ASIDE: ADAPTIVE AND SEPARABLE INTERVENTIONAL DYNAMICS VIA PROGRESSIVE META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

To inform decisions about changing the future trajectories of a dynamics system, it is important to predict not only the intrinsic dynamics of the system but also its response to external interventions. While notable progress has been made in learning intervention effects over time, existing research has prioritized the challenge of time-varying confounding in observational data. Significant challenges however remain in aspects related to the modeling and inference of latent dynamics. A first and foremost challenge lies in the need to separate, from a composite observation, the natural temporal evolution of intrinsic dynamics from its response to external interventions. This challenge is further exacerbated by the need to integrate rich history information into these latent dynamics. In this paper, we present a novel framework of adaptive and separable interventional dynamics (ASIDE) to overcome these challenges. First, we leverage inductive bias and progressive learning to allow separable modeling and inference of the intrinsic dynamics and its responses to external interventions at the latent space. This is in contrast to existing approaches that model and infer the composite dynamics as a black box. Second, we leverage meta-learning to enable these latent dynamics to adapt to context examples in past history, addressing both inter- and intra-subject variabilities. This is in contrast to existing approaches that use history only to initialize a *one-size-fit-all* forecasting function. On synthetic and real benchmarks, we demonstrate the advantage of ASIDE in improving forecasting accuracy for both intrinsic and interventional dynamics, in settings with or without time-varying confounding.

1 INTRODUCTION

Across diverse domains, high-dimensional time-series observations are becoming increasingly abundant. This trend underscores the growing importance of time-series modeling as a foundation for enabling prediction and optimal control of observed systems (Krishnan et al., 2015). While forecasting the *intrinsic dynamics* native to a system is important for predicting its future trajectories, to inform optimal decisions that can influence such trajectories requires predicting the effect of *external interventions* on the system's intrinsic dynamics (Krishnan et al., 2015; Gwak et al., 2020). Using medicine as an example, longitudinal multi-modal data of an individual can be leveraged to predict the progression of an underlying health condition, while the decision of what interventions (*e.g.*, medication, life-style, etc.) to best improve such progression requires an ability to model and predict the effect of these interventions on the individual's native health condition over time.

Significant advances have been made in developing deep learning models for modeling the latent dynamics underlying high-dimensional time-series data (Chung et al., 2015; Krishnan et al., 2015; Fraccaro et al., 2017; Botev et al., 2021). However, most of these developments are focused on the intrinsic dynamics of a system, with limited consideration about the effects of external interventions (Krishnan et al., 2015; Gwak et al., 2020; Brouwer et al., 2022). In parallel, there has been a rising interest in modeling intervention effects over time from observational data: however, with a priority on addressing the challenge of time-varying confounding, dynamics modeling in these works have mostly leveraged established techniques such as LSTM (Lim, 2018; Bica et al., 2020; Berrevoets et al., 2021), transformers (Melnychuk et al., 2022), and neural ordinary differential equations (ODEs) (Brouwer et al., 2022). At the intersection of these two vibrant research areas,

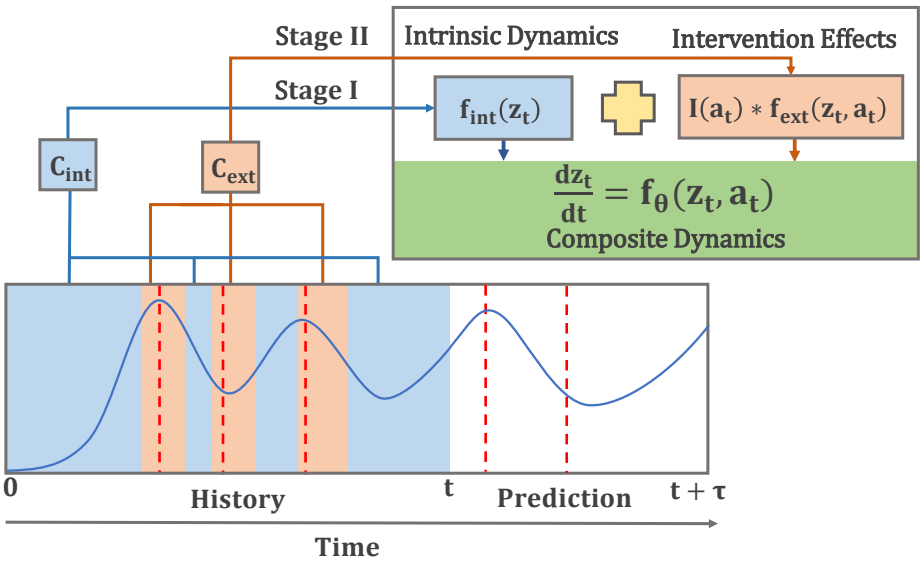


Figure 1: Illustration of ASIDE with separate modeling and inference of intrinsic dynamics and its response to external interventions, both adaptive from context samples in the history as enabled by progressive meta-learning.

significant gaps remain in tackling the challenges of modeling and inferring latent dynamics under external interventions.

A first and foremost challenge lies in the need to separate, from a composite observation, the natural temporal evolution of intrinsic dynamics from its response to external interventions. Consider time-series covariate data $\{\mathbf{x}_t\}_{t=1}^T$ and its corresponding latent dynamics states $\{\mathbf{z}_t\}_{t=1}^T$. Assume that the temporal evolution of $\mathbf{z}_t$ under external interventions $\mathbf{a}_t$ is governed by a dynamics function $\frac{d\mathbf{z}_t}{dt} = f(\mathbf{z}_t, \mathbf{a}_t)$, which describes both how $\mathbf{z}_t$ naturally evolves (*intrinsic dynamics*) and its response to the external intervention $\mathbf{a}_t$. Unfortunately, these two mechanisms are not separately observed. Existing works in longitudinal intervention-effect modeling partially addresses this challenge from the perspective of causal-inference, *e.g.*, by extracting from the history a latent representation $\mathbf{z}_t$ that is minimally predictive of $\mathbf{a}_t$ when predicting the effect of $\mathbf{z}_t$ and $\mathbf{a}_t$ on an outcome variable $\mathbf{y}_{t+1}$. These however do not explicitly disentangle intrinsic dynamics and its response to external interventions in $f(\mathbf{z}_t, \mathbf{a}_t)$. Limited works have attempted this disentangling. In (Gwak et al., 2020), two separate neural ODEs were used to respectively model intrinsic dynamics $f(\mathbf{z}_x)$ and its effect from external interventions $f(\mathbf{z}_a)$. In (Brouwer et al., 2022), the latent dynamics is modeled as a neural controlled differential equation (CDE) where the dynamics $f(\mathbf{z}_t)$ is modulated by incoming treatments $\mathbf{a}_t$'s. While explicit separation in $f(\mathbf{z}_t, \mathbf{a}_t)$ improves its interpretability compared to a black box, effective inference strategies to ensure such separation remain an open problem.

