IDENTIFYING NEURAL DYNAMICS USING NONLINEAR INTERVENTIONAL STATE SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

State-space models (SSMs) recover low-dimensional latent dynamics from high-dimensional observations, but latents are generally not identifiable from observational data alone. Interventional state-space models (iSSMs) use perturbations to break observational symmetries and identify parameters up to permutation, scaling, and shift. However, prior work restricts identifiability to linear latent dynamics and piecewise-linear emissions, excluding many behaviors characteristic of neural circuits. We extend iSSMs to a substantially broader model class: we prove that (i) affine identifiability holds under general injective emission functions and Gaussian latent transitions, and (ii) with interventions, identifiability up to permutation, scaling, and shift is preserved under general nonlinear latent dynamics. We validate our theory on simulated FitzHugh-Nagumo dynamics, and then demonstrate two applications: adjudicating between competing models of decision-making; and predicting responses to unseen micro-stimulation in macaque and human neural recordings, with proof-of-concept generalization to a held-out stimulation site in the latter. Finally, simulations yield practical guidance for designing stimulation protocols that improve identifiability.

1 INTRODUCTION

In biology, and especially neuroscience, experimental interventions are central to answering causal questions. Neuroscientists routinely perturb circuits using pharmacology (Bolz et al., 1989), electrophysiology (Cohen & Newsome, 2004; Histed et al., 2013), photostimulation (Fenno et al., 2011; Adesnik & Abdeladim, 2021), and more recently brain-machine interfaces (Motiwala et al., 2026).

Classical perturbation experiments establish causal involvement – e.g., by silencing a brain region and measuring behavioral deficits. Many core questions in systems neuroscience, however, require mechanistic explanations that specify not only which variables are involved, but also how they interact to generate observed dynamics. Such explanations could resolve longstanding debates about circuit motifs of working memory (Javadzadeh et al., 2024; Daie et al., 2023; Inagaki et al., 2019), e.g., or dynamical regimes underlying perceptual decision-making (Galgali et al., 2023).

A standard approach to such mechanistic questions is to fit state-space models (SSMs) to neural recordings. SSMs posit a low-dimensional latent state with explicit dynamics, and an emission map linking latents to observed activity. Without identifiability, however, distinct mechanisms can produce identical observations, so a fitted SSM cannot generally be viewed as the true causal mechanism (Qian et al., 2024; Maheswaranathan et al., 2019; Lipshutz et al., 2024; Galgali et al., 2023).

Recent work by Nejatbakhsh & Wang (2025) addressed this gap with interventional state-space models (iSSMs), a framework for fitting SSMs to interventional data with theoretical guarantees of unique recovery up to benign symmetries. Although prior work has fitted SSMs to perturbation data, conventional SSMs remain subject to observational symmetries that permit spurious mechanisms. iSSM breaks those symmetries by fixing targeted latents to external inputs via do-interventions (Fig. 1a), severing causal dependencies from past latents and enabling identification up to permutation, coordinate-wise scaling, and shift (Ahuja et al. (2023) for a review of interventional inference).

While iSSM marked an important step, its guarantees only apply when the true system lies within the model's hypothesis class. Its identifiability results cover linear latent dynamics with piecewise-linear emission functions – restrictive for neural circuits, which commonly exhibit nonlinear dy-

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.

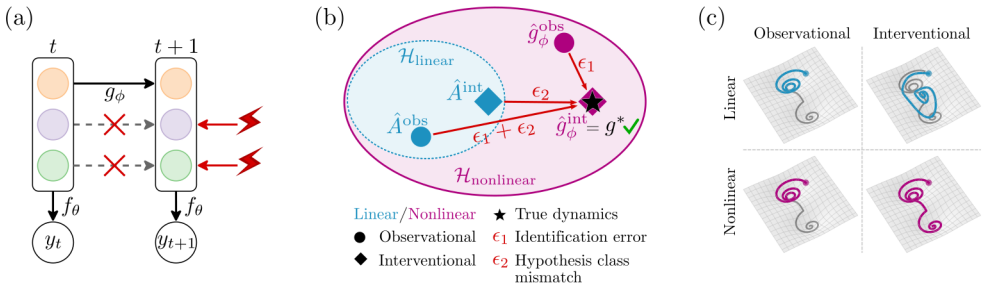

Figure 1: **Overview: ingredients for identifying latent dynamics in state-space models. (a)** iSSM leverages do-interventions to identify latent dynamics up to permutation, coordinate-wise scaling, and shift. **(b)** Two independent failure modes prevent SSMs from identifying the true dynamics g^*: identification error ϵ_1, resolved by interventions, and hypothesis class mismatch ϵ_2, resolved by a nonlinear model class. **(c)** Latent trajectory recovery across model class (rows) and data regime (columns): only a nonlinear model trained on interventional data recovers the true trajectory (gray).

