi}_1} - \frac{\pi_2}{\pi_1}\right)\big(Y_i - \mu_1\big)\right] \\
&= A_i\frac{\hat{\pi}_2}{\hat{\pi}_1}\big(\mu_1 - \hat{\mu}_1\big) + A_i\left[\frac{\hat{\pi}_2 - \pi_2}{\hat{\pi}_1} + \pi_2\left(\frac{1}{\hat{\pi}_1} - \frac{1}{\pi_1}\right)\right]\big(Y_i - \mu_1\big) \\
&= A_i\frac{\hat{\pi}_2}{\hat{\pi}_1}\big(\mu_1 - \hat{\mu}_1\big) + A_i\frac{\hat{\pi}_2 - \pi_2}{\hat{\pi}_1}\big(Y_i - \mu_1\big) + A_i\pi_2\left(\frac{1}{\hat{\pi}_1} - \frac{1}{\pi_1}\right)\big(Y_i - \mu_1\big).
\end{aligned}
$$

We next expand $Z_{0,i}$ in (S5.2). Again, by direct algebra,

$$
\begin{aligned}
Z_{0,i} &= \hat{\mu}_{10} - \mu_{10} + \frac{1 - A_i}{1 - \hat{\pi}_1}\big(\hat{\mu}_1 - \hat{\mu}_{10}\big) - \frac{1 - A_i}{1 - \pi_1}\big(\mu_1 - \mu_{10}\big) \\
&= \hat{\mu}_{10} - \mu_{10} + (1 - A_i)\left[\frac{\hat{\mu}_1 - \hat{\mu}_{10}}{1 - \hat{\pi}_1} - \frac{\mu_1 - \mu_{10}}{1 - \pi_1}\right] \\
&= \hat{\mu}_{10} - \mu_{10} + (1 - A_i)\left[\frac{\hat{\mu}_1 - \hat{\mu}_{10}}{1 - \hat{\pi}_1} - \frac{\mu_1 - \mu_{10}}{1 - \hat{\pi}_1} + \frac{\mu_1 - \mu_{10}}{1 - \hat{\pi}_1} - \frac{\mu_1 - \mu_{10}}{1 - \pi_1}\right] \\
&= \hat{\mu}_{10} - \mu_{10} + (1 - A_i)\left[\frac{\hat{\mu}_1 - \hat{\mu}_{10} - \mu_1 + \mu_{10}}{1 - \hat{\pi}_1} + (\mu_1 - \mu_{10})\left(\frac{1}{1 - \hat{\pi}_1} - \frac{1}{1 - \pi_1}\right)\right] \\
&= \hat{\mu}_{10} - \mu_{10} + (1 - A_i)\left[\frac{\hat{\mu}_1 - \mu_1}{1 - \hat{\pi}_1} - \frac{\hat{\mu}_{10} - \mu_{10}}{1 - \hat{\pi}_1} + (\mu_1 - \mu_{10})\left(\frac{1}{1 - \hat{\pi}_1} - \frac{1}{1 - \pi_1}\right)\right] \\
&= \hat{\mu}_{10} - \mu_{10} - \frac{1 - A_i}{1 - \hat{\pi}_1}\big(\hat{\mu}_{10} - \mu_{10}\big) + \frac{1 - A_i}{1 - \hat{\pi}_1}\big(\hat{\mu}_1 - \mu_1\big) \\
&\qquad + (1 - A_i)(\mu_1 - \mu_{10})\left(\frac{1}{1 - \hat{\pi}_1} - \frac{1}{1 - \pi_1}\right) \\
&= \left[1 - \frac{1 - A_i}{1 - \hat{\pi}_1}\right](\hat{\mu}_{10} - \mu_{10}) + \frac{1 - A_i}{1 - \hat{\pi}_1}(\hat{\mu}_1 - \mu_1) \\
&\qquad + (1 - A_i)(\mu_1 - \mu_{10})\left(\frac{1}{1 - \hat{\pi}_1} - \frac{1}{1 - \pi_1}\right) \\
&= \frac{A_i - \hat{\pi}_1}{1 - \hat{\pi}_1}(\hat{\mu}_{10} - \mu_{10}) + \frac{1 - A_i}{1 - \hat{\pi}_1}(\hat{\mu}_1 - \mu_1) \\
&\qquad + (1 - A_i)(\mu_1 - \mu_{10})\left(\frac{1}{1 - \hat{\pi}_1} - \frac{1}{1 - \pi_1}\right).
\end{aligned}
$$

Substituting the expansions of $Z_{1,i}$ and $Z_{0,i}$ into (S5.2), we obtain

$$
\begin{aligned}
&\hat{\theta}_0^{\mathrm{IF}} - \theta_0 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Psi_i
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n} A_i \frac{\hat{\pi}_2}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)+\frac{1}{n}\sum_{i=1}^{n} A_i \frac{\hat{\pi}_2-\pi_2}{\hat{\pi}_1}(Y_i-\mu_1)+\frac{1}{n}\sum_{i=1}^{n} A_i\pi_2\left(\frac{1}{\hat{\pi}_1}-\frac{1}{\pi_1}\right)(Y_i-\mu_1)\\
&+\frac{1}{n}\sum_{i=1}^{n} \frac{A_i-\hat{\pi}_1}{1-\hat{\pi}_1}(\hat{\mu}_{10}-\mu_{10})+\frac{1}{n}\sum_{i=1}^{n} \frac{1-A_i}{1-\hat{\pi}_1}(\hat{\mu}_1-\mu_1)\\
&+\frac{1}{n}\sum_{i=1}^{n}(1-A_i)(\mu_1-\mu_{10})\left(\frac{1}{1-\hat{\pi}_1}-\frac{1}{1-\pi_1}\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\Psi_i\\
&+\frac{1}{n}\sum_{i=1}^{n} A_i\pi_2\left(\frac{1}{\hat{\pi}_1}-\frac{1}{\pi_1}\right)(Y_i-\mu_1)\\
&+\frac{1}{n}\sum_{i=1}^{n} A_i \frac{\hat{\pi}_2-\pi_2}{\hat{\pi}_1}(Y_i-\mu_1)\\
&+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-A_i}{1-\hat{\pi}_1}-A_i\frac{\hat{\pi}_2}{\hat{\pi}_1}\right](\hat{\mu}_1-\mu_1)\\
&+\frac{1}{n}\sum_{i=1}^{n}(1-A_i)(\mu_1-\mu_{10})\left(\frac{1}{1-\hat{\pi}_1}-\frac{1}{1-\pi_1}\right)\\
&+\frac{1}{n}\sum_{i=1}^{n} \frac{A_i-\hat{\pi}_1}{1-\hat{\pi}_1}(\hat{\mu}_{10}-\mu_{10}).
\end{aligned}
$$

Comparing with the definitions of R_0, R_1, R_2, R_3, R_4, we obtain

$$
\hat{\theta}_0^{\mathrm{IF}}-\theta_0=\frac{1}{n}\sum_{i=1}^{n}\Psi_i+R_0+R_1+R_2+R_3+R_4. \tag{S5.3}
$$

Reduction to the remainder terms. To establish the asymptotic linearity of $\hat{\theta}_0^{\mathrm{IF}}$, it suffices to show that $\sqrt{n}\,R_j=o_p(1),\quad j=0,\ldots,4$. Multiplying the decomposition in Equation (S5.3) by $\sqrt{n}$ yields

$$
\begin{aligned}
\sqrt{n}\big(\hat{\theta}_0^{\mathrm{IF}}-\theta_0\big)&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_i+\sqrt{n}\,R_0+\sqrt{n}\,R_1+\sqrt{n}\,R_2+\sqrt{n}\,R_3+\sqrt{n}\,R_4\\
&=:\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_i+T_1+T_2+T_3+T_4+T_5.
\end{aligned}
$$

Thus, it remains to show that $T_j=o_p(1)$ for $j=1,\ldots,5$, in which case

$$
\sqrt{n}\big(\hat{\theta}_0^{\mathrm{IF}}-\theta_0\big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_i+o_p(1).
$$

We now write the remainder terms explicitly:

$$
\begin{aligned}
T_1&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_i\pi_2(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i})\left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}-\frac{1}{\pi_1(\boldsymbol{f}_{X,i})}\right)\big(Y_i-\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i})\big),\\
T_2&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_i\frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i},\hat{\boldsymbol{f}}_{M,i})-\pi_2(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\big(Y_i-\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i})\big),\\
T_3&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1-A_i}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}-A_i\frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i},\hat{\boldsymbol{f}}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\right]\big(\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i},\hat{\boldsymbol{f}}_{M,i})-\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i})\big),
\end{aligned}
$$

$$T_4 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - A_i)\big(\mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big) \left(\frac{1}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{1 - \pi_1(\boldsymbol{f}_{X,i})} \right),$$
$$T_5 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{A_i - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} \big(\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big).$$

Before proceeding with the analysis, we define the following sigma-algebras:

$$\begin{aligned}
\mathcal{G}_n^{T_1} &:= \sigma\Big(\{\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}, \boldsymbol{u}_{X,i}, \boldsymbol{u}_{M,i} : i \in \mathcal{I}_4\}, \hat{\pi}_1\Big), \\
\mathcal{G}_n^{T_2} &:= \sigma\Big(\{\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}, \boldsymbol{u}_{X,i}, \boldsymbol{u}_{M,i} : i \in \mathcal{I}_4\}, \hat{\pi}_1, \hat{\rho}\Big), \\
\mathcal{G}_n^{T_3} &:= \sigma\Big(\{\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}, \boldsymbol{u}_{X,i}, \boldsymbol{u}_{M,i} : i \in \mathcal{I}_4\}, \hat{\pi}_1, \hat{\rho}, \hat{\mu}_1\Big), \\
\mathcal{G}_n^{T_4} = \mathcal{G}_n^{T_5} &:= \sigma\Big(\{\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}, \boldsymbol{u}_{X,i}, \boldsymbol{u}_{M,i} : i \in \mathcal{I}_4\}, \hat{\pi}_1, \hat{\mu}_{10}\Big).
\end{aligned}$$

Analysis of T_1. We consider a single fold $\mathcal{I}_4$ for evaluation and $|\mathcal{I}_4| = n$, and all summations below are taken over $i \in \mathcal{I}_4$. We first study the conditional expectation of T_1:

$$\begin{aligned}
&\mathbb{E}[T_1 \mid \mathcal{G}_n^{T_1}] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} A_i \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})} \right) \big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \,\middle|\, \mathcal{G}_n^{T_1} \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})} \right) \mathbb{E}\big[A_i\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \mid \mathcal{G}_n^{T_1}\big] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})} \right) \\
&\qquad \cdot \mathbb{E}\big[A_i \mathbb{E}[Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \mid A_i = 1, \boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}] \mid \mathcal{G}_n^{T_1}\big] \\
&= 0.
\end{aligned}$$

Therefore, $\mathbb{E}[T_1 \mid \mathcal{G}_n^{T_1}] = 0$. We next study the conditional variance of T_1. Using conditional independence across observations given $\mathcal{G}_n^{T_1}$, we have:

$$\begin{aligned}
&\mathrm{Var}(T_1 \mid \mathcal{G}_n^{T_1}) \\
&= \frac{1}{n} \sum_{i \in \mathcal{I}_4} \mathrm{Var}\left(A_i \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})} \right) \big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \,\middle|\, \mathcal{G}_n^{T_1} \right) \\
&\leq \frac{1}{n} \sum_{i \in \mathcal{I}_4} \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})^2 \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})} \right)^2 \\
&\qquad \cdot \mathbb{E}\Big[A_i\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big)^2 \,\Big|\, \mathcal{G}_n^{T_1}\Big] \\
&\leq C \frac{1}{n} \sum_{i \in \mathcal{I}_4} \left(\frac{1}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{\pi_1(\boldsymbol{f}_{X,i})} \right)^2 \\
&= C \frac{1}{n} \sum_{i \in \mathcal{I}_4} \left| \frac{\pi_1(\boldsymbol{f}_{X,i}) - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}) \pi_1(\boldsymbol{f}_{X,i})} \right|^2 \\
&\leq C \frac{1}{n} \sum_{i \in \mathcal{I}_4} \big(\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}) - \pi_1(\boldsymbol{f}_{X,i})\big)^2 \\
&= C \left\| \hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X) \right\|_{L_2(P_n)}^2
\end{aligned}$$

$$
\begin{aligned}
&= C\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) + \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X)\right\|_{L_2(P_n)}^2 \\
&\le C\Big(\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|_{L_2(P_n)} + \left\|\pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X)\right\|_{L_2(P_n)}\Big)^2 \\
&\le C\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|_{L_2(P_n)}^2 + C\left\|\pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X)\right\|_{L_2(P_n)}^2.
\end{aligned}
$$

