iASIDE: IDENTIFIABLE, ADAPTABLE, AND SEPARABLE INTERVENTIONAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning latent causal dynamics offers an important approach for predicting a dynamics system's response to hypothetical external interventions, especially in systems with high-dimensional observations. This presents two related but distinct challenges: identifiability of the intervention effect in the presence of potential intervention assignment bias, and identifiability of the latent dynamics process that is not directly observed. While the latter challenge persists regardless of the former, it has received limited attention in existing works of intervention response prediction. We bridge this gap with iASIDE, a framework for attaining *identifiable, adaptable* and *separable* interventional dynamics. We formulate a latent process with non-stationary parameters and establish its *theoretical* identifiability by conditioning its non-stationary parameters on the variable history. We then address the *algorithmic* challenge of reaching the optimal identifiable solution — caused by the difficulty of separating native and response dynamics from their composite observations — by a separable formulation for both the latent dynamics and the conditional priors of its non-stationary parameters. We realize amortized variational inference of this latent dynamics with feed-forward meta-learning. Experiments on synthetic and real data showed that iASIDE significantly improved the accuracy of intervention response predictions, especially over increased prediction horizons and data heterogeneity.

1 INTRODUCTION

To make optimal intervention decisions on a temporally-evolving dynamics systems require predicting its response to hypothetical *external interventions* (Krishnan et al., 2015). As observations grow in dimensionality and modality, it is becoming increasingly common to model the temporal processes in an abstract latent space, broadly known as *latent dynamics modeling*.

More specifically, consider observed time series of covariates $\mathcal{X}_T = \{\mathbf{x}_t\}_{t=1}^{T}$, external interventions $\mathcal{A}_{T-1} = \{\mathbf{a}_t\}_{t=1}^{T-1}$, and outcomes $\mathcal{Y}_T = \{\mathbf{y}_t\}_{t=1}^{T}$. We are interested in predicting the future outcome trajectory $\{\mathbf{y}_t\}_{t=T+1}^{T+\tau}$ for a horizon of τ, given hypothetical interventions $\{\mathbf{a}_t\}_{t=T}^{T+\tau-1}$ and history $\mathcal{H}_T = \{\mathcal{X}_T, \mathcal{A}_{T-1}, \mathcal{Y}_T\}$ as: $\mathbb{E}[\mathbf{y}_{T+1:T+\tau}|\mathbf{a}_{T:T+\tau-1}, \mathcal{H}_T]$. In the latent dynamics framework, this can be represented by a latent causal dynamics process $\frac{d\mathbf{z}_t}{dt} = f(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c})$ for latent dynamics state $\mathbf{z}_t$ and its emission to $\mathbf{y}_t = g(\mathbf{z}_t)$. As illustrated in Figure 1A, f describes how $\mathbf{z}_{t+1}$ naturally evolves from $\mathbf{z}_t$ (*native dynamics*) along with its response to external intervention $\mathbf{a}_t$ (*response dynamics*).

Learning this latent process from observed data faces two challenges. First, when the collection of data is not controlled, the assignment of treatment $\mathbf{a}_t$ may be influenced by $\mathbf{z}_t$ as shown in red in Figure 1A, leading to confounding that complicates the estimation of its effect on $\mathbf{z}_{t+1}$. Second, this process is characterized by an additional latent exogenous variable $\mathbf{c}$ which, as shown in blue in Figure 1A, captures the inter- or intra-subject variations in f: in the example of tumor growth, $\mathbf{c}$ would represent heterogeneous tumor growth and decay rates under interventions. These lead to two related but distinct *challenges of identifiability:* 1) identifiability of the *latent* dynamics since neither $\mathbf{z}_t$ nor $\mathbf{c}$ is observed, and 2) identifiability of the causal effect of $\mathbf{a}_t$ in the presence of confounding.

Identifiability of latent dynamics process: Dynamics processes are naturally causal in time. When learning such a temporal process in the latent space as $f(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c})$, the absence of direct supervision on $\mathbf{z}_t$ leads to an important question: even when $p(\hat{\mathbf{y}}_t)$ is perfectly optimized to match observations, what can we say about its underlying $\hat{\mathbf{z}}_t$ (and $\hat{\mathbf{c}}$) in relation to their truths? This question has received

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

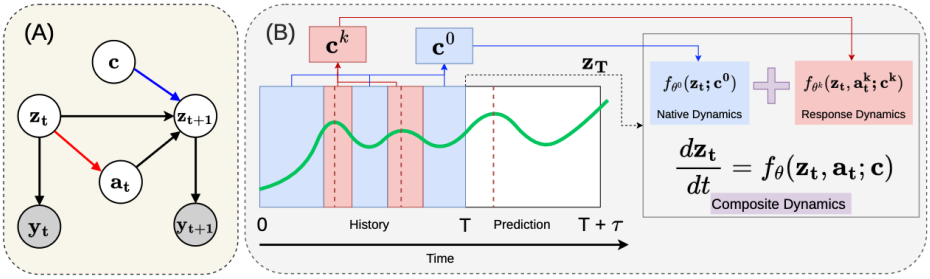

Figure 1: A: Illustration of the latent causal dynamics process under external interventions $\mathbf{a}_t$, with non-stationarity governed by the exogenous variable $\mathbf{c}$ and potential confounding due to intervention assignment bias $\mathbf{z}_t \rightarrow \mathbf{a}_t$. B: Overview of iASIDE that learns *separable* and *adaptable* interventional dynamics, made identifiable via separate conditioning on relevant context samples in the history.

significant interest built on the theory of nonlinear independent component analysis (ICA) (Hyvarinen et al., 2019; Khemakhem et al., 2020). Most existing works have focused on the identifiability of $\mathbf{z}_t$, leveraging information about the non-stationarity of f (Hälvä & Hyvarinen, 2020; Yao et al., 2022b;a). Two recent works take a step further to establish identifiability of the latent parameter $\mathbf{c}$ that characterizes the non-stationarity of f, showing its importance for f's forecasting ability (Ye et al., 2026; Hı zlıet al., 2025). Intuitively, this property is also fundamental to the prediction of dynamic intervention responses. Unfortunately, identifiability of latent dynamics has mainly been examined in systems of native dynamics and received little attention in interventional dynamics.

Identifiability of causal effect with confounding: In contrast, the challenge of identifying interventional effect due to confounding has received central attention in the context of interventional dynamics. Representative approaches include 1) propensity-weighting (Lim, 2018; Hess & Feuerriegel, 2025), 2) learning intervention-invariant representations (Bica et al., 2020; Melnychuk et al., 2022; Wang et al., 2024), and 3) iterative G-computation (Li et al., 2021; Xiong et al., 2024; Hess et al., 2026). In these approaches, the temporal process has been modeled autoregressively with LSTMs (Lim, 2018; Bica et al., 2020; Berrevoets et al., 2021) and transformers (Melnychuk et al., 2022), or with latent dynamics models such as neural/controlled ordinary differential equations (O/CDEs) (Brouwer et al., 2022; Seedat et al., 2022; Vanderschueren et al., 2023; Hess & Feuerriegel, 2025). Although the latter is gaining popularity, identifiability of the latent process has not received attention in this context, and its impact on the accuracy of outcome prediction remains unclear.

Summary of Gap: The joint presence of two identifiability challenges as described above has an interesting analogy to the field of *causal representation learning* (Komanduri et al., 2024), where the identifiability of latent representations (even without causal relations) and the identifiability of causal graph/mechanisms (even when causal variables are directly observed) must be addressed together. Similarly, when interventional dynamics becomes latent, both the identifiability of latent dynamics (even without confounding) and the identifiability of causal effects (even when directly observed) must be addressed together – the former, however, has received limited attention. This paper is poised to address this gap: as the first step towards this direction, we *prioritize the identifiability of latent interventional dynamics* and address confounding by integration with existing methods.

Contribution: We pinpoint two sources of difficulties for learning identifiable latent interventional dynamics: theoretical and algorithmic. Theoretically, learning a latent dynamics function $f(\mathbf{z}_t, \mathbf{a}_t)$ without direct supervision or additional assumptions will result in a non-identifiable model (Hyvarinen et al., 2019). The practice of conditioning the prediction of $\mathbf{y}_{T+1:T+\tau}$ on history $\mathcal{H}_T$ offers an excellent opportunity to leverage $\mathcal{H}_T$ as the auxiliary information to establish identifiability – a missed opportunity not yet explored in existing works. Inspired by a recent work of (Ye et al., 2026), we introduce a novel *identifiable* formulation of $f_\theta(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c}|\mathcal{H}_T)$, with latent non-stationary parameter $\mathbf{c}$ conditioned on history $\mathcal{H}_T$ through a conditionally independent prior $p(\mathbf{c}|\mathcal{H}_T)$, assuming $\mathbf{c}$ locally stationary within the prediction horizon of τ. While we do not contribute new theoretical results over (Ye et al., 2026), this is the first time identifiability is established for latent interventional dynamics. Moreover, compared to existing approaches that condition the prediction of $\mathbf{y}_{T+1:T+\tau}$ on an initial state $\hat{\mathbf{z}}_T$ encoded from $\mathcal{H}_T$, our conditional prior $p(\mathbf{c}|\mathcal{H}_T)$ affects each rollout step of

the prediction horizon, offering a history-adaptable $f_\theta(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c}|\mathcal{H}_T)$ instead of a *one-size-fits-all* $f(\mathbf{z}_t, \mathbf{a}_t)$. This *identifiable and history-adaptable formulation* of latent intervention dynamics is our main contribution over existing works in dynamic intervention response prediction.

Algorithmically, while identifiable in theory, the optimal solution may be difficult to reach in practice because the *response* dynamics is always observed as a composite with *native* dynamics. To facilitate their separate identification, we further introduce a *separable* formulation of our $f_\theta(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c}|\mathcal{H}_T)$, with 1) *separate* components of native dynamics $f_{\theta^0}(\mathbf{z}_t; \mathbf{c}^0)$ and its response dynamics $f_{\theta^k}(\mathbf{z}_t, a_t^k; \mathbf{c}^k)$ to each intervention $a_t^k \in \mathbf{a}_t$ ($k \geq 1$), and 2) factoring $p(\mathbf{c}|\mathcal{H}_T)$ into $p(\mathbf{c}^k|\mathcal{H}_T^k)$'s where $\mathcal{H}_T^0$ represents intervention-free segments in the history and $\mathcal{H}_T^k$ when $k \geq 1$ represents history segments under the effects of intervention k. We realize amortized variational inference of this latent dynamics with feed-forward meta-learning, and demonstrate the importance of our separable formulation compared to its various non-separated counterparts. This *separable adaptive formulation* represents our main contribution to bring existing works in identifiable latent dynamics to the interventional space.

We evaluated our framework, iASIDE (*identifiable*, *adaptable* and *separable* interventional dynamics), in 1) lower-dimensional benchmark datasets used in the literature of dynamic intervention response prediction (synthetic tumor growth (Geng et al., 2017) and real MIMIC-III data (Johnson et al., 2023)), and 2) high-dimensional datasets used in the literature of identifiable latent dynamics (synthetic pendulum image sequences (Ye et al., 2026)). We considered settings with and without intervention assignment bias in synthetic datasets: when confounding exists, we augmented iASIDE with the common de-confounding strategy of invariant representations (Bica et al., 2020). Our experimental results provided strong evidence that, regardless of whether confounding exists in the data, addressing the identifiability of latent dynamics played a significant role in improving the accuracy of intervention outcome predictions over time, especially as the prediction horizon or non-stationarity increased.

2 Related Works

Identifiability of latent dynamics process: Identifiability of latent dynamics models has attracted significant attention, with successes mainly built on the nonlinear ICA theory (Hyvarinen et al., 2019; Khemakhem et al., 2020). Existing works have considered three types of latent dynamics: stationary (Hyvarinen & Morioka, 2017; Klindt et al., 2021), non-stationary with time-varying latent parameter $\mathbf{c}_t$ that follows a hidden Markov process (Song et al., 2023; Li et al., 2025; Hälvä & Hyvarinen, 2020; Hälvä et al., 2021; Balsells-Rodas et al., 2024), and locally stationary processes with latent parameter $\mathbf{c}$ governing a time window of τ (Khemakhem et al., 2020; Yao et al., 2022b;a; Hı zlıet al., 2025). iASIDE is most related to the last category, where most existing works establish the identifiability of $\mathbf{z}_t$'s by conditioning the latent process on observed labels of non-stationarity. Two recent works have taken the next step to establish identifiability for the latent parameter $\mathbf{c}$ characterizing this non-stationarity, leveraging pre-defined labels of $\mathbf{c}$ (Hı zlıet al., 2025) or few-shot context samples generated under $\mathbf{c}$ (Ye et al., 2026). Both provided empirical evidence that the identifiability of $\mathbf{c}$ is critical for the prediction accuracy of the learned latent dynamics (Ye et al., 2026; Hı zlıet al., 2025).

These efforts provide strong motivation for addressing the identifiability of latent interventional dynamics for predicting intervention responses. However, they were developed for native dynamics and, as we will show empirically in Table 1, face the optimization challenge of separating native and interventional dynamics from their composite observations if directly applied. iASIDE addresses this with a novel separable formulation, extending identifiability to interventional dynamics.

Identifiability of interventional effects under confounding: Because the intervention assignment mechanism is unknown in observational data, existing works in predicting dynamic intervention response have prioritized the challenge of time-varying confounding. Successes are represented by propensity weighting (Lim, 2018; Hess & Feuerriegel, 2025), treatment-invariant representations (Bica et al., 2020; Melnychuk et al., 2022; Seedat et al., 2022; Wang et al., 2024), and iterative G-computation (Li et al., 2021; Xiong et al., 2024; Hess et al., 2026). While the last category is dedicated to estimating the final outcome $\mathbf{y}_{T+\tau}$, the first two, like iASIDE, are able to predict the dynamic outcome trajectory throughout the horizon of τ. To this end, existing works mostly adopt an encoder-decoder formulation, where an *encoder* is trained to extract latent representations $\hat{\mathbf{z}}_T$ from history $\mathcal{H}_T$, and a *decoder* is trained to use $\hat{\mathbf{z}}_T$ to initialize the prediction for a horizon of τ under given interventions. Architectures used include RNNs (Lim, 2018; Bica et al., 2020), transformers (Melnychuk et al., 2022; Wang et al., 2024), and O/CDEs (Brouwer et al., 2022; Seedat et al., 2022;

Hess & Feuerriegel, 2025; Vanderschueren et al., 2023). Some of these are learned in two stages (Lim, 2018; Bica et al., 2020; Seedat et al., 2022; Hess & Feuerriegel, 2025) while others end-to-end (Melnychuk et al., 2022; Brouwer et al., 2022; Vanderschueren et al., 2023; Wang et al., 2024). Common to all these approaches is a temporal *decoder* that 1) is a composite of native and response dynamics, and 2) has *one-size-fits-all* parameters where $\mathcal{H}_T$ goes in as the initializing $\hat{\mathbf{z}}_T$.

iASIDE complements these works by addressing the identifiability of latent dynamics—a challenge that exists regardless of confounding but yet addressed. This identifiable formulation comes with the additional benefit of allowing the latent dynamics to change with history. We demonstrate the complementarity of iASIDE to these existing works by integrating it with a representative approach, intervention-invariant representations (Bica et al., 2020), to address confounding as needed.

Separable dynamics: Explicit separation of control and intervention outcomes is not uncommon in intervention effect modeling, but mostly in the static setting (Künzel et al., 2019; Shi et al., 2019). In the temporal setting, Gwak et al. (2020) achieved this by using two separate neural ODEs that are then combined in a third neural ODE to generate the observed composite trajectories; in (Seedat et al., 2022), a neural CDE is used where the native dynamics is modulated by incoming treatments $\mathbf{a}_t$'s. Neither has addressed the adaptivity or identifiability of the separable latent dynamics modeled.

Adaptable dynamics: Allowing latent dynamics to adapt to inferred non-stationarity parameters has gained increasing interest (Auzina et al., 2023; Wang et al., 2022; Jiang et al., 2023; Kirchmeyer et al., 2022), among which identifiability is only addressed in one recent work (Ye et al., 2026). iASIDE builds on and extends this work to interventional dynamics with a novel separable formulation.

3 Identifiable, Adaptable, & Separable Interventional Dynamics

Given observed history $\mathcal{H}_T = \{\mathcal{X}_T, \mathcal{A}_{T-1}, \mathcal{Y}_T\}$, we predict the trajectory of future potential outcome $\mathbf{y}_{T+1:T+\tau}$ under a hypothetical intervention sequence $\mathbf{a}_{T:T+\tau-1}$, over a prediction horizon of τ. We consider d_a different interventions $\mathbf{a}_t$ where each $a_t^k \in \{0, 1\}$ is binary. We approach this with 1) a separable and adaptable formulation of the latent dynamics (Section 3.1), and 2) its identifiable and separable conditioning on $\mathcal{H}_T$ (Section 3.2). iASIDE is realized with feed-forward Bayesian meta-learning and underpinned by the identifiability of its latent dynamics (Section 3.3).