namics such as thresholding, bistability, and limit cycles. Specifically, we distinguish two inference failure modes (Fig. 1b): (1) identification error ϵ_1 from observational symmetries, resolved by interventional data; and (2) hypothesis class mismatch ϵ_2 from model misspecification, requiring a sufficiently expressive nonlinear model class. Fig. 1c illustrates both errors on a bistable-well system: without interventions, even a nonlinear model fails to recover transitions between attractors (ϵ_1); with interventions, a linear model still misspecifies the true dynamics (ϵ_2).

In this work, we generalize iSSM to nonlinear latent dynamics and general injective emission functions, addressing the principal open limitation of prior theory, and mitigating both inference failure modes – ϵ_1 through interventions and ϵ_2 through a nonlinear hypothesis class. We validate the framework in simulations, demonstrate applications to macaque and human neural recordings under microstimulation, and distill practical guidance for experimental design to improve identifiability.

2 Related Work

This paper builds on the literature on identifiability in time series models. Broadly, this literature can be divided into theoretical guarantees obtained with or without interventional data.

Identifiability in time series models without interventions. One line of work establishes identifiability purely from the statistical structure of latent processes. Early nonlinear ICA results exploited changing source distributions across time segments (Hyvarinen & Morioka, 2016) or latent discrete Markov states (Hälvä & Hyvarinen, 2020; Hälvä et al., 2021), but treated latents as non-interacting; Yao et al. (2022) later allowed general nonlinear time-delayed coupling, identifying latents up to permutation and component-wise transformations when transition distributions vary sufficiently with the past. In a parallel line on switching dynamical systems, Balsells-Rodas et al. (2024) identified coupled latents with regime-switching nonlinear Gaussian transitions up to an affine map under piecewise-linear emissions, and Balsells-Rodas et al. (2026) extended this to multi-lag dependencies and refined the map to permutation, scaling, and shift when noise variances differ across regimes. Across these observational approaches, identifiability relies on non-interacting latents or transition noise that is non-Gaussian or state-/regime-dependent; for coupled dynamics with additive Gaussian noise of constant variance – the standard SSM assumption – temporal structure alone cannot resolve rotational symmetries (Yao et al., 2022; Lippe et al., 2022).

Identifiability in time series models via interventions. A second line obtains identifiability from interventions or exogenous inputs acting on latents. CITRIS (Lippe et al., 2022) and iCITRIS (Lippe et al., 2023) identify latent causal variables with general nonlinear stochastic dynamics from intervened sequences, but only their intervention-dependent parts, up to component-wise transformations, and only when intervention targets are known. Zhang & Xie (2025) identify latents up to scaling under deterministic linear or control-affine dynamics in a known controllable canonical form with sparse input matrices. Interventional state-space models (iSSMs) (Nejatbakhsh & Wang, 2025) make

the target map learnable and identify full latents up to permutation, scaling, and shift, yet only for linear latent dynamics and piecewise-linear emissions inherited from Balsells-Rodas et al. (2024). The above approaches to interventional guarantees for latent dynamics therefore either require known intervention targets or restrict the dynamics to linear or control-affine forms and – in the case of iSSM – piecewise-linear emissions.

Building on this literature, we relax iSSM's piecewise-linear emission and linear-dynamics restrictions while retaining permutation-scale-shift identifiability under hard latent interventions.