Secondly, consider $\mathbf{z}_{t+1:t+\tau}$ generated by $f(\mathbf{z}_t, \mathbf{a}_t)$ over any time window τ and its associated co-variates $\mathbf{x}_{t+1:t+\tau}$ and outcomes $\mathbf{y}_{t+1:t+\tau}$. All existing intervention-effect models attempt to describe this by a good initial estimate $\hat{\mathbf{z}}_t$ (from history) along with a *global* dynamics function that applies to all data samples. This is typically achieved in a two-stage encoding-decoding framework: in the first stage, a sequential encoder is trained to extract a latent representation $\hat{\mathbf{z}}_t$ from past history; in the second stage, a forecasting decoder is then trained to use $\hat{\mathbf{z}}_t$ to forecast ahead for a window length of τ. This popular approach has two limitations. First, the training of the encoder, *i.e.*, the extraction of $\hat{\mathbf{z}}_t$, is unaware of the primary forecasting objective in the second stage. Second and more importantly, a global $f(\mathbf{z}_t, \mathbf{a}_t)$ can be limited in its ability to describe the heterogeneity in both the intrinsic dynamics and responses to external interventions – a variability that can exist both among individual systems (*e.g.,* different patient subgroups) and within the same system over time (*e.g.*, different disease stages of the same patient). While $\hat{\mathbf{z}}_t$ as an estimated initial condition may capture such heterogeneity from past history, the ability for such information to pass forward in a *one-size-fit-all* dynamics function is not clear, especially as the forecasting horizon increases.

In this paper, we overcome these two challenges in a novel framework of ASIDE to achieve adaptive and separable interventional dynamics. As outlined in Fig. 1, ASIDE has two major innovations.

- We introduce a novel strategy for separately modeling and inferring the mechanisms of intrinsic dynamics $f_{\text{int}}(\mathbf{z}_t)$ and its response to external interventions $f_{\text{ext}}(\mathbf{z}_t, \mathbf{a}_t)$. This includes 1) explicit separation of these components in the model of $f(\mathbf{z}_t, a_t)$, along with 2) a progressive learning strategy to first estimate f_{int} from intervention-free history and, with which, to isolate f_{ext} from their composite observations. In contrast to black-box learning of $f(\mathbf{z}_t, a_t)$, we show that this progressive learning of separate dynamics improves the accuracy of forecasting for both intrinsic dynamics and its response to external interventions.
- We further allow both dynamics to vary as $f_{\text{int}}(\mathbf{z}_t; \mathbf{c}_{\text{int}})$ and $f_{\text{ext}}(\mathbf{z}_t, a_t; \mathbf{c}_{\text{ext}})$, where, $\mathbf{c}_{\text{int}}$ and $\mathbf{c}_{\text{ext}}$ are *context* embeddings extracted from respective samples from the past history. This is enabled in a novel end-to-end meta-learning framework to allow rapid feed-forward extraction of context embeddings $\mathbf{c}_{\text{int}}$/ $\mathbf{c}_{\text{ext}}$ and its adaptation of the latent dynamics function, enabling adaptive and individualized forecasting that can address both inter- and intra-subject variations. In contrast to existing approaches that uses history information to only initialize a global *one-size-fit-all* dynamics function, we show that this adaptive dynamics function will improve forecasting accuracy, especially in long forecasting horizons, and, increasing heterogeneity.

We first evaluated ASIDE in a synthetic benchmark of tumor growth under radiation and chemotherapies (Geng et al., 2017): to isolate the effect of the dynamics modeling strategies introduced in ASIDE, we considered data settings without treatment assignment bias. We then evaluated ASIDE on a real dataset from MIMIC-III (Johnson et al., 2023), which are inherently associated with unknown treatment assignment bias inherent in observational data. In both settings, we demonstrated the improved performance of ASIDE for time-series forecasting under dynamic external interventions in comparison to contemporary baselines.

2 Related Works

Modeling intervention effects over time: Rapid progress has been made in learning intervention effects over time (Lim, 2018; Bica et al., 2020; Melnychuk et al., 2022; Seedat et al., 2022; Berrevoets et al., 2021). Most of these works focus on addressing the challenge of time-varying confounding and treatment assignment bias, represented by propensity weight (Lim, 2018) and invariant representation (Bica et al., 2020; Melnychuk et al., 2022; Seedat et al., 2022) based approaches. In terms of dynamics modeling, earlier works have mostly adopted a two-stage learning process, where the first-stage trains an *encoder* to extract latent representations $\mathbf{z}_t$ from past history $\mathcal{H}_{1:t}$ and, with this fixed, a second-stage *decoder* learns to take the encoded $\mathbf{z}_t$ to predict treatment outcome given intervention $\mathbf{a}_t$ within a horizon of τ. Different types of neural architectures have been used, such as recurrent neural networks (Lim, 2018; Bica et al., 2020) and controlled differential equations (Seedat et al., 2022). Recognizing the limitations associated with such two-stage training, especially that the encoding from history is not made aware of the forecasting objectives, growing recent works have attempted learning the encoding-decoding process end-to-end (Melnychuk et al., 2022), using neural ODEs (Brouwer et al., 2022) and transformers (Melnychuk et al., 2022).

Orthogonal to the contribution of addressing time-varying confounding, ASIDE aims to advance intervention-effect modeling by enabling the learning of separable and adaptive interventional dynamics, realized in a novel progressive (for separable) meta-learning (for adaptive) framework.

Separating intrinsic dynamics and interventional effects: There have been limited works that share our motivation in separating the intrinsic and interventional dynamics from observed time-series. In (Gwak et al., 2020), this is achieved by using two separate neural ODEs which are then combined in a third neural ODE to generate the observed composite trajectories. In (Seedat et al., 2022), the latent dynamics is modeled as a neural CDE where the intrinsic dynamics $f(\mathbf{z}_t)$ is modulated by incoming treatments $\mathbf{a}_t$'s. This explicit separation of intrinsic dynamics and intervention effects improves the interpretability of the latent dynamics models compared to a black box. However, effective inference strategies to ensure such separation remain an open problem. Furthermore, neither of these works have considered adaptive latent dynamics.

3 METHODOLOGY

As outlined in Fig. 1, ASIDE includes two key innovations. First, we leverage inductive bias to enable separable modeling and inference of the intrinsic dynamics and its responses to external interventions at the latent space. This is in contrast to existing approaches that model and infer the composite dynamics as a black box. Second, we leverage meta-learning to enable these latent dynamics functions to adapt to context examples in past history. This is in contrast to existing approaches that use history to only initialize a *one-size-fit-all* forecasting function. Below, we describe ASIDE by its adaptable latent dynamics models (Section 3.1), the extraction of context embedding from history to adapt the dynamics (Section 3.2), and the progressive meta-learning scheme (3.3).

3.1 GENERATION PROCESS: ADAPTIVE AND SEPARABLE LATENT DYNAMICS

While our model is agnostic to the type of functions used to describe latent dynamics, in this paper we choose a neural ODE to describe the latent dynamics as a *continuous* process:

$$\frac{d\mathbf{z}_t}{dt} = f_\theta(\mathbf{z}_t, \mathbf{a}_t) = f_{\theta_{\text{int}}}(\mathbf{z}_t; \mathbf{c}_{\text{int},t}) + \sum_k \mathcal{I}^k(a_t^k) f_{\theta^k_{\text{ext}}}(\mathbf{z}_t, a_t^k; \mathbf{c}^k_{\text{ext},t});\ \mathbf{y}_{t+1} = g_\eta(\mathbf{z}_t, \mathbf{a}_t) \tag{1}$$

where $f_{\theta_{\text{int}}}$ models the intrinsic dynamics parameterized by θ_{int}. Multiple intervention mechanisms can exist, where $a_t^k \in \mathbf{a}_t$ represents intervention k and the indicator function $\mathcal{I}(a_t^k) = 1$ flags its presence at time t. The response of $\mathbf{z}_t$ to the intervention k is modeled by $f_{\theta^k_{\text{ext}}}$ parameterized by θ^k_{ext}. g describes the effect of $\mathbf{z}_t$ and $\mathbf{a}_t$ parameterized by η. Instead of learning fixed functions for $f_{\theta_{\text{int}}}$ and $f_{\theta^k{}_{\text{ext}}}$'s, we allow them to change with time-varying embeddings $\mathbf{c}_{\text{int},t}$ and $\mathbf{c}^k_{\text{ext},t}$'s that are separately identified from history $\mathcal{H}_{1:t} = \{\mathbf{x}_{1:t}, \mathbf{a}_{1:t-1}, \mathbf{y}_{1:t}\}$ up to time t. While different adaptation mechanisms exist, here we consider a simple conditioning for additive adaptation.