3.1 Adaptable and Separable Latent Dynamics

We model the data generation process for $\mathbb{E}[\mathbf{y}_{T+1:T+\tau}|\mathbf{a}_{T:T+\tau-1}, \mathcal{H}_T]$ as illustrated in Figure 1A, including the latent causal dynamics process for the latent state $\mathbf{z}_t \in \mathbb{R}^{d_z}$ and its emission to the observed outcome $\mathbf{y}_t \in \mathbb{R}^{d_y}$ as: $\frac{d\mathbf{z}_t}{dt} = f_\theta(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c}), \mathbf{y}_t = g_\eta(\mathbf{z}_t) + \epsilon, \epsilon \perp\!\!\!\perp g_\eta(\mathbf{z}_t)$, where $\mathbf{c} \in \mathbb{R}^{d_c}$ characterizes the non-stationarity of f_θ. Because of binary interventions, f_θ can be decomposed into an additive function of its native dynamics and responses to each intervention $k (k \geq 1)$, if multiplicative interactions between interventions are neglected (derivation in Appendix A):

$$\frac{d\mathbf{z}_t}{dt} = f_\theta(\mathbf{z}_t, \mathbf{a}_t; \mathbf{c}) = f_{\theta^0}(\mathbf{z}_t; \mathbf{c}^0) + \sum\nolimits_{k=1}^{d_a} a_t^k f_{\theta^k}(\mathbf{z}_t, a_t^k; \mathbf{c}^k), \quad \mathbf{c} \coloneqq \left[\mathbf{c}^0, \mathbf{c}^1, ..., \mathbf{c}^{d_a}\right] \tag{1}$$

where f_{θ^0} describes native dynamics and f_{θ^k} $(k \geq 1)$ describes response dynamics to intervention a^k. f_{θ^k}'s and g_η are neural networks parameterized by θ^k's and η, respectively. Importantly, instead of assuming θ^k to be globally applicable for all data samples, we allow them to vary with latent variables $\mathbf{c}^k$'s to capture non-stationary latent dynamics. We assume the stochasticity within f_θ to be fully determined by $\mathbf{c}$. Consider a prediction horizon $T+1:T+\tau$ given history $\mathcal{H}_T$ and intervention sequence $\mathbf{a}_{T:T+\tau-1}$, $p(\mathbf{y}_{T+1:T+\tau}|\mathbf{a}_{T:T+\tau-1}, \mathcal{H}_T)$ following standard practice would be:

$$p(\mathbf{y}_{T+1:T+\tau}|\mathbf{a}_{T:T+\tau-1}, \mathcal{H}_T) = \int_{\mathbf{c}} \prod_{t=1}^{\tau} p_\eta(\mathbf{y}_{T+t}|\mathbf{z}_{T+t})|_{\mathbf{z}_{T+t}=\mathcal{F}_\theta(\mathbf{z}_{T+t-1}, \mathbf{a}_{T+t-1}; \mathbf{c})} \underline{p(\mathbf{c})} d\mathbf{c} \tag{2}$$

$$\text{where} \quad \mathcal{F}_\theta(\mathbf{z}_{T+t-1}, \mathbf{a}_{T+t-1}; \mathbf{c}) = \mathbf{z}_{T+t-1} + \int_{s=t-1}^{t} f_\theta(\mathbf{z}_{T+s}, \mathbf{a}_{T+t-1}; \mathbf{c}) ds, \underline{\mathbf{z}_T = \Phi(\mathcal{H}_T)} \tag{3}$$

As underscored in Equation (3), $\mathcal{H}_T$ enters the latent dynamics as an initial condition $\mathbf{z}_T$ through an encoder $\Phi(\mathcal{H}_T)$. Unfortunately, based on the nonlinear ICA theory, the unconditional $p(\mathbf{c})$ underscored in Equation (2) is not identifiable (Khemakhem et al., 2020).

3.2 IDENTIFIABLE FORMULATION OF LATENT NATIVE & RESPONSE DYNAMICS

To construct an alternative identifiable formulation for Equation (2), we build on recent advances in nonlinear ICA theory (Hyvarinen et al., 2019) and change the unconditional $p(\mathbf{c})$ to a conditionally independent $p_\phi(\mathbf{c}|\mathcal{H}_T)$ given the auxiliary information in history $\mathcal{H}_T$. To facilitate the separation of native and response dynamics from their composite observations, we further factorize $p_\phi(\mathbf{c}|\mathcal{H}_T)$ as:

$$p_\phi(\mathbf{c}|\mathcal{H}_T) = \prod_{k=0}^{d_a} p_{\phi^k}(\mathbf{c}^k|\mathcal{H}_T^k), \quad p_{\phi^k}(\mathbf{c}^k|\mathcal{H}_T^k) \sim \mathcal{N}\big(\boldsymbol{\mu}^k, \boldsymbol{\sigma}^k\big), \quad \{\boldsymbol{\mu}^k, \boldsymbol{\sigma}^k\} = \frac{1}{l}\sum\nolimits_{\mathbf{s}_m^k \in \mathcal{H}_T \otimes \mathcal{I}_T^k} \Psi_{\phi^k}(\mathbf{s}_m^k) \tag{4}$$

where we condition $\mathbf{c}^k$'s for native dynamics ($k = 0$) and each response dynamics ($k \geq 1$) on their relevant history context $\mathcal{H}_T^k$'s. We assume $\mathbf{c}^k$ to be shared by l number of segments $\mathbf{s}_m^k \in \mathcal{H}_T \otimes \mathcal{I}_T^k$, where $\mathcal{I}_T^k$ selects timeframes following an assignment of intervention k (or intervention-free for $k = 0$) in history $\mathcal{H}_T$, the length m of $\mathbf{s}_m^k$ is tunable, and l grows with time. Ψ_{ϕ^k} is a neural network parameterized by ϕ^k with two output heads. We now formulate an alternative for Equation (2) as:

$$p(\mathbf{y}_{T+1:T+\tau}|\mathbf{a}_{T:T+\tau-1}, \mathcal{H}_T) = \int_{\mathbf{c}} \prod_{t=1}^{\tau} p_\eta(\mathbf{y}_{T+t}|\mathbf{z}_{T+t})|_{\mathbf{z}_{T+t}=\mathcal{F}_\theta(\mathbf{z}_{T+t-1}, \mathbf{a}_{T+t-1}; \mathbf{c})} \underline{p_\phi(\mathbf{c}|\mathcal{H}_T)} d\mathbf{c} \tag{5}$$

$$\mathcal{F}_\theta(\mathbf{z}_{T+t-1}, \mathbf{a}_{T+t-1}; \mathbf{c}) = \mathbf{z}_{T+t-1} + \int_{s=t-1}^{t} f_\theta(\mathbf{z}_{T+s}, \mathbf{a}_{T+t-1}; \mathbf{c}) ds, \quad \underline{\mathbf{z}_T = \Phi_\zeta(\mathcal{H}_{T:T-l_i})} \tag{6}$$

In contrast to Equations (2) to (3), $\mathcal{H}_T$ now enters the likelihood through $p_\phi(\mathbf{c}|\mathcal{H}_T)$ in Equation (5) to govern the non-stationarity of the latent dynamics. Accordingly, as underscored in Equation (6), the initial latent state $\mathbf{z}_T$ is now extracted from l_i frames of observations preceding the prediction horizon ($l_i \ll \tau$; *e.g.*, $l_i = 3$ in our experiments), via a neural network Φ_ζ parameterized by ζ. These change the generative model fundamentally. Mathematically, in Equations (2) to (3), $\mathcal{H}_T$ affects the prediction only by its initial condition $\mathbf{z}_T$; in Equations (5) to (6), $\mathcal{H}_T$ affects each rollout step throughout the horizon of τ, which should intuitively improve prediction as τ or the non-stationarity of f_θ increases. Theoretically, this changes a non-identifiable latent process into an identifiable one.

3.3 VARIATIONAL INFERENCE VIA BAYESIAN META-LEARNING

To enable variational inference over Equations (5) to (6), we approximate the conditional posterior density of $\mathbf{c}$ as $q_\phi(\mathbf{c}|\mathcal{H}_T \cup \mathbf{y}_{T+1:T+\tau})$, sharing the same networks as defined in Equation (4). Given N training sequences of length L, $\{\mathcal{H}_L^n\}_{n=1}^N$, prediction of $p(\mathbf{y}_{T+1:T+\tau}^n|\mathbf{a}_{T:T+\tau-1}^n, \mathcal{H}_T^n)$ is performed on a sliding window of $\mathbf{y}_{T+1:T+\tau}^n$ over each sequence n given an increasing history $\mathcal{H}_T^n$, for $T = \tau : L - \tau$ for all sequences. We construct an evidence lower bound loss (Appendix B) as:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} \quad = \sum_{n=1}^{N} \sum_{T=\tau}^{L-\tau} \{ \sum_{t=1}^{\tau} \log p_\eta(\mathbf{y}_{T+t}^n|\mathbf{z}_{T+t})|_{\mathbf{z}_{T+t}=\mathcal{F}_\theta(\mathbf{z}_{T+t-1}, \mathbf{a}_{T+t-1}; \mathbf{c})} \\ - \text{KL}\big(q_\phi(\mathbf{c}|\mathcal{H}_T^n \cup \mathbf{y}_{T+1:T+\tau}^n) \,||\, p_\phi(\mathbf{c}|\mathcal{H}_T^n)\big)\}\end{aligned} \tag{7}$$

This optimization can be viewed as a Bayesian meta-learning objective: $p_\phi(\mathbf{c}|\mathcal{H}_T^n)$ represents a feed-forward meta-model that extracts $\mathbf{c}$ from context samples in $\mathcal{H}_T^n$ and uses it to adapt the latent causal dynamics process to be specific to the observed history. The accumulated loss on the prediction horizon $\mathbf{y}_{T+1:T+\tau}^n$ across $n = 1 : N$ sequences represents the meta-loss to optimize the meta-model.

Identifiability theory: Our identifiability results build primarily on the recent work of (Ye et al., 2026) which, with simple extension, allows us to establish identifiability for $p(\mathbf{c}|\mathcal{H}_T)$ up to affine transformations (Theorem 1) and identifiability of $\mathbf{z}_t$ up to permutation and component-wise invertible transformations (Theorem 2). We include both theorems and proofs in Appendix C. While we do not introduce new theoretical results, iASIDE's main contribution is algorithmic: in (Ye et al., 2026), identifiable latent dynamics is constructed by conditioning its *latent* non-stationarity parameter on few-shot examples from the same *environment*, considering only *native dynamics*. We 1) extend this formulation to separable native and interventional dynamics, and 2) introduce a novel separable conditioning of their non-stationarity on their respective history contexts. With this, we extend identifiable latent dynamics to interventional response prediction for the first time.

Table 1: Ablation results for key elements in iASIDE on $heterogenity = 3$ and $\tau = 5$.

		Model 1	Model 2	Model 3	Model 4	Model 5 (iASIDE)
Component	Separated Dynamics	✗	✓	✗	✓	✓
	Separated Conditional Prior	✗	✗	✗	✗	✓
	Adaptability/Identifiability	✗	✗	✓	✓	✓
Metrics	RMSE (↓)	$1.64_{0.11}$	$1.68_{0.06}$	$1.21_{0.03}$	$1.23_{0.05}$	$\mathbf{1.17_{0.02}}$
	MCC(↑) growth	-	-	$0.66_{0.01}$	$0.63_{0.01}$	$\mathbf{0.94_{0.01}}$
	MCC(↑) radio	-	-	$0.84_{0.01}$	$0.84_{0.01}$	$\mathbf{0.90_{0.01}}$
	MCC(↑) chemo	-	-	$0.01_{0.01}$	$0.01_{0.01}$	$\mathbf{0.76_{0.01}}$

4 EXPERIMENTS

Because counterfactual outcomes are rarely observed in real-world data, synthetic data are important for evaluating predicted intervention effect/outcomes (Lim, 2018; Bica et al., 2020; Seedat et al., 2022; Melnychuk et al., 2022). In Section 4.2, we considered synthetic image sequences of physics systems commonly used for evaluating identifiable latent dynamics models. In Section 4.1 and Section 4.3, we considered lower-dimensional synthetic and real-data benchmarks commonly used for evaluating dynamic intervention response prediction.

Baselines: We considered representative baselines able to predict dynamic outcome trajectories $\mathbf{y}_{T+1:T+\tau}$ (and excluded those focused on a final outcome $\mathbf{y}_{T+\tau}$): MSM (Hernan et al., 2000) as a classical causal model not based on neural networks, RMSN (Lim, 2018) and CRN (Bica et al., 2020) as two-stage encoder-decoder RNNs strengthened with propensity weighting and invariant representations, respectively, CausalTransformers (CT) (Melnychuk et al., 2022) and ACTIN (Wang et al., 2024) as end-to-end transformers combined with invariant representations, and SCIP (Hess & Feuerriegel, 2025) and TESAR-CDE (Vanderschueren et al., 2023) as CDE-based approaches combined with propensity weighting or invariance. For a fair comparison, in datasets with no confounding, we disabled intervention-invariant representations in all relevant baselines. When there was confounding, we tested iASIDE with intervention-invariant representations implemented through gradient reversal from an intervention classifier (Bica et al., 2020). All models were matched in parameter counts. Additional architecture details are included in Appendix E.3.

4.1 SYNTHETIC PHARMACOKINETIC-PHARMACODYNAMIC (PK-PD) TUMOR GROWTH

Data: We considered an established PK-PD model of tumor growth that includes chemotherapy and radiotherapy (Geng et al., 2017). Without confounding, we adopted a random intervention assignment probability of $p_0 = 0.1$ to avoid killing the tumor too quickly. To introduce confounding, we adopted $p(a_t) = p_0\ \sigma(\gamma(\sqrt[3]{V_t/V_{max}} - 1/2)$ from (Lim, 2018) with $p_0 = 0.2$ and $\gamma = 1, 2, 3, 4$ controlling increasing levels of confounding. We considered four levels of heterogeneity in growth rate and intervention-effect parameters as summarized in Table 4. More details are provided in Appendix E.2.

Metrics: We evaluated the accuracy of outcome prediction with root mean squared error (RMSE) normalized by the maximum tumor volume (death threshold), following standard practice (Lim, 2018). To verify *identifiability*, we adopted the mean correlation coefficient (MCC) between the inferred $\mathbf{c}$ and true growth/intervention parameters following (Khemakhem et al., 2020). All baselines were run on multiple seeds. Metrics for iASIDE were tested against the one-sided null hypothesis that they are not better than the best-performing baseline, using Welch's t-test.

Ablating the importance of *adaptability* and *separability*: We started by presenting ablation evidence for the key components of iASIDE: *adaptability* realized via conditioning $\mathbf{c}$ on history, and *separability* realized jointly by the separate dynamics and separate conditional priors for $\mathbf{c}^k$'s, which work together to reach *identifiability* both in theory and in practice. We started with a single latent neural ODE that is neither separable nor adaptable (Model 1). As we decomposed this single neural ODE into separate native and response dynamics similar to Equation (1) but without $\mathcal{H}_T$-conditioning (Model 2), there was no improvement. Conversely, conditioning this single model on $\mathcal{H}_T$ without separate dynamics (Model 3) brought notable improvements in prediction RMSE, although MCCs of the identified parameters were suboptimal: this demonstrated both the importance of theoretical identifiability, and the optimization challenge in reaching this theoretical solution in practice. Further decomposing the dynamics alone (Model 4) did not help, while also separating the conditional prior of $\mathbf{c}^k$'s (Model 5) brought statistically significant improvements, especially in MCC of identified parameters. This ablation confirms the most important innovation within iASIDE: beyond conditional

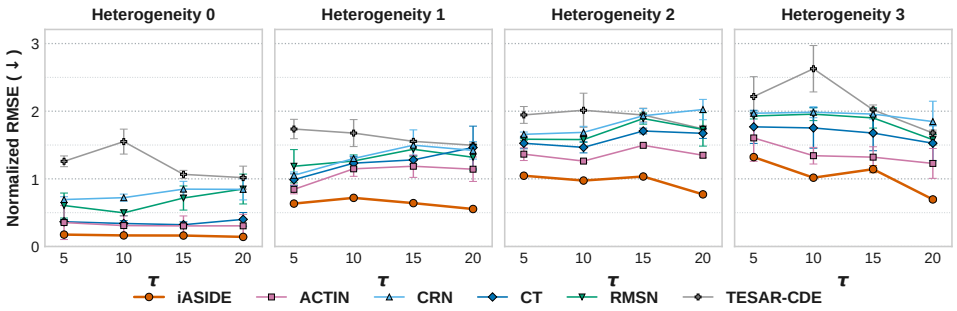

Figure 2: Normalized RMSEs (↓) over increasing projection horizons and different levels of data heterogeneity, without confounding. Full quantitative results are included in Sections F.2.1 and F.2.2.

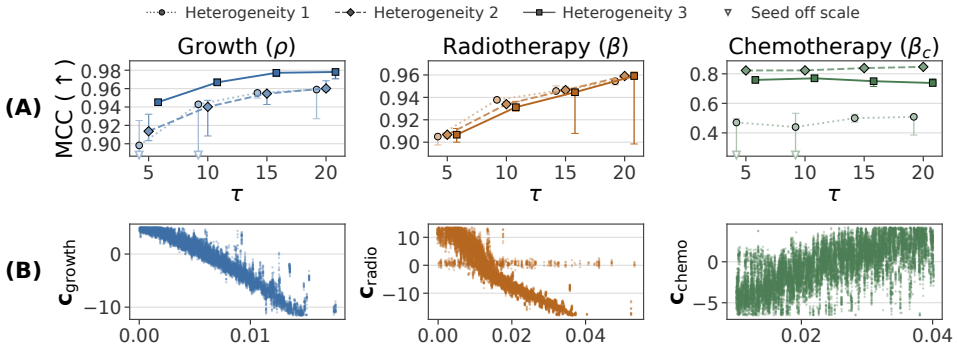

Figure 3: (A) Spearman-based MCC (↑) between learned embedding $\mathbf{c}^k$ and true parameters: mean and standard deviation from 5 seeds. (B) Example results of learned (y-axis) versus true parameters (x-axis), one seed at heterogeneity 3, $\tau = 20$. Top and bottom 5th percentiles excluded as outliers.

prior to establish theoretical identifiability, the factorization of history-conditional priors for native and response dynamics is critical to reach the identifiable solution in practical optimization.