3 METHODS

3.1 GENERAL INTERVENTIONAL STATE SPACE MODELS

We consider a general class of interventional state-space models (iSSMs) that capture latent dynamical systems under interventions. Let $x_t \in \mathbb{R}^D$ denote latent variables and $y_t \in \mathbb{R}^N$ the observations at time t. The model is defined as:

$$y_t \sim P(y_t \mid f_\theta(x_t)), \tag{1}$$

$$x_{t+1} = m_t \odot g_\phi(x_t) + Bu_t + \epsilon_t, \tag{2}$$

where f_θ is an injective emission function, g_ϕ is the latent transition function, $u_t \in \mathbb{R}^M$ denotes interventional inputs, and $\epsilon_t \sim \mathcal{N}(0, Q)$ is Gaussian process noise with diagonal covariance Q. The matrix $B \in \mathbb{R}^{D\times M}$ maps interventions to latent variables. The inputs u_t are exogenous and observed, i.e., independent of the latent states and of the process noise. The binary mask $m_t \in \{0,1\}^D$ encodes which latent variables are not intervened on at time t, i.e., it is 1 where $(Bu_t)_i = 0$ and otherwise 0. This masking mechanism implements *hard interventions* in latent space, where intervened variables are disconnected from their causal parents in the past. While we use a linear input map Bu_t in all our experiments, we note that our theory extends to arbitrary deterministic input maps $h(u_t)$, with $m_t = \mathbf{1}\{h(u_t) = 0\}$, under a mild support condition on h (App. A.5).

The above formulation reduces to the original iSSM when $g_\phi(x_t) = Ax_t$ is linear and the emission function satisfies additional structural constraints.

3.2 THEORETICAL RESULTS

We provide sufficient conditions for identifiability of the latent dynamics in the proposed nonlinear iSSM. The goal is to recover the latent transition and emission functions up to permutation, scaling, and shift. The main challenge is establishing this under nonlinear latent dynamics and general injective emissions.

Our approach follows the original iSSM proof and proceeds in three steps: (1) identify the distribution of noiseless latent representations from observed data; (2) identify latent variables and transition dynamics up to an affine transformation; and (3) use interventional structure to further restrict the equivalence class to permutation, scaling, and shift. We keep Step (1) from prior work and extend Steps (2) and (3) to general injective emission functions and nonlinear latent dynamics.

We begin by identifying the distribution of the noiseless latent representations from the observed data. Let $\bar{y}_t = f_\theta(x_t)$ denote the noiseless signal. Under the following standard regularity condition, this distribution can be recovered from the observed data.

Assumption 1 (Bounded completeness)**.** The conditional distribution $P(y_t \mid \bar{y}_t)$ is bounded complete in the following measure-theoretic sense: for any finite signed measure μ on $\mathbb{R}^N$,

$$\int P(y_t \mid \bar{y})\, d\mu(\bar{y}) = 0 \quad \text{for all } y_t \quad \Rightarrow \quad \mu = 0.$$

This condition is satisfied by many commonly used observation models, including Gaussian additive noise and Poisson rate observations.

Proposition 1 (Recovery of noiseless observations (Nejatbakhsh & Wang, 2025))**.** *Under Assumption 1, the distribution of $\{\bar{y}_t\}_{t=1}^T$ is uniquely identifiable from the observed distribution of $\{y_t\}_{t=1}^T$.*

Proposition 1 reduces identifiability to recovering latent variables from noiseless observations. This is needed because Theorem 2 requires noiseless observations $\{\bar{y}_t\}$, but experimental neural data is usually noisy.

For Step (2), affine identifiability, prior work establishes this under injective piecewise-linear emissions (Nejatbakhsh & Wang, 2025). We show the same conclusion for general injective emissions under Gaussian latent transitions and mild regularity.

Theorem 2 (Affine identifiability)**.** *Consider the distribution of $\{\bar{y}_t\}_{t=1}^T$, where $\bar{y}_t = f_\theta(x_t)$. Assume the following conditions hold:*

(i) ***Observation model and regularity.*** *The emission function f_θ is injective and continuously differentiable, with continuously differentiable inverse on its image. The latent transition function g_ϕ is continuously differentiable.*

(ii) ***Gaussian latent transitions.*** *The latent dynamics are Gaussian and follow a first-order Markov process:*

$$x_t \mid x_{t-1} \sim \mathcal{N}(g_\phi(x_{t-1}), Q), \quad Q \succ 0.$$

(iii) ***Local non-degeneracy of dynamics.*** *There exists $x^* \in \mathbb{R}^D$ such that $J_{g_\phi}(x^*)$ is full rank.*

Let $(f_{\hat{\theta}}, g_{\hat{\phi}}, \hat{Q})$ be another model parametrisation under assumptions (i-iii) that induces the same distribution on $\{\bar{y}_t\}_{t=1}^T$, where $\bar{y}_t = f_{\hat{\theta}}(\hat{x}_t)$. Then there exists an invertible affine transformation $\psi(x) = Mx + c$ such that $\hat{x}_t = \psi(x_t)$ for all t, where $M \in \mathbb{R}^{D\times D}$ is invertible and $c \in \mathbb{R}^D$.