The last inequality follows from Young's inequality, equivalently $(a+b)^2 \le 2a^2 + 2b^2$. The second inequality above uses $A_i \le 1$, boundedness of π_2, π_1 and $\hat{\pi}_1$, and the bounded conditional second moment in Assumption 4.1. It remains to bound the first empirical $L_2(P_n)$ term. We first add and subtract its population counterpart:

$$
\begin{aligned}
&\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|_{L_2(P_n)}^2 \\
&\quad = \left\|\hat{\pi}_1 - \pi_1 \circ v_X^{-1}\right\|_{L_2(P)}^2 + \left\{\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|_{L_2(P_n)}^2 - \left\|\hat{\pi}_1 - \pi_1 \circ v_X^{-1}\right\|_{L_2(P)}^2\right\} \\
&\quad = O_p(\delta_{n,\pi_1}^2 + n^{-1/2}),
\end{aligned}
$$

where the first term follows from the population L_2 rate in Theorem 4.1, and the second term is obtained via Lemma S5.1 applied to the empirical average over the evaluation fold. Combining the above bounds, we obtain $\mathrm{Var}(T_1 \mid \mathcal{G}_n^{T_1}) = O_p(\delta_{n,\pi_1}^2 + n^{-1/2} + \delta_{n,f}^2) = o_p(1)$. Together with $\mathbb{E}[T_1 \mid \mathcal{G}_n^{T_1}] = 0$, Chebyshev's inequality gives, for any $\varepsilon > 0$,

$$
\mathbb{P}(|T_1| > \varepsilon \mid \mathcal{G}_n^{T_1}) \le \frac{\mathrm{Var}(T_1 \mid \mathcal{G}_n^{T_1})}{\varepsilon^2} = o_p(1).
$$

Since $0 \le \mathbb{P}(|T_1| > \varepsilon \mid \mathcal{G}_n^{T_1}) \le 1$, the dominated convergence argument gives $\mathbb{P}(|T_1| > \varepsilon) = \mathbb{E}\big[\mathbb{P}(|T_1| > \varepsilon \mid \mathcal{G}_n^{T_1})\big] \to 0$. Hence $T_1 = o_p(1)$.

Analysis of T_2. Recall that the density ratio $\hat{\pi}_2$ is represented as $\hat{\pi}_2(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) = \frac{1-\hat{\rho}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M)}{\hat{\rho}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M)} \cdot \frac{\hat{\pi}_1(\hat{\boldsymbol{f}}_X)}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_X)}$. We first study the conditional expectation of T_2 given $\mathcal{G}_n^{T_2}$:

$$
\begin{aligned}
\mathbb{E}[T_2 \mid \mathcal{G}_n^{T_2}] &= \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} A_i \frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} \big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \,\middle|\, \mathcal{G}_n^{T_2}\right] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} \mathbb{E}\big[A_i\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \mid \mathcal{G}_n^{T_2}\big] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} \\
&\qquad \cdot \mathbb{E}\big[A_i\, \mathbb{E}[Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \mid A_i = 1, \boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}] \mid \mathcal{G}_n^{T_2}\big] \\
&= 0.
\end{aligned}
$$

We next study the conditional variance of T_2:

$$
\begin{aligned}
&\mathrm{Var}(T_2 \mid \mathcal{G}_n^{T_2}) \\
&= \frac{1}{n} \sum_{i \in \mathcal{I}_4} \mathrm{Var}\left(A_i \frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} \big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big) \,\middle|\, \mathcal{G}_n^{T_2}\right) \\
&\le \frac{1}{n} \sum_{i \in \mathcal{I}_4} \left(\frac{\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\right)^2 \mathbb{E}\Big[A_i\big(Y_i - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big)^2 \,\Big|\, \mathcal{G}_n^{T_2}\Big] \\
&\le C \frac{1}{n} \sum_{i \in \mathcal{I}_4} \big(\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\big)^2.
\end{aligned}
$$

The second inequality uses $A_i \le 1$, positivity of $\hat{\pi}_1$, and the bounded conditional second moment in Assumption 4.1. Now, using the ratio representation of π_2 together with the uniform boundedness of $(\pi_1, \hat{\pi}_1, \rho, \hat{\rho})$, we obtain

$$\begin{aligned}
&\left|\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right| \\
&\le C\Big(\left|\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \rho \circ v^{-1}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})\right| + \left|\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i})\right| \\
&\qquad + \left|\rho \circ v^{-1}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right| + \left|\pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) - \pi_1(\boldsymbol{f}_{X,i})\right| \Big),
\end{aligned}$$

where we use that the maps $x \mapsto (1-x)/x$ and $x \mapsto x/(1-x)$ are Lipschitz. Therefore, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i \in \mathcal{I}_4} \left(\hat{\pi}_2(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \pi_2(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right)^2 \\
&\le C \left\|\hat{\rho}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) - \rho \circ v^{-1}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M)\right\|_{L_2(P_n)}^2 + C \left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|_{L_2(P_n)}^2 \\
&\quad + C \left\|\rho \circ v^{-1}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) - \rho(\boldsymbol{f}_X, \boldsymbol{f}_M)\right\|_{L_2(P_n)}^2 + C \left\|\pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X)\right\|_{L_2(P_n)}^2 \\
&= O_p(\delta_{n,\rho}^2 + n^{-1/2}) + O_p(\delta_{n,\pi_1}^2 + n^{-1/2}) + O_p(\delta_{n,f}^2) + O_p(\delta_{n,f}^2) \\
&= O_p(\delta_{n,\rho}^2 + \delta_{n,\pi_1}^2 + \delta_{n,f}^2 + n^{-1/2}).
\end{aligned}$$

where the inequality follows $(a+b+c+d)^2 \le C(a^2+b^2+c^2+d^2)$. The first two rates follow by adding and subtracting their population $L_2(P)$ counterparts, using the nuisance rates in Theorem 4.1, and applying Lemma S5.1. The last two rates follow from the Lipschitz continuity of ρ and π_1 together with the factor-recovery rate in Theorem 4.1. Combining the above bounds, we obtain $\mathrm{Var}(T_2 \mid \mathcal{G}_n^{T_2}) = O_p(\delta_{n,\rho}^2 + \delta_{n,\pi_1}^2 + \delta_{n,f}^2 + n^{-1/2}) = o_p(1)$. Together with $\mathbb{E}[T_2 \mid \mathcal{G}_n^{T_2}] = 0$, Chebyshev's inequality gives, for any $\varepsilon > 0$,

$$\mathbb{P}(|T_2| > \varepsilon \mid \mathcal{G}_n^{T_2}) \le \frac{\mathrm{Var}(T_2 \mid \mathcal{G}_n^{T_2})}{\varepsilon^2} = o_p(1).$$

Hence $T_2 = o_p(1)$ by the similar argument as in proof of T_1.

Analysis of T_3. Consider T_3, which can be expressed as

$$\begin{aligned}
T_3 &= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{(1-A_i) - A_i \frac{1-\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}{\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} \left(\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right) \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{(1-A_i) - A_i \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}}{1 - \pi_1(\boldsymbol{f}_{X,i})} \left(\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \left[\frac{(1-A_i) - A_i \frac{1-\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}{\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}}{1 - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{(1-A_i) - A_i \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}}{1 - \pi_1(\boldsymbol{f}_{X,i})} \right] \\
&\qquad \cdot \left(\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right) \\
&=: T_{31} + T_{32}.
\end{aligned}$$

We first study T_{31}. Then

$$\begin{aligned}
\mathbb{E}[T_{31} \mid \mathcal{G}_n^{T_3}] &= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{1 - \pi_1(\boldsymbol{f}_{X,i})} \\
&\qquad \cdot \mathbb{E}\left[(1-A_i) - A_i \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})} \;\middle|\; \mathcal{G}_n^{T_3}\right] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{1 - \pi_1(\boldsymbol{f}_{X,i})}
\end{aligned}$$

$$
\begin{aligned}
&\quad \cdot \mathbb{E}\left[\mathbb{E}\left[(1-A_i) - A_i \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})} \;\middle|\; \boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}\right] \;\middle|\; \mathcal{G}_n^{T_3}\right] \\
&= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} \frac{\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{1-\pi_1(\boldsymbol{f}_{X,i})} \cdot 0 \\
&= 0,
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbb{E}\left[(1-A_i) - A_i \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})} \;\middle|\; \boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}\right] \\
&= 1 - \rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) - \rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})} \\
&= 0.
\end{aligned}
$$

Therefore, $\mathbb{E}[T_{31} \mid \mathcal{G}_n^{T_3}] = 0$. Next, Given $\mathcal{G}_n^{T_3}$,

$$
\begin{aligned}
&\mathrm{Var}(T_{31} \mid \mathcal{G}_n^{T_3}) \\
&\le C \frac{1}{n} \sum_{i \in \mathcal{I}_4} \left(\hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})\right)^2 \\
&= C \left\| \hat{\mu}_1(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) - \mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) \right\|_{L_2(P_n)}^2 \\
&\le C \left\| \hat{\mu}_1(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) - \mu_1 \circ v^{-1}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) \right\|_{L_2(P_n)}^2 + C \left\| \mu_1 \circ v^{-1}(\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M) - \mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) \right\|_{L_2(P_n)}^2 \\
&= O_p(\delta_{n,\mu_1}^2 + n^{-1/2}) + O_p(\delta_{n,f}^2) \\
&= O_p(\delta_{n,\mu_1}^2 + \delta_{n,f}^2 + n^{-1/2}) = o_p(1).
\end{aligned}
$$

Together with $\mathbb{E}[T_{31} \mid \mathcal{G}_n^{T_3}] = 0$, Chebyshev's inequality and the same dominated-convergence argument as in the proof of T_1 imply $T_{31} = o_p(1)$.

We next study T_{32}. Recall that

$$
T_{32} = \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} Z_{1,i} Z_{2,i},
$$

where

$$
\begin{aligned}
Z_{1,i} &:= \frac{(1-A_i) - A_i \frac{1-\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}{\hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i})}}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{(1-A_i) - A_i \frac{1-\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}{\rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i})}}{1-\pi_1(\boldsymbol{f}_{X,i})}, \\
Z_{2,i} &:= \hat{\mu}_1(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}).
\end{aligned}
$$

By Cauchy–Schwarz,

$$
|T_{32}| \le \frac{1}{\sqrt{n}} \left(\sum_{i \in \mathcal{I}_4} Z_{1,i}^2 \right)^{1/2} \left(\sum_{i \in \mathcal{I}_4} Z_{2,i}^2 \right)^{1/2} = \sqrt{n} \left(\frac{1}{n} \sum_{i \in \mathcal{I}_4} Z_{1,i}^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i \in \mathcal{I}_4} Z_{2,i}^2 \right)^{1/2}.
$$

By positivity and Lipschitz continuity of the maps $x \mapsto (1-x)/x$ and $x \mapsto 1/(1-x)$,

$$
\begin{aligned}
|Z_{1,i}| \le\; & C \left| \hat{\rho}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \rho \circ v^{-1}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) \right| + C \left| \rho \circ v^{-1}(\hat{\boldsymbol{f}}_{X,i}, \hat{\boldsymbol{f}}_{M,i}) - \rho(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) \right| \\
& + C \left| \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) \right| + C \left| \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) - \pi_1(\boldsymbol{f}_{X,i}) \right|.
\end{aligned}
$$

Therefore, using $(a+b+c+d)^2 \le C(a^2+b^2+c^2+d^2)$, Lemma S5.1, and Assumptions in Theorem 4.1, we have:

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i\in\mathcal{I}_4} Z_{1,i}^2\right)^{1/2} &\le C\left\|\hat{\rho}(\hat{\boldsymbol{f}}_X,\hat{\boldsymbol{f}}_M) - \rho\circ v^{-1}(\hat{\boldsymbol{f}}_X,\hat{\boldsymbol{f}}_M)\right\|_{L_2(P_n)} \\
&\quad + C\left\|\rho\circ v^{-1}(\hat{\boldsymbol{f}}_X,\hat{\boldsymbol{f}}_M) - \rho(\boldsymbol{f}_X,\boldsymbol{f}_M)\right\|_{L_2(P_n)} \\
&\quad + C\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1\circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|_{L_2(P_n)} \\
&\quad + C\left\|\pi_1\circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X)\right\|_{L_2(P_n)} \\
&= O_p(\delta_{n,\rho}+n^{-1/2}) + O_p(\delta_{n,f}) + O_p(\delta_{n,\pi_1}+n^{-1/2}) + O_p(\delta_{n,f}) \\
&= O_p(\delta_{n,\rho}+\delta_{n,\pi_1}+\delta_{n,f}+n^{-1/2})\,.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i\in\mathcal{I}_4} Z_{2,i}^2\right)^{1/2} &\le C\left\|\hat{\mu}_1(\hat{\boldsymbol{f}}_X,\hat{\boldsymbol{f}}_M) - \mu_1\circ v^{-1}(\hat{\boldsymbol{f}}_X,\hat{\boldsymbol{f}}_M)\right\|_{L_2(P_n)} \\
&\quad + C\left\|\mu_1\circ v^{-1}(\hat{\boldsymbol{f}}_X,\hat{\boldsymbol{f}}_M) - \mu_1(\boldsymbol{f}_X,\boldsymbol{f}_M)\right\|_{L_2(P_n)} \\
&= O_p(\delta_{n,\mu_1}+n^{-1/2}) + O_p(\delta_{n,f}) \\
&= O_p(\delta_{n,\mu_1}+\delta_{n,f}+n^{-1/2}),
\end{aligned}$$

again by Lemma S5.1. Combining the above bounds gives

$$\begin{aligned}
|T_{32}| &= O_p\Big[\sqrt{n}(\delta_{n,\rho}+\delta_{n,\pi_1}+\delta_{n,f}+n^{-1/2})(\delta_{n,\mu_1}+\delta_{n,f}+n^{-1/2})\Big] \\
&= o_p(1),
\end{aligned}$$

Combining the results for T_{31} and T_{32}, we conclude $T_3 = o_p(1)$.

Analysis of T_4. We first study the conditional expectation of T_4 given $\mathcal{G}_n^{T_4}$:

$$\begin{aligned}
\mathbb{E}[T_4 \mid \mathcal{G}_n^{T_4}] &= \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\left(\frac{1}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{1-\pi_1(\boldsymbol{f}_{X,i})}\right) \\
&\qquad \cdot \mathbb{E}\big[(1-A_i)\big(\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big) \mid \mathcal{G}_n^{T_4}\big] \\
&= \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\left(\frac{1}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{1-\pi_1(\boldsymbol{f}_{X,i})}\right) \\
&\qquad \cdot \mathbb{E}\big[\mathbb{E}\big[(1-A_i)\big(\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big) \mid A_i, \boldsymbol{f}_{X,i}\big] \mid \mathcal{G}_n^{T_4}\big] \\
&= 0.
\end{aligned}$$

The last equality follows from the definition $\mu_{10}(\boldsymbol{f}_X) = \mathbb{E}[\mu_1(\boldsymbol{f}_X,\boldsymbol{f}_M) \mid A=0, \boldsymbol{f}_X]$, which implies

$$\begin{aligned}
&\mathbb{E}\big[(1-A_i)\big(\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big) \mid A_i, \boldsymbol{f}_{X,i}\big] \\
&\quad = (1-A_i)\mathbb{E}[\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i}) \mid A_i, \boldsymbol{f}_{X,i}] = 0.
\end{aligned}$$

We next study the conditional variance:

$$\begin{aligned}
&\mathrm{Var}(T_4 \mid \mathcal{G}_n^{T_4}) \\
&\le \frac{1}{n}\sum_{i\in\mathcal{I}_4}\mathbb{E}\left[(1-A_i)\big(\mu_1(\boldsymbol{f}_{X,i},\boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i})\big)^2\left(\frac{1}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{1}{1-\pi_1(\boldsymbol{f}_{X,i})}\right)^2 \,\middle|\, \mathcal{G}_n^{T_4}\right]
\end{aligned}$$

$$
\begin{aligned}
&\le C\frac{1}{n}\sum_{i\in\mathcal{I}_4}\left(\frac{1}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}-\frac{1}{1-\pi_1(\boldsymbol{f}_{X,i})}\right)^2\\
&\le C\frac{1}{n}\sum_{i\in\mathcal{I}_4}\big(\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})-\pi_1(\boldsymbol{f}_{X,i})\big)^2\\
&= C\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X)-\pi_1(\boldsymbol{f}_X)\right\|^2_{L_2(P_n)}\\
&\le C\left\|\hat{\pi}_1(\hat{\boldsymbol{f}}_X)-\pi_1\circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|^2_{L_2(P_n)}+C\left\|\pi_1\circ v_X^{-1}(\hat{\boldsymbol{f}}_X)-\pi_1(\boldsymbol{f}_X)\right\|^2_{L_2(P_n)}\\
&= O_p(\delta^2_{n,\pi_1}+n^{-1/2})+O_p(\delta^2_{n,f})\\
&= O_p(\delta^2_{n,\pi_1}+\delta^2_{n,f}+n^{-1/2})=o_p(1).
\end{aligned}
$$

The second inequality uses $1-A_i\le 1$, boundedness of π_1 and $\hat{\pi}_1$, and the bounded conditional second moment in Assumption 4.1. Together with $\mathbb{E}[T_4\mid\mathcal{G}_n^{T_4}]=0$, Chebyshev's inequality and the same bounded-convergence-in-probability argument as in the proof of T_1 imply $T_4=o_p(1)$.

Analysis of T_5. Consider T_5, which can be expressed as

$$
\begin{aligned}
T_5 &= \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\frac{A_i-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}\big(\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i})-\mu_{10}(\boldsymbol{f}_{X,i})\big)\\
&= \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\frac{A_i-\pi_1(\boldsymbol{f}_{X,i})}{1-\pi_1(\boldsymbol{f}_{X,i})}\big(\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i})-\mu_{10}(\boldsymbol{f}_{X,i})\big)\\
&\quad+\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\left[\frac{A_i-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}-\frac{A_i-\pi_1(\boldsymbol{f}_{X,i})}{1-\pi_1(\boldsymbol{f}_{X,i})}\right]\\
&\qquad\cdot\big(\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i})-\mu_{10}(\boldsymbol{f}_{X,i})\big)\\
&=: T_{51}+T_{52}.
\end{aligned}
$$

We first study T_{51}. Then

$$
\begin{aligned}
\mathbb{E}[T_{51}\mid\mathcal{G}_n^{T_5}] &= \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\frac{\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i})-\mu_{10}(\boldsymbol{f}_{X,i})}{1-\pi_1(\boldsymbol{f}_{X,i})}\mathbb{E}\big[A_i-\pi_1(\boldsymbol{f}_{X,i})\mid\mathcal{G}_n^{T_5}\big]\\
&= \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}_4}\frac{\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i})-\mu_{10}(\boldsymbol{f}_{X,i})}{1-\pi_1(\boldsymbol{f}_{X,i})}\mathbb{E}\big[\mathbb{E}[A_i-\pi_1(\boldsymbol{f}_{X,i})\mid\boldsymbol{f}_{X,i}]\mid\mathcal{G}_n^{T_5}\big]\\
&= 0.
\end{aligned}
$$

We next study the conditional variance:

$$
\begin{aligned}
&\mathrm{Var}(T_{51}\mid\mathcal{G}_n^{T_5})\\
&\le C\frac{1}{n}\sum_{i\in\mathcal{I}_4}\big(\hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i})-\mu_{10}(\boldsymbol{f}_{X,i})\big)^2\\
&= C\left\|\hat{\mu}_{10}(\hat{\boldsymbol{f}}_X)-\mu_{10}(\boldsymbol{f}_X)\right\|^2_{L_2(P_n)}\\
&\le C\left\|\hat{\mu}_{10}(\hat{\boldsymbol{f}}_X)-\mu_{10}\circ v_X^{-1}(\hat{\boldsymbol{f}}_X)\right\|^2_{L_2(P_n)}+C\left\|\mu_{10}\circ v_X^{-1}(\hat{\boldsymbol{f}}_X)-\mu_{10}(\boldsymbol{f}_X)\right\|^2_{L_2(P_n)}\\
&= O_p(\delta^2_{n,\mu_{10}}+n^{-1/2})+O_p(\delta^2_{n,f})\\
&= O_p(\delta^2_{n,\mu_{10}}+\delta^2_{n,f}+n^{-1/2})=o_p(1).
\end{aligned}
$$

Together with $\mathbb{E}[T_{51}\mid\mathcal{G}_n^{T_5}]=0$, Chebyshev's inequality and the same bounded-convergence-in-probability argument as in the proof of T_1 imply $T_{51}=o_p(1)$.

We next study T_{52}. Let $Z_{1,i} := \frac{A_i - \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})}{1-\hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i})} - \frac{A_i - \pi_1(\boldsymbol{f}_{X,i})}{1-\pi_1(\boldsymbol{f}_{X,i})}$, and $Z_{2,i} := \hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i}) - \mu_{10}(\boldsymbol{f}_{X,i})$. Then, we write

$$T_{52} = \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_4} Z_{1,i} Z_{2,i}.$$

By Cauchy–Schwarz,

$$|T_{52}| \le \frac{1}{\sqrt{n}} \left(\sum_{i \in \mathcal{I}_4} Z_{1,i}^2 \right)^{1/2} \left(\sum_{i \in \mathcal{I}_4} Z_{2,i}^2 \right)^{1/2} = \sqrt{n} \left(\frac{1}{n} \sum_{i \in \mathcal{I}_4} Z_{1,i}^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i \in \mathcal{I}_4} Z_{2,i}^2 \right)^{1/2}.$$

By positivity and Lipschitz continuity of the map $x \mapsto (A_i - x)/(1-x)$ on the relevant range,

$$\begin{aligned} |Z_{1,i}| &\le C \left| \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}) - \pi_1(\boldsymbol{f}_{X,i}) \right| \\ &\le C \left| \hat{\pi}_1(\hat{\boldsymbol{f}}_{X,i}) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) \right| + C \left| \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) - \pi_1(\boldsymbol{f}_{X,i}) \right|. \end{aligned}$$

Therefore,

$$\begin{aligned} \left(\frac{1}{n} \sum_{i \in \mathcal{I}_4} Z_{1,i}^2 \right)^{1/2} &\le C \left\| \hat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) \right\|_{L_2(P_n)} + C \left\| \pi_1 \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X) \right\|_{L_2(P_n)} \\ &= O_p(\delta_{n,\pi_1} + n^{-1/2}) + O_p(\delta_{n,f}) \\ &= O_p(\delta_{n,\pi_1} + \delta_{n,f} + n^{-1/2}). \end{aligned}$$

Similarly,

$$|Z_{2,i}| \le \left| \hat{\mu}_{10}(\hat{\boldsymbol{f}}_{X,i}) - \mu_{10} \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) \right| + \left| \mu_{10} \circ v_X^{-1}(\hat{\boldsymbol{f}}_{X,i}) - \mu_{10}(\boldsymbol{f}_{X,i}) \right|,$$

and hence

$$\begin{aligned} &\left(\frac{1}{n} \sum_{i \in \mathcal{I}_4} Z_{2,i}^2 \right)^{1/2} \\ &\le C \left\| \hat{\mu}_{10}(\hat{\boldsymbol{f}}_X) - \mu_{10} \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) \right\|_{L_2(\mathbb{P}_n)} + C \left\| \mu_{10} \circ v_X^{-1}(\hat{\boldsymbol{f}}_X) - \mu_{10}(\boldsymbol{f}_X) \right\|_{L_2(\mathbb{P}_n)} \\ &= O_p(\delta_{n,\mu_{10}} + n^{-1/2}) + O_p(\delta_{n,f}) \\ &= O_p(\delta_{n,\mu_{10}} + \delta_{n,f} + n^{-1/2}). \end{aligned}$$

Combining the above bounds gives

$$|T_{52}| = O_p\Big[\sqrt{n} (\delta_{n,\pi_1} + \delta_{n,f} + n^{-1/2})(\delta_{n,\mu_{10}} + \delta_{n,f} + n^{-1/2}) \Big] = o_p(1).$$

we obtain $T_{52} = o_p(1)$. Together with $T_{51} = o_p(1)$, this implies $T_5 = o_p(1)$.