Results without confounding: Figure 2 summarizes the normalized RMSE for all models considered, under different levels of heterogeneity and different prediction horizons when there was no confounding: SCIP was removed due to their much larger RMSEs; complete results can be found in Table 8. Notably, iASIDE (red) provided significant margins of improvement over the baselines across all prediction horizons and all levels of heterogeneity, especially when τ or heterogeneity was high. Figure 3 summarizes the MCC metrics between estimated $\mathbf{c}^k$'s and the true parameters, excluding heterogeneity level 0 because it violated the sufficient variability condition in Theorem 1: As shown, parameter ρ for native growth dynamics and β for radiation effect were in general well identified, while β_c for chemotherapy was comparatively more difficult to precisely identify. The lower MCC results for ρ at heterogeneity level 1 and $\tau = 5$–10 and the poor MCC for β_c at heterogeneity level 1 implied potentially insufficient data variability for establishing identifiability. Figure 8 in Appendix F.3 provides examples of prediction results on individual trajectories.

Appendix F.4-F.5 provide additional results of testing the identified models in selecting optimal treatment types (F.4) and predicting with out-of-distribution dynamics parameters (F.5), both showing significant improvements over baselines. These results provided strong evidence that addressing the identifiability of latent dynamics is critical even when there is no confounding.

Results with confounding: Table 2 summarizes the RMSE of iASIDE *vs*. all baselines at varying levels of intervention assignment bias, for prediction horizon $\tau = 5$ and the highest heterogeneity level. Because the CDE-based variants delivered similar performance, we only included their best representative TESAR-CDE. As shown, even without any additional strategies for addressing time-varying confounding ($\alpha = 0$), iASIDE showed significantly better performance than all baselines across all levels of confounding. The additional use of invariant representation further improved performance, demonstrating iASIDE's complementarity to existing de-confounding works.

Table 2: Normalized RMSE (↓) at different treatment-assignment confounding coefficient γ, ($\tau = 5$, heterogeneity 3) with mean and standard deviation (subscript) over 5 training seeds (except for CT at $\gamma = 2$ for which a bad outlier seed was removed). α is the multiplier for learning treatment-invariant representations following (Bica et al., 2020). **Bold** is the lowest value and *italic* the second lowest in a row. * marks $p < 0.05$ under a one-sided Welch t-test as mentioned in section 4.1.

γ	ACTIN	CRN	CT	RMSN	SCIP	TESAR-CDE	iASIDE α=0	iASIDE α=1
1	$\mathit{1.888}_{0.452}$	$2.510_{0.169}$	$2.364_{0.052}$	$2.493_{0.130}$	$4.626_{0.625}$	$2.564_{0.246}$	$1.525_{0.101}$	$\mathbf{1.444}^{*}_{0.049}$
2	$\mathit{2.170}_{0.511}$	$2.781_{0.067}$	$2.692^{(4)}_{0.031}$	$2.748_{0.201}$	$5.394_{0.518}$	$3.341_{0.433}$	$1.845_{0.201}$	$\mathbf{1.661}^{*}_{0.037}$
3	$\mathit{2.309}_{0.070}$	$3.514_{0.272}$	$3.199_{0.179}$	$3.097_{0.126}$	$5.364_{0.375}$	$3.933_{0.168}$	$1.936_{0.268}$	$\mathbf{1.897}^{*}_{0.080}$
4	$\mathit{2.694}_{0.214}$	$4.019_{0.297}$	$3.733_{0.219}$	$3.480_{0.125}$	$6.396_{0.671}$	$4.732_{0.301}$	$2.621_{0.432}$	$\mathbf{2.279}^{*}_{0.089}$

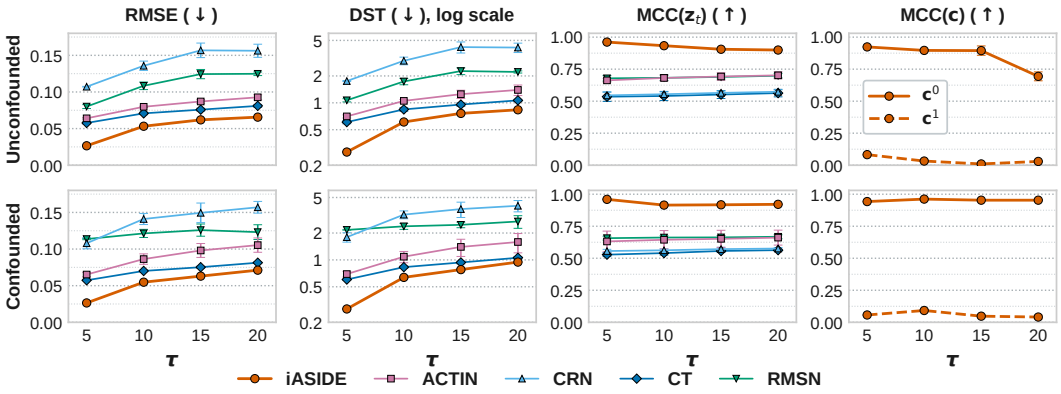

Figure 4: Representative results on Pendulum without (A) and with (B) confounding.

4.2 High-dimensional Image Sequences of Physics Systems

Data: To test iASIDE in high-dimensional data, we considered the damped Pendulum (Ye et al., 2026) under external forces: native dynamics was governed by heterogeneous frequency coefficients, and response dynamics governed by varying force-pulse magnitudes. Confounding was added by influencing the probability of force application with pendulum positions, at a marginal pulse rate matched to the unconfounded setting. For each parameter combination, we generated image sequences with varying initial conditions. We used prediction horizons of $\tau = 5, 10, 15, 20$, conditioned on histories of length $T \in [\tau,\ 59 - \tau]$. Additional details are in Appendix E.1.

Metrics: For the accuracy of *predicting intervention outcomes*, we considered the MSE of predicted images and the distance (DST) between the true and predicted location of the pendulum. Identifiability of $\mathbf{c}$ and $\mathbf{z}_t$ was measured by MCC. We included the subset of baselines suitable for image datasets.

Results: Fig. 4 summarizes representative results both without (A) and with (B) confounding, where iASIDE continued to outperform included baselines in both the prediction of outcomes (*i.e.*, images) and the recovery of $\mathbf{z}_t$'s across horizons. iASIDE also achieved strong MCC metrics for the identification of the pendulum's native frequency coefficient ($\mathbf{c}^0$), while the MCC metrics for the force magnitude (interventional) were lower suggesting increased challenge for optimization. Complete quantitative results and visual examples of image sequences are provided in Appendix F.1.

4.3 Real Data Experiments

Finally, to test feasibility in real-world settings, we conducted experiments on MIMIC-III (Johnson et al., 2023), an electronic health record dataset with inherent real-world intervention assignment bias (Bica et al., 2020). Following Melnychuk et al. (2022), we considered covariates $\mathbf{x}_t$ as 25 vitals and 3 static features, outcome $\mathbf{y}_t$ as diastolic blood pressure (DBP), and treatment $\mathbf{a}_t$ as vasopressor and ventilation. We selected only the most competitive baselines in this experiment. Test RMSE results in Table 3 show that iASIDE was able to deliver state-of-the-art performance across all prediction horizons, although without statistical significance between the top-performing models (Welch's t-tests). It is noteworthy that, while all models' RMSE increased as τ increased, iASIDE

	$\tau = 5$	$\tau = 10$	$\tau = 15$	$\tau = 20$
ACTIN	$9.63_{0.13}$	$10.08_{0.10}$	$10.44_{0.15}$	$10.71_{0.17}$
CRN	$9.87_{0.12}$	$10.57_{0.15}$	$11.11_{0.22}$	$11.62_{0.20}$
CT	$\mathbf{9.57}_{0.13}$	$\mathit{10.02}_{0.14}$	$\mathit{10.37}_{0.22}$	$\mathit{10.64}_{0.29}$
RMSN	$10.63_{0.41}$	$12.12_{0.66}$	$12.96_{0.32}$	$13.05_{0.31}$
SCIP	$13.06_{2.14}$	$14.20_{1.67}$	$12.66_{0.78}$	$13.15_{1.53}$
iASIDE	$\mathit{9.61}_{0.14}$	$\mathbf{10.01}_{0.12}$	$\mathbf{10.30}_{0.19}$	$\mathbf{10.51}_{0.23}$

Table 3: Test RMSE (↓) of DBP (mmHg) on held-out MIMIC-III patients across prediction horizon τ; mean and standard deviation (subscript) over 5 seeds.

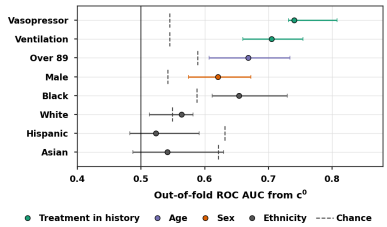

Figure 5: Subgroup analysis of learned embedding $\mathbf{c}^0$ on MIMIC-III, showing Linear-probe results on $\mathbf{c}^0$ against historical intervention presence and **unseen** demographic variables.

demonstrated the least deterioration (9%) compared to the baselines (ranging from 11% to 22%). To understand what heterogeneity iASIDE may have captured, we tested a linear probe on the learned $\mathbf{c}^0$ against several demographic variables (not provided in training data) along with the presence of vasopressor or ventilation in history. Results in Figure 5 suggest that iASIDE may have learned to adapt the native dynamics to subgroups indicated by historical interventions and age.

4.4 Additional Ablation Results

Ablation of data-generating assumptions: In the two synthetic datasets, the data generation dynamics followed an additive separation of native and response dynamics. In Appendix G.1, we created additional toy datasets of nonlinear native and interventional dynamics 1) with no shared components or 2) with shared components but non-additive responses. Results demonstrated that, under binary interventions, iASIDE is generally applicable to any nonlinear native and response dynamics as long as there are no *multiplicative interactions among interventions.*

Sample efficiency: In Appendix G.2, we provided additional results comparing iASIDE with baselines as the amount of training data decreased, demonstrating its improved sample efficiency potentially due to the inductive bias associated with the separable dynamics.

Computational cost: Table 14 in Appendix H compares the parameter count and computation time of iASIDE *vs*. representative baselines in Pendulum. As shown, iASIDE did incur additional training overhead due to the use of an iterative ODE solver for latent dynamics and multiple meta-models for adapting separable dynamics. This overhead brought significant forecasting improvements, with test-time efficiency in between RNN/CDE- and transformer-based baselines.

5 Conclusions and Discussion

We present iASIDE, a novel framework for learning identifiable, adaptable, and separable interventional dynamics, and demonstrate its improved accuracy in predicting intervention outcomes over increased prediction horizons and increased data heterogeneity. **Limitations:** As a first step towards addressing the identifiability of latent interventional dynamics, iASIDE was not designed with a new mechanism to address the separate challenge of confounding. Although its performance remained strong in the presence of time-varying confounding, a necessary next step is to investigate, from a theoretical perspective, the impact of such confounding on iASIDE's identifiability theory and the corresponding solutions. Secondly, while iASIDE showed strength in adapting to non-stationary dynamics systems, such non-stationarity might also arise over time within the same system. This in principle can be incorporated into iASIDE but requires the design of feed-forward meta-models to replace the aggregation mechanisms currently used in Equation (4) with, for instance, weighted averaging, time decay, or attention mechanisms to find similar dynamics in the history. Thirdly, the additive decomposition of response dynamics to multiple interventions currently does not address multiplicative interactions among interventions. Finally, identifiability results were suboptimal for some interventional dynamics (*e.g.*, chemotherapy). Potential reasons, such as violation of the identifiability conditions or optimization challenges, need to be examined in future work.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and, interpreting results. We have not used generative AI tools for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or assisting with translation. Additionally, we used generative AI tools for creating and modifying scientific figures or images, suggesting experimental parameters, creating and editing software code, creating artifacts, sourcing/searching for information, and format references. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We share our anonymzed source code. Refer to Appendix I.

REFERENCES

Ilze Amanda Auzina, Çağatay Yı ldız, Sara Magliacane, Matthias Bethge, and Efstratios Gavves. Modulated neural odes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems, volume 36, pp. 44572–44594. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2023/file/8bc74514d554a90c996576f6c373f5f3-Paper-Conference.pdf.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 2639–2672. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/balsells-rodas24a.html.

Jeroen Berrevoets, Alicia Curth, Ioana Bica, Eoin McKinney, and Mihaela van der Schaar. Disentangled counterfactual recurrent networks for treatment effect inference over time. CoRR, abs/2112.03811(arXiv:2112.03811), Dec 2021. doi: 10.48550/arXiv.2112.03811. URL http://arxiv.org/abs/2112.03811. arXiv:2112.03811 [cs].

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Aleksandar Botev, Andrew Jaegle, Peter Wirnsberger, Daniel Hennes, and Irina Higgins. Which priors matter? benchmarking models for learning latent dynamics. In Advances in Neural Information Processing Systems, 2021.

Edward De Brouwer, Javier Gonzalez, and Stephanie Hyland. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, pp. 4705–4722. PMLR, May 2022. URL https://proceedings.mlr.press/v151/de-brouwer22a.html.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo- and radiation therapy for non-small cell lung cancer patients using a biomathematical model. Scientific Reports, 7(1), October 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13646-z. URL http://dx.doi.org/10.1038/s41598-017-13646-z.

Daehoon Gwak, Gyuhyeon Sim, Michael Poli, Stefano Massaroli, Jaegul Choo, and Edward Choi. Neural ordinary differential equations for intervention modeling. CoRR, abs/2010.08304 (arXiv:2010.08304), Oct 2020. doi: 10.48550/arXiv.2010.08304. URL http://arxiv.org/abs/2010.08304. arXiv:2010.08304 [cs].

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In Jonas Peters and David Sontag (eds.), Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), volume 124 of Proceedings of Machine Learning Research, pp. 939–948. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/halva20a.html.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), Advances in Neural Information Processing Systems, volume 34, pp. 1624–1633. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0cdbb4e65815fbaf79689b15482e7575-Paper.pdf.

MA Hernan and JM Robins. Causal Inference: What If. Chapman & Hall/CRC, 2020. URL https://miguelhernan.org/whatifbook.

Miguel Angel Hernan, Babette Brumback, and James M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. Epidemiology, 11(5): 561–570, September 2000. ISSN 1044-3983. doi: 10.1097/00001648-200009000-00012. URL http://dx.doi.org/10.1097/00001648-200009000-00012.

Konstantin Hess and Stefan Feuerriegel. Stabilized neural prediction of potential outcomes in continuous time. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=aN57tSd5Us.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. IGC-net for conditional average potential outcome estimation over time. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=ZmhpqpKzAT.

Çağlar Hı zlı, Çağatay Yı ldız, Matthias Bethge, ST John, and Pekka Marttinen. Identifying latent state transitions in non-linear dynamical systems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), International Conference on Learning Representations, volume 2025, pp. 71286–71315, 2025. URL https://proceedings.iclr.cc/paper%5Ffiles/paper/2025/file/b192f601432dff4e6777d2d06d52cfa5-Paper-Conference.pdf.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In Aarti Singh and Jerry Zhu (eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pp. 460–469. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/hyvarinen17a.html.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of Proceedings of Machine Learning Research, pp. 859–868. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Xiajun Jiang, Ryan Missel, Zhiyuan Li, and Linwei Wang. Sequential latent variable models for few-shot high-dimensional time-series forecasting. In The Eleventh International Conference on Learning Representations, Feb 2023. URL https://openreview.net/forum?id=7C9aRX2nBf2.

Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database, 2023. URL https://physionet.org/content/mimiciii/1.4/.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In Advances in Neural Information Processing Systems, volume 33, pp. 6696–6707. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Matthieu Kirchmeyer, Yuan Yin, Jeremie Dona, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 11283–11301. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kirchmeyer22a.html.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In ICLR, 2021. URL https://openreview.net/forum?id=EbIDjBynYJ8.

Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PUpZXvNqmb.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters, 2015. URL https://arxiv.org/abs/1511.05121.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10):4156–4165, 2019.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei Lehman. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer (eds.), Proceedings of Machine Learning for Health, volume 158 of Proceedings of Machine Learning Research, pp. 282–299. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/li21a.html.

Zijian Li, Ruichu Cai, Zhenhui Yang, Haiqin Huang, Guangyi Chen, Yifan Shen, Zhengming Chen, Xiangchen Song, and Kun Zhang. Nonstationary time series forecasting via unknown distribution adaptation, 2025. URL https://arxiv.org/abs/2402.12767.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/56e6a93212e4482d99c84a639d254b67-Abstract.html.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In Proceedings of the 39th International Conference on Machine Learning, pp. 15293–15329. PMLR, Jun 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

James M. Robins, Sander Greenland, and Fu-Chang Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. Journal of the American Statistical Association, 94(447):687–700, Sep 1999. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1999.10474168.

Donald B Rubin. Causal inference using potential outcomes. Journal of the American Statistical Association, 100(469):322–331, 2005. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In Proceedings of the 39th International Conference on Machine Learning, pp. 19497–19521. PMLR, Jun 2022. URL https://proceedings.mlr.press/v162/seedat22b.html.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. Advances in neural information processing systems, 32, 2019.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems, volume 36, pp. 8092–8113. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2023/file/19a567abaec3990cb40d7a013556fecd-Paper-Conference.pdf.

Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), Advances in Neural Information Processing Systems, volume 34, pp. 14809–14821. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7ca57a9f85a19a6e4b9a248c1daca185-Paper.pdf.

Toon Vanderschueren, Alicia Curth, Wouter Verbeke, and Mihaela Van Der Schaar. Accounting for informative sampling when learning to forecast treatment outcomes over time. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 34855–34874. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/vanderschueren23a.html.

Rui Wang, Robin Walters, and Rose Yu. Meta-learning dynamics forecasting using task inference. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems, volume 35, pp. 21640–21653. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2022/file/87f476af4053961667c2c08e9f4b850e-Paper-Conference.pdf.

Xin Wang, Shengfei Lyu, Lishan Yang, Yibing Zhan, and Huanhuan Chen. A dual-module framework for counterfactual estimation over time. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 51063–51085. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24au.html.