The generation process as described in equation 1 differ from existing works in two aspects. First, the latent dynamics is explicitly decomposed into intrinsic dynamics and its response to interventions. Second, latent dynamics functions are adaptable instead of fixed for all training samples.

3.2 INFERENCE PROCESS: ADAPTATION VIA DYNAMICS-SPECIFIC CONTEXT EMBEDDING

To optimize the generation process in equation 1, existing works mostly focus on first learning an *encoder* to extract a latent representation $\mathbf{z}_t$ from the history $\mathcal{H}_{1:t}$, which is then utilized to initialize a fixed forecasting function optimized by predicting forward for a duration of τ. ASIDE differs in the following aspects. First, because the trajectory of $\mathbf{z}_t$'s is *governed* by f_θ, we shift the focus of learning to f_θ: this results in a fundamentally-different inference formulation where information from a patient's history $\mathcal{H}_{1:t}$ is used to adapt f_θ to capture inter- and intra-subject variability, optimized by forecasting forward for a window of τ using a simple $\mathbf{z}_t$ estimated from recent observations. Second, to encourage the separation of intrinsic dynamics and responses to interventions, we design the extraction of $\mathbf{c}_{\text{int},t}$ and $\mathbf{c}^k_{\text{ext},t}$ from 1) different portions of the history data $\mathcal{H}_{1:t}$ depending on their relevance to f_{int} and f_{ext} and 2) via different extraction mechanisms.

Inference of initial latent states: To shift the learning focus to a strong forecasting function f_θ that is able to capture history information, we use a simple strategy to infer $\mathbf{z}_t$ from the first frames of $\mathbf{x}_{t-l-1:t-1} (l < \tau)$ via a neural encoder Ψ_{ϕ_z} with weight parameters ϕ_z: $\hat{\mathbf{z}}_t = \Psi_{\phi_z}(\mathbf{x}_{t-l-1:t-1})$.

Inference of intrinsic dynamics: To infer $\mathbf{c}_{\text{int},t}$ for adapting $f_{\theta_{\text{int}}}$ at time t, we assume that it may be shared by past trajectories of natural dynamics of the same individual. To remove confounding interventional effects in past history, we further consider only the trajectories in history $\mathcal{H}_{1:t}$ that are free from any intervention, denoted by $\mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}$ with $\mathcal{I}^0_{1:t}$ indicating the absence of intervention $(= 1)$ or not $(= 0)$ at each time instant t. We further divide available $\mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}$ into l_{int} number of segments $\mathbf{s}_{\text{int}}$'s of duration τ, where l_{int} varies as the history grows. Each individual segment $\mathbf{s}_{\text{int}}$ is fed into an encoder $\Psi_{\phi_{\text{int}}}(\cdot)$ to extract an embedding, which are then aggregated across segments as:

$$\mathbf{c}_{\text{int,t}} = \mathcal{M}_{\text{int}}(\mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}) = \frac{1}{l_{\text{int}}} \sum\nolimits_{\mathbf{s}_{\text{int}} \in \mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}} \Psi_{\phi_{\text{int}}}(\mathbf{s}_{\text{int}}) \tag{2}$$

where we adopt a simple averaging here to extract the shared embedding by the context samples $\mathbf{s}_{\text{int}} \in \mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}$. Additional weighted averaging or time decay can be added to share with only recent history, or using attention mechanism to find similar dynamics in the history.

Inference of response to external interventions: The inference of $\mathbf{c}^k_{\text{ext},t}$ for adapting $f_{\theta_{\text{ext}}}$ is more challenging as, unlike intrinsic dynamics, the history will never have *intrinsic-free* trajectories. Instead, for any intervention k, only its composite effect with intrinsic dynamics can be observed. To separate out the latter, we leverage the concept of *counterfactuals*: for any factual composite trajectory under the effect of intervention k, we introduce an *intervention-free* counterfactual for an encoder $\Psi^k_{\phi_{\text{ext}}}(\cdot)$ to compare and extract an embedding that accounts for the difference due to intervention k. This is realized in two strategies. First, for each segment $\mathbf{s}^k_{\text{ext}} \in \mathcal{H}_{1:t} * \mathcal{I}^k_{1:t}$ with $\mathcal{I}^k_{1:t}$ indicating the presence of intervention type k $(= 1)$ or not $(= 0)$, we synthesize its counterfactual $\mathbf{s}^k_{\text{ext,CF}}$ with our learned $f_{\theta_{\text{int}}}$ and the initial latent state estimate, projecting what the segment would have looked like if intervention k was not applied. The pair of factual and counterfactual samples are fed into an encoder $\Psi_{\phi^k_{\text{ext}}}(\cdot)$ to extract an embedding, which are then aggregated across the segments as:

$$\mathbf{c}^k_{\text{ext},t} = \mathcal{M}^k_{\text{ext}}(\mathcal{H}_{1:t} * \mathcal{I}^k_{1:t}) = \frac{1}{l^k_{\text{ext}}} \sum\nolimits_{\mathbf{s}^k_{\text{ext}} \in \mathcal{H}_{1:t} * \mathcal{I}^k_{1:t}} \Psi_{\phi^k_{\text{ext}}}(\mathbf{s}^k_{\text{ext}}, \mathbf{s}^k_{\text{ext,CF}}) \tag{3}$$

where l^k_{ext} represents the number of history segments with intervention k. Similarly, a simple averaging is used here, although more advanced aggregation strategy can be used depending on prior knowledge about the intra-subject variability in an individual's response to intervention k.

Alternatively, we can include factual intervention-free segments in the history $\mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}$ in addition to the interventional segments $\mathcal{H}_{1:t} * \mathcal{I}^k_{1:t}$, along with the mask $\mathbb{I}(a_k = 1)$ that indicates the presence or absence of a_k. While not paired, this can be considered as comparing interventional and intervention-free data at a distribution level. We realize this with an convolutional architecture over these segments concatenated with intervention masks:

$$\mathbf{c}^k_{\text{ext},t} = \mathcal{M}^k_{\text{ext}}([\mathcal{H}_{1:t} * \mathcal{I}^k_{1:t}, \mathcal{H}_{1:t} * \mathcal{I}^0_{1:t}; \mathbb{I}(a_k = 1)]) \tag{4}$$