Asymptotics. Combining the bounds for $T_1, \ldots, T_5$, we obtain

$$\sqrt{n}\big(\hat{\theta}_0^{\mathrm{IF}} - \theta_0\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_i + o_p(1).$$

Since $\mathbb{E}[\Psi_i] = 0$ and $\mathrm{Var}(\Psi_i) = \sigma_{\mathrm{eff}}^2 < \infty$ by the central limit theorem,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_i \xrightarrow{d} \mathcal{N}(0, \sigma_{\mathrm{eff}}^2).$$

Applying Slutsky's theorem yields

$$\sqrt{n}\big(\hat{\theta}_0^{\mathrm{IF}} - \theta_0\big) \xrightarrow{d} \mathcal{N}(0, \sigma_{\mathrm{eff}}^2).$$

Lemma S5.1. *Suppose $f : \mathcal{X} \to [-B, B]$ where $\mathcal{X} \subseteq \mathbb{R}^d$. Let $X_1, \ldots, X_n \sim P$ for some probability measure P on $\mathcal{X}$. Define $L_2(\mathbb{P}_n)$ and $L_2(P)$ norm of f as:*

$$\|f\|_n := \|f\|_{L_2(\mathbb{P}_n)} = \sqrt{\frac{1}{n}\sum_i f^2(X_i)}, \qquad \|f\|_2 = \|f\|_{L_2(P)} = \sqrt{\int f^2(x)\, dP(x)}\,.$$

Then, we have:

$$\mathbb{P}\left(|\|f\|_n - \|f\|_2| > t\right) \leq \frac{B^2}{nt^2}\,.$$

As a consequence, $|\|f\|_n - \|f\|_2| = O_p(n^{-1/2})$.

S5.2 Proof of Lemma S5.1

From definition we have:

$$|\|f\|_n - \|f\|_2| = \frac{\left|\|f\|_n^2 - \|f\|_2^2\right|}{\|f\|_n + \|f\|_2} \leq \frac{\left|\|f\|_n^2 - \|f\|_2^2\right|}{\|f\|_2}$$

As a consequence, we have:

$$\begin{aligned}
\mathbb{P}\left(|\|f\|_n - \|f\|_2| > t\right) &\leq \mathbb{P}\left(\left|\|f\|_n^2 - \|f\|_2^2\right| > t\|f\|_2\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (f^2(X_i) - \mathbb{E}[f^2(X)])\right| > t\|f\|_2\right) \\
&\leq \frac{\mathsf{var}(f^2(X))}{nt^2\|f\|_2^2} \\
&\leq \frac{\mathbb{E}(f^4(X))}{nt^2\|f\|_2^2} \\
&\leq \frac{B^2\mathbb{E}(f^2(X))}{nt^2\|f\|_2^2} = \frac{B^2}{nt^2}\,.
\end{aligned}$$

The last inequality follows from the fact that $\|f\|_\infty \leq B$. This completes the proof.

S5.3 Proof of Theorem 4.2

The proof mostly follows using a similar argument as in the proof of Theorem 1 of Fan and Gu (2023). We highlight the key parts here. As the proof for $\hat{\pi}_1$ and $\hat{\rho}$ are analogous, we skip the latter for brevity and present the proof of the other three nuisances here.

S5.3.1 Rate for $\widehat{\mu}_1$ under the transformed factor representation

Recall our notation $\boldsymbol{f} = (\boldsymbol{f}_X, \boldsymbol{f}_M)$ denotes the unobserved factors, and $\hat{\boldsymbol{f}} = (\hat{\boldsymbol{f}}_X, \hat{\boldsymbol{f}}_M)$ denotes the estimated factors, i.e. $\hat{\boldsymbol{f}}_X = \hat{g}_X(X)$ and $\hat{\boldsymbol{f}}_M = \hat{g}_M(M)$. By the regression model, we have:

$$Y_i = \mu_1(\boldsymbol{f}_i) + \epsilon_i$$

where $\mathbb{E}[\epsilon_i \mid \boldsymbol{f}_i, \mathbf{u}_{X,i}, \mathbf{u}_{M,i}, A] = 0$ and ϵ_i is sub-gaussian random variable. As $\hat{\boldsymbol{f}}$ is a function of $(\boldsymbol{f}, \mathbf{u}_X, \mathbf{u}_X)$, and the encoder is estimated from a separate dataset, it is immediate that $\mathbb{E}[\epsilon \mid \hat{\boldsymbol{f}}, A] = 0$. We use the notation $\eta = \mu_1 \circ \nu^{-1}$. As per Theorem 4.2, η is bounded, Lipschitz, and moreover $\eta \in \mathcal{H}(d_\eta, \ell_\eta, \mathcal{P})$ with DAS γ_η^*. Let $\mathcal{I}_1 = \{i \in \mathcal{I}_{\text{nuis}} : A_i = 1\}$. As per our assumption,

$$\|\hat{\boldsymbol{f}} - \nu(\boldsymbol{f})\|_{L_2(P_1)} = O_p(\delta_{n,f}),$$

Let $\mathcal{G}_{\eta,n} := \mathcal{G}(L_\eta, d_\eta, 1, N_\eta, M_\eta, B_\eta)$ be a collection of truncated ReLU network class used to estimate η_1, where $d_\eta = \widetilde{p} + \widetilde{q}$ be the input dimension, L_η is the depth, N_η is the width, M_η is the truncation level, and B_η is the weight bound. We take M_η larger than $\|\eta\|_\infty$. The estimator $\hat{\eta}_1$ is defined as:

$$\widehat{\eta} \in \arg\min_{g \in \mathcal{G}_{\eta,n}} \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \left\{ Y_i - g(\hat{\boldsymbol{f}}_i) \right\}^2.$$

By Theorem 4 of Fan and Gu (2024), since $\eta \in \mathcal{H}(d_\eta, \ell_\eta, \mathcal{P})$ and is supported on a bounded domain, there exist constants $C_1, C_2, C_3, C_4 > 0$, depending only on the HCM parameters, such that for any integer $N_\mu \geq 2$, there exists a ReLU network

$$\eta^\dagger \in \mathcal{G}(C_1, d_\eta, 1, C_2 N_\eta, \infty, C_3 N_\eta^{C_4})$$

satisfying

$$\|\eta^\dagger - \eta\|_\infty \leq C N_\eta^{-2\gamma_\eta^\star}.$$

After truncation at level M_η, the approximation error does not increase, provided $M_\eta > \|\eta\|_\infty$. Therefore, by enlarging constants, we may regard $\eta^\dagger \in \mathcal{G}_{\eta,n}$ and retain the same approximation bound. Consequently,

$$\|\eta^\dagger - \eta\|_\infty^2 \leq C N_\eta^{-4\gamma_\eta^\star}. \tag{S5.4}$$

However, our estimator is evaluated at the learned representation $\hat{\boldsymbol{f}}$, whereas the target is $\mu_1(\boldsymbol{f}) = \eta(\nu(\boldsymbol{f}))$. Thus, the relevant oracle error is

$$\|\eta^\dagger(\hat{\boldsymbol{f}}) - \eta(\nu(\boldsymbol{f}))\|_{L_2(P_1)}^2.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\mathbb{E}_1 \left[\left\{ \eta^\dagger(\hat{\boldsymbol{f}}) - \eta(\nu(\boldsymbol{f})) \right\}^2 \right] \leq 2\mathbb{E}_1 \left[\left\{ \eta^\dagger(\hat{\boldsymbol{f}}) - \eta(\hat{\boldsymbol{f}}) \right\}^2 \right] + 2\mathbb{E}_1 \left[\left\{ \eta(\hat{\boldsymbol{f}}) - \eta(\nu(\boldsymbol{f})) \right\}^2 \right].$$

The first term is controlled by the uniform approximation bound:

$$\mathbb{E}_1 \left[\left\{ \eta^\dagger(\hat{\boldsymbol{f}}) - \eta(\hat{\boldsymbol{f}}) \right\}^2 \right] \leq \|\eta^\dagger - \eta\|_\infty^2 \leq C N_\eta^{-4\gamma_\eta^\star}.$$

For the second term, since η is Lipschitz,

$$\mathbb{E}_1 \left[\left\{ \eta(\hat{\boldsymbol{f}}) - \eta(\nu(\boldsymbol{f})) \right\}^2 \right] \leq L_\eta^2 \mathbb{E}_1 \|\hat{\boldsymbol{f}} - \nu(\boldsymbol{f})\|_2^2 = O_p(\delta_{n,f}^2).$$

Therefore,

$$\|\eta^\dagger(\hat{\boldsymbol{f}}) - \eta(\nu(\boldsymbol{f}))\|_{L_2(P_1)}^2 = O_p\left(N_\eta^{-4\gamma_\eta^\star} + \delta_{n,f}^2 \right). \tag{S5.5}$$

Next, we bound the stochastic error. For any $g \in \mathcal{G}_{\eta,n}$, define $r_g(i) := g(\hat{\boldsymbol{f}}_i) - \eta(\nu(\boldsymbol{f}_i))$. In particular,

$$r_{\widehat{\eta}}(i) = \widehat{\eta}(\hat{\boldsymbol{f}}_i) - \eta(\nu(\boldsymbol{f}_i)), \qquad r_{\eta^\dagger}(i) = \eta^\dagger(\hat{\boldsymbol{f}}_i) - \eta(\nu(\boldsymbol{f}_i)).$$

Since $Y_i = \eta(\nu(\boldsymbol{f}_i)) + \varepsilon_i$, the empirical loss can be written as

$$\frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \{Y_i - g(\hat{\boldsymbol{f}}_i)\}^2 = \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \{\varepsilon_i - r_g(i)\}^2.$$

By the empirical-risk minimizing property of $\widehat{\eta}$,

$$\frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \{\varepsilon_i - r_{\widehat{\eta}_1}(i)\}^2 \leq \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \{\varepsilon_i - r_{\eta^\dagger}(i)\}^2.$$

Expanding both sides and canceling the common term $n^{-1}\sum_i \varepsilon_i^2$, we obtain the basic inequality

$$\begin{aligned}\|r_{\widehat{\eta}_1}\|_{n,1}^2 &\le \|r_{\eta^\dagger}\|_{n,1}^2 + \frac{2}{n_1}\sum_{i\in\mathcal{I}_1}\varepsilon_i\left\{\widehat{\eta}_1(\hat{\boldsymbol{f}}_i) - \eta^\dagger(\hat{\boldsymbol{f}}_i)\right\},\\ &\le \|r_{\eta^\dagger}\|_{n,1}^2 + 2\sup_{g\in\mathcal{G}_{\eta,n}}\left|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\varepsilon_i\left\{g(\hat{\boldsymbol{f}}_i) - \eta^\dagger(\hat{\boldsymbol{f}}_i)\right\}\right|, \end{aligned} \tag{S5.6}$$

where

$$\|h\|_{n,1}^2 := \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} h_i^2.$$