Hong Xiong, Feng Wu, Leon Deng, Megan Su, and Li-wei H. Lehman. G-transformer: Counterfactual outcome prediction under dynamic and time-varying treatment regimes. In Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary Lipton, Rajesh Ranganath, and Iñigo Urteaga (eds.), Proceedings of the 9th Machine Learning for Healthcare Conference, volume 252 of Proceedings of Machine Learning Research. PMLR, 16–17 Aug 2024. URL https://proceedings.mlr.press/v252/xiong24a.html.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems, volume 35, pp. 26492–26503. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2022/file/a938292feb86b94ebe3e6200ff7786ef-Paper-Conference.pdf.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In International Conference on Learning Representations, 2022b. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Yubo Ye, Maryam Toloubidokhti, Sumeet Vadhavkar, Xiajun Jiang, Huafeng Liu, and Linwei Wang. On the identifiability of hybrid deep generative models: Meta-learning as a solution. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 7714–7735. Curran Associates, Inc., 2024. doi: 10.52202/079017-0247. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2024/file/0e8909cae8248c98279f6cd82074aa6d-Paper-Conference.pdf.

Yubo Ye, Sweekar Piya, Xiajun Jiang, and Linwei Wang. Meta-ilad: Identifiable latent dynamics via meta-learning of dynamics environments. In Forty-third International Conference on Machine Learning, 2026. URL https://openreview.net/forum?id=yQJFBNKbvm.

TABLE OF CONTENTS

A Separable Latent Dynamics under Binary Interventions **16**

B ELBO Derivation **17**

C Identifiability Theorems and Proofs **18**

C.1 Identifiability of Latent Exogenous Variable $\mathbf{c}$ for Non-stationarity 18

C.2 Identifiability of Latent Dynamics States $\mathbf{z}_t$ 20

D Causal Identifiability Assumptions **22**

E Additional Experimental Details **23**

E.1 Details on Pendulum Data 23

E.2 Details on PK-PD Data of Tumor Growth 23

E.3 Implementation details 25

E.4 Description of Baselines 26

F Additional Results **27**

F.1 Complete Results on Pendulum 27

F.2 Complete Results on Synthetic Tumor Growth Data 30

F.3 Visual examples 32

F.4 Optimal Treatment Selection Accuracy 33

F.5 Prediction in out-of-distribution dynamics parameters 34

G Additional Ablation Results **35**

G.1 Ablation of Data Generating Assumptions 35

G.2 Ablation of Sample Efficiency 35

H Computation Cost **36**

I Source Code **36**

A SEPARABLE LATENT DYNAMICS UNDER BINARY INTERVENTIONS

Assume the dynamics of a system in the absence of intervention and in the presence of interventions are two completely separate functions f_0 and f_1, and thus the dynamics can be written as:

$$\frac{d\mathbf{z}_t}{dt} = \begin{cases} f_0(\mathbf{z}_t, \mathbf{c}), \text{ when } \mathbf{a}_t = 0 \\ f_1(\mathbf{z}_t, \mathbf{c}), \text{ when } \mathbf{a}_t = 1 \end{cases} \tag{8}$$

The above equation in eq. (8) can also be written in the form:

$$\begin{aligned} \frac{d\mathbf{z}_t}{dt} &= f_0(\mathbf{z}_t, \mathbf{c}) + \mathbf{a}_t \times f_\tau(\mathbf{z}_t, \mathbf{c}) \\ \text{where} \quad f_\tau(\mathbf{z}_t, \mathbf{c}) &= f_1(\mathbf{z}_t, \mathbf{c}) - f_0(\mathbf{z}_t, \mathbf{c}) \end{aligned} \tag{9}$$

Now, for d_a different binary interventions and assuming no multiplicative interactions among different interventions, eq. (9) will be of the form:

$$\frac{d\mathbf{z}_t}{dt} = f_0(\mathbf{z}_t, \mathbf{c}) + \mathbf{a}_t^1 \times f_\tau^1(\mathbf{z}_t, \mathbf{c}^1) + ... + \mathbf{a}_t^{d_a} \times f_\tau^{d_a}(\mathbf{z}_t, \mathbf{c}^{d_a}) \tag{10}$$

$$\frac{d\mathbf{z}_t}{dt} = f_0(\mathbf{z}_t, \mathbf{c}) + \sum_{k=1}^{d_a} \mathbf{a}_t^k \times f_\tau^k(\mathbf{z}_t, \mathbf{c}^k) \tag{11}$$

Irrespective of the functional form of different dynamics of f_0 or $f^1, ..., f^{d_a}$, the combined effect can be written in the form of Equation (11). In other words, this additive decomposition between native and response dynamics follows naturally from the binary nature of $\mathbf{a}_t$ and **does not** induce any additional structural assumptions. The additive formulation among d_a interventions, however, does omit multiplicative interactions among interventions.

B ELBO DERIVATION

For clarity, let $\mathcal{F}_\theta(\mathbf{c})$ be a shorthand for $\mathcal{F}_\theta(\mathbf{z}_{T+t-1}, \mathbf{a}_{T+t-1}; \mathbf{c})$ in the derivation below.

$$
\begin{aligned}
&\sum_{n=1}^{N}\sum_{T=\tau}^{L-\tau} \log p\big(\mathbf{y}^n_{T+1:T+\tau} | \mathbf{a}_{T:T+\tau-1}, \mathcal{H}^n_T\big) \\
&= \sum_{n=1}^{N}\sum_{T=\tau}^{L-\tau} \log \int_{\mathbf{c}} \prod_{t=1}^{\tau} p_\eta(\mathbf{y}^n_{T+t} | \mathbf{z}_{T+t})|_{\mathcal{F}_\theta(\mathbf{c})} p_\phi(\mathbf{c}|\mathcal{H}^n_T) d\mathbf{c} \\
&= \sum_{n=1}^{N}\sum_{T=\tau}^{L-\tau} \log \int_{\mathbf{c}} \prod_{t=1}^{\tau} p_\eta(\mathbf{y}^n_{T+t} | \mathbf{z}_{T+t})|_{\mathcal{F}_\theta(\mathbf{c})} \frac{p_\phi(\mathbf{c}|\mathcal{H}^n_T)}{q_\phi(\mathbf{c}|\mathcal{H}^n_T \cup \mathbf{y}_{T+1:T+\tau})} q_\phi(\mathbf{c}|\mathcal{H}^n_T \cup \mathbf{y}_{T+1:T+\tau}) d\mathbf{c} \\
&= \sum_{n=1}^{N}\sum_{T=\tau}^{L-\tau} \log \mathbb{E}_{q(\mathbf{c})}\Big[\prod_{t=1}^{\tau} p_\eta(\mathbf{y}^n_{T+t} | \mathbf{z}_{T+t})|_{\mathcal{F}_\theta(\mathbf{c})} \frac{p_\phi(\mathbf{c}|\mathcal{H}^n_T)}{q_\phi(\mathbf{c}|\mathcal{H}^n_T \cup \mathbf{y}_{T+1:T+\tau})} \Big] \\
&\geq \sum_{n=1}^{N}\sum_{T=\tau}^{L-\tau} \mathbb{E}_{q(\mathbf{c})}\Big[\sum_{t=1}^{\tau} \log p_\eta(\mathbf{y}^n_{T+t} | \mathbf{z}_{T+t})|_{\mathcal{F}_\theta(\mathbf{c})} - \log \Big[\frac{q_\phi(\mathbf{c}|\mathcal{H}^n_T \cup \mathbf{y}_{T+1:T+\tau})}{p_\phi(\mathbf{c}|\mathcal{H}^n_T)} \Big] \Big] \\
&= \sum_{n=1}^{N}\sum_{T=\tau}^{L-\tau} \Big[\sum_{t=1}^{\tau} \mathbb{E}_{q(\mathbf{c})}[\log p_\eta(\mathbf{y}^n_{T+t} | \mathbf{z}_{T+t})|_{\mathcal{F}_\theta(\mathbf{c})}] - \mathrm{KL}\big(q_\phi(\mathbf{c}|\mathcal{H}^n_T \cup \mathbf{y}^n_{T+1:T+\tau}) \,||\, p_\phi(\mathbf{c}|\mathcal{H}^n_T)\big) \Big]
\end{aligned}
\tag{12}
$$

C IDENTIFIABILITY THEOREMS AND PROOFS

C.1 IDENTIFIABILITY OF LATENT EXOGENOUS VARIABLE $\mathbf{c}$ FOR NON-STATIONARITY

To begin, we define $\mathbf{y}_{T+1:T+\tau} = \mathcal{G}(\mathbf{a}_{T:T+\tau-1}, \mathbf{c}) + \epsilon$ where $\mathcal{G}$ summarizes the generation process described in Equation (1). We first introduce the exponential family distribution and related properties that form the theoretical foundation of our identifiability proof.

Definition 1 (Exponential family)**.** *A univariate exponential family is a set of distributions whose probability density function can be written as*

$$p(x) = Q(x)\, Z(\theta)\, e^{\langle \mathbf{T}(x),\, \theta \rangle} \tag{13}$$

where $\mathbf{T} : \mathbb{R} \to \mathbb{R}^k$ is the sufficient statistic, $\theta \in \mathbb{R}^k$ is the natural parameter, $Q : \mathbb{R} \to \mathbb{R}$ is the base measure, and $Z(\theta)$ is the normalization constant. The dimension $k \in \mathbb{N} \setminus \{0\}$ is always considered to be minimal, meaning that the density cannot be rewritten in the above form with a smaller $k' < k$. We call k the size of p.

Definition 2 (Strongly exponential distribution)**.** *We say that an exponential family distribution is strongly exponential if for any subset $\mathcal{X} \subseteq \mathbb{R}$ the following holds:*

$$\left(\exists\, \theta \in \mathbb{R}^k \mid \forall x \in \mathcal{X},\ \langle \mathbf{T}(x), \theta \rangle = \text{const}\right) \implies \left(\ell(\mathcal{X}) = 0 \ \text{ or } \ \theta = \mathbf{0}\right) \tag{14}$$

where ℓ denotes the Lebesgue measure.

Lemma 1. *Consider a strongly exponential distribution of size $k \geq 2$ with sufficient statistic $\mathbf{T}(x) = (T_1(x), \ldots, T_k(x))$. Further assume that $\mathbf{T}$ is differentiable almost everywhere. Then there exist k distinct values $x_1, \ldots, x_k$ such that $(\mathbf{T}'(x_1), \ldots, \mathbf{T}'(x_k))$ are linearly independent in $\mathbb{R}^k$.*

Proof. Suppose for contradiction that for any choice of k such points, the family $(\mathbf{T}'(x_1), \ldots, \mathbf{T}'(x_k))$ is never linearly independent. That means $\mathbf{T}'(\mathbb{R})$ is contained in a subspace of $\mathbb{R}^k$ of dimension at most $k-1$. Let θ be a nonzero vector orthogonal to $\mathbf{T}'(\mathbb{R})$. Then for all $x \in \mathbb{R}$, $\langle \mathbf{T}'(x), \theta \rangle = 0$. Integrating, $\langle \mathbf{T}(x), \theta \rangle = \text{const}$ for all $x \in \mathbb{R}$ with $\theta \neq \mathbf{0}$, which contradicts the assumption that the distribution is strongly exponential. □

We now assume $p_\phi(\mathbf{c} \mid \mathcal{H}_T)$ to be a distribution of the exponential family where each dimension of $\mathbf{c}$ is independent of the others when conditioned on $\mathcal{H}_T$, with sufficient statistic $\mathbf{T}(\mathbf{c})$ and natural parameters $\boldsymbol{\lambda}_\phi(\mathcal{H}_T)$, as parameterized by the prior encoders ϕ^k defined in Equation (4). Note that the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is a strongly exponential distribution of size $k = 2$ with sufficient statistic $\mathbf{T}_i(c_i) = (c_i, c_i^2)$ and natural parameter

$$\boldsymbol{\lambda}_{\phi_i}(\mathcal{H}_T) = \left(\frac{\mu_{\phi_i}(\mathcal{H}_T)}{\sigma^2_{\phi_i}(\mathcal{H}_T)},\ \frac{-1}{2\sigma^2_{\phi_i}(\mathcal{H}_T)} \right) \tag{15}$$

where $\mu_{\phi_i}(\mathcal{H}_T)$ and $\sigma^2_{\phi_i}(\mathcal{H}_T)$ are the outputs of the prior encoder networks Ψ_{ϕ^k}'s as defined in Equation (4).

Theorem 1. *Assume that we observed data $\mathbf{y}_{T+1:T+\tau}$ generated following Equations (5) to (6) with parameters $\Phi = \{\theta, \eta, \phi, \zeta\}$ and latent variable $\mathbf{c}$. Assume a model with parameters $\hat{\Phi} = \{\hat{\theta}, \hat{\eta}, \hat{\phi}, \hat{\zeta}\}$ and latent variable $\hat{\mathbf{c}}$ such that $p_{\hat{\Phi}}(\mathbf{y}_{T+1:T+\tau})$ matches $p_\Phi(\mathbf{y}_{T+1:T+\tau})$ everywhere. Assume the following hold:*

1. *The set $\{\mathbf{y}_{T+1:T+\tau} \in \mathcal{Y}^\tau \mid \varphi(\mathbf{y}_{T+1:T+\tau}) = 0\}$ has measure zero, where φ is the characteristic function of the density p_ϵ.*
2. *The function $\mathcal{G}$ is injective.*
3. *The sufficient statistics $\mathbf{T}$ are differentiable almost everywhere, and the elements of $\mathbf{T}$ are linearly independent on any subset of $\mathcal{Y}^\tau$ of measure greater than zero.*
4. *There exist $2d_c + 1$ distinct $\mathcal{H}^0, \ldots, \mathcal{H}^{2d_c}$ such that the $2d_c \times 2d_c$ matrix*

$$\boldsymbol{L} = \left(\boldsymbol{\lambda}_\phi(\mathcal{H}^1_T) - \boldsymbol{\lambda}_\phi(\mathcal{H}^0_T),\ \ldots,\ \boldsymbol{\lambda}_\phi(\mathcal{H}^{2d_c}_T) - \boldsymbol{\lambda}_\phi(\mathcal{H}^0_T) \right) \tag{16}$$

is invertible, meaning that there is sufficient variability within the latent values of $\mathbf{c}$ observed in the history data.

Then there exists an invertible matrix $\bar{\mathbf{A}}$ and vector $\bar{\mathbf{b}}$, such that

$$\mathbf{T}(\mathbf{c}) = \bar{\mathbf{A}}\,\mathbf{T}(\hat{\mathbf{c}}) + \bar{\mathbf{b}} \tag{17}$$

In other words, the sufficient statistics for $\mathbf{c}$ are identifiable up to an affine transformation.

Proof. Suppose we have two sets of parameters $\Phi = (\theta, \eta, \phi, \zeta)$ and $\hat{\Phi} = (\hat{\theta}, \hat{\eta}, \hat{\phi}, \hat{\zeta})$ such that:

$$p_\Phi(\mathbf{y}_{T+1:T+\tau}|\mathcal{H}_T) = p_{\hat{\Phi}}(\mathbf{y}_{T+1:T+\tau}|\mathcal{H}_T) \tag{18}$$

$$\int p_{\eta,\theta}(\mathbf{y}_{T+1:T+\tau}|\mathbf{c})\, p_\phi(\mathbf{c}|\mathcal{H}_T)\, d\mathbf{c} = \int p_{\hat{\eta},\hat{\theta}}(\mathbf{y}_{T+1:T+\tau}|\hat{\mathbf{c}})\, p_{\hat{\phi}}(\hat{\mathbf{c}}|\mathcal{H}_T)\, d\hat{\mathbf{c}} \tag{19}$$

Following (Ye et al., 2024), the above can be shown to reduce to:

$$p_\phi(\mathbf{c}|\mathcal{H}_T)\left|\boldsymbol{J}_{\mathcal{G}^{-1}}(\mathbf{y}_{T+1:T+\tau})\right| = p_{\hat{\phi}}(\hat{\mathbf{c}}|\mathcal{H}_T)\left|\boldsymbol{J}_{\hat{\mathcal{G}}^{-1}}(\mathbf{y}_{T+1:T+\tau})\right| \tag{20}$$

where $\boldsymbol{J}$ is the Jacobian of $\mathcal{G}$. Now consider two different values of $\mathcal{H}$: $\mathcal{H}_T^0$ and $\mathcal{H}_T^i$ for $i = 1, \ldots, 2d_c$. Using the exponential family form of $p_\phi(\mathbf{c}|\mathcal{H}_T)$ with natural parameters $\boldsymbol{\lambda}_\phi(\mathcal{H}_T)$ and sufficient statistics $\mathbf{T}(\mathbf{c})$, taking the log ratio of Equation (20) evaluated at $\mathcal{H}_T^i$ and $\mathcal{H}_T^0$ gives:

$$\log\frac{Z_\phi(\mathcal{H}_T^0)}{Z_\phi(\mathcal{H}_T^i)} + \left\langle \mathbf{T}(\mathbf{c}),\, \boldsymbol{\lambda}_\phi(\mathcal{H}_T^i) - \boldsymbol{\lambda}_\phi(\mathcal{H}_T^0)\right\rangle = \log\frac{Z_{\hat{\phi}}(\mathcal{H}_T^0)}{Z_{\hat{\phi}}(\mathcal{H}_T^i)} + \left\langle \mathbf{T}(\hat{\mathbf{c}}),\, \boldsymbol{\lambda}_{\hat{\phi}}(\mathcal{H}_T^i) - \boldsymbol{\lambda}_{\hat{\phi}}(\mathcal{H}_T^0)\right\rangle \tag{21}$$

Rearranging:

$$\left\langle \mathbf{T}(\mathbf{c}),\, \boldsymbol{\lambda}_\phi(\mathcal{H}_T^i) - \boldsymbol{\lambda}_\phi(\mathcal{H}_T^0)\right\rangle = \left\langle \mathbf{T}(\hat{\mathbf{c}}),\, \boldsymbol{\lambda}_{\hat{\phi}}(\mathcal{H}_T^i) - \boldsymbol{\lambda}_{\hat{\phi}}(\mathcal{H}_T^0)\right\rangle + b_i \tag{22}$$

Stacking all $2d_c$ equations ($i = 1, \ldots, 2d_c$) into matrix form:

$$\boldsymbol{L}^\top \mathbf{T}(\mathbf{c}) = \hat{\boldsymbol{L}}^\top \mathbf{T}(\hat{\mathbf{c}}) + \mathbf{b} \tag{23}$$

By assumption 4, $\boldsymbol{L}$ is invertible. Left-multiplying by $\boldsymbol{L}^{-\top}$:

$$\mathbf{T}(\mathbf{c}) = \underbrace{\boldsymbol{L}^{-\top}\hat{\boldsymbol{L}}^\top}_{\bar{\mathbf{A}}}\,\mathbf{T}(\hat{\mathbf{c}}) + \underbrace{\boldsymbol{L}^{-\top}\mathbf{b}}_{\bar{\mathbf{b}}} \tag{24}$$

We now establish that $\bar{\mathbf{A}}$ is invertible. By definition of $\mathbf{T}$ and according to assumption 3, its Jacobian $J_\mathbf{T}$ exists and is a $d_c k \times d_c$ matrix of rank d_c. This implies that the Jacobian of $\mathbf{T}(\hat{\mathcal{G}}^{-1} \circ \mathcal{G}(\mathbf{c}))$ with respect to $\mathbf{c}$ exists and is of rank d_c, and so is $\bar{\mathbf{A}}$ in terms of rank.