The proof of Theorem 2 is in App. A.3.

We now show that the intervention-based reduction from affine to block identifiability extends to the general nonlinear iSSM. This step follows the same high-level strategy as prior work on iSSMs Nejatbakhsh & Wang (2025). The extension from linear to nonlinear latent dynamics is achieved by the additional assumption of independent latent noise, i.e., diagonal process noise covariance.

Theorem 3 (Block identifiability under interventions)**.** *Assume the conditions of Theorem 2 hold, so that the latent variables are identifiable up to an affine transformation. Suppose additionally that the following conditions hold:*

(i) ***Faithfulness.*** *The latent dynamical system is faithful, i.e., there do not exist non-zero vectors $v, w \in \mathbb{R}^D$ such that $v^\top x_{t+1} \perp w^\top x_t$ for all t.*

(ii) ***Independent process noise with non-vanishing characteristic function.*** *The process noise ε_t has independent coordinates, and its distribution has a non-vanishing characteristic function.*

Then the affine transformation $\psi(x) = Mx + c$ is block-diagonal up to permutation, i.e., each estimated latent coordinate depends only on a subset of true latent coordinates corresponding to an intervention block.

The non-vanishing characteristic function assumption is satisfied by many common noise distributions, including Gaussian, Poisson, and Gamma.

Assumption 2 (Unordered pairs condition)**.** A set of interventions satisfies the unordered pairs condition on the latents if, for any unordered pair of latent coordinates (i, j) with $i \neq j$, there exists an intervention environment in which exactly one of the two coordinates is intervened on.

Corollary 4 (Permutation, scaling, and shift identifiability)**.** *Under the conditions of Theorem 3, assume further that the unordered pairs condition holds. Then the affine transformation $\psi(x) = Mx + c$ reduces to a permutation, scaling, and shift of the latent coordinates. That is, M is a scaled permutation matrix, and the latent variables are identifiable up to permutation, scaling, and shift.*

Permutation-scale-shift identifiability is desirable because it enables prediction of intervention combinations never seen in training. Additionally, Assumption 2 implies that experimenters can partially control latent disentanglement by adding intervention channels, which increases the chances of latent pairs for which the unordered-pairs condition holds.

3.3 Evaluation metrics

Reconstruction accuracy. The encoder infers latents from the full observed sequence and decodes them to observations. Reconstruction accuracy is the Pearson correlation between predicted and true observations on held-out trials. Prior work argues that reconstruction-based metrics poorly reflect capture of underlying dynamics (Versteeg et al., 2025; Lampinen et al., 2026; Sedler et al., 2023); we therefore report two additional metrics.

Rollout accuracy. We take only the initial latent state $\boldsymbol{x}_0$ from the encoder and roll out the learned dynamics open-loop without further observations. On neural recordings, the encoder sees only pre-stimulation observations when inferring $\boldsymbol{x}_0$ (App. G.4). Rollout accuracy is the Pearson correlation between predicted and true observations over the horizon on held-out trials.

Latent accuracy. When ground-truth latents are available, we apply the same open-loop rollout in latent space. Since iSSM identifies latents only up to permutation, coordinate-wise scaling, and shift (Corollary 4), we align inferred latents to ground truth before evaluation. We use the Hungarian algorithm to find the optimal permutation matching columns of inferred and true B matrices, then fit per-dimension affine transforms (scaling and shift) from true to inferred latents by ordinary least squares on test trajectories. Latent accuracy is the Pearson correlation between aligned predicted and true latent trajectories on held-out trials.

4 Results

We first validate our theoretical results on simulated FitzHugh-Nagumo dynamics with known ground truth (Sec. 4.1). We then demonstrate the framework's practical value across increasingly realistic use cases: adjudicating between competing models of decision-making under idealized conditions (Sec. 4.2); and predicting responses to unseen micro-stimulation in macaque dlPFC and human intracranial recordings, with proof-of-concept generalization to a held-out stimulation site in the human data (Secs. 4.3–4.4). Finally, we return to controlled simulations to ask how future experiments can improve identifiability through the design of their stimulation protocols (Sec. 4.5).