Thus, it remains to control the supremum of the empirical process. By Lemma 4 of Fan and Gu (2024), the suprema can be bounded in terms of the pseudo-dimension of the network class. Let $V_n = \mathrm{Pdim}(\mathcal{G}_{\eta,n})$ denotes the pseudo-dimension of $\mathcal{G}_{\eta,n}$. Then with probability at least $1-e^{-t}$,

$$\left|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\varepsilon_i\left\{g(\hat{\boldsymbol{f}}_i) - \eta^\dagger(\hat{\boldsymbol{f}}_i)\right\}\right| \le C\left(\|g(\hat{\boldsymbol{f}}) - \eta^\dagger(\hat{\boldsymbol{f}})\|_{n,1} + \sqrt{\frac{V_n\log n_1}{n_1}}\right)\sqrt{\frac{V_n\log n_1}{n_1}+\frac{t}{n}}$$

simultaneously for all $g\in\mathcal{G}_{\eta,n}$. Applying this inequality to $g=\widehat{\eta}$ gives

$$\begin{aligned}\left|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\varepsilon_i\left\{\widehat{\eta}_1(\hat{\boldsymbol{f}}_i) - \eta^\dagger(\hat{\boldsymbol{f}}_i)\right\}\right| &\le C\left(\|\widehat{\eta}_1(\hat{\boldsymbol{f}}) - \eta^\dagger(\hat{\boldsymbol{f}})\|_{n,1} + \sqrt{\frac{V_n\log n_1}{n_1}}\right)\sqrt{\frac{V_n\log n_1}{n_1}+\frac{t}{n}}\\ &\le C\left(\|r_{\widehat{\eta}_1}\|_{n,1} + \|r_{\eta^\dagger}\|_{n,1} + \sqrt{\frac{V_n\log n_1}{n_1}}\right)\sqrt{\frac{V_n\log n_1}{n_1}+\frac{t}{n}}\end{aligned}$$

where the last line follows from the fact that:

$$\|\widehat{\eta}_1(\hat{\boldsymbol{f}}) - \eta^\dagger(\hat{\boldsymbol{f}})\|_{n,1} \le \|r_{\widehat{\eta}_1}\|_{n,1} + \|r_{\eta^\dagger}\|_{n,1}\,.$$

This, along with Equation (S5.6) yields:

$$\|r_{\widehat{\eta}_1}\|_{n,1}^2 \le \|r_{\eta^\dagger}\|_{n,1}^2 + C\left(\|r_{\widehat{\eta}_1}\|_{n,1} + \|r_{\eta^\dagger}\|_{n,1} + \sqrt{\frac{V_n\log n_1}{n_1}}\right)\sqrt{\frac{V_n\log n_1}{n_1}+\frac{t}{n}}\,.$$

We then apply Young's inequality and change sides to obtain:

$$\|r_{\widehat{\eta}_1}\|_{n,1}^2 \le C\left(\|r_{\eta^\dagger}\|_{n,1}^2 + \frac{V_n\log n_1}{n_1} + \frac{t}{n}\right).$$

Conditional on the representation-learning fold, $r_{\eta^\dagger}$ is a fixed bounded function of the nuisance-fold observation. Therefore, by Bernstein's inequality applied to the fixed oracle $r_{\eta^\dagger}$,

$$\|r_{\eta^\dagger}\|_{n,1}^2 \le C\|r_{\eta^\dagger}\|_{L_2(P_1)}^2 + C\frac{t}{n}$$

with probability at least $1-e^{-t}$. This, along with the previous equation, yields with probability $\ge 1-2e^{-t}$:

$$\|r_{\widehat{\eta}_1}\|_{n,1}^2 \lesssim \left(\|r_{\eta^\dagger}\|_{L_2(P_1)}^2 + \frac{V_n\log n_1}{n_1} + \frac{t}{n}\right) \tag{S5.7}$$

It remains to bound $V_\mu = \mathrm{Pdim}(\mathcal{G}_{\mu,n})$. By Theorem 7 of Bartlett et al. (2019), we have:

$$V_n \le CW_\eta L_\eta \log W_\eta \le C_1 N_\eta^2 L_\eta^2 \log\left(N_\eta^2 L_\eta\right).$$

Choosing $L_\eta \asymp 1$ and $N_\eta \asymp n^{1/(4\gamma_\eta^*+2)}$, we conclude from Equation (S5.5) and (S5.7):

$$\|r_{\widehat{\eta}_1}\|_{n,1}^2 = O_p\left(n^{-\frac{2\gamma_\eta^*}{2\gamma_\eta^*+1}}(\log n)^a + \delta_{n,f}^2\right).$$

Finally, applying the empirical-to-population norm comparison, as in Lemma 3 and the last step of the proof of Theorem 1 of Fan and Gu (2024), gives the same bound for the population norm:

$$\|r_{\widehat{\eta}_1}\|_{L_2(P_1)}^2 = O_p\left(n^{-\frac{2\gamma_\eta^*}{2\gamma_\eta^*+1}}(\log n)^a + \delta_{n,f}^2\right).$$

Remark S5.1. *As $n_1 \sim n$ (by positivity assumption $\mathbb{P}(A=1)>0$) we use n instead of n_1 in the rate.*

S5.3.2 Rate for $\widehat{\pi}_1$ under the transformed factor representation

We now prove the rate for the treatment propensity nuisance $\pi_1(\boldsymbol{f}_X) := \mathbb{P}(A=1 \mid \boldsymbol{f}_X)$. Define as before, the relevant nuisance parameter,

$$\eta := \pi_1 \circ \nu^{-1}, \qquad \eta(\nu(\boldsymbol{f}_X)) = \pi_1(\boldsymbol{f}_X).$$

Let $\mathcal{I}_\pi = \mathcal{I}_{\text{nuis}}$ be the nuisance-estimation fold and write $n = |\mathcal{I}_\pi|$ for simplicity. Assume

$$\|\hat{\boldsymbol{f}}_X - \nu(\boldsymbol{f}_X)\|_{L_2(P)} = O_p(\delta_{n,f}).$$

From our assumption in Theorem 4.1, the propensity score is uniformly bounded away from 0 and 1, i.e., for some $\kappa \in (0, 1/2)$,

$$\kappa \le \eta(\nu(\boldsymbol{f}_X)) \le 1-\kappa$$

almost surely. We restrict the neural-network estimators to take values in $[\kappa/2, 1-\kappa/2]$ by clipping. This does not affect the rate. Suppose η admits a bounded Lipschitz HCM extension to a compact set containing both $\nu(\boldsymbol{f}_X)$ and $\hat{\boldsymbol{f}}_X$ with probability tending to one. Let $\eta \in \mathcal{H}(d_\eta, \ell_\eta, \mathcal{P})$ with DAS γ_η^*, where $d_\eta = \widetilde{p}$ is the input dimension. Let

$$\mathcal{G}_{\eta,n} := \mathcal{G}(L_\eta, d_\eta, 1, N_\eta, M_\eta, B_\eta)$$

be the clipped ReLU network class. As per Line 7 of Algorithm 1

$$\widehat{\eta} \in \arg\min_{g \in \mathcal{G}_{\eta,n}} \frac{1}{n}\sum_{i \in \mathcal{I}_\pi} \ell(A_i, g(\hat{\boldsymbol{f}}_{X,i})) := \arg\min_{g \in \mathcal{G}_{\eta,n}} \mathbb{P}_n(\ell(A, g(\hat{\boldsymbol{f}}_X))$$

where

$$\ell(a,u) := -a\log u - (1-a)\log(1-u), \qquad u \in [\kappa/2, 1-\kappa/2].$$

As in the previous subsection, we can use the approximation theorem (e.g., Theorem 4 of Fan and Gu (2024)), there exists

$$\eta^\dagger \in \mathcal{G}(C_1, d_\eta, 1, C_2 N_\eta, \infty, C_3 N_\eta^{C_4})$$

such that

$$\|\eta^\dagger - \eta\|_\infty \le C N_\eta^{-2\gamma_\eta^\star}.$$

After clipping, we may take $\eta^\dagger \in \mathcal{G}_{\eta,n}$ and retain

$$\|\eta^\dagger - \eta\|_\infty^2 \le C N_\eta^{-4\gamma_\eta^\star}. \tag{S5.8}$$

As in the proof for $\widehat{\mu}_1$, the relevant oracle error is

$$\|\eta^\dagger(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2.$$

Using the same decomposition as in the previous subsection, and the Lipschitzness of η,

$$\|\eta^\dagger(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 \lesssim \|\eta^\dagger - \eta\|_\infty^2 + \|\hat{\boldsymbol{f}}_X - \nu(\boldsymbol{f}_X)\|_{L_2(P)}^2.$$

Therefore,

$$\|\eta^\dagger(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 = O_p\left(N_\eta^{-4\gamma_\eta^\star} + \delta_{n,f}^2\right). \tag{S5.9}$$

By empirical risk minimization,

$$\mathbb{P}_n \ell(A, \widehat{\eta}(\hat{\boldsymbol{f}}_X)) \le \mathbb{P}_n \ell(A, \eta^\dagger(\hat{\boldsymbol{f}}_X)).$$

Adding and subtracting population risks gives

$$\begin{aligned} P\{\ell(A, \widehat{\eta}(\hat{\boldsymbol{f}}_X)) - \ell(A, \eta(\nu(\boldsymbol{f}_X)))\} \le\ & P\{\ell(A, \eta^\dagger(\hat{\boldsymbol{f}}_X)) - \ell(A, \eta(\nu(\boldsymbol{f}_X)))\} \\ & + (P - \mathbb{P}_n)\{\ell(A, \widehat{\eta}(\hat{\boldsymbol{f}}_X)) - \ell(A, \eta^\dagger(\hat{\boldsymbol{f}}_X))\}. \end{aligned} \tag{S5.10}$$

Since $\hat{\boldsymbol{f}}_X$ is a function of $(\boldsymbol{f}_X, \mathbf{u}_X)$ and A is indepndent of $\mathbf{u}_X$,

$$\mathbb{P}(A = 1 \mid \nu(\boldsymbol{f}_X), \hat{\boldsymbol{f}}_X) = \pi_1(\boldsymbol{f}_X) = \eta(\nu(\boldsymbol{f}_X)).$$

Hence, for any $g \in \mathcal{G}_{\eta,n}$,

$$P\{\ell(A, g(\hat{\boldsymbol{f}}_X)) - \ell(A, \eta(\nu(\boldsymbol{f}_X)))\} = \mathbb{E}\left[\mathrm{KL}\{\mathrm{Bern}(\eta(\nu(\boldsymbol{f}_X))) \,||\, \mathrm{Bern}(g(\hat{\boldsymbol{f}}_X))\}\right].$$

By overlap and clipping, this KL divergence is equivalent to squared error:

$$c_\kappa\{g(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\}^2 \le \mathrm{KL}\{\cdot,\cdot\} \le C_\kappa\{g(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\}^2.$$

Therefore, applying the lower bound to $\widehat{\eta}$ and the upper bound to $\eta^\dagger$, (S5.10) yields

$$\begin{aligned} \|\widehat{\eta}(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 \lesssim\ & \|\eta^\dagger(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 \\ & + (P - \mathbb{P}_n)\{\ell(A, \widehat{\eta}(\hat{\boldsymbol{f}}_X)) - \ell(A, \eta^\dagger(\hat{\boldsymbol{f}}_X))\}. \end{aligned} \tag{S5.11}$$

The logistic loss is Lipschitz on $[\kappa/2, 1-\kappa/2]$ since

$$|\partial_u \ell(a,u)| \le \frac{2}{\kappa}.$$

Thus, by the contraction inequality and the pseudo-dimension bound used in Lemma 4 of Fan and Gu (2024), if

$$V_n := \mathrm{Pdim}(\mathcal{G}_{\eta,n}),$$

then with probability at least $1 - e^{-t}$,

$$\begin{aligned} &(P - \mathbb{P}_n)\{\ell(A, \widehat{\eta}(\hat{\boldsymbol{f}}_X)) - \ell(A, \eta^\dagger(\hat{\boldsymbol{f}}_X))\} \\ &\quad \le \zeta\|\widehat{\eta}(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 + C_\zeta\|\eta^\dagger(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 + C_\zeta\left(\frac{V_n \log n}{n} + \frac{t}{n}\right). \end{aligned}$$

Choosing ζ small enough and combining this with (S5.11), we obtain

$$\|\widehat{\eta}(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 \lesssim \|\eta^\dagger(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 + \frac{V_n \log n}{n} + \frac{t}{n}.$$

Using (S5.9),

$$\|\widehat{\eta}(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 = O_p\left(N_\eta^{-4\gamma_\eta^\star} + \delta_{n,f}^2 + \frac{V_n \log n}{n}\right). \tag{S5.12}$$

Finally, as in the proof for $\widehat{\mu}_1$,

$$V_n \lesssim \left(L_\eta^2 N_\eta^2 + d_\eta L_\eta N_\eta\right)\log(L_\eta N_\eta d_\eta).$$