For $k \geq 2$, define $\bar{\mathbf{c}} = \mathcal{G}^{-1}(\mathbf{y}_{T+1:T+\tau})$ and $\mathbf{T}_i(\bar{c}_i) = (T_{i,1}(\bar{c}_i), \ldots, T_{i,k}(\bar{c}_i))$ for each dimension $i \in [1, \ldots, d_c]$. According to Lemma 1, for each i there exist k points $\bar{c}_i^1, \ldots, \bar{c}_i^k$ such that $(\mathbf{T}_i'(\bar{c}_i^1), \ldots, \mathbf{T}_i'(\bar{c}_i^k))$ are linearly independent. Collect those points into k vectors $(\bar{\mathbf{c}}^1, \ldots, \bar{\mathbf{c}}^k)$ where $\bar{\mathbf{c}}^l = (\bar{c}_1^l, \ldots, \bar{c}_{d_c}^l)^\top$, and concatenate the k Jacobians $J_\mathbf{T}(\bar{\mathbf{c}}^l)$ evaluated at each of those vectors horizontally into the $d_c k \times d_c k$ matrix:

$$Q = \left(J_\mathbf{T}(\bar{\mathbf{c}}^1),\ \ldots,\ J_\mathbf{T}(\bar{\mathbf{c}}^k)\right) \tag{25}$$

and similarly define $\tilde{Q}$ as the concatenation of the Jacobians of $\mathbf{T}\left(\hat{\mathcal{G}}^{-1} \circ \mathcal{G}(\bar{\mathbf{c}})\right)$ evaluated at those same k points. Then the matrix Q is invertible through a combination of Lemma 1 and the fact that each component $\mathbf{T}_i$ of $\mathbf{T}$ is univariate: since $J_\mathbf{T}(\bar{\mathbf{c}}^l)$ is block-diagonal with blocks $\mathbf{T}_i'(\bar{c}_i^l)$, the concatenated matrix Q has the same block structure, and the linear independence of $(\mathbf{T}_i'(\bar{c}_i^1), \ldots, \mathbf{T}_i'(\bar{c}_i^k))$ for each i ensures Q is invertible. By differentiating Equation (24) for each $\bar{\mathbf{c}}^l$, we get in matrix form:

$$Q = \bar{\mathbf{A}}\,\tilde{Q} \tag{26}$$

The invertibility of Q implies the invertibility of both $\bar{\mathbf{A}}$ and $\tilde{Q}$, establishing the affine identifiability of the sufficient statistics. □

Corollary 1. *When $p_\phi(\mathbf{c} \mid \mathcal{H}_T)$ is a Normal distribution, for each dimension $i \in [1, d_c]$ with sufficient statistic $\mathbf{T}_i(c_i) = (c_i, c_i^2)$ and natural parameters*

$$\boldsymbol{\lambda}_{\phi_i}(\mathcal{H}_T) = \left(\frac{\mu_{\phi_i}(\mathcal{H}_T)}{\sigma^2_{\phi_i}(\mathcal{H}_T)},\ \frac{-1}{2\sigma^2_{\phi_i}(\mathcal{H}_T)}\right) \tag{27}$$

Equation (24) *reduces to:*

$$\mathbf{c} = A \cdot P(\hat{\mathbf{c}}) + \mathbf{b} \tag{28}$$

where $A = \mathrm{diag}(a_1, \ldots, a_{d_c})$ with $a_i \neq 0$, P is a permutation matrix, and $\mathbf{b} = (b_1^{(1)}, \ldots, b_{d_c}^{(1)})^\top$.

Proof. Since $\mathbf{T}_i(c_i) = (c_i, c_i^2)$, the relationship from Equation (24) becomes

$$\begin{pmatrix} \mathbf{c} \\ \mathbf{c}^2 \end{pmatrix} = \bar{\mathbf{A}} \begin{pmatrix} \hat{\mathbf{c}} \\ \hat{\mathbf{c}}^2 \end{pmatrix} + \bar{\mathbf{b}} \tag{29}$$

where squaring is applied element-wise. Writing $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$ in block form:

$$\bar{\mathbf{A}} = \begin{pmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(2)} \\ \mathbf{A}^{(3)} & \mathbf{A}^{(4)} \end{pmatrix}, \qquad \bar{\mathbf{b}} = \begin{pmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \end{pmatrix} \tag{30}$$

Then for each dimension i:

$$c_i = \sum_j A_{ij}^{(1)} \hat{c}_j + \sum_j A_{ij}^{(2)} \hat{c}_j^2 + b_i^{(1)} \tag{31}$$

$$c_i^2 = \sum_j A_{ij}^{(3)} \hat{c}_j + \sum_j A_{ij}^{(4)} \hat{c}_j^2 + b_i^{(2)} \tag{32}$$

Squaring the $\mathbf{A}^{(2)}$ term in Equation (31):

$$\left(\sum_j A_{ij}^{(2)} \hat{c}_j^2 \right)^2 = \sum_j \left(A_{ij}^{(2)} \right)^2 \hat{c}_j^4 + \sum_{j \neq j'} A_{ij}^{(2)} A_{ij'}^{(2)} \hat{c}_j^2 \hat{c}_{j'}^2, \tag{33}$$

The $\hat{c}_j^4$ terms match no term in Equation (32), so we must set $A_{ij}^{(2)} = 0$ for all i, j. Equation (31) simplifies to $c_i = \sum_j A_{ij}^{(1)} \hat{c}_j + b_i^{(1)}$. Squaring this:

$$\left(\sum_j A_{ij}^{(1)} \hat{c}_j \right)^2 = \sum_j \left(A_{ij}^{(1)} \right)^2 \hat{c}_j^2 + \sum_{j \neq j'} A_{ij}^{(1)} A_{ij'}^{(1)} \hat{c}_j \hat{c}_{j'} \tag{34}$$

The cross terms $\hat{c}_j \hat{c}_{j'}$ ($j \neq j'$) match no term in Equation (32), so we must set $A_{ij}^{(1)} A_{ij'}^{(1)} = 0$ for all $j \neq j'$. Hence each row of $\mathbf{A}^{(1)}$ has at most one nonzero entry. Since $\bar{\mathbf{A}}$ has full rank and $\mathbf{A}^{(2)} = 0$, each row must have exactly one nonzero entry, with no two rows sharing the same nonzero column. By Lemma 1, this is consistent with the strongly exponential property of the Gaussian distribution. Therefore $\mathbf{A}^{(1)} = A \cdot P$ where $A = \operatorname{diag}(a_1, \ldots, a_{d_c})$ with $a_i \neq 0$ and P is a permutation matrix, giving:

$$\mathbf{c} = A \cdot P(\hat{\mathbf{c}}) + \mathbf{b} \tag{35}$$

□

C.2 Identifiability of Latent Dynamics States $\mathbf{z}_t$

Again, we define the latent dynamics of the latent state $\mathbf{z}_t$ and its emission to observations $\mathbf{y}_t$ following Equation (1).

Theorem 2. *Suppose there exists an invertible function $\hat{g}_\eta^{-1}$ such that $\hat{\mathbf{z}}_t = \hat{g}_\eta^{-1}(\mathbf{y}_t)$ and the conditional distribution $p(z_{kt}|\mathbf{z}_{t-1}, \mathcal{H}_T)$ changes across M values of the history $\mathcal{H}_T$, denoted by $\mathcal{H}_T^1, \mathcal{H}_T^2, \ldots, \mathcal{H}_T^M$. Assume the components of $\hat{\mathbf{z}}_t$ are also mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$ and $\mathcal{H}_T$. Let $\eta_{kt}(\mathcal{H}_T) = \log p(z_{kt}|\mathbf{z}_{t-1}, \mathcal{H}_T)$ and assume $\eta_{kt}(\mathcal{H}_T)$ is twice differentiable in z_{kt} and differentiable in $z_{l,t-1}$, $l = 1, 2, \ldots, d_z$. Also let*

$$\mathbf{s}_{kt} = \left(\frac{\partial^2 \eta_{kt}(\mathcal{H}_T^2)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(\mathcal{H}_T^1)}{\partial z_{kt}^2}, \ \ldots, \ \frac{\partial^2 \eta_{kt}(\mathcal{H}_T^M)}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(\mathcal{H}_T^{M-1})}{\partial z_{kt}^2} \right)^\top \tag{36}$$

$$\mathring{\mathbf{s}}_{kt} = \left(\frac{\partial \eta_{kt}(\mathcal{H}_T^2)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(\mathcal{H}_T^1)}{\partial z_{kt}}, \ \ldots, \ \frac{\partial \eta_{kt}(\mathcal{H}_T^M)}{\partial z_{kt}} - \frac{\partial \eta_{kt}(\mathcal{H}_T^{M-1})}{\partial z_{kt}} \right)^\top \tag{37}$$

If for each value of $\mathbf{z}_t$, the $2d_z$ function vectors $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, with $k = 1, 2, \ldots, d_z$, are linearly independent, then $\hat{\mathbf{z}}_t$ is a permuted invertible component-wise transformation of $\mathbf{z}_t$:

$$\hat{g}_\eta^{-1}(\mathbf{y}_t) = C \circ P \circ g_\eta^{-1}(\mathbf{y}_t) \ \Leftrightarrow \ \hat{\mathbf{z}}_t = C \circ P(\mathbf{z}_t) \tag{38}$$

where C is a component-wise invertible transformation and P is a permutation.

Proof. Using g_η and $\hat{g}_\eta$, we can relate the ground-truth and estimated latent states $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$:

$$\mathbf{z}_t = g_\eta^{-1}(\mathbf{y}_t) = g_\eta^{-1} \circ \hat{g}_\eta(\hat{\mathbf{z}}_t) = \mathbf{h}(\hat{\mathbf{z}}_t) \tag{39}$$

where $\mathbf{h} = g_\eta^{-1} \circ \hat{g}_\eta$. Using $\mathbf{z}_t = \mathbf{h}(\hat{\mathbf{z}}_t)$, we perform change of variables on the conditional latent density $\log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathcal{H}_T)$:

$$\log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathcal{H}_T) = \log p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathcal{H}_T) + \log |\mathbf{H}_t| \tag{40}$$

where $\mathbf{H}_t = \mathbf{J}_\mathbf{h}(\hat{\mathbf{z}}_t)$ is the Jacobian matrix of $\mathbf{h}$ evaluated at $\hat{\mathbf{z}}_t$. Since the components of $\mathbf{z}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_{t-1}$ and $\mathcal{H}_T$, for any $i \neq j$, $\hat{z}_{it}$ and $\hat{z}_{jt}$ are conditionally independent given $\hat{\mathbf{z}}_{t-1} \cup \{\hat{\mathbf{z}}_t \setminus \{\hat{z}_{it}, \hat{z}_{jt}\}\}$, such that:

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathcal{H}_T)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0 \tag{41}$$

Let $\log p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathcal{H}_T) = \sum_{k=1}^{d_z} \eta_{kt}(\mathcal{H}_T)$, $\eta_{kt}(\mathcal{H}_T) = \log p(z_{kt}|\mathbf{z}_{t-1}, \mathcal{H}_T)$, then

$$\frac{\partial^2 \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathcal{H}_T)}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = \sum_{k=1}^{d_z} \left(\frac{\partial^2 \eta_{kt}(\mathcal{H}_T)}{\partial z_{kt}^2} \mathbf{H}_{kit} \mathbf{H}_{kjt} + \frac{\partial \eta_{kt}(\mathcal{H}_T)}{\partial z_{kt}} \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \right) - \frac{\partial^2 \log |\mathbf{H}_t|}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} = 0 \tag{42}$$

Taking the difference of this equation across different values of $\mathcal{H}_T$ gives

$$\begin{aligned} &\frac{\partial^2 \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathcal{H}_T^{r+1})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} - \frac{\partial^2 \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathcal{H}_T^{r})}{\partial \hat{z}_{it} \partial \hat{z}_{jt}} \\ &= \sum_{k=1}^{d_z} \left[\left(\frac{\partial^2 \eta_{kt}(\mathcal{H}_T^{r+1})}{\partial z_{kt}^2} - \frac{\partial^2 \eta_{kt}(\mathcal{H}_T^{r})}{\partial z_{kt}^2} \right) \mathbf{H}_{kit} \mathbf{H}_{kjt} + \left(\frac{\partial \eta_{kt}(\mathcal{H}_T^{r+1})}{\partial z_{kt}} - \frac{\partial \eta_{kt}(\mathcal{H}_T^{r})}{\partial z_{kt}} \right) \frac{\partial \mathbf{H}_{kit}}{\partial \hat{z}_{jt}} \right] = 0 \end{aligned} \tag{43}$$

Therefore, if $\mathbf{s}_{kt}$ and $\mathring{\mathbf{s}}_{kt}$, for $k = 1, 2, \ldots, d_z$, are linearly independent, $\mathbf{H}_{kit} \mathbf{H}_{kjt}$ must be zero for all k and $i \neq j$. That is, in each row of $\mathbf{H}$, there is only one non-zero entry. Since $\mathbf{h}$ is invertible, $\hat{\mathbf{z}}_t$ must be a permuted invertible component-wise transformation of $\mathbf{z}_t$. □

D CAUSAL IDENTIFIABILITY ASSUMPTIONS

We refer to section A of Figure 1 for the DGP we assume for this section. Building on the potential outcome framework (Rubin, 2005; Robins et al., 1999), the following assumptions are necessary to properly model the objective for $\boldsymbol{E}[p(\mathbf{y}_{t+1:t+\tau}|\mathbf{a}_{t+1:t+\tau})]$.

1. **Assumption 1: Consistency** If the treatment assignment $\bar{\mathbf{a}}_{t:t+\tau} = \bar{\mathbf{a}}_{t:t+\tau}$, then the observed outcome is equal to the potential outcome under this assignment, that is, $\mathbf{y}_{t:t+\tau}[\bar{\boldsymbol{a}}_{t:t+\tau}] = \mathbf{y}_{t:t+\tau}$.
2. **Assumption 2: Sequential Overlap** At each timestep, for all histories $\mathcal{H}_t$, there is a non-zero probability of receiving all potential treatment assignments: $\forall\, \mathbf{a}_t\ Pr(\mathbf{a}_t = \mathbf{a}_t|\mathcal{H}_t) > 0$.
3. **Assumption 3: Ignorability** When conditioned on the observed history $\mathcal{H}_t$, the space of potential outcomes is independent of assigned treatment $\forall \mathbf{a},\ \mathbf{a}_t \perp \boldsymbol{Y_{t+1}}|\mathcal{H}_t$.

E Additional Experimental Details

E.1 Details on Pendulum Data

Following (Ye et al., 2026; Takeishi & Kalousis, 2021), we generate sequences of images of a pendulum whose dynamics is defined by a second-order differential equation (Equation (44)). The parameters of the equation are the natural frequency ω, the damping coefficient α and the force A. The $\text{Bin}(p(t))$ is a random variable that decides whether the force is applied. In effect, the force acts as a pulse with the pulse arrival rate decided by the random treatment assignment $\text{Bin}(p(t))$.

$$\frac{d^2\varphi}{dt^2} + \omega^2 \sin\varphi + 2\alpha\frac{d\varphi}{dt} = A \times \text{Bin}(p(t)) \tag{44}$$

We generate 4000 sequences of 60 steps with $\varphi(t=0) \in [-1.0, 1.0]$, $\dot{\varphi}(t=0) \in [-0.5, 0.5]$, $\omega \in [0.4, 0.7]$ and $A \in [0.033, 0.1]$, with ω and A paired so that they are uncorrelated. The damping is held fixed at $\alpha = 0.03845$, so the native dynamics of a sequence is described by ω alone and $\mathbf{c}^0$ is one-dimensional here.

Both settings use the same assignment policy, so that they differ in confounding and in nothing else:

$$p(t) = \sigma(\gamma\,(\bar{\varphi}_t - \varphi_{\text{ref}}) + b)\,, \qquad \bar{\varphi}_t = \frac{1}{W}\sum_{s=t-W+1}^{t} |\varphi_s| \tag{45}$$

where σ is the standard sigmoid or expit function, $W = 5$ and φ_{ref} is the median of $|\varphi|$ over a reference rollout. For the unconfounded experiments $\gamma = 0$, which makes $p(t)$ constant and the assignment independent of the position of the pendulum; for the confounding experiments $\gamma = -20$, so that a pendulum nearer the bottom of its swing is more likely to be forced. We take γ negative because the opposite sign is positive feedback — a pulse raises $|\varphi|$, which raises the probability of the next pulse — and drives trajectories out of the oscillating regime. In both settings the intercept b is calibrated so that the marginal pulse rate is 0.2: the two therefore deliver the same dose (11.20 pulses per sequence on average) and differ only in the dependence of assignment on the state, $|\text{corr}(a_t, |\varphi_t|)| = 0.002$ against 0.327.