4.1 Identifying FitzHugh-Nagumo dynamics

The FitzHugh-Nagumo (FHN) system is a canonical nonlinear model of neuronal excitability with thresholding and stable limit-cycle dynamics (FitzHugh, 1961; ichi Nagumo et al., 1962). Because these dynamics lie outside the linear iSSM hypothesis class, FHN provides a suitable test of whether nonlinear iSSM can identify nonlinear latent dynamics from observations.

We simulated FHN dynamics

$$\frac{dv}{dt} = v - \frac{v^3}{3} - w + I_{\text{ext}}, \qquad \tau \frac{dw}{dt} = v + a - bw, \tag{3}$$

with $a = 0.7$, $b = 0.8$, $\tau = 12.5$, and $I_{\text{ext}} = 0.5$ ($D = 2$ latent dimensions). Latents were mapped to $N = 4$ observations through a fixed random injective projection and pointwise $\tanh$ nonlinearity. Training used $K = 200$ trials of length $T = 50\,\text{s}$ at $\Delta t = 0.1\,\text{s}$, with hard interventions arriving as a Poisson process at rate $\lambda = 0.2\,\text{s}^{-1}$ and amplitudes in $[-2, 2]$. We trained 10 seeds per model class and evaluated on 100 held-out trajectories with novel interventions; full details are in App. D.

The nonlinear iSSM substantially outperforms the linear iSSM on both rollout and latent accuracy (Fig. 2a,b; both $p = 0.0039$). Consistent with this performance gap, nonlinear models recover the characteristic relaxation-type limit cycle whereas linear models produce an elliptic spiral attractor (Fig. 2c); both recover the input map $\boldsymbol{B}$ (Fig. 2d). Finally, a latent-dimension sweep over $D \in \{1, 2, 3, 4\}$ selects the true dimension ($D = 2$; Fig. 2e). We replicate this analysis on the Van der Pol oscillator in App. E. Overall, these results support Corollary 4: nonlinear iSSM identifies the latent dynamics up to permutation, scaling, and shift, while linear iSSM cannot because the true system lies outside its hypothesis class.