Taking $L_\eta \asymp 1$ and balancing

$$N_\eta^{-4\gamma_\eta^\star} \qquad \text{and} \qquad \frac{N_\eta^2 \log N_\eta \log n}{n}$$

gives

$$N_\eta \asymp n^{1/(4\gamma_\eta^\star+2)}$$

up to logarithmic factors. Therefore,

$$\|\widehat{\eta}(\hat{\boldsymbol{f}}_X) - \eta(\nu(\boldsymbol{f}_X))\|_{L_2(P)}^2 = O_p\left(n^{-2\gamma_\eta^\star/(2\gamma_\eta^\star+1)}(\log n)^a + \delta_{n,f}^2\right).$$

Equivalently,

$$\|\widehat{\pi}_1(\hat{\boldsymbol{f}}_X) - \pi_1(\boldsymbol{f}_X)\|_{L_2(P)}^2 = O_p\left(n^{-2\gamma_\eta^\star/(2\gamma_\eta^\star+1)}(\log n)^a + \delta_{n,f}^2\right).$$

S5.3.3 Rate for $\widehat{\mu}_{10}$ under the transformed factor representation

We now prove the rate for $\mu_{10}(\boldsymbol{f}_X) := \mathbb{E}[\mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M) \mid A = 0, \boldsymbol{f}_X]$. Define the relevant nuisance parameter $\eta := \mu_{10} \circ \nu_X^{-1}$ so that $\eta(\nu_X(\boldsymbol{f}_X)) = \mu_{10}(\boldsymbol{f}_X)$. Let us also define

$$\mathcal{I}_0 := \{i \in \mathcal{I}_{\text{nuis}} : A_i = 0\}, \qquad n_0 := |\mathcal{I}_0|.$$

The estimator $\widehat{\eta}$ is obtained by regressing the generated outcome $\widehat{\mu}_1(\widehat{\boldsymbol{f}}_{X,i}, \widehat{\boldsymbol{f}}_{M,i})$ on $\widehat{\boldsymbol{f}}_{X,i}$:

$$\widehat{\eta} \in \arg\min_{g \in \mathcal{G}_{\eta,n}} \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} \left\{\widehat{\mu}_1(\widehat{\boldsymbol{f}}_{X,i}, \widehat{\boldsymbol{f}}_{M,i}) - g(\widehat{\boldsymbol{f}}_{X,i})\right\}^2.$$

Assume

$$\|\widehat{\boldsymbol{f}}_X - \nu_X(\boldsymbol{f}_X)\|_{L_2(P_0)} = O_p(\delta_{n,f}),$$

where P_0 denotes the distribution conditional on $A = 0$. Assume also that η admits a bounded Lipschitz HCM extension to a compact set containing both $\nu_X(\boldsymbol{f}_X)$ and $\widehat{\boldsymbol{f}}_X$ with probability tending to one, and $\eta \in \mathcal{H}(d_\eta, \ell_\eta, \mathcal{P})$ with DAS γ_η^*. As before, let

$$\mathcal{G}_{\eta,n} := \mathcal{G}(L_\eta, d_\eta, 1, N_\eta, M_\eta, B_\eta)$$

be the truncated ReLU class. By Theorem 4 of Fan and Gu (2024), there exists

$$\eta^\dagger \in \mathcal{G}(C_1, d_\eta, 1, C_2 N_\eta, \infty, C_3 N_\eta^{C_4})$$

such that

$$\|\eta^\dagger - \eta\|_\infty \le C N_\eta^{-2\gamma_\eta^\star}.$$

After truncation, we may regard $\eta^\dagger \in \mathcal{G}_{\eta,n}$ and retain

$$\|\eta^\dagger - \eta\|_\infty^2 \le C N_\eta^{-4\gamma_\eta^\star}.$$

As in the proof of $\widehat{\mu}_1$, the relevant oracle error is

$$\|\eta^\dagger(\widehat{\boldsymbol{f}}_X) - \eta(\nu_X(\boldsymbol{f}_X))\|_{L_2(P_0)}^2.$$

Using the same decomposition as before and Lipschitzness of η,

$$\|\eta^\dagger(\widehat{\boldsymbol{f}}_X) - \eta(\nu_X(\boldsymbol{f}_X))\|_{L_2(P_0)}^2 = O_p\left(N_\eta^{-4\gamma_\eta^\star} + \delta_{n,f}^2\right). \tag{S5.13}$$

Now, for $i \in \mathcal{I}_0$, define

$$\xi_i := \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}) - \mu_{10}(\boldsymbol{f}_{X,i}),$$
$$\Delta_i := \widehat{\mu}_1(\widehat{\boldsymbol{f}}_{X,i}, \widehat{\boldsymbol{f}}_{M,i}) - \mu_1(\boldsymbol{f}_{X,i}, \boldsymbol{f}_{M,i}).$$

Then the generated outcome satisfies

$$\widehat{\mu}_1(\widehat{\boldsymbol{f}}_{X,i}, \widehat{\boldsymbol{f}}_{M,i}) = \eta(\nu_X(\boldsymbol{f}_{X,i})) + \xi_i + \Delta_i.$$

Here ξ_i is the intrinsic second-stage regression noise, while Δ_i is the first-stage error from estimating μ_1. By independent of $(\boldsymbol{f}_X, \boldsymbol{f}_M)$ with $(\mathbf{u}_X, \mathbf{u}_M)$, we have:

$$\mathbb{E}[\xi_i \mid \boldsymbol{f}_{X,i}, \mathbf{u}_{X,i}, A_i = 0] = 0,$$

so that, conditional on the representation-learning fold,

$$\mathbb{E}[\xi_i \mid \widehat{\boldsymbol{f}}_{X,i}, A_i = 0] = 0.$$

For any $g \in \mathcal{G}_{\eta,n}$, define

$$r_g(i) := g(\widehat{\boldsymbol{f}}_{X,i}) - \eta(\nu_X(\boldsymbol{f}_{X,i})).$$

By empirical risk minimization,

$$\frac{1}{n_0}\sum_{i\in\mathcal{I}_0}\{\xi_i + \Delta_i - r_{\widehat{\eta}}(i)\}^2 \le \frac{1}{n_0}\sum_{i\in\mathcal{I}_0}\{\xi_i + \Delta_i - r_{\eta^\dagger}(i)\}^2.$$

Expanding both sides and canceling common terms gives

$$\|r_{\widehat{\eta}}\|_{n,0}^2 \le \|r_{\eta^\dagger}\|_{n,0}^2 + \frac{2}{n_0}\sum_{i\in\mathcal{I}_0}\xi_i\left\{\widehat{\eta}(\widehat{\boldsymbol{f}}_{X,i}) - \eta^\dagger(\widehat{\boldsymbol{f}}_{X,i})\right\} + \frac{2}{n_0}\sum_{i\in\mathcal{I}_0}\Delta_i\left\{\widehat{\eta}(\widehat{\boldsymbol{f}}_{X,i}) - \eta^\dagger(\widehat{\boldsymbol{f}}_{X,i})\right\}. \tag{S5.14}$$

The first stochastic term is handled exactly as in the proof of $\widehat{\mu}_1$, using Lemma 4 of Fan and Gu (2024). The second term is new and comes from the generated outcome.

By Cauchy's inequality and Young's inequality,

$$\begin{aligned}\left|\frac{2}{n_0}\sum_{i\in\mathcal{I}_0}\Delta_i\left\{\widehat{\eta}(\widehat{\boldsymbol{f}}_{X,i}) - \eta^\dagger(\widehat{\boldsymbol{f}}_{X,i})\right\}\right| &\le 2\|\Delta\|_{n,0}\|\widehat{\eta}(\widehat{\boldsymbol{f}}_X)\eta^\dagger(\widehat{\boldsymbol{f}}_X)\|_{n,0} \\ &\le \zeta\|r_{\widehat{\eta}}\|_{n,0}^2 + C_\zeta\|r_{\eta^\dagger}\|_{n,0}^2 + C_\zeta\|\Delta\|_{n,0}^2.\end{aligned}$$

Combining this with (S5.14), and applying the same empirical process bound as in the proof of $\widehat{\mu}_1$, we obtain

$$\|r_{\widehat{\eta}}\|_{n,0}^2 \lesssim \|r_{\eta^\dagger}\|_{n,0}^2 + \frac{V_n\log n}{n} + \|\Delta\|_{n,0}^2 + \frac{t}{n},$$

where

$$V_n := \mathrm{Pdim}(\mathcal{G}_{\eta,n}).$$

By Bernstein's inequality for the fixed oracle and the already established rate for $\widehat{\mu}_1$ under P_0,

$$\|\Delta\|_{n,0}^2 = O_p(r_{\mu_1,0,n}^2),$$

where

$$r_{\mu_1,0,n}^2 := \|\widehat{\mu}_1(\widehat{\boldsymbol{f}}_X, \widehat{\boldsymbol{f}}_M) - \mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M)\|_{L_2(P_0)}^2.$$

Therefore,

$$\|\widehat{\eta}(\widehat{\boldsymbol{f}}_X) - \eta(\nu_X(\boldsymbol{f}_X))\|_{L_2(P_0)}^2 = O_p\left(N_\eta^{-4\gamma_\eta^\star} + \delta_{n,f}^2 + \frac{V_n\log n}{n} + r_{\mu_1,0,n}^2\right). \tag{S5.15}$$

Finally, as before,

$$V_n \lesssim \left(L_\eta^2 N_\eta^2 + d_\eta L_\eta N_\eta\right)\log(L_\eta N_\eta d_\eta).$$

Taking $L_\eta \asymp 1$ and balancing

$$N_\eta^{-4\gamma_\eta^\star} \qquad \text{and} \qquad \frac{N_\eta^2\log N_\eta\log n}{n}$$

gives

$$N_\eta \asymp n^{1/(4\gamma_\eta^\star+2)}$$

up to logarithmic factors. Hence,

$$\|\widehat{\eta}(\widehat{\boldsymbol{f}}_X) - \eta(\nu_X(\boldsymbol{f}_X))\|_{L_2(P_0)}^2 = O_p\left(n^{-2\gamma_\eta^\star/(2\gamma_\eta^\star+1)}(\log n)^a + \delta_{n,f}^2 + r_{\mu_1,0,n}^2\right).$$

Equivalently,

$$\|\widehat{\mu}_{10}(\widehat{\boldsymbol{f}}_X) - \mu_{10}(\boldsymbol{f}_X)\|_{L_2(P_0)}^2 = O_p\left(n^{-2\gamma_\eta^\star/(2\gamma_\eta^\star+1)}(\log n)^a + \delta_{n,f}^2 + r_{\mu_1,0,n}^2\right).$$

S5.4 Proof of Proposition S3.1

By condition (S3.1) and (S3.2), we have:

$$d_{\mathcal{V}}^2(\hat{z}) \le C\{\mathcal{L}_\lambda(\hat{z}) - \mathcal{L}_\lambda^\star\} + C(\delta_{X,\text{app}}^2 + \delta_{M,\text{app}}^2) = O_p(r_n^2 + \delta_{X,\text{app}}^2 + \delta_{M,\text{app}}^2).$$

Therefore,

$$d_{\mathcal{V}}(\hat{z}) = O_p(r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}}).$$

By the definition of $d_{\mathcal{V}}$, for every $\varepsilon > 0$ there exists $\nu_\varepsilon = (\nu_{X,\varepsilon}, \nu_{M,\varepsilon}) \in \mathcal{V}$ such that

$$\begin{aligned}&\|\hat{g}_X(X) - \nu_{X,\varepsilon}(\boldsymbol{f}_X)\|_{L_2(P)}^2 + \|\hat{g}_M(M) - \nu_{M,\varepsilon}(\boldsymbol{f}_M)\|_{L_2(P)}^2 \\ &\quad \le d_{\mathcal{V}}^2(\hat{z}) + \varepsilon.\end{aligned}$$

Taking, for instance, $\varepsilon = n^{-1}$ and writing $\nu_n := \nu_{1/n}$ gives

$$\|\hat{g}_X(X) - \nu_{X,n}(\boldsymbol{f}_X)\|_{L_2(P)}^2 + \|\hat{g}_M(M) - \nu_{M,n}(\boldsymbol{f}_M)\|_{L_2(P)}^2 = O_p(r_n^2 + \delta_{X,\text{app}}^2 + \delta_{M,\text{app}}^2).$$

Taking square roots and using $\sqrt{x+y+z} \le \sqrt{x} + \sqrt{y} + \sqrt{z}$ yields

$$\Big(\|\hat{g}_X(X) - \nu_{X,n}(\boldsymbol{f}_X)\|_{L_2(P)}^2 + \|\hat{g}_M(M) - \nu_{M,n}(\boldsymbol{f}_M)\|_{L_2(P)}^2\Big)^{1/2} = O_p(r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}}).$$

Consequently,

$$\|\hat{g}_X(X) - \nu_{X,n}(\boldsymbol{f}_X)\|_{L_2(P)} + \|\hat{g}_M(M) - \nu_{M,n}(\boldsymbol{f}_M)\|_{L_2(P)} = O_p(r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}}),$$

where the last step follows because the sum of two nonnegative terms is bounded by $\sqrt{2}$ times their Euclidean norm.