E.2 Details on PK-PD Data of Tumor Growth

The dynamic model for the growth of non-small cell lung cancer under radiotherapy and chemotherapy interventions is given in (Geng et al., 2017). This model assumes that the intrinsic growth follows a Gompertz growth model, the effect of radiotherapy follows a linear-quadratic (LQ) model and the chemotherapy effect follows a log-cell-kill model. This results in the tumor volume V evolving as:

$$\frac{dV}{dt} = V * (\rho \log(\frac{K}{V}) - \beta_c C(t) - (\alpha R(t) + \beta R(t)^2)) \tag{46}$$

For the Gompertz growth model, ρ represents the cell-specific growth rate and K is the carrying capacity. A constant value of $K = 14137.167$ is used, whereas ρ is varied based on the distribution suggested in (Geng et al., 2017) with changes for different heterogeneity settings as mentioned in Table 4.

For the chemotherapy part of the model, the value of $C(t)$ represents the drug concentration and β_c controls the effect of chemotherapy. Once administered, the drug is assumed to decay with a half-life of 1 timestep. Based on the clinical practice of administering vinblastine at $5\ mg/m^2$ per week, the value of $C(t) = 5$ is introduced at any time the treatment assignment dictates chemotherapy administration.

For radiotherapy, the parameters $\alpha, \beta : \alpha/\beta = 10$ are made to affect the cell volume at only the time of administration, and the effects disappear immediately. $R(t) = 2\ Gy$ at the time of administration is kept constant to simulate practice.

Treatment assignment bias: To isolate the benefits of the dynamics modeling strategies introduced by iASIDE, we first considered random treatment assignment with a probability of $p_0 = 0.2$ regardless

of tumor volume: the relatively low assignment probability was chosen to avoid generating numerous time series where the tumor was killed within a small time window.

To include treatment assignment bias, we then adopted the mechanism for treatment assignment in (Melnychuk et al., 2022): $p(a_t) = p_0\ \sigma(\gamma(\sqrt[3]{V_t/V_{max}} - 1/2)$ with $p_0 = 0.2$, considering increasing levels of treatment bias as controlled by $\gamma = 1, 2, 3, 4$. For a fair comparison, for all baselines that use treatment-invariant representations to address confounding, we disabled their adversarial training element when there was no confounding. We tested iASIDE with and without the addition of treatment-invariant representations when considering confounding.

Heterogeneity levels: To examine the importance of *adaptable* and *identifiable* latent dynamics, we created training data with four different levels of heterogeneity in the key parameters (ρ, β_c, α) by controlling their ranges as summarized in Table 4.

Table 4: Parameter ranges for different levels of data heterogeneity. lo indicates lower bound and hi indicates upper bound.

Heterogeneity Level	Growth (ρ)		Radio (α)		Chemo (β_c)	
	low	high	low	high	low	high
0	7e-5	7e-5	.039	.039	.028	.028
1	0.0	7e-3	0.0	.21	.027	.029
2	0.0	7e-3	0.0	.51	.026	.030
3	0.0	21e-3	0.0	.51	.026	.030

Distribution of tumor volume across heterogeneity and projection horizons: The distribution of volume across different heterogeneity levels and different projection horizons varies greatly, as shown in Figure 6. Since the ground-truth volumes differ, the RMSE calculated for these varying settings will also be different and not easily comparable. As clearly seen, volume distribution is higher for heterogeneity level 1, so the RMSE is expected to be higher for this setting across all τ.

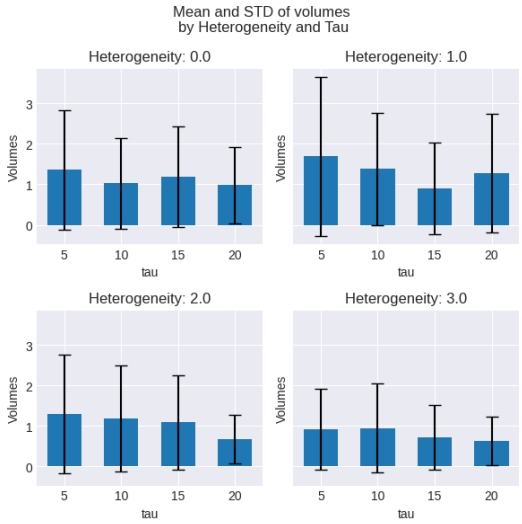

Figure 6: Volume distribution across τ for different heterogeneity levels

E.3 IMPLEMENTATION DETAILS

Table 5: Model components and specifications

Component	**Layer**	**In features**	**Out features**	**Param Counts**
Encoder	Linear + ELU Linear + ELU Linear	$3 \cdot F$ 8 8	8 8 $d(\mathbf{z}_{init}) = 16$	26K
Meta-encoder (native)	Linear + ELU (Linear + ELU)*8 Linear	d_{in} $H = 16$ $H = 16$	$H = 16$ $H = 16$ $d(\mathbf{c}_{int}) = 16$	18K
Meta-encoder (intervention)	Conv1D + ELU (Conv1D + ELU)*8 Linear	$d_{in} + d_{mask}$ $H = 16$ $H = 16$	$H = 16$ $H = 16$ $d(c_{ext}) = 16$	18K
Native Dynamics	Linear + ELU (Linear + ELU)*8 Linear	$c_{int} + z_{init} = 32$ $H = 16$ $H = 16$	$H = 16$ $H = 16$ $d(\mathbf{z}_{init}) = 16$	25K
Extrinsic Dynamics	Linear + ELU (Linear + ELU)*8 Linear	$c_{ext} + z_{init} + d_{int}$ $H = 16$ H	$H = 16$ $H = 16$ $d(\mathbf{z}_{init}) = 16$	25K
Decoder	Linear + ELU (Linear + ELU) Linear	$d(\mathbf{z}_{init}) = 16$ $H = 16$ H	$H = 16$ $H = 16$ $d(\mathbf{y})$	26K

Model	Submodule	Parameter Counts
RMSN	Propensity History	13K
	Propensity Treatment	144K
	Encoder	178K
	Decoder	170K
CRN	Encoder	155K
	Decoder	115K
CT	Transformer	132K
iASIDE	Encoder	26K
	Decoder	26K
	Meta-Encoder	36K
	Dynamics	50K

E.4 DESCRIPTION OF BASELINES

Marginal Structural Model (MSM) MSM is a classic model based on (Hernan et al., 2000) and extended further in (Hernan & Robins, 2020). The idea here is to learn a structured model (here linear) for first forecasting treatment from covariates: $f_{p_t}(\boldsymbol{a_t}|\boldsymbol{a_{1:t-1}})$ which is used to calculate inverse probability of treatment weights (IPTW). These IPTW are used to rescale the inputs to learn another structured (again linear) model: $y_t = f_o(\boldsymbol{x_{t:t-l}}, \boldsymbol{a_{t:t-l}})$

Recurrent Marginal Structural Network (RMSN): Based on work by (Lim, 2018), RMSN is built completely on RNNs. It is a simple extension to linear marginal structural models described in (Hernan & Robins, 2020) using RNN. First, the inverse probability weights are learned. One RNN models: $f_{p_t}(\boldsymbol{a_t}|\boldsymbol{a_{1:t-1}})$ and another RNN models $f_{p_h}(\boldsymbol{a_t}|\boldsymbol{a_{1:t-1}}, \mathcal{H}_{1:t-1})$. Then, the weight is calculated to be: $w_i = \frac{f_{p_t}(x_i)}{f_{p_h}(x_i)}$. After learning the weights, an encoder is trained with the task of learning representations from data as: $e_{i,t} = f_e(w_i, \mathcal{H}_{1:t-1})$. The representation e_i learned above is then used to make autoregressive predictions for $y_{t+1:t+\tau}$ and is modeled by another RNN: $y_{i,t+1:t+\tau} = f_d(e_{i,t}, w_i, y_{i,t})$

Counterfactual Recurrent Network (CRN): CRN by (Bica et al., 2020) is also an encoder-decoder architecture. However, unlike RMSN, in CRN, the bias handling is done by an adversarial loss component instead of learned weights, which reduces the need for separately modeling the weights by RNNs. The idea behind CRN is to learn balanced representations $r_i(\mathcal{H}^i_{1:t-1})$ which are invariant to treatment assignment. This is done by building two separate heads to the RNN models, one to predict the treatment $\boldsymbol{a_{t+1}} = G_A(r_i(\mathcal{H}^i_{1:t-1}))$ and another to predict outcome $Y_{t+1} = G_Y(r_i(\mathcal{H}^i_{1:t-1}), \boldsymbol{a_t})$ at each time point. An adversarial loss is used to learn this model.

In the open-source implementation provided by Bica et al. (2020), this loss is implemented using a gradient reversal layer.

Causal Transformer (CT): CT (Melnychuk et al., 2022) extends the idea of CRN using a transformer to make the pipeline end-to-end and replace the two-stage learning process in previous models. They use multi-headed self- and cross-attention mechanisms common in transformer-based approaches to model the sequences, which then learns to attend to different parts of the history to make future predictions. As in CRN, they learn the transformer by an adversarial component on top of the transformer. The transformer is tasked to learn a balanced representation $r_i(\mathcal{H}^i_{1:t-1})$ such that it is invariant to treatment predictions. Two heads are added on top of the transformer: one to predict the treatment $\boldsymbol{a_{t+1}} = G_A(r_i(\mathcal{H}^i_{1:t-1}))$ and another to predict outcome $Y_{t+1} = G_Y(r_i(\mathcal{H}^i_{1:t-1}), \boldsymbol{a_t})$ at each time point. An adversarial loss known as Counterfactual Domain Confusion (CDC) loss is used to learn the balanced representation.

TESAR-CDE: (Vanderschueren et al., 2023) is similar in concept to CRN/CT, apart from the base architecture, which is a Controlled Differential Equation (CDE) following (Kidger et al., 2020) instead of a Recurrent Neural Network or a Transformer. The major difference between our work and the use of controlled differential equations is that CDEs need to be "controlled" by an observation pathway which in turn makes the model inherently good for reconstruction tasks but not so much for forecasting tasks. This makes the CDE models equivalent to continuous latent space extensions to recurrent networks such as RNNs and LSTMs. This continuous nature helps in solving the problem of informative treatment sampling presented in (Seedat et al., 2022; Vanderschueren et al., 2023), but becomes much less relevant when data is regularly sampled. Moreover, the inherent need to have a control pathway makes these models not very reliable for forecasting tasks. These models are trained end to end encoder-decoder architecture.

SCIP: (Hess & Feuerriegel, 2025) takes inspiration from TE-CDE for using CDEs, but unlike the two-stage architecture and domain-invariance for bias elimination, they do a single-stage end-to-end training of the encoder-decoder. For bias elimination, they have a second CDE which is trained to predict treatment assignment (similar to RMSN and MSM) and use it to calculate IPTW weights for weighting the dataset.

F Additional Results

F.1 Complete Results on Pendulum

F.1.1 Complete quantitative results for unconfounded settings

Table 6 includes complete results on the Pendulum dataset in the unconfounded setting. In addition to MSE and DST reported in the main text, we further include the metric of valid prediction time (VPT) that measures how long the predicted object's trajectory remains close to the ground truth trajectory based on the MSE (VPT-MSE) (Botev et al., 2021) or DST (VPT-DST).

Table 6: Complete results on the unconfounded Pendulum data across projection horizons τ: mean, standard deviation (subscript) over 3 seeds. Numbers in bold indicate that iiASIDE's improvement over the best baseline is statistically significant (Welch's t-test, $\alpha = 0.05$). Baselines' MCC is the same statistic on their representation at the last observed step. †: not above the 95th percentile of its permutation null, i.e. at chance.

Metric	Model	5	10	15	20
RMSE ↓	ACTIN	$0.064_{0.001}$	$0.080_{0.001}$	$0.087_{0.003}$	$0.093_{0.004}$
	CRN	$0.107_{0.000}$	$0.136_{0.006}$	$0.157_{0.010}$	$0.156_{0.009}$
	CT	$0.058_{0.001}$	$0.071_{0.002}$	$0.076_{0.002}$	$0.081_{0.002}$
	RMSN	$0.080_{0.003}$	$0.108_{0.005}$	$0.124_{0.006}$	$0.125_{0.002}$
	iASIDE	$\mathbf{0.027_{0.000}}$	$\mathbf{0.053_{0.000}}$	$\mathbf{0.062_{0.000}}$	$\mathbf{0.066_{0.000}}$
DST ↓	ACTIN	$0.704_{0.028}$	$1.051_{0.038}$	$1.250_{0.132}$	$1.394_{0.188}$
	CRN	$1.753_{0.022}$	$2.958_{0.290}$	$4.209_{0.597}$	$4.159_{0.501}$
	CT	$0.608_{0.025}$	$0.843_{0.038}$	$0.956_{0.042}$	$1.065_{0.029}$
	RMSN	$1.067_{0.044}$	$1.734_{0.140}$	$2.267_{0.198}$	$2.218_{0.062}$
	iASIDE	$\mathbf{0.281_{0.002}}$	$\mathbf{0.610_{0.001}}$	$\mathbf{0.761_{0.002}}$	$\mathbf{0.836_{0.010}}$
VPT-MSE ↑	ACTIN	$0.915_{0.007}$	$0.872_{0.013}$	$0.896_{0.031}$	$0.887_{0.030}$
	CRN	$0.625_{0.013}$	$0.518_{0.096}$	$0.402_{0.091}$	$0.473_{0.093}$
	CT	$0.936_{0.003}$	$0.908_{0.007}$	$0.931_{0.007}$	$0.925_{0.002}$
	RMSN	$0.789_{0.018}$	$0.664_{0.020}$	$0.637_{0.052}$	$0.740_{0.030}$
	iASIDE	$\mathbf{0.999_{0.000}}$	$\mathbf{0.965_{0.001}}$	$\mathbf{0.956_{0.001}}$	$\mathbf{0.964_{0.001}}$
VPT-DST ↑	ACTIN	$1.000_{0.000}$	$0.998_{0.001}$	$0.995_{0.004}$	$0.991_{0.011}$
	CRN	$0.998_{0.000}$	$0.971_{0.006}$	$0.923_{0.020}$	$0.910_{0.014}$
	CT	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.001}$	$0.999_{0.001}$
	RMSN	$1.000_{0.000}$	$1.000_{0.000}$	$0.998_{0.002}$	$0.997_{0.004}$
	iASIDE	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$
MCC($\mathbf{z}_t$) ↑	ACTIN	$0.663_{0.013}$	$0.681_{0.024}$	$0.693_{0.029}$	$0.702_{0.027}$
	CRN	$0.545_{0.026}$	$0.554_{0.017}$	$0.565_{0.014}$	$0.573_{0.014}$
	CT	$0.535_{0.037}$	$0.540_{0.036}$	$0.552_{0.031}$	$0.563_{0.029}$
	RMSN	$0.678_{0.021}$	$0.682_{0.021}$	$0.689_{0.015}$	$0.699_{0.017}$
	iASIDE	$\mathbf{0.961_{0.000}}$	$\mathbf{0.933_{0.001}}$	$\mathbf{0.905_{0.003}}$	$\mathbf{0.899_{0.004}}$
MCC($\mathbf{c}^0$) ↑	ACTIN	$0.099_{0.003}$	$0.086_{0.003}$	$0.102_{0.007}$	$0.108_{0.012}$
	CRN	$0.084_{0.006}$	$0.062_{0.003}$	$0.095_{0.003}$	$0.104_{0.014}$
	CT	$0.087_{0.005}$	$0.068_{0.007}$	$0.101_{0.011}$	$0.117_{0.021}$
	RMSN	$0.092_{0.003}$	$0.074_{0.006}$	$0.099_{0.006}$	$0.121_{0.003}$
	iASIDE	$\mathbf{0.924_{0.003}}$	$\mathbf{0.896_{0.003}}$	$\mathbf{0.895_{0.036}}$	$\mathbf{0.694_{0.035}}$
MCC($\mathbf{c}^1$) ↑	ACTIN	$0.047_{0.002}$	$0.033^{\dagger}_{0.004}$	$0.036^{\dagger}_{0.003}$	$0.057^{\dagger}_{0.005}$
	CRN	$0.048_{0.002}$	$0.032^{\dagger}_{0.004}$	$0.036^{\dagger}_{0.002}$	$0.053^{\dagger}_{0.002}$
	CT	$0.042_{0.001}$	$0.027^{\dagger}_{0.001}$	$0.032^{\dagger}_{0.004}$	$0.051^{\dagger}_{0.004}$
	RMSN	$0.050_{0.002}$	$0.034^{\dagger}_{0.003}$	$0.045^{\dagger}_{0.001}$	$0.066^{\dagger}_{0.004}$
	iASIDE	$\mathbf{0.083_{0.002}}$	$0.032^{\dagger}_{0.007}$	$0.010^{\dagger}_{0.008}$	$0.029^{\dagger}_{0.001}$

F.1.2 COMPLETE QUANTITATIVE RESULTS FOR CONFOUNDED SETTINGS

Table 7 includes complete results on the Pendulum dataset under confounding. In addition to MSE and DST reported in the main text, we further include the metric of valid prediction time (VPT) that measures how long the predicted object's trajectory remains close to the ground truth trajectory based on the MSE (VPT-MSE) (Botev et al., 2021) or DST (VPT-DST).

Table 7: Complete results on the confounded Pendulum data across projection horizons τ: mean, standard deviation (subscript) over 3 seeds. Numbers in bold indicate that iiASIDE's improvement over the best baseline is statistically significant (Welch's t-test, $\alpha = 0.05$). Baselines' MCC is the same statistic on their representation at the last observed step. †: not above the 95th percentile of its permutation null, i.e. at chance.