3.3 Progressive Meta-Leaerning

Learn-to-adapt meta-objectives: Given a dataset consisting of N unique time-series, we consider the forecasting task of predicting the values of $\mathbf{y}^i_{t+1:t+\tau}$ given the values of history to the point $\mathcal{H}^i_{1:t} = \{\mathbf{x}^i_{1:t}, \mathbf{a}^i_{1:t-1}, \mathbf{y}^i_{1:t}\}$ and intervention assignment $\mathbf{a}^i_t$, where $i = 1 : N$ indicates the i-th time-series in the dataset. For each $\mathbf{y}^i_{t+1:t+\tau}$, the predicted $\hat{\mathbf{y}}^i_{t+1:t+\tau}$ is generated as described by equation 1, with $\mathbf{z}^i_t$ estimated by the initial state encoder, and $f_{\theta_{int}}$ and $f^k_{\theta_{ext}}$'s respectively adapted by $\mathcal{M}_{\text{int}}$ and $\mathcal{M}^k_{\text{ext}}$'s as described in equation 2 – equation 4. The mean-squared-error (MSE) loss between $\hat{\mathbf{y}}^i_{t+1:t+\tau}$ and $\mathbf{y}^i_{t+1:t+\tau}$ is used to optimize the weight parameters of the latent dynamics functions θ_{int} and θ^k_{ext}'s, their corresponding encoders for adaptation ϕ_{int} and ϕ^k_{ext}'s, along with that for the initial state encoder ϕ_z and emission function η.

$$\min_{\phi^k_{\text{ext}}, \phi_{\text{int}}, \phi_z, \theta_{\text{int}}, \theta^k_{\text{ext}}, \eta} \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T-\tau} \|\mathbf{y}^i_{t+1:t+\tau} - \hat{\mathbf{y}}^i_{t+1:t+\tau}(\mathcal{H}_{1:t}, \mathbf{a}_t)\|^2_2 \quad k = 1, \cdots, K \tag{5}$$

where K is the maximum numbers of intervention types considered in the dataset.

The optimization of equation 5 thus corresponds to a meta-learning objective when treating the prediction of each $\mathbf{y}^i_{t+1:t+\tau}$ as its own task, with context samples selected from the history $\mathcal{H}_{1:t}$ as described in Section 3.2, to adapt the intrinsic dynamics $f_{\theta_{\text{int}}}$ and interventional dynamics $f^k_{\theta_{\text{ext}}}$'s to the specific task. $\mathcal{M}_{\text{int}}$ and $\mathcal{M}^k_{\text{ext}}$ as described in equation 2 – equation 4 thus represent feedforward meta-models to extract task-specific embedding for rapid adaptation of latent dynamics models.

Progressive meta-learning to separate intrinsic and interventional dynamics: While the optimization in equation 5 in theory can be carried out simultaneously for all parameters involved, due to the composite observation of intrinsic and multiple interventional dynamics, their separation cannot be guranteed while optimized simultaneously (see ablation in Section 4.3). Instead, we adopt a progressive training strategy where different componnets of the model are estimated at a schedule.

More specifically, we first optimize ϕ_{int} and θ_{int} related to intrisinc dynamics, *i.e.*, the intrinsic dynamics function $f_{\theta_{\text{int}}}$ and the meta-encoder $\mathcal{M}_{\text{int}}$ used to adapt it. Note that this optimization only involve intervention-free observations, removing the challenge of separating composite effects. With the optimized ϕ_{int} and θ_{int} fixed, we then simultaneously optimize ϕ^k_{ext}'s and θ^k_{int}'s related

to responses to interventions, *i.e.*, interventional dynamics functions $f_{\theta^k_{\text{ext}}}$'s and the meta-encoders $\mathcal{M}^k_{\text{int}}$ used to adapt them, for all intervention types k. While this stage of the training does involve interventional data with composite effects, leveraging the optimized intrinsic dynamics facilitates the separation of interventional dynamics. Finally, all model parameters are finetuned together as a fully-integrated model. The initial state encoder Ψ_{ϕ_z} and emission g_η are trained throughout. This progressive training is achieved by turning off the gradient flow to the parameters not involved in training at different stages, each with its respective optimization and learning hyperparameters. Transition between the stages is determined by when the loss for the previous step plateaus or when a max epoch limit is reached.

4 EXPERIMENTS

Counterfactual outcomes are not commonly observed for real-world data, due to which synthetic data have become important for evaluating intervention effect models (Lim, 2018; Bica et al., 2020; Seedat et al., 2022; Melnychuk et al., 2022). We thus first evaluated ASIDE on the well-established benchmark generated by a pharmacokinetic-pharmacodynamic (PK-PD) model of tumor growth in lung cancer patients that includes the effects of chemotherapy and radiotherapy (Geng et al., 2017). To isolate the effect of dynamics models, we considered experimental settings without time-varying confoudning due to treatment assignment bias. To test feasibility in real world settings, we then conducted experiments on the MIMIC-III (Johnson et al., 2023), an electronic health record dataset with inherent real-world treatment assignment bias in observational data (Bica et al., 2020).

Baselines: On both datasets, we considered representative baselines in intervention effect modeling over time, including: 1) RMSN (Lim, 2018), 2) CRN (Bica et al., 2020), and CausalTransformers (CT) (Melnychuk et al., 2022), In terms of the underlying dynamics models, RMSN and CRN use the two-stage training strategies as described earlier while CT is end-to-end. In terms of model architectures, RMSN and CRN are based on RNNs whereas CT is based on transformers.

Metrics: We evaluated the performance of all models by their accuracy in predicting counterfactual outcomes over time, measured by rooted-mean-square-error (RMSE). Following standard practice, the RMSE is normalized by the maximum volume of the tumor (death threshold defined in Lim (2018)) fon the synthetic dataset.

4.1 SYNTHETIC DATA EXPERIMENTS

4.1.1 DATA

Following the PK-PD model in (Geng et al., 2017) for non-small cell lung cancer, we used the following Gompertz model to describe the grwoth of tumor volume with a starting volume of V_0:

$$V_{t+1} = V_t(1 + \rho \log(\tfrac{K}{V}) \ - \beta_c C_t - (\alpha r_t + \beta r_t^2) + \epsilon) \tag{6}$$

where parameters ρ and K control natural growth dynamics, β_c controls the effect of chemotherapy with dose C_t, and β, α control the effect of radiotherapy dose of r_t. The parameter K is the carrying capacity of the model. More details are provided in Appendix B.1

Treatment assignment bias: To isolate the benefits of the dynamics modeling strategies introduced by ASIDE, we considered random treatment assignment with a probability of $p_0 = 0.1$ 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. For a fair comparison, for CRN and CT, that uses treatment-invariant representations to address confounding, we disabled their adversarial training element as it was not needed on the dataset.

Heterogeneity levels: To examine the importance of *adaptive* latent dynamics, we created training data with four different levels of heterogeneity in the key parameters ($\rho, \beta_c, \alpha, \beta$) in Equation (6). This was achieved by controlling the sampling of the parameters to be within a preset range from the mean as summarized in Table 1.

Heterogeneity	Intrinsic Growth (ρ)		Radio effect (α)		Chemo effect (β_c)	
	lower	upper	lower	upper	lower	upper
0	7×10^{-5}	7×10^{-5}	.0398	.0398	.028	.028
1	0.0	7×10^{-3}	0.0	.208	.0273	.0287
2	0.0	7×10^{-3}	0.0	.508	.0259	.0301
3	0.0	21×10^{-3}	0.0	.508	.0259	.0301

Table 1: Parameter ranges for different levels of heterogeneity in data generation.