Finally, by the definition of

$$\hat{\boldsymbol{f}} := (\hat{g}_X(X), \hat{g}_M(M)), \qquad \nu_n(\boldsymbol{f}) := (\nu_{X,n}(\boldsymbol{f}_X), \nu_{M,n}(\boldsymbol{f}_M)),$$

we obtain

$$\|\hat{\boldsymbol{f}} - \nu_n(\boldsymbol{f})\|_{L_2(P)} = O_p(r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}}).$$

Thus, if $r_n + \delta_{X,\text{app}} + \delta_{M,\text{app}} = o(n^{-1/4})$, then

$$\|\hat{\boldsymbol{f}} - \nu_n(\boldsymbol{f})\|_{L_2(P)} = o_p(n^{-1/4}),$$

which is the desired factor-recovery condition.

S6 More on Simulation and Ablation Study

All experiments were conducted in a cluster computing environment on a CPU compute node. The computational environment used Python 3.10.12 with 32 CPU cores. A typical full experimental run with $B = 200$ replications took approximately 84 hours to complete. The tables below, together with Tables 1 and 2 in the main manuscript, report the performance of different methods and ablation study compared with MediEncoder across varying sample sizes and covariate/mediator dimensions. The ablation study further evaluates the performance of MediEncoder after removing the alignment component in the loss function (3.1).

n	**Estimator**	**SD**	**RMSE**	**CI Length**	**Coverage**
100	Projection	16.859	16.952	66.086	0.980
	Autoencoder	47.032	47.072	184.363	0.985
	VAE	20.089	20.081	78.749	0.990
	MediEncoder	16.935	16.940	66.383	0.980
300	Projection	1.816	1.822	7.119	0.935
	Autoencoder	2.555	2.566	10.015	0.960
	VAE	2.092	2.092	8.199	0.950
	MediEncoder	1.998	2.007	7.831	0.968
800	Projection	0.674	0.679	2.641	0.950
	Autoencoder	0.703	0.702	2.757	0.963
	VAE	0.629	0.634	2.467	0.948
	MediEncoder	0.621	0.621	2.434	0.960
1200	Projection	0.587	0.594	2.299	0.948
	Autoencoder	0.595	0.595	2.332	0.960
	VAE	0.512	0.521	2.006	0.950
	MediEncoder	0.506	0.507	1.983	0.943
2000	Projection	0.565	0.564	2.213	0.955
	Autoencoder	0.604	0.604	2.368	0.955
	VAE	0.491	0.496	1.925	0.943
	MediEncoder	0.512	0.512	2.007	0.953
3000	Projection	0.605	0.619	2.373	0.943
	Autoencoder	0.561	0.562	2.198	0.943
	VAE	0.510	0.510	1.999	0.960
	MediEncoder	0.502	0.502	1.967	0.960

Table S1: Estimator performance comparison by sample size under the nonlinear wavelet DGP with $p = 800$, $q = 200$, $\sigma_X = 2$, $\sigma_M = 1$, $\sigma_Y = 1$, $\bar{p} = \bar{q} = 5$, $\tilde{p} = \tilde{q} = 10$, and $B = 200$ replications.

	SD		**RMSE**		**CI Length**	
	$\lambda_3 = 0$	Tuning	$\lambda_3 = 0$	Tuning	$\lambda_3 = 0$	Tuning
100	33.415	16.935	33.577	16.940	130.986	66.383
300	2.113	1.998	2.122	2.007	8.283	7.831
800	0.621	0.621	0.623	0.621	2.436	2.434
1200	1.252	0.506	1.251	0.507	4.909	1.983
2000	0.574	0.512	0.577	0.512	2.250	2.007
3000	0.580	0.502	0.579	0.502	2.272	1.967

Table S2: Ablation study on the alignment term (λ_3) in `MediEncoder` with $p = 800$, $q = 200$, $\sigma_X = 2$, $\sigma_M = 1$, $\sigma_Y = 1$, $\bar{p} = \bar{q} = 5$, and $\tilde{p} = \tilde{q} = 10$.

n	**Estimator**	**SD**	**RMSE**	**CI Length**	**Coverage**
100	Projection	12.446	12.416	48.788	0.975
	Autoencoder	20.185	20.148	79.123	0.975
	VAE	10.013	9.989	39.251	0.945
	MediEncoder	33.241	33.269	130.304	0.990
300	Projection	2.147	2.150	8.415	0.940
	Autoencoder	3.131	3.160	12.272	0.945
	VAE	1.232	1.240	4.831	0.960
	MediEncoder	3.645	3.707	14.290	0.965
800	Projection	0.945	0.944	3.704	0.935
	Autoencoder	1.481	1.499	5.805	0.960
	VAE	0.928	0.932	3.639	0.965
	MediEncoder	0.772	0.781	3.027	0.950
1200	Projection	0.809	0.807	3.172	0.950
	Autoencoder	0.797	0.802	3.125	0.960
	VAE	0.801	0.801	3.138	0.975
	MediEncoder	0.885	0.883	3.467	0.975
2000	Projection	0.618	0.619	2.424	0.930
	Autoencoder	0.659	0.659	2.585	0.950
	VAE	0.520	0.519	2.037	0.935
	MediEncoder	0.591	0.591	2.318	0.950
3000	Projection	0.538	0.550	2.108	0.930
	Autoencoder	0.657	0.657	2.575	0.945
	VAE	0.568	0.569	2.228	0.965
	MediEncoder	0.494	0.492	1.934	0.940

Table S3: Estimator performance comparison by sample size under the nonlinear wavelet DGP with $p = 2500$, $q = 2500$, $\sigma_X = 1.5$, $\sigma_M = 1$, $\sigma_Y = 1$, $\bar{p} = \bar{q} = 5$, $\tilde{p} = \tilde{q} = 10$, and $B = 200$ replications.

	SD		**RMSE**		**CI Length**	
	$\lambda_3 = 0$	Tuning	$\lambda_3 = 0$	Tuning	$\lambda_3 = 0$	Tuning
100	67.968	33.241	67.836	33.269	266.429	130.304
300	4.450	3.645	4.471	3.707	17.445	14.290
800	1.295	0.772	1.301	0.781	5.078	3.027
1200	0.627	0.885	0.626	0.883	2.459	3.467
2000	0.642	0.591	0.646	0.591	2.516	2.318
3000	0.611	0.494	0.610	0.492	2.397	1.934

Table S4: Ablation study on the alignment term (λ_3) in `MediEncoder` with $p = 2500$, $q = 2500$, $\sigma_X = 1.5$, $\sigma_M = 1$, $\sigma_Y = 1$, $B = 200$, $\bar{p} = \bar{q} = 5$, and $\tilde{p} = \tilde{q} = 10$.

n	**Estimator**	**SD**	**RMSE**	**CI Length**	**Coverage**
100	Projection	30.554	30.532	119.769	0.995
	Autoencoder	44.535	44.944	174.573	0.975
	VAE	12.740	12.727	49.941	0.960
	MediEncoder	15.041	15.091	58.959	0.985
300	Projection	2.643	2.664	10.361	0.965
	Autoencoder	5.398	5.470	21.160	0.960
	VAE	10.925	10.899	42.825	0.990
	MediEncoder	3.187	3.238	12.492	0.985
800	Projection	1.096	1.104	4.296	0.950
	Autoencoder	0.967	0.965	3.790	0.940
	VAE	0.841	0.856	3.295	0.945
	MediEncoder	0.798	0.798	3.127	0.965
1200	Projection	1.067	1.065	4.183	0.950
	Autoencoder	0.665	0.664	2.607	0.955
	VAE	0.634	0.634	2.486	0.960
	MediEncoder	0.601	0.600	2.356	0.970
2000	Projection	0.638	0.638	2.502	0.960
	Autoencoder	0.687	0.685	2.693	0.960
	VAE	0.486	0.492	1.904	0.955
	MediEncoder	0.541	0.542	2.122	0.960
3000	Projection	0.576	0.589	2.259	0.930
	Autoencoder	0.562	0.564	2.203	0.950
	VAE	0.473	0.472	1.854	0.960
	MediEncoder	0.445	0.445	1.744	0.960

Table S5: Estimator performance comparison by sample size under the nonlinear wavelet DGP with $p = 5000$, $q = 5000$, $\sigma_X = 1.5$, $\sigma_M = 1$, $\sigma_Y = 1$, $\bar{p} = \bar{q} = 5$, $\tilde{p} = \tilde{q} = 10$, and $B = 200$ replications.

S7 Comparisons

S7.1 Comparison with Nath et al. (2023)

The work of Nath et al. (2023) is related to ours at a broad level, as both approaches use machine learning to construct low-dimensional representations for high-dimensional mediation analysis. However, the underlying statistical models, causal targets, and inferential guarantees are substantially different. Specifically, Nath et al. (2023) learns a scalar representation $z_i = \Phi(\boldsymbol{m}_i)$ from the high-dimensional mediator and inserts it into the linear structural equation model (SEM)

$$z_i = \alpha_0 + x_i\alpha + \epsilon_i^{(1)},$$
$$y_i = \beta_0 + z_i\beta + x_i\gamma + \epsilon_i^{(2)}.$$

Their algorithm alternates between updating the unknown map Φ and fitting the path coefficients (α, β, γ), to estimate the natural indirect effect (given by $\alpha\beta$). The main distinctions from our work are as follows:

1. **Measurements versus the mediation mechanism.** In Nath et al. (2023), M is treated as the high-dimensional mediators, whereas we treat M as noisy, high-dimensional measurements of a true latent mediator vector $\boldsymbol{f}_M$ through a nonlinear factor model $M = \phi_M(\boldsymbol{f}_M) + \boldsymbol{u}_M$. This distinction is important in many biomedical applications, where molecular, epigenetic, or imaging measurements are not themselves the underlying biological mechanism, but rather indirect and noisy measurements of it.
2. **Pre-treatment covariates and interactions.** The core model in Nath et al. (2023) contains an exposure, a scalar mediator score, and an outcome, but does not include pre-treatment covariates, which can limit its applicability. Our framework includes latent pre-treatment covariates $\boldsymbol{f}_X$, observed through high-dimensional surrogates X, which may confound the treatment–mediator, treatment–outcome, and mediator–outcome relationships. Moreover, we allow treatment–covariate interactions in the mediator regression model and treatment–covariate and treatment–mediator interactions in the outcome regression model.
3. **Nonparametric model flexibility.** Although the representation map z in Nath et al. (2023) may be a nonlinear function of m, the outcome regression model is linear in (x, z) and the mediator regression models are linear in x and therefore are at risk of potential model misspecification. Our framework imposes no linearity on any component of the model: the measurement maps ϕ_X and ϕ_M, the mediator regressions μ_0^* and μ_1^*, and the outcome regressions μ_0 and μ_1 are all nonlinear functions.
4. **Multivariate and unknown latent dimensions.** The formulation in Nath et al. (2023) compresses the high-dimensional mediator into a single scalar z_i, which can be restrictive. In our framework, both $\boldsymbol{f}_X \in \mathbb{R}^{\bar{p}}$ and $\boldsymbol{f}_M \in \mathbb{R}^{\bar{q}}$ may be multivariate. Furthermore, the exact latent dimensions $(\bar{p}, \bar{q})$ need not be known. One may use over-specified working dimensions $(\widetilde{p} \geq \bar{p}, \widetilde{q} \geq \bar{q})$ and our approach would yield a valid $\sqrt{n}$-CAN estimator.
5. **Rigorous inferential guarantees.** The focus of Nath et al. (2023) is primarily algorithmic and empirical, and it does not provide analogous theoretical guarantees for the consistency or asymptotic distribution of the estimated indirect effect after learning Φ. In contrast, our approach is theoretically grounded as we prove under standard assumptions our estimator is a $\sqrt{n}$-CAN estimator of causal parameters of interest and consequently permits valid inference under a flexible nonparametric framework.