Metric	Model	5	10	15	20
RMSE ↓	RMSN	$0.114_{0.001}$	$0.121_{0.006}$	$0.126_{0.008}$	$0.123_{0.010}$
	CRN	$0.108_{0.008}$	$0.141_{0.008}$	$0.149_{0.013}$	$0.157_{0.008}$
	CT	$0.057_{0.000}$	$0.070_{0.001}$	$0.075_{0.001}$	$0.081_{0.001}$
	ACTIN	$0.065_{0.003}$	$0.086_{0.007}$	$0.098_{0.009}$	$0.105_{0.010}$
	iASIDE	$\mathbf{0.026_{0.000}}$	$\mathbf{0.055_{0.000}}$	$\mathbf{0.063_{0.001}}$	$\mathbf{0.071_{0.000}}$
DST ↓	RMSN	$2.168_{0.009}$	$2.377_{0.180}$	$2.471_{0.178}$	$2.693_{0.436}$
	CRN	$1.798_{0.216}$	$3.215_{0.324}$	$3.709_{0.723}$	$4.039_{0.586}$
	CT	$0.603_{0.008}$	$0.829_{0.026}$	$0.936_{0.032}$	$1.056_{0.037}$
	ACTIN	$0.694_{0.051}$	$1.086_{0.157}$	$1.398_{0.305}$	$1.587_{0.387}$
	iASIDE	$\mathbf{0.282_{0.001}}$	$\mathbf{0.637_{0.001}}$	$\mathbf{0.779_{0.012}}$	$\mathbf{0.945_{0.009}}$
VPT-MSE ↑	RMSN	$0.566_{0.005}$	$0.688_{0.061}$	$0.779_{0.039}$	$0.833_{0.081}$
	CRN	$0.637_{0.053}$	$0.502_{0.063}$	$0.553_{0.169}$	$0.604_{0.150}$
	CT	$0.940_{0.003}$	$0.913_{0.004}$	$0.929_{0.005}$	$0.920_{0.005}$
	ACTIN	$0.911_{0.015}$	$0.854_{0.034}$	$0.860_{0.046}$	$0.861_{0.047}$
	iASIDE	$\mathbf{0.999_{0.000}}$	$\mathbf{0.956_{0.002}}$	$\mathbf{0.963_{0.000}}$	$\mathbf{0.947_{0.007}}$
VPT-DST ↑	RMSN	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$
	CRN	$0.998_{0.001}$	$0.966_{0.007}$	$0.952_{0.015}$	$0.921_{0.026}$
	CT	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$
	ACTIN	$1.000_{0.000}$	$0.998_{0.002}$	$0.991_{0.008}$	$0.984_{0.015}$
	iASIDE	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$	$1.000_{0.000}$
MCC($\mathbf{z}_t$) ↑	RMSN	$0.657_{0.015}$	$0.662_{0.014}$	$0.663_{0.015}$	$0.668_{0.015}$
	CRN	$0.556_{0.013}$	$0.562_{0.011}$	$0.571_{0.014}$	$0.576_{0.008}$
	CT	$0.528_{0.015}$	$0.540_{0.016}$	$0.557_{0.012}$	$0.563_{0.008}$
	ACTIN	$0.631_{0.081}$	$0.644_{0.072}$	$0.652_{0.067}$	$0.661_{0.058}$
	iASIDE	$\mathbf{0.960_{0.000}}$	$\mathbf{0.915_{0.001}}$	$\mathbf{0.918_{0.005}}$	$\mathbf{0.921_{0.001}}$
MCC($\mathbf{c}^0$) ↑	RMSN	$0.095_{0.008}$	$0.079_{0.003}$	$0.096_{0.004}$	$0.098_{0.006}$
	CRN	$0.089_{0.002}$	$0.068_{0.001}$	$0.085_{0.004}$	$0.084_{0.008}$
	CT	$0.093_{0.009}$	$0.070_{0.009}$	$0.099_{0.012}$	$0.102_{0.016}$
	ACTIN	$0.095_{0.007}$	$0.078_{0.013}$	$0.083_{0.003}$	$0.084_{0.013}$
	iASIDE	$\mathbf{0.943_{0.001}}$	$\mathbf{0.962_{0.003}}$	$\mathbf{0.953_{0.001}}$	$\mathbf{0.954_{0.004}}$
MCC($\mathbf{c}^1$) ↑	RMSN	$0.067_{0.000}$	$0.082_{0.008}$	$0.068_{0.008}$	$0.080_{0.004}$
	CRN	$0.062_{0.006}$	$0.072_{0.008}$	$0.064_{0.011}$	$0.088_{0.006}$
	CT	$0.065_{0.002}$	$0.084_{0.006}$	$0.069_{0.004}$	$0.087_{0.001}$
	ACTIN	$0.057_{0.007}$	$0.073_{0.006}$	$0.061_{0.010}$	$0.079_{0.008}$
	iASIDE	$0.057_{0.004}$	$0.092_{0.006}$	$0.048_{0.012}$	$0.041^{\dagger}_{0.014}$

F.1.3 VISUAL EXAMPLES

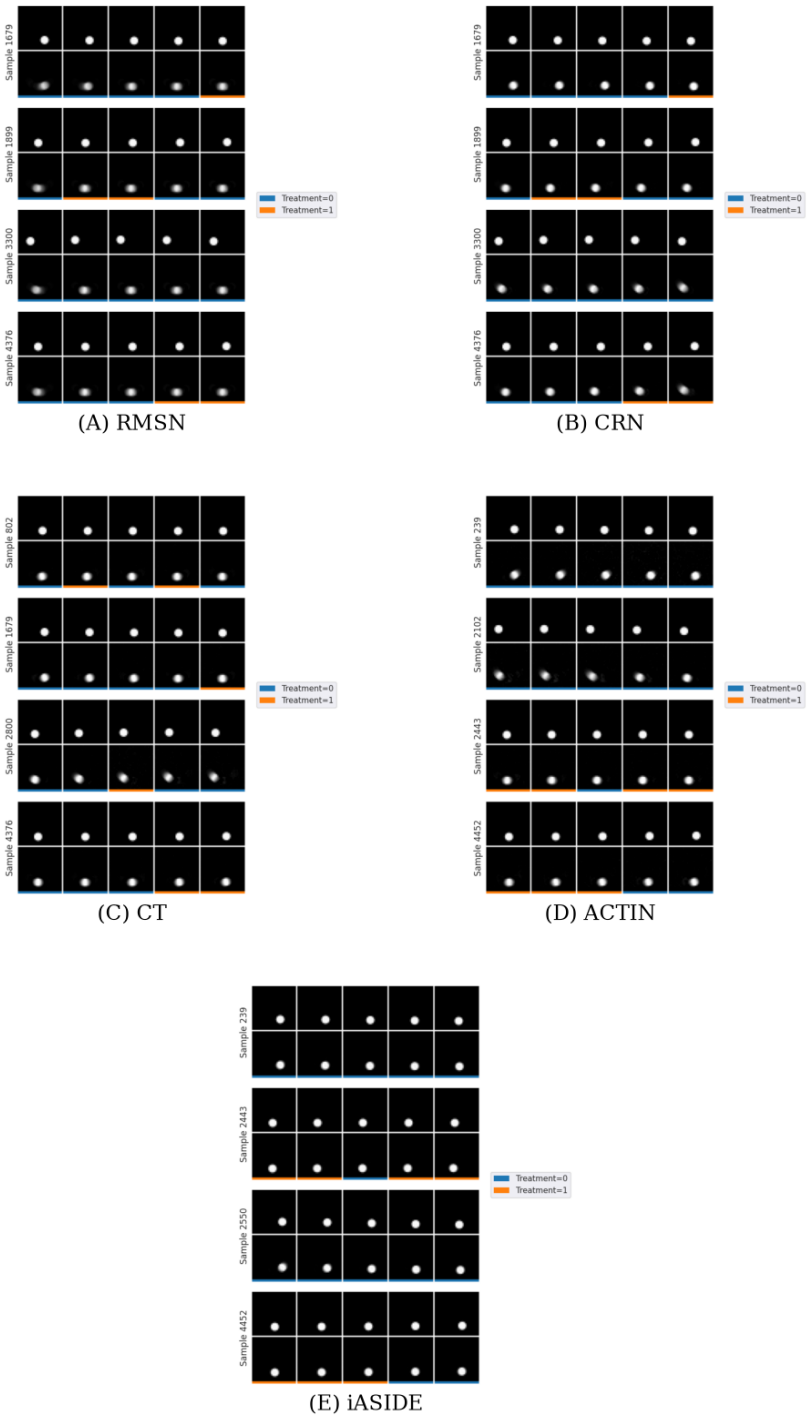

Figure 7: Visual examples of ground truth and predicted trajectories for different models for confounded pendulum data. For each sample, the top row represents the ground truth and the bottom row represents the predictions.

F.2 Complete Results on Synthetic Tumor Growth Data

There are d_a^τ possible potential outcomes of $\{\mathbf{y}\}_{t+1}^{t+\tau}$, which grows exponentially with the projection horizon. For a more feasible testing scenario we follow (Melnychuk et al., 2022) and adopt two alternative schemes:

1. *Random assignment*: Simulate τ different trajectories where the treatment assignment is completely random and thus unconfounded. This is the setting presented in the main paper in Figure 2. Comprehensive results are presented in Section F.2.1.
2. *Sliding treatment*: Simulate counterfactual $\{\mathbf{y}\}_{t+1}^{t+\tau}$ with a single treatment but where the treatments are iteratively moved over a window, resulting in τ different trajectories for each intervention, and therefore a total of $d_a \times \tau$ trajectories. Comprehensive results for this setting are presented in Section F.2.2.

F.2.1 Complete Results for random treatment assignment

		τ			
Heterogeneity	Model	5	10	15	20
0	ACTIN	$\mathit{0.354}_{0.251}$	$\mathit{0.309}_{0.143}$	$\mathit{0.304}_{0.147}$	$\mathit{0.305}_{0.170}$
	CRN	$0.694_{0.042}$	$0.722_{0.053}$	$0.848_{0.116}$	$0.845_{0.156}$
	CT	$0.366_{0.032}$	$0.340_{0.043}$	$0.322_{0.044}$	$0.401_{0.102}$
	RMSN	$0.605_{0.185}$	$0.498_{0.043}$	$0.717_{0.178}$	$0.850_{0.221}$
	SCIP	$8.702_{8.162}$	$4.942_{0.954}$	$10.903_{4.286}$	$8.209_{1.925}$
	TESAR-CDE	$1.259_{0.080}$	$1.551_{0.184}$	$1.067_{0.059}$	$1.018_{0.170}$
	iASIDE	$\mathbf{0.176}_{0.001}$	$\mathbf{0.164}^{*}_{0.004}$	$\mathbf{0.162}^{*}_{0.001}$	$\mathbf{0.142}^{*}_{0.003}$
1	ACTIN	$\mathit{0.845}_{0.066}$	$\mathit{1.148}_{0.111}$	$\mathit{1.186}_{0.165}$	$\mathit{1.143}_{0.181}$
	CRN	$1.052_{0.049}$	$1.301_{0.038}$	$1.498_{0.226}$	$1.421_{0.129}$
	CT	$0.988_{0.086}$	$1.232_{0.098}$	$1.281_{0.066}$	$1.404_{0.304}$
	RMSN	$1.186_{0.246}$	$1.264_{0.090}$	$1.437_{0.109}$	$1.319_{0.189}$
	SCIP	$4.955_{1.903}$	$6.344_{2.299}$	$8.183_{3.485}$	$13.963_{5.500}$
	TESAR-CDE	$1.737_{0.144}$	$1.676_{0.199}$	$1.554_{0.016}$	$1.496_{0.033}$
	iASIDE	$\mathbf{0.634}^{*}_{0.009}$	$\mathbf{0.719}^{*}_{0.014}$	$\mathbf{0.641}^{*}_{0.010}$	$\mathbf{0.554}^{*}_{0.008}$
2	ACTIN	$\mathit{1.365}_{0.094}$	$\mathit{1.263}_{0.020}$	$\mathit{1.496}_{0.021}$	$\mathit{1.349}_{0.035}$
	CRN	$1.658_{0.034}$	$1.687_{0.090}$	$1.933_{0.105}$	$2.025_{0.150}$
	CT	$1.526_{0.081}$	$1.466_{0.082}$	$1.707_{0.042}$	$1.673_{0.075}$
	RMSN	$1.586_{0.015}$	$1.581_{0.040}$	$1.891_{0.083}$	$1.731_{0.247}$
	SCIP	$4.220_{0.892}$	$4.752_{1.220}$	$4.262_{0.701}$	$5.525_{1.644}$
	TESAR-CDE	$1.945_{0.124}$	$2.014_{0.252}$	$1.944_{0.101}$	$1.738_{0.053}$
	iASIDE	$\mathbf{1.046}^{*}_{0.005}$	$\mathbf{0.975}^{*}_{0.013}$	$\mathbf{1.035}^{*}_{0.003}$	$\mathbf{0.771}^{*}_{0.015}$
3	ACTIN	$\mathit{1.605}_{0.345}$	$\mathit{1.342}_{0.123}$	$\mathit{1.321}_{0.154}$	$\mathit{1.228}_{0.221}$
	CRN	$1.970_{0.022}$	$1.985_{0.081}$	$1.958_{0.075}$	$1.847_{0.301}$
	CT	$1.769_{0.245}$	$1.752_{0.299}$	$1.696_{0.230}$	$1.527_{0.258}$
	RMSN	$1.930_{0.044}$	$1.956_{0.096}$	$1.901_{0.150}$	$1.579_{0.077}$
	SCIP	$5.663_{0.851}$	$5.801_{1.223}$	$5.590_{1.399}$	$5.047_{0.433}$
	TESAR-CDE	$2.216_{0.294}$	$2.628_{0.343}$	$2.020_{0.072}$	$1.678_{0.043}$
	iASIDE	$\mathbf{1.322}_{0.023}$	$\mathbf{1.017}^{*}_{0.016}$	$\mathbf{1.143}^{*}_{0.053}$	$\mathbf{0.696}^{*}_{0.026}$

Table 8: n-step RMSE on the held-out counterfactual sequences, as a percentage of the tumour death threshold. Each entry is the mean over training seeds with standard deviation as a subscript. **Bold** marks the lowest value in a row and *italic* the second lowest. * marks a gap to the runner-up that is significant under a one-sided Welch's t-test.

F.2.2 Complete Results for sliding treatment

Table 9: n-step RMSE on the sliding-treatment test set, as a percentage of the tumour death threshold: mean and standard deviation over training seeds. Lowest mean per heterogeneity level and τ in bold, second lowest in italics.

Heter.	Model	τ = 5	τ = 10	τ = 15	τ = 20
0	ACTIN	$\mathit{0.344}_{0.243}$	$\mathit{0.329}_{0.093}$	$\mathit{0.341}_{0.132}$	$\mathit{0.400}_{0.174}$
	CRN	$0.787_{0.099}$	$0.891_{0.086}$	$0.988_{0.132}$	$1.044_{0.045}$
	CT	$0.354_{0.016}$	$0.437_{0.034}$	$0.352_{0.044}$	$0.424_{0.075}$
	RMSN	$0.572_{0.177}$	$0.646_{0.134}$	$0.575_{0.131}$	$0.695_{0.082}$
	SCIP	$4.058_{1.093}$	–	–	–
	TESAR-CDE	$0.744_{0.033}$	$0.790_{0.061}$	$0.621_{0.006}$	$0.597_{0.017}$
	iASIDE	$\mathbf{0.197}_{0.003}$	$\mathbf{0.236}_{0.004}$	$\mathbf{0.210}_{0.003}$	$\mathbf{0.223}_{0.005}$
1	ACTIN	$\mathit{0.866}_{0.048}$	$\mathit{0.923}_{0.137}$	$\mathit{1.045}_{0.204}$	$\mathit{1.162}_{0.268}$
	CRN	$1.094_{0.060}$	$1.139_{0.051}$	$1.464_{0.530}$	$1.576_{0.290}$
	CT	$1.009_{0.039}$	$1.035_{0.124}$	$1.221_{0.197}$	$1.404_{0.208}$
	RMSN	$1.172_{0.200}$	$1.100_{0.144}$	$1.172_{0.152}$	$1.272_{0.178}$
	SCIP	$4.124_{1.431}$	–	–	–
	TESAR-CDE	$1.313_{0.096}$	$1.262_{0.024}$	$1.329_{0.017}$	$1.435_{0.049}$
	iASIDE	$\mathbf{0.663}_{0.006}$	$\mathbf{0.549}_{0.008}$	$\mathbf{0.525}_{0.009}$	$\mathbf{0.575}_{0.006}$
2	ACTIN	$\mathit{1.098}_{0.080}$	$\mathit{0.966}_{0.013}$	$\mathit{1.030}_{0.007}$	$\mathit{1.164}_{0.016}$
	CRN	$1.362_{0.027}$	$1.440_{0.104}$	$1.567_{0.212}$	$1.842_{0.151}$
	CT	$1.239_{0.102}$	$1.174_{0.129}$	$1.292_{0.200}$	$1.522_{0.263}$
	RMSN	$1.268_{0.029}$	$1.217_{0.036}$	$1.247_{0.093}$	$1.451_{0.166}$
	SCIP	$3.359_{0.608}$	–	–	–
	TESAR-CDE	$1.479_{0.074}$	$1.434_{0.048}$	$1.430_{0.030}$	$1.583_{0.015}$
	iASIDE	$\mathbf{0.863}_{0.004}$	$\mathbf{0.697}_{0.017}$	$\mathbf{0.674}_{0.018}$	$\mathbf{0.694}_{0.034}$
3	ACTIN	$\mathit{1.289}_{0.307}$	$\mathit{1.179}_{0.190}$	$\mathit{1.335}_{0.384}$	$\mathit{1.625}_{0.594}$
	CRN	$1.665_{0.014}$	$1.562_{0.060}$	$1.911_{0.241}$	$2.077_{0.139}$
	CT	$1.481_{0.209}$	$1.446_{0.138}$	$1.573_{0.191}$	$1.931_{0.280}$
	RMSN	$1.604_{0.059}$	$1.474_{0.073}$	$1.610_{0.124}$	$1.895_{0.101}$
	SCIP	$4.225_{0.519}$	–	–	–
	TESAR-CDE	$1.937_{0.198}$	$1.710_{0.087}$	$1.709_{0.015}$	$2.003_{0.035}$
	iASIDE	$\mathbf{1.036}_{0.028}$	$\mathbf{0.858}_{0.012}$	$\mathbf{0.718}_{0.019}$	$\mathbf{0.933}_{0.041}$

A notable observation in Table 9 was that for most baselines, their forecasting accuracy actually improved as τ or the heterogeneity level increased (from level 1 to levels 2–3). This was potentially because a higher heterogeneity level in treatment effect parameters or time-window τ has resulted in a dataset with smaller tumor volumes (and hence smaller RMSE) (see fig. 6 in Appendix E.2).