τ	Heterogeneity	RMSN	CRN	CT	ASIDE	min gain (%)
5	0	0.75 (0.26)	0.62 (0.09)	0.87 (0.21)	**0.32 (0.05)**	93.75
	1	1.80 (0.53)	1.35 (0.06)	1.57 (0.12)	**0.80 (0.16)**	68.75
	2	1.19 (0.06)	1.09 (0.04)	1.31 (0.29)	**0.66 (0.12)**	65.15
	3	1.14 (0.03)	1.16 (0.02)	1.29 (0.06)	**0.68 (0.02)**	67.65
10	0	0.91 (0.61)	0.69 (0.10)	0.87 (0.15)	**0.23 (0.01)**	200.00
	1	1.84 (0.36)	1.19 (0.02)	1.69 (0.16)	**0.59 (0.04)**	101.69
	2	1.40 (0.32)	1.20 (0.14)	1.37 (0.19)	**0.60 (0.03)**	100.00
	3	1.31 (0.22)	1.26 (0.11)	1.44 (0.11)	**0.57 (0.05)**	121.05
15	0	1.74 (0.47)	0.93 (0.25)	1.20 (0.24)	**0.31 (0.02)**	200.00
	1	1.38 (0.65)	0.98 (0.29)	1.11 (0.12)	**0.39 (0.05)**	151.28
	2	1.28 (0.26)	1.15 (0.22)	1.34 (0.17)	**0.58 (0.07)**	98.28
	3	1.00 (0.10)	1.02 (0.07)	1.35 (0.11)	**0.35 (0.01)**	185.71
20	0	1.60 (0.92)	1.11 (0.05)	1.02 (0.15)	**0.25 (0.01)**	308.00
	1	1.89 (1.02)	1.21 (0.13)	1.66 (0.20)	**0.52 (0.02)**	132.69
	2	1.38 (0.33)	1.20 (0.33)	1.29 (0.18)	**0.39 (0.02)**	207.69
	3	1.23 (0.17)	1.06 (0.13)	1.45 (0.13)	**0.37 (0.02)**	186.49

Table 2: RMSE for different heterogeneity levels and projection horizon in tumor dataset. min gain = |RMSE of ASIDE - RMSE of the second best model| / RMSE of the second best model.

Prediction horizon: Compared to existing works that integrate history into an initial condition for forecasting in time, ASIDE's use of history to adapt the dynamics function is expected to be able to carry heterogeneity information forward for a longer prediction horizon. To delineate this benefit, we considered prediction horizons of different lengths ($\tau = 5 - 20$). Furthermore, within a given prediction horizon, we examined the per-step prediction RMSE over time in addition to the commonly-considered average RMSE.

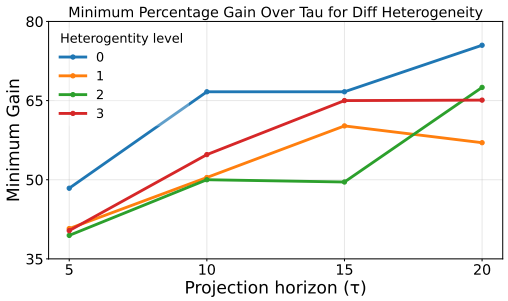
4.1.2 RESULTS

Table 2 summarizes the results for the normalized RMSE for all models considered, under different levels of heterogeneity and over different prediction horizons: the last column further summarizes the % gain of improvements obtained by ASIDE over the next best models in each experiment setting. As shown, ASIDE was able to provide significant margins of improvement over the included baselines across all prediction horizons and all levels of heterogeneity. As further highlighted in the last column of Table 2 and summarized in Fig. 2a, this improvement overall improved as τ increased for all levels of heterogeneity. Fig. 2a further shows the per-step RMSE prediction accuracy of ASIDE when trained to predict for a horizon of $\tau = 20$ at the highest level of heterogeneity, where ASID consistently attained the lowest prediction accuracy over time *versus* all baseline models. This provided strong evidence that 1) ingesting history into the latent dynamics is stronger than ingesting history into an initial condition for predicting over longer horizons, and 2) adaptive dynamics is important for addressing data heterogeneity.

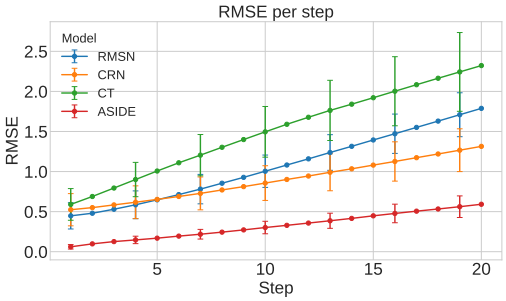
To further demonstrate the benefits of separable dynamics enabled by ASIDE, we examined the RMSE for intervention-free and interventional segments of the test samples separately. As shown in the example in Table 3, ASIDE delivered significantly improved RMSE for both predicting intrinsic dynamics and its changes under radiation or chemotherapies. While most baseline methods experienced an increase in error when predicting the effect of radiation therapies, potentially due to its relatively small effect compared to the growth of tumor due to intrinsic dynamics, ASIDE was able

τ	Heterogeneity	RMSE on	RMSN	CRN	CT	ASIDE
20	3	intrinsic growth	1.30 (0.18)	1.15 (0.13)	1.14 (0.17)	0.42 (0.02)
		chemo steps only	1.21 (0.18)	1.04 (0.10)	1.55 (0.12)	0.44 (0.03)
		radio steps only	1.50 (0.25)	1.38 (0.19)	1.65 (0.22)	0.44 (0.02)

Table 3: Separate RMSEs for different dynamics for all models on datasets with heterogenity level 3 and a projection horizon of 20.



(a) Percentage gain over the second-best model across τ for different heterogeneity levels.



(b) Stepwise RMSE for all the models when trained for heterogeneity level 3 and $\tau = 20$

to maintain consistently prediction errors across the various dynamics. Fig. 3 provides examples of individual trajectories.

A notable observation in Table 2 was that for most baselines, their forecasting accuracy actually improved as τ or the heterogeneity level increased (from level 1 to level 2-3). This was potentially because a higher heterogeneity level in treatment effect parameters or time-window τ have resulted in a dataset with smaller tumor volumes (and hence small RMSE) (see Fig. 4 in Appendix B.1.1).

4.2 Real Data Experiments

Data: MIMIC-III (Johnson et al., 2023) contains electronic health records of ICU patients. Following Melnychuk et al. (2022), we considered covariates $\mathbf{x}_t$ as 25 vitals and 3 static features, outcome y_t as diastolic blood pressure, and treatment $\mathbf{a}_t$ as vasopressor and ventilation. Because of the treatment bias inheret within observational data, all baselines were included with their original mechanisms for addressing confounding on.

Results: Table 4 summarizes the test RMSE results by all models in MIMIC-III. As shown, even without any specific mechanisms to addressing time-varying confounding, ASIDE was able to deliver an margin of improvement that was statistically significant over the second best models across all prediction horizons, except at $\tau = 5$ where $p = 0.104$ (*paired-t* tests). Similarly, while all models RMSE increased as τ increased, ASIDE demonstrated the least deterioration (10%) compared to the baselines (ranging 13% to 22%). Fig. 5 provides the RMSE metric at each step over the prediction horizon.