The main differences between Nath et al. (2023) and MediEncoder are summarized below.

Summary of differences:

	Nath et al. (2023)	MediEncoder (ours)
Mediator structure	Observed $\boldsymbol{m}$ is the mediator	Observed M is a noisy proxy of true mediators $\boldsymbol{f}_M$
Pre-treatment covariates	None	High-dimensional X (proxy of $\boldsymbol{f}_X$)
Treatment–mediator interaction	Not modeled	Fully modeled (Eq. (2.1))
Outcome model	Linear: $y = \beta_0 + \beta z + \gamma x$	Nonparametric: $\mu_1(\boldsymbol{f}_X, \boldsymbol{f}_M)$
Estimator	OLS on path coefficients	EIF-based, multiply robust
Robustness	Requires both path equations correct	Consistent under partial nuisance misspecification

S7.2 Comparison with Jiang et al. (2023)

The work of Jiang et al. (2023) is closely related to ours, as both methods use nonlinear representation learning for mediation analysis with multi-dimensional and indirectly observed mediators. However, the two approaches differ substantially in their treatment of pre-treatment covariates, representation-learning objectives, effect estimators, and theoretical guarantees.

Using the notation of Jiang et al. (2023), let T denote the treatment, W the observed pre-treatment covariates, $Z \in \mathbb{R}^d$ the latent mediator, $X \in \mathbb{R}^D$ its high-dimensional measurement, and Y the outcome. Their IMAVAE model is based on

$$p_\vartheta(x, z, y \mid u) = p_f(x \mid z)\, p_h(y \mid z, u)\, p_{S,\lambda}(z \mid u), \qquad u = t \text{ or } u = (w, t).$$

They jointly learn an encoder $q_\phi(z \mid x, u)$, a decoder $p_f(x \mid z)$, a conditional prior $p_{S,\lambda}(z \mid u)$, and an outcome predictor $g_\gamma(z, u)$ by minimizing

$$(\widehat{\vartheta}, \widehat{\phi}, \widehat{\gamma}) \in \arg\min_{\vartheta,\phi,\gamma} \{\alpha \mathcal{L}_{\text{rec}} - \beta \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{out}}\}.$$

After fitting the model, they sample latent mediators from the fitted treatment-specific distribution and evaluate the fitted outcome predictor under different treatment levels to estimate the average mediation, direct, and total effects. The main distinctions from our work are as follows:

1. **Latent covariates and mediators.** In Jiang et al. (2023), only the mediator Z is latent and indirectly observed through the high-dimensional feature X, whereas the pre-treatment covariates W are directly observed. In our framework, both the covariates and mediators are latent:

   $$X = \phi_X(\boldsymbol{f}_X) + \boldsymbol{u}_X, \qquad M = \phi_M(\boldsymbol{f}_M) + \boldsymbol{u}_M.$$

   Thus, our method accommodates nonlinear measurement error in both the confounders and the mediators.

2. **Structural coupling of the learned representations.** In Jiang et al. (2023), treatment and observed covariates enter the conditional prior, encoder, and outcome predictor through the auxiliary variable $u = (W, T)$. This allows their model to capture nonlinear dependence on W and T, but it does not learn a latent representation of the covariates or explicitly connect covariate and mediator representations. In contrast, MediEncoder jointly learns $g_X(X)$ and $g_M(M)$ and uses

   $$g_{XM}\big(A, g_X(X)\big)$$

   to align the mediator representation with the treatment and covariate representation.

3. **Identification requirements and latent dimensions.** The identifiability result of Jiang et al. (2023) relies on the identifiable-VAE structure, including a conditionally factorial exponential-family prior, injective measurement and outcome maps, and sufficient variation in the auxiliary variable. Their latent dimension d is specified in the model. Our theory does not require a particular generative prior or unique identification of the latent coordinates. It only requires recovery of an informative left-invertible transformation of the latent factors and permits over-specified working dimensions

   $$\widetilde{p} \geq \bar{p}, \qquad \widetilde{q} \geq \bar{q}.$$

4. **Plug-in versus influence-function-based estimation.** The estimator in Jiang et al. (2023) is a Monte Carlo plug-in estimator based on the fitted latent-mediator distribution and outcome predictor. Therefore, estimation errors in these fitted components enter the final causal-effect estimator directly. MediEncoder instead uses the learned representations in a cross-fitted efficient influence function-based estimator, which is multiply robust and reduces the first-order effect of nuisance-estimation errors.

5. **Inferential guarantees.** The main theoretical result of Jiang et al. (2023) establishes identifiability of the latent generative model up to an equivalence transformation. It does not provide analogous consistency or asymptotic-normality guarantees for the estimated mediation effects after learning the representation. In contrast, we establish that the MediEncoder estimator is $\sqrt{n}$-consistent and asymptotically normal, thereby permitting valid inference for the natural indirect, direct, and total effects.

The main differences between Jiang et al. (2023) and MediEncoder are summarized below.

Summary of differences:

	Jiang et al. (2023)	**MediEncoder (ours)**
Mediator structure	Latent Z measured by high-dimensional X	Latent $\boldsymbol{f}_M$ measured by high-dimensional M
Pre-treatment covariates	Directly observed W	Latent $\boldsymbol{f}_X$ measured by high-dimensional X
Representation learning	Conditional iVAE $q_\phi(z \mid x, u)$	Coupled encoders with $g_{XM}(A, g_X(X))$
Effect modification	Implicit through conditioning on $u = (W, T)$	Explicit through arm-specific nonparametric functions
Latent dimension	Prespecified d	Over-specified $(\widetilde{p}, \widetilde{q})$ allowed
Estimator	Monte Carlo plug-in	Cross-fitted EIF-based estimator
Robustness	Relies on fitted generative and outcome models	Multiply robust under partial misspecification
Theory	Latent-model identifiability	$\sqrt{n}$-CAN inference for causal effects

S7.3 Comparison with Existing Machine-Learning-Based Mediation Analysis

The primary objective of MediEncoder is accurate estimation and valid inference for the natural direct and indirect effects, rather than discovery of individual latent factors. Accordingly, we conduct a numerical comparison with three machine-learning-based mediation methods, namely Nath et al. (2023), Wang and Huang (2024) and Jiang et al. (2023). Our model is substantially more flexible than these models, as it not only allows ultra-high-dimensional surrogates for covariates and mediators, but also their interaction with treatment in both the outcome regression and the (latent) mediation regression models. To our knowledge, no existing method addresses both these modeling aspects (UHD surrogates and treatment interaction) simultaneously. Therefore, we consider a simpler no-interaction model for fair comparison:

$$X = \phi_X(\boldsymbol{f}_X) + \boldsymbol{u}_X, \qquad M = \phi_M(\boldsymbol{f}_M) + \boldsymbol{u}_M$$

$$\boldsymbol{f}_M = \mu^*(\boldsymbol{f}_X) + A\,\boldsymbol{\alpha}_M + \boldsymbol{u}_{XM}, \qquad Y = \mu(\boldsymbol{f}_X, \boldsymbol{f}_M) + A\,\alpha_Y + \epsilon.$$

We follow a similar approach to Section 5; the generative model of $(A, \boldsymbol{f}_X)$ remains unchanged. The generation of $\boldsymbol{f}_M$ and Y changes because we no longer consider the treatment interaction, and are given by

$$\begin{aligned} \boldsymbol{f}_M &= \boldsymbol{\delta}_0 \boldsymbol{f}_X^{\circ 2} + A\boldsymbol{\alpha}_M + \mathbf{u}_{XM}, \qquad \mathbf{u}_{XM} \sim \mathcal{N}(0, \Sigma_U), \\ Y &= \sin(5\boldsymbol{f}_X)^\top \beta_0 + \log\big(1 + \exp(\boldsymbol{f}_X^\top \kappa_0)\big) + \boldsymbol{f}_M^\top \gamma_0 + A\,\alpha_Y + \epsilon_Y, \end{aligned}$$

where $(\boldsymbol{\delta}_0, \beta_0, \kappa_0, \gamma_0, \Sigma_U)$ are as in Section 5, $\boldsymbol{\alpha}_M = c\mathcal{U}(0.5, 1.5)^{\bar{q}}$ and $\alpha_Y = c\mathcal{U}(0.5, 1.5)$ with $c = 0.5$, and $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$. As in Section 5, all coefficients are redrawn in every Monte Carlo replication and the target is recomputed for that draw, so the reported truth is an average over the parameter distribution rather than a single fixed value.

The main simulation in Section 5.1 uses Haar-wavelet loadings, whose construction is given in Section S2 of the Supplementary Material. For this comparison, the measurement maps ϕ_X, ϕ_M use additive polynomial loadings of degree $D = 3$. Each observed coordinate is a sum of univariate polynomials

of the corresponding latent coordinates, with terms up to degree three. Under this design the natural effects admit closed forms, NIE $= \boldsymbol{\alpha}_M^\top \gamma_0$, NDE $= \alpha_Y$. We take $p = q = 500$, $\bar{p} = \bar{q} = 3$, $\tilde{p} = \tilde{q} = 7$, $\sigma_X = \sigma_M = \sigma_Y = 1$, and $n \in \{500, 800, 1500, 2000\}$. We use $B = 200$ Monte Carlo replications for each setting.

We then briefly summarize the implementation of the comparison methods. The alternating estimation procedure of Nath et al. (2023) is described in the preceding section and is not repeated here. To accommodate the UHD covariate surrogates in our setting, we augment their method with a learned covariate summary $z_X = \Phi_X(X)$, which enters both the mediator and outcome regressions linearly. Wang and Huang (2024) model the treatment and observed mediators linearly, while using ReLU neural networks to adjust for the observed covariates in both the mediator and outcome regressions. They estimate the direct effect from a SCAD-penalized partially linear outcome model and obtain the indirect effect as the difference between the total and direct effects. Jiang et al. (2023) uses an identifiable VAE to learn a latent mediator representation z_M from the observed mediator surrogates M, while A and X are used as conditioning inputs for the latent-mediator distribution and the outcome model. They estimate the natural effects by sampling latent mediators under each treatment level and plugging these samples into the fitted outcome predictor.

Method	SD	RMSE	CI Len.	Cov.
$n = 500$				
Ours	2.176	2.268	8.530	**0.920**
Nath et al. (2023)	0.625	0.687	2.450	0.810
Wang and Huang (2024)	0.963	1.866	3.774	0.600
Jiang et al. (2023)	0.379	1.515	1.486	0.010
$n = 800$				
Ours	1.067	1.139	4.181	**0.960**
Nath et al. (2023)	0.506	0.618	1.984	0.710
Wang and Huang (2024)	0.925	1.767	3.626	0.620
Jiang et al. (2023)	0.379	1.539	1.486	0.000
$n = 1500$				
Ours	0.828	0.833	3.244	**0.953**
Nath et al. (2023)	0.344	0.448	1.348	0.750
Wang and Huang (2024)	0.908	1.731	3.560	0.650
Jiang et al. (2023)	0.378	1.463	1.482	0.031
$n = 2000$				
Ours	0.245	0.304	0.960	**0.947**
Nath et al. (2023)	0.174	0.850	0.681	0.000
Wang and Huang (2024)	0.363	0.930	1.423	0.253
Jiang et al. (2023)	0.651	1.114	2.553	0.740

Table S6: Natural indirect effect (NIE) with $p = q = 500$.

The results in the table above demonstrate the advantage of MediEncoder in both point estimation and uncertainty quantification **even for this simplified setup without treatment interaction**. MediEncoder consistently maintains coverage close to the nominal 95% level across all sample sizes, whereas the substantially shorter intervals produced by the competing methods are accompanied by severe undercoverage, which indicates that the resulting estimates are severely biased. This is evident because Nath et al. (2023) and Wang and

Huang (2024) impose linear mediation structures and consequently suffer from model misspecification, whereas Jiang et al. (2023), although accommodating indirectly observed mediators, does not jointly learn the nonlinear latent covariate structure and model its structural dependence with the latent mediator representation, and its plug-in estimator is sensitive to model misspecification as well. Therefore, this demonstrates the efficacy of our methodology for estimating NDE and NIE in a non-linear structural equation model.