F.3 VISUAL EXAMPLES

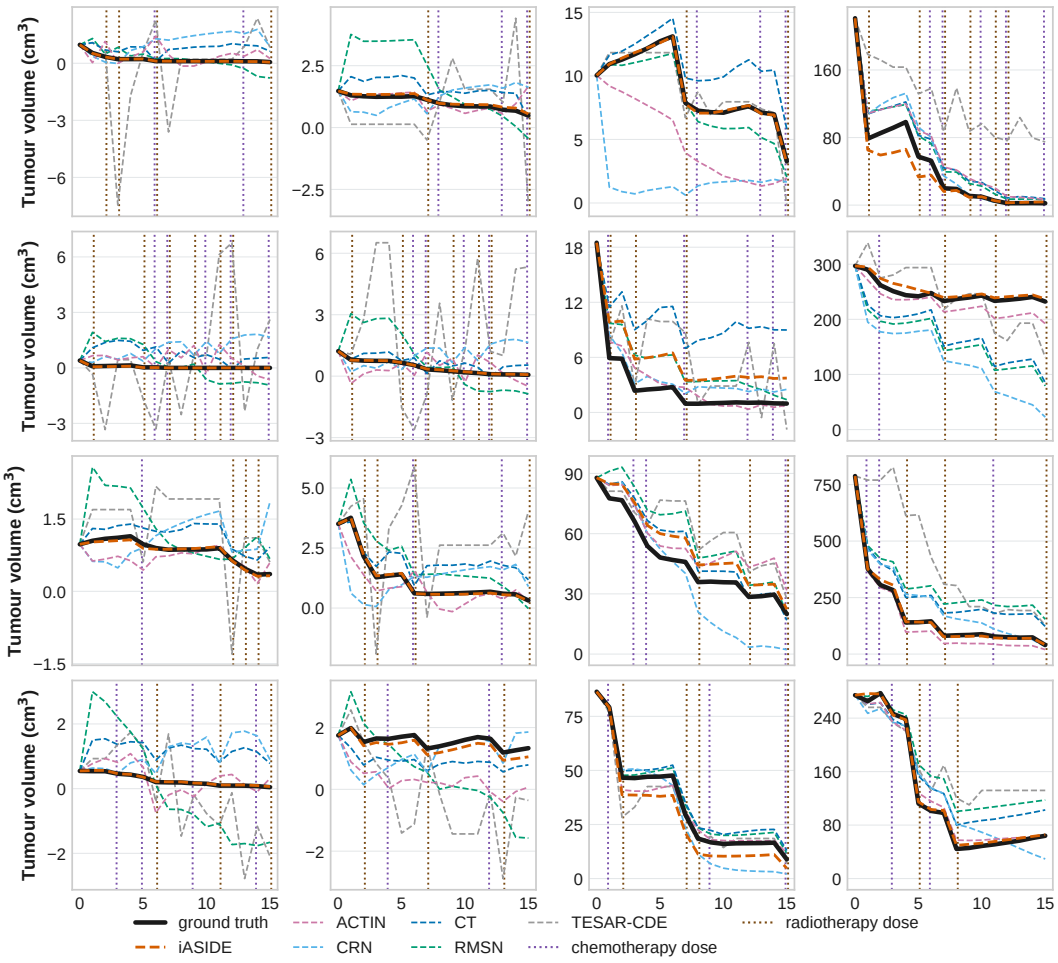

Figure 8: Some visual examples of forecast data for the different models

F.4 Optimal Treatment Selection Accuracy

We evaluated optimal treatment selection on simulated tumour-growth data at all four heterogeneity levels. For each test patient simulated under sliding treatment described in Appendix F.2, the test set holds 2τ plans, each one dose of chemotherapy or radiotherapy at one of τ projected steps ($\tau \in \{5, 10, 15, 20\}$). The treatment plan which results in lowest tumor volume at the τ-step is then considered optimal treatment. Each trained model, predicts the tumour volume under every plan and selects the plan minimising predicted final volume. Against the plan minimising true final volume, we score agreement in treatment type and the results are presented in Table 10 and Figure 9.

From the table and the plots clearly indicate that the task of predicting optimal treatment gets harder as heterogenity increases from 0 to 1. All models detoriate. However, iASIDE is significantly better than all other baselines.

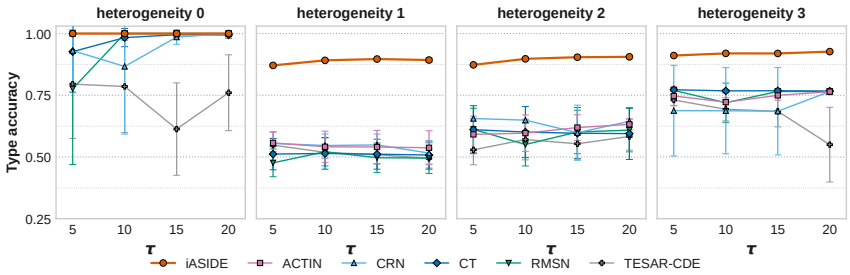

Figure 9: Treatment accuracies across different heterogeneity and τ

Heterogeneity	Model	τ = 5	τ = 10	τ = 15	τ = 20
0	ACTIN	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$
	CRN	$\mathit{0.93}^{\dagger}_{0.14}$	$0.87_{0.27}$	$\mathit{0.99}_{0.029}$	$\mathbf{1.0}_{0.00016}$
	CT	$\mathit{0.93}_{0.16}$	$\mathit{0.98}_{0.037}$	$\mathbf{1.0}_{0.011}$	$\mathit{0.99}_{0.013}$
	RMSN	$0.78_{0.31}$	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$
	TESAR-CDE	$0.80_{0.22}$	$0.79_{0.19}$	$0.61_{0.19}$	$0.76_{0.15}$
	iASIDE	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$	$\mathbf{1.0}_{0}$
1	ACTIN	$\mathit{0.56}_{0.044}$	$0.54_{0.063}$	$0.54_{0.067}$	$\mathit{0.54}_{0.069}$
	CRN	$0.55_{0.048}$	$\mathit{0.55}_{0.046}$	$\mathit{0.55}_{0.044}$	$0.52^{\dagger}_{0.044}$
	CT	$0.51^{\dagger}_{0.064}$	$0.51_{0.064}$	$0.51_{0.061}$	$0.51_{0.057}$
	RMSN	$0.48_{0.056}$	$0.52_{0.057}$	$0.50_{0.060}$	$0.50_{0.062}$
	TESAR-CDE	$0.55_{0.026}$	$0.52_{0.026}$	$0.51_{0.038}$	$0.49_{0.039}$
	iASIDE	$\mathbf{0.87}^{*}_{0.0045}$	$\mathbf{0.89}^{*}_{0.0031}$	$\mathbf{0.90}^{*}_{0.0015}$	$\mathbf{0.89}^{*}_{0.0089}$
2	ACTIN	$0.59_{0.074}$	$0.60_{0.074}$	$\mathit{0.62}_{0.052}$	$0.63_{0.023}$
	CRN	$\mathit{0.66}_{0.0011}$	$\mathit{0.65}_{0.0014}$	$0.60_{0.11}$	$\mathit{0.64}_{0.0088}$
	CT	$0.61_{0.097}$	$0.60_{0.10}$	$0.60_{0.10}$	$0.59_{0.10}$
	RMSN	$0.61_{0.086}$	$0.55_{0.087}$	$0.60_{0.088}$	$0.61_{0.088}$
	TESAR-CDE	$0.53_{0.061}$	$0.57_{0.082}$	$0.55_{0.059}$	$0.58_{0.055}$
	iASIDE	$\mathbf{0.87}^{*}_{0.0021}$	$\mathbf{0.90}^{*}_{0.0013}$	$\mathbf{0.90}^{*}_{0.0019}$	$\mathbf{0.91}^{*}_{0.0021}$
3	ACTIN	$0.75_{0.039}$	$0.72_{0.037}$	$0.75_{0.019}$	$0.76_{0.012}$
	CRN	$0.69_{0.18}$	$0.69_{0.17}$	$0.69_{0.18}$	$\mathit{0.77}_{0}$
	CT	$\mathit{0.77}_{0.0057}$	$\mathit{0.77}_{0.0058}$	$\mathit{0.77}_{0.0064}$	$\mathit{0.77}_{0.0045}$
	RMSN	$\mathit{0.77}_{0.000024}$	$0.72_{0.080}$	$\mathit{0.77}_{0}$	$\mathit{0.77}_{0}$
	TESAR-CDE	$0.73_{0.022}$	$0.69_{0.047}$	$0.69_{0.063}$	$0.55_{0.15}$
	iASIDE	$\mathbf{0.91}^{*}_{0.0011}$	$\mathbf{0.92}^{*}_{0.0031}$	$\mathbf{0.92}^{*}_{0.0025}$	$\mathbf{0.93}^{*}_{0.0043}$

Table 10: Treatment-type selection accuracy (↑): mean and std (subscript) of 5 seeds, on the sliding-treatment testset. **Bold** marks the highest and *italic* the second highest. * marks $p < 0.05$ under a one-sided Welch's t-test of the null that iASIDE's mean accuracy is no higher than the highest baselines.

F.5 PREDICTION IN OUT-OF-DISTRIBUTION DYNAMICS PARAMETERS

We evaluated generalisation to a shift in patient parameters on simulated tumour-growth data. Models trained at heterogeneity 0, 1 or 2 were all scored, without retraining, on counterfactual test sequences with random treatment plans simulated at heterogeneity 3, where growth and treatment-response parameters vary most. For $\tau \in \{5, 10, 15, 20\}$ we report normalized RMSE in Table 11. On doing so, some seeds clearly diverged and were suboptimal for use in out-of-distribution (OOD) dataset. To avoid those outliers from affecting the analysis, we exclude any seed whose normalized rmse was more than twice the median. The results in Table 11, are clearly worse than results in Table 8 for all models, which is expected as the data contains OOD samples for DGP parameters. Despite of that difference between train and test, we can see iASIDE performing decently specially when trained on heterogenty 1 or 2. In heterogenity 0, the parameter diversity is poor and model trained on such dataset is not expected to do generalize to more diverse parameters. However, when parameter diversity increases even slightly, iASIDE can learn to generalize to an unseen out-of-distribution parameter range better than the baselines.

		τ			
Trained at	Model	5	10	15	20
0	ACTIN	$\mathbf{1.5}_{0.036}$	$\mathbf{1.6}_{0.0087}$	$\mathbf{1.6}_{0.022}$	$\mathbf{1.6}_{0.0096}$
	CRN	$2.1^{\dagger}_{0.11}$	$\mathit{1.9}_{0.058}$	$1.9_{0.043}$	$1.9_{0.029}$
	CT	$2.0_{0.11}$	$\mathit{1.9}_{0.095}$	$\mathit{1.7}_{0.054}$	$\mathit{1.7}_{0.075}$
	RMSN	$\mathit{1.6}_{0.095}$	$\mathbf{1.6}_{0.049}$	$\mathbf{1.6}_{0.024}$	$\mathbf{1.6}_{0.021}$
	TESAR-CDE	$1.8_{0.044}$	$\mathit{1.9}_{0.065}$	$1.9_{0.035}$	$1.8_{0.036}$
	iASIDE	$\mathit{1.6}^{\ddagger}_{0.020}$	$\mathbf{1.6}^{\ddagger}_{0.017}$	$\mathbf{1.6}^{\ddagger}_{0.0048}$	$\mathbf{1.6}_{0.020}$
1	ACTIN	$\mathit{1.8}_{0.042}$	$\mathit{1.7}_{0.060}$	$\mathit{1.4}_{0.072}$	$\mathit{1.3}_{0.11}$
	CRN	$1.9_{0.024}$	$1.8_{0.015}$	$1.5_{0.10}$	$1.5^{\dagger}_{0.082}$
	CT	$\mathit{1.8}^{\dagger}_{0.092}$	$\mathit{1.7}_{0.061}$	$\mathit{1.4}_{0.039}$	$1.4_{0.022}$
	RMSN	$1.9_{0.079}$	$\mathit{1.7}_{0.036}$	$\mathit{1.4}_{0.024}$	$1.4_{0.055}$
	TESAR-CDE	$2.4_{0.34}$	$2.3_{0.20}$	$1.8_{0.043}$	$1.6_{0.032}$
	iASIDE	$\mathbf{1.4}^{*\ddagger}_{0.024}$	$\mathbf{1.3}^{*\ddagger}_{0.015}$	$\mathbf{0.97}^{*}_{0.031}$	$\mathbf{0.88}^{*\ddagger}_{0.024}$
2	ACTIN	$\mathit{1.6}_{0.11}$	$\mathit{1.5}_{0.053}$	$\mathit{1.2}_{0.043}$	$\mathit{1.2}_{0.012}$
	CRN	$1.8_{0.046}$	$1.8_{0.031}$	$1.5_{0.062}$	$1.5_{0.071}$
	CT	$1.7_{0.086}$	$1.7_{0.082}$	$1.3_{0.033}$	$1.3_{0.017}$
	RMSN	$1.8_{0.031}$	$1.7_{0.051}$	$1.4_{0.021}$	$1.4_{0.068}$
	TESAR-CDE	$2.2_{0.23}$	$2.4_{0.25}$	$1.9_{0.16}$	$1.6_{0.031}$
	iASIDE	$\mathbf{1.2}^{*}_{0.028}$	$\mathbf{1.1}^{*}_{0.013}$	$\mathbf{0.81}^{*}_{0.043}$	$\mathbf{0.87}^{*}_{0.043}$

Table 11: RMSE on a test split simulated at heterogeneity 3 for models trained at a lower level, over different τ. Models are trained at heterogeneity 0, 1 or 2 without confounding and scored without retraining on counterfactual test sequences simulated at heterogeneity 3, where the growth and treatment-response parameters vary most across patients. **Bold** marks the lowest RMSE and *italic* the second lowest. * marks $p < 0.05$ under a one-sided Welch's t-test of the null that iASIDE's mean error is no lower than the lowest baselines. ‡A run whose RMSE on the heterogeneity-3 test split exceeds 2 times the median of its cell's seeds is left out as outlier, in every column: 6 runs, all of them iASIDE's and one seed per cell.

G Additional Ablation Results

G.1 Ablation of Data Generating Assumptions

We ran experiments on data generated from the two synthetic data-generating processes below, and present the results in Table 12. The two settings are similar to the PK-PD model in every other way except for the dynamics functions. The native dynamics, henceforth represented with f_n and external intervention represented by f_e were changed to the ones given below for two different synthetic settings, and the experiments were conducted on some selected baselines.

Setting 1:

$$f_n(\mathbf{z}) = \rho\, \mathbf{z}\frac{(K-\mathbf{z})}{K}, \qquad f_e(\mathbf{z}) = \alpha\sqrt{\mathbf{z}} \tag{47}$$

Setting 2:

$$f_n(\mathbf{z}) = \rho\sqrt{K^2-\mathbf{z}^2}, \qquad f_e(\mathbf{z}) = \beta\sqrt{K^2-\mathbf{z}^2} \tag{48}$$

Table 12: RMSE of selected models on data generated under the two settings in Appendix G.1

	RMSN	CRN	CT	TESAR-CDE	iASIDE
Setting 1	1.02 (0.34)	.88 (.12)	.66 (.13)	.59 (.02)	.21 (.01)
Setting 2	12.26 (.37)	13.83 (.40)	11.01 (.04)	–	4.79 (.99)

G.2 Ablation of Sample Efficiency

We reran experiments on the PK-PD dataset, training iASIDE and a subset of baseline models with a reduced number of training sequences. When the total number of training sequences was reduced from 1000 to 100, we observed that the RMSE worsened the least for iASIDE at 19.7%.

Table 13: Percentage decrease in RMSE on PK-PD tumor dataset when number of training samples is decreased.

Model	$n_{\text{train}} = 1000$	$n_{\text{train}} = 100$	% decrease
RMSN	1.20(.04)	1.51(.04)	25.8%
CRN	1.23(.02)	2.25(.63)	82.9%
CT	1.36(.10)	3.94(.26)	189.7%
iASIDE	1.57(.08)	1.88(.14)	19.7%

H COMPUTATION COST

For the Pendulum dataset without confounding, the complexity of the models and their training times are given in Table 14.

Table 14: Comparison of parameter counts as well as training and test computation time.

			Training			Inference	
Model	Parameters	Epochs	s/epoch	min	Device	min	Device
ACTIN	6 142	150	$7.5^{(2)}$	19	T4	$29^{(2)}$	T4
CRN	42 774	134	$110^{(4)}$	250	Grace CPU	$8.2^{\dagger}$	A100
CT	33 909	72	$95^{(4)}$	110	Grace CPU	$18^{\dagger}$	A100
RMSN	22 079	81	$33^{(2)}$	44	Grace CPU	$16^{\dagger}$	A100
SCIP	84 021	68	$42^{(2)}$	48	Grace CPU	$22^{(3)}$	x86 CPU
TESAR-CDE	40 773	20	$250^{(2)}$	83	Grace CPU	$4.0^{(2)}$	Grace CPU
iASIDE	19 016	50	340	280	GH200	4.2	GH200

I SOURCE CODE

Source code can be found here: `https://anonymous.4open.science/r/iaside-iclr`