4.3 Ablation

To further understand the significant gain of ASIDE over considered baselines in its prediction accuracy, we conducted an ablation study on the key components of ASIDE: separated latent dynamics components, meta-learning for adaptive dynamics, and progressive learning to ensure separation. This set of experiments was conducted considering only radiation therapy without the presence of chemotherapy: in the absence of the latter, the average tumor value in the dataset had increased which resulted in an increased RMSE in all models, as shown in Table 5 . For ablation, we started with a global neural ODE (Model 1) in which a single global ODE function was learned at the latent space. that was not conditioned on any history $\mathcal{H}$. This model delivered comparable performance to the other baselines considered. As we decomposed this single neural ODE into a formulation with separated intrinsic and interventional dynamics similar to equation 1 but without the adaptive

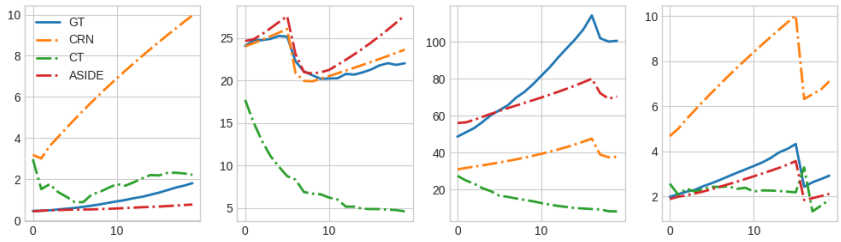


Figure 3: Some visual examples of forecast data for the different models

Model	$\tau = 5$	$\tau = 10$	$\tau = 15$	$\tau = 20$
RMSN	12.28 (1.88)	12.58 (0.87)	14.93 (1.40)	14.09 (0.93)
CRN	9.82 (0.10)	10.46 (0.15)	11.00 (0.12)	11.46 (0.11)
CT	9.66 (0.11)	10.23 (0.15)	10.58 (0.19)	10.92 (0.22)
ASIDE	9.55 (0.11)	10.02 (0.13)*	10.31 (0.15)*	10.56 (0.16)*

Table 4: Test RMSEs for different prediction projection horizon in MIMIC-III dataset.

component (Model 2), a moderate improvement was obtained without statistical significance. The addition of meta-learning to allow adaptation of the latent dynamics, and the addition of progressive training strategy (Model 3) achieved significant reduction of RMSE.

5 Conclusions & Discussion

We present ASIDE, a novel framework for learning adaptive and separable interventional dynamics with progressive meta-learning. In contrast to black-box based approaches for modelling, this approach leverages inductive biases and meta-learning based approaches to learn latent dynamics and adapt it to pass history. We validate our models on simulation data along with a real world dataset. The major contributions were also validated using an ablation study.

This model does have some limitations. The model is not particularly designed for the handling of treatment selection biases. Incorporating challenges brought forth by such biases is a future work necessary. Apart from that, ASIDE is good at handling heterogeneity across samples, however, heterogeneity might also be caused by parameters of the dynamics changing with time. Handling such heterogeneity is another important next step for future.

Model	Separate Dynamics	Meta Learning with Progressive Training	RMSE (avg)
RMSN	✗	✗	2.25
CRN	✗	✗	2.24
CT	✗	✗	2.27
Model 1	✗	✗	2.17 (0.05)
Model 2	✓	✗	1.96 (0.15)
Model 3	✓	✓	1.06 (0.08)

Table 5: Ablation results for setting with only radiotherapy, $heterogenity = 3$ and $\tau = 5$

REFERENCES

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 Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021. URL https://openreview.net/forum?id=qBl8hnwR0px.

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.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/b618c3210e934362ac261db280128c22-Abstract.html.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.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].

MA Hernan and JM Robins. Causal Inference: What If. Chapman & Hall/CRC, 2020.

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

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

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.

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.

A SUMMARY OF MATHEMATICAL NOTATIONS

Symbol	Definition
$\mathbf{x_t}$	Observed covariates at time t
$\mathbf{z_t}$	Latent state at time t
$\mathbf{a_t}$	External intervention(s) applied at time t
$\mathbf{y_t}$	Observed outcome at time t
$f_\theta(\mathbf{z_t}, \mathbf{a_t})$	Composite latent dynamics function
$\hat{\mathbf{z}}_t$	Estimated latent state from encoder given short past history
$\hat{\mathbf{y}}_t$	Predicted outcome at time t
$f_{\theta_{\text{int}}}(\mathbf{z}_t; \mathbf{c}_{\text{int},t})$	Intrinsic dynamics function
$f_{\theta^k_{\text{ext}}}(\mathbf{z}_t, a^k_t; \mathbf{c}^k_{\text{ext},t})$	Effect of intervention k on dynamics
$\mathbf{c}_{\text{int},t}$	Context embedding for intrinsic dynamics
$\mathbf{c}^k_{\text{ext},t}$	Context embedding for intervention k
$\mathcal{I}^k(a^k_t)$	Indicator function for intervention k at time t
$g_\eta(\mathbf{z}_t, \mathbf{a}_t)$	Emission function to observable outcome
τ	Forecasting horizon length
$\mathcal{H}_{1:t}$	History of covariates, interventions, and outcomes up to time t
$\mathbf{s}_{\text{int}}$	History segment without interventions
$\mathbf{s}^k_{\text{ext}}$	History segment with intervention k
$\mathbf{s}^k_{\text{ext,CF}}$	Counterfactual segment without intervention k
Ψ_{ϕ_z}	Neural encoder function for initial latent states
$\Psi_{\phi_{\text{int}}}$	Neural encoder function for intrinsic dynamics
$\Psi_{\phi^k_{\text{ext}}}$	Neural encoder function for external intervention
$\mathcal{M}_{\text{int}}$	Meta-encoder for intrinsic dynamics
$\mathcal{M}^k_{\text{int}}$	Meta-encoder for intervention type k
V_t	Tumor volume at time t
ρ	Intrinsic tumor growth rate
K	Carrying capacity of the tumor growth model
β_c	Effect of chemotherapy
C_t	Chemotherapy dose at time t
α	Linear effect of radiotherapy
β	Quadratic effect of radiotherapy
r_t	Radiotherapy dose at time t
ϵ	Random noise term in tumor dynamics

Table 6: Summary of Mathematical Notations

B DETAIL OF EXPERIMENTS AND ADDITIONAL RESULTS

B.1 DATASET DETAILS

Geng et al. (2017) provide the dynamic model for the growth of non-small cell lung cancer under radiotherapy and chemotherapy interventions. This model is based on assuming that the intrinsic growth follows a Gompertz growth model. Similarly, effect of radiothrapy is taken to be linear quadratic (LQ) model and for chemotherapy is a log-cell kill model. This results in the trajectory of the tumor volume V to be defined by a differential equation defined as:

$$\frac{dV}{dt} = V * (\rho \log(\frac{K}{V}) - \beta_c C(t) - (\alpha R(t) + \beta R(t)^2)) \tag{7}$$

For the growth model, Gompertz is a common way to model a more general form of logistic growth. ρ represents the cell specific growth rate and K is the carrying capacity. The value of K is kept constant at $K = 14137.167$ across the simulation whereas ρ is varied based on distribution suggested in (Geng et al., 2017) with changes for different heterogeneity setting as mentioned in 4.1.1.

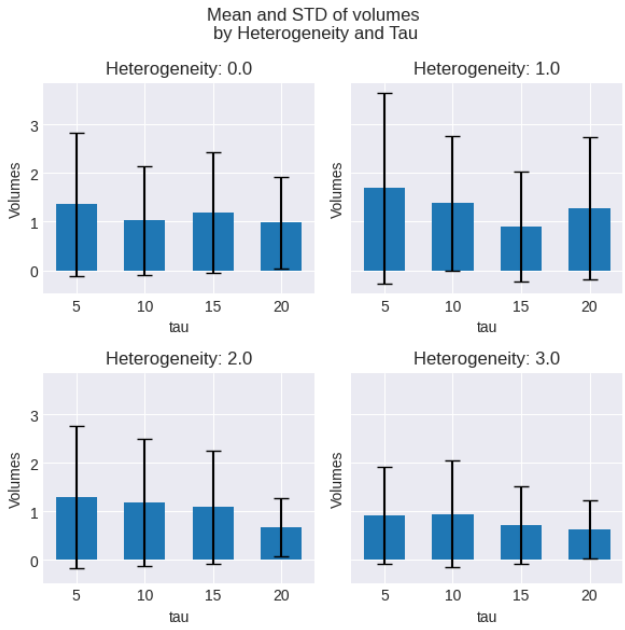


Figure 4: Volume distribution across τ for different heterogeneity

For the chemotherapy part of the model, the value of $C(t)$ represents the drug concentration. Drug once administered is assumed to decay with an half life of 1 timestep. based on the clinical practice of administering vinblastine at 5 mg/m^3 per week, the value of $C(t) = 5$ is introduced at anytime the treatment assignment dictates chemotherapy administration. The drug constantly decays with half life of 1 timestep.

For radiotherapy, the parameters $\alpha, \beta : \alpha/\beta = 10$ are made to effect the cell volume at only the time of administration, and effects disappear immediately disappears. $R(t) = 5\ Gy$ at time of administration is kept constant to simulate practice.

B.1.1 Distribution of volume across Heterogeneity and Projection Horizon

The distribution of volume across different stages and different projection horizon vary greatly, as shown in 4. Since the volume of ground truth are different, the RMSE calculated for these varying setting will also be different and not easily comparable. As clearly seen, volume distribution is higher for $heterogenity = 1$, thus RMSE are expected to be higher for this setting across all τ.

B.2 Implementation details

- Latent Encoder: $\Psi_{\phi_z}(\mathbf{x}_{t-l-1:t-1})$
  - Input: $x_{t-l:t-1}, a_{t-l:t-1}$ with $l = 3$
  - Model structure: MLP with 3 layers
  - Output: z_t
- Metamodel for intrinsic dynamics equation 2
  - Input: $\mathbf{s}_{\text{int}}$
  - Model structure: MLP with 8 layers
  - Output: $\mathbf{c}_{\text{int,t}}$
- Metamodel for intervention dynamics equation 4
  - Input: $\mathcal{H}_{1:t}, \mathcal{I}^k_{1:t}$
  - Model structure: CNN with 8 layers
  - Output: $\mathbf{c}^{\mathbf{k}}_{\text{ext,t}}$

- Dynamics Intrinsic Function
  - Input: $z_t,\ \mathbf{c}_{\text{int,t}}$
  - Model: MLP 9 layer, 16 hidden units
  - Output: $\frac{dz}{dt} = f_{\text{int}}$
- Dynamics Intervention Function
  - Input: $z_t,\ \mathbf{c}^{\mathbf{k}}_{\text{ext,t}}$
  - Model: MLP 9 layer, 16 hidden units
  - Output: $\frac{dz}{dt} = f^k_{\text{ext}}$
- Emission decoder
  - Input: z_t
  - Model: MLP 2 layer, 16 hidden units
  - Output: $y_{t+1:t+\tau}$

Encoder		
Layer	**Output Dim**	**Details**
Linear + ELU	8	in_features = $3 \cdot F$, out_features = 8
Linear + ELU	8	in_features = 8, out_features = 8
Linear	8	in_features = 8, out_features = z_{init}

Meta-encoder (intrinsic)		
Layer	**Output Dim**	**Details**
Linear + ELU	H	in_features = d_{in}, out_features = H
(Linear + ELU)*N	H	in_features = H, out_features = H
Linear	c_{int}	in_features = H, out_features = c_{int}

Meta-encoder (intervene)		
Layer	**Output Dim**	**Details**
Conv1D + ELU	H	in_features = $d_{in} + d_{mask}$, out_features = H
(Conv1D + ELU)*N	H	in_features = H, out_features = H
Linear	c_{int}	in_features = H, out_features = c_{ext}

Intrinsic Dynamics		
Layer	**Output Dim**	**Details**
Linear + ELU	H	in_features = $c_{int} + z_{init}$, out_features = H
(Linear + ELU)*N	H	in_features = H, out_features = H
Linear	z_t	in_features = H, out_features = z_t

Extrinsic Dynamics		
Layer	**Output Dim**	**Details**
Linear + ELU	H	in_features = $c_{ext} + z_{init} + d_{int}$, out_features = H
(Linear + ELU)*N	H	in_features = H, out_features = H
Linear	z_t	in_features = H, out_features = z_t

Table 7: Architecture of the encoder, meta-encoders and dynamics networks.

Architecture of codes

B.3 Description of Baselines

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: $f_{p_t}(\mathbf{a_t}|\mathbf{a_{1:t-1}})$ and another RNN models $f_{p_h}(\mathbf{a_t}|\mathbf{a_{1:t-1}}, \mathcal{H}_{1:t-1})$. 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 learnt above is then used to make autoregressive predictions for $y_{t+1:\tau}$ and is modeled by another RNN: $y_{i,t+1:\tau} = f_d(e_{i,t}, w_i, y_{i,t})$

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 modelling the weights by RNNs. The idea behind CRN, is to learn balanced representations $r_i(\mathcal{H}^i_{1:t-1})$ which is invariant to treatment assignment. This is done by building to separate heads to the RNN models, one to predict the treatment $\mathbf{a_{t+1}} = G_A(r_i(\mathcal{H}^i_{1:t-1}))$ and another to predict outcome $Y_{t+1} = G_Y(r_i(\mathcal{H}^i_{1:t-1}), \mathbf{a_t})$ at each time point. An adversarial loss is used to learn this model.

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

Causal Transformer (CT): CT (Melnychuk et al., 2022) extends the idea of CRN using 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 part of history to make future predictions. As in CRN, they learn the transformer by an adversarial component on top of transformer. The transformer is tasked to learn a balanced represenattion $r_i(\mathcal{H}^i_{1:t-1})$ such that it is invariant of treatment predictions. Two heads are added on top of transformer: one to predict the treatment $\mathbf{a_{t+1}} = G_A(r_i(\mathcal{H}^i_{1:t-1}))$ and another to predict outcome $Y_{t+1} = G_Y(r_i(\mathcal{H}^i_{1:t-1}), \mathbf{a_t})$ at each time point. An adversarial loss known as Counterfactual Domain Confusion (CDC) loss is used to learn the balanced representation.

C Additional Results

C.1 Separate RMSE

Separate RMSE across intrinsic dynamcis and extrinsic dynamics can be found in table 8 for RMSN (Lim, 2018), table 9 for CRN (**?**), table 10 for CT (Melnychuk et al., 2022), and table 11 for ASIDE

RMSE Type	Heterogeneity	5	10	15	20
Growth	0	0.86 (0.30)	1.01 (0.70)	1.96 (0.54)	1.78 (1.03)
	1	2.03 (0.64)	2.11 (0.44)	1.52 (0.74)	2.16 (1.18)
	2	1.30 (0.08)	1.58 (0.38)	1.42 (0.30)	1.55 (0.39)
	3	1.20 (0.03)	1.47 (0.27)	1.07 (0.12)	1.30 (0.18)
Chemo	0	0.84 (0.30)	0.95 (0.67)	1.86 (0.52)	1.72 (1.05)
	1	1.58 (0.71)	1.85 (0.45)	1.39 (0.74)	1.97 (1.22)
	2	0.87 (0.12)	1.34 (0.52)	1.36 (0.36)	1.46 (0.41)
	3	0.73 (0.07)	1.20 (0.38)	1.02 (0.15)	1.21 (0.18)
Radio	0	0.86 (0.30)	0.97 (0.67)	1.96 (0.54)	1.80 (1.11)
	1	3.02 (0.54)	2.85 (0.45)	1.87 (0.77)	2.59 (1.10)
	2	1.98 (0.09)	2.11 (0.24)	1.70 (0.28)	1.85 (0.34)
	3	2.04 (0.05)	2.09 (0.11)	1.23 (0.13)	1.50 (0.25)

Table 8: Separate RMSE for the different segments for RMSN

C.2 Per Step RMSE for MIMIC-III

Fig 5 shows per-step RMSE for MIMIC-III

RMSE Type	Heterogeneity	5	10	15	20
Growth	0	0.68 (0.10)	0.73 (0.12)	1.04 (0.28)	1.24 (0.05)
	1	1.49 (0.06)	1.36 (0.04)	1.10 (0.34)	1.39 (0.17)
	2	1.18 (0.04)	1.34 (0.14)	1.29 (0.26)	1.36 (0.38)
	3	1.22 (0.03)	1.40 (0.12)	1.10 (0.08)	1.15 (0.13)
Chemo	0	0.65 (0.12)	0.72 (0.14)	0.91 (0.28)	1.18 (0.15)
	1	0.99 (0.09)	0.98 (0.04)	0.94 (0.38)	1.16 (0.18)
	2	0.77 (0.07)	1.00 (0.23)	1.19 (0.35)	1.16 (0.39)
	3	0.77 (0.06)	1.08 (0.15)	1.06 (0.11)	1.04 (0.10)
Radio	0	0.64 (0.11)	0.72 (0.14)	0.99 (0.30)	1.23 (0.11)
	1	2.53 (0.15)	2.19 (0.07)	1.38 (0.14)	1.96 (0.22)
	2	1.78 (0.06)	1.84 (0.10)	1.54 (0.20)	1.67 (0.34)
	3	2.06 (0.03)	2.04 (0.04)	1.24 (0.10)	1.38 (0.19)

Table 9: Separate RMSE for the different segments for CRN

RMSE Type	Heterogeneity	5	10	15	20
Growth	0	1.00 (0.24)	0.98 (0.17)	1.36 (0.27)	1.14 (0.17)
	1	1.80 (0.14)	1.96 (0.18)	1.27 (0.14)	1.95 (0.24)
	2	1.43 (0.34)	1.56 (0.23)	1.53 (0.19)	1.49 (0.21)
	3	1.38 (0.07)	1.63 (0.13)	1.49 (0.12)	1.63 (0.15)
Chemo	0	0.95 (0.26)	0.97 (0.18)	1.34 (0.27)	1.13 (0.15)
	1	1.32 (0.15)	1.67 (0.18)	1.11 (0.13)	1.79 (0.22)
	2	1.10 (0.27)	1.29 (0.20)	1.28 (0.21)	1.37 (0.18)
	3	0.92 (0.07)	1.34 (0.13)	1.46 (0.10)	1.55 (0.12)
Radio	0	0.98 (0.22)	1.00 (0.18)	1.39 (0.29)	1.16 (0.16)
	1	2.54 (0.18)	2.47 (0.40)	1.37 (0.20)	2.23 (0.44)
	2	2.11 (0.38)	2.02 (0.20)	1.67 (0.19)	1.73 (0.16)
	3	2.19 (0.08)	2.18 (0.10)	1.50 (0.18)	1.65 (0.22)

Table 10: Separate RMSE for the different segments for CT

D Societal Impact

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

E LLM Usage

LLM tools such as ChatGPT, DeepSeek were used in limited capacity for refining this paper. Tasks for which LLMs were used are mostly spellchecks and grammar checks. Apart from that, LLMs were also used to generate table templates and filler codes for some plots in this paper.

F Source Code

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

RMSE Type	Heterogeneity	5	10	15	20
Growth	0	0.36 (0.06)	0.26 (0.01)	0.35 (0.03)	0.28 (0.00)
	1	0.92 (0.22)	0.68 (0.05)	0.44 (0.06)	0.60 (0.02)
	2	0.73 (0.15)	0.68 (0.04)	0.65 (0.08)	0.45 (0.02)
	3	0.70 (0.03)	0.64 (0.06)	0.39 (0.01)	0.42 (0.02)
Chemo	0	0.37 (0.07)	0.26 (0.01)	0.32 (0.03)	0.29 (0.02)
	1	0.61 (0.16)	0.57 (0.12)	0.39 (0.03)	0.53 (0.03)
	2	0.49 (0.04)	0.42 (0.04)	0.50 (0.06)	0.39 (0.02)
	3	0.24 (0.02)	0.47 (0.09)	0.40 (0.02)	0.44 (0.03)
Radio	0	0.33 (0.03)	0.25 (0.01)	0.32 (0.02)	0.29 (0.01)
	1	1.30 (0.02)	0.93 (0.06)	0.55 (0.03)	0.79 (0.04)
	2	1.00 (0.26)	0.99 (0.07)	0.87 (0.17)	0.56 (0.03)
	3	1.38 (0.05)	1.02 (0.13)	0.38 (0.04)	0.44 (0.02)

Table 11: Separate RMSE for the different segments for ASIDE

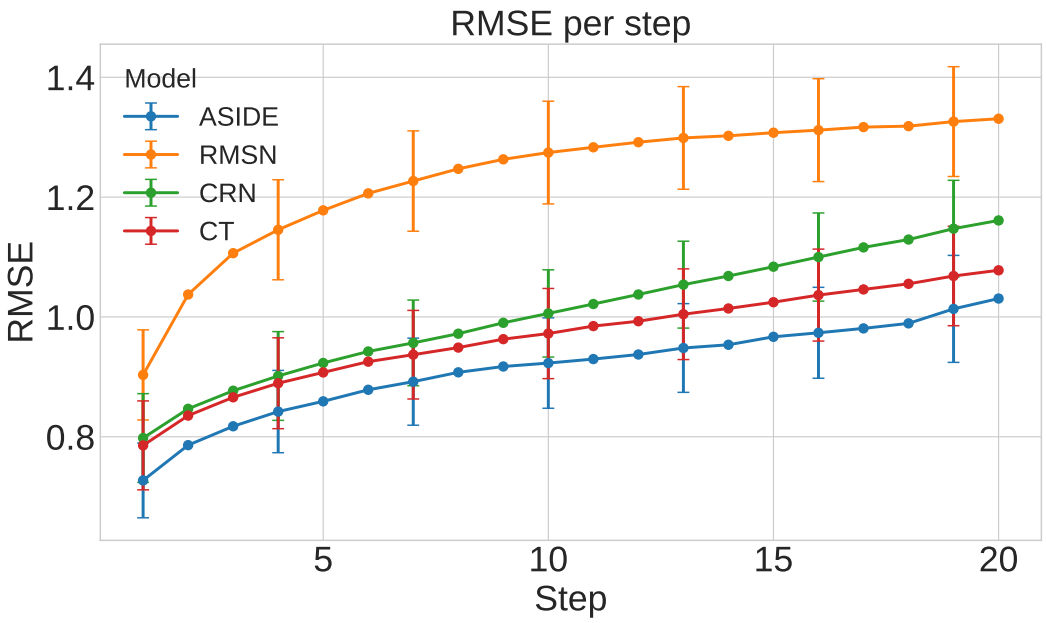


Figure 5: RMSE per step for MIMIC-III for different models