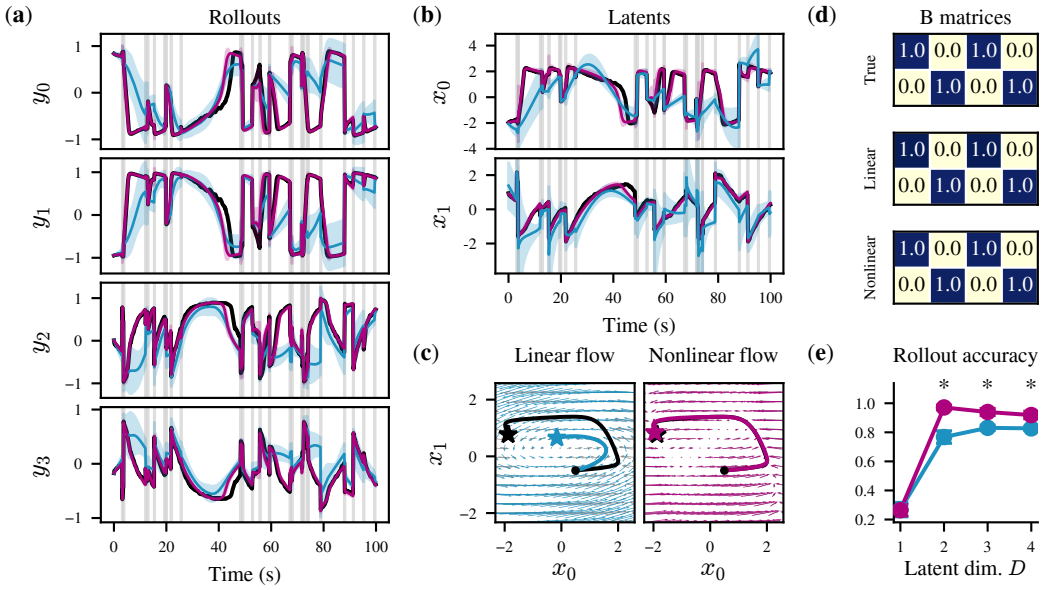

Figure 2: **Nonlinear iSSM identifies FitzHugh-Nagumo latent dynamics.** (a,b) Open-loop rollout and aligned latent trajectories on a test trial: true (black), linear iSSM (blue), nonlinear iSSM (magenta); mean ± s.d. across 10 seeds. Gray lines mark interventions. (c) Latent flow fields from the seed with highest Jacobian similarity to ground truth. Gray/black: true flow (arrows) and trajectory (line); blue/magenta: inferred linear/nonlinear flow and trajectories. (d) Median inferred input map $\boldsymbol{B}$ after permutation-scale alignment. (e) Latent-dimension sweep over $D \in \{1, 2, 3, 4\}$ shows that model selection picks out the true dimension ($D = 2$).

4.2 Identifying latent dynamics of decision-making models

We next consider a scientific use case under idealized conditions: adjudicating between three competing models of decision making. These posit (i) input-driven stable decay toward a point attractor, (ii) perfect integration along a line attractor, or (iii) autonomous nonlinear integration in a bistable well. Despite their distinct recurrent mechanisms, these systems can generate observationally indistinguishable trajectories (Galgali et al., 2023, Fig. 1).

We simulated the three systems ($D = 2$),

$$\dot{\mathbf{x}} = \mathrm{diag}(a_1, a_2)\,\mathbf{x} \qquad \text{(point attractor)}, \tag{4}$$

$$\dot{\mathbf{x}} = A\mathbf{x} \qquad \text{(line attractor)}, \tag{5}$$

$$\dot{x}_1 = ax_1^3 + bx_1 + c_{\mathrm{oh}}, \qquad \dot{x}_2 = cx_2 + c_{\mathrm{oh}} \qquad \text{(bistable well)}, \tag{6}$$

with $(a_1, a_2) = (-0.5, -1.0)$, $A = M(\theta)R\Lambda R^{-1}$ ($\lambda_1 = -1.0$, $\theta = 45°$), and $(a, b, c, c_{\mathrm{oh}}) = (-0.6, 2.0, -1.0, 0.1)$. Following the protocol in Sec. 4.1, latents were mapped to $N = 10$ observations; training used $K = 40$ trials of length $T = 20\,\mathrm{s}$ at $\Delta t = 0.1\,\mathrm{s}$, with hard interventions at rate $\lambda = 0.75\,\mathrm{s}^{-1}$ and amplitudes in $[-3, 3]$ on $M = 2$ input channels. We trained 10 seeds per model class and evaluated on 100 held-out trajectories; full details are in App. F.

As expected, the linear iSSM recovers the point- and line-attractor dynamics, which lie within its hypothesis class (Fig. 3, left). However for the bistable-well system, its learned flow collapses the two wells into a single fixed point and fails to reproduce trajectories on both sides of the separatrix. In contrast, the nonlinear iSSM recovers the flow fields and trajectories of all three systems (Fig. 3, right); and on the bistable system, it achieves significantly higher latent accuracy ($p = 0.0044$).

Interventional data can disambiguate competing decision models, but only within the assumed dynamics class. Nonlinear iSSM extends that class to include bistable attractor dynamics that linear iSSM inherently cannot represent. We next show that this also matters for real neural recordings.

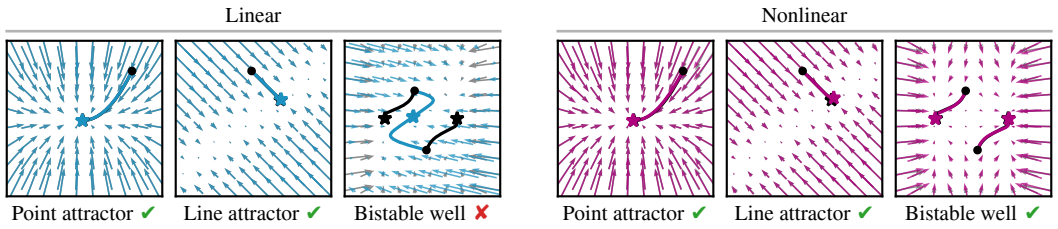

Figure 3: **Nonlinear iSSM identifies a broader class of decision-making dynamics.** Latent flow fields and trajectories for linear (left, blue) and nonlinear (right, magenta) iSSMs trained on interventional data from point-, line-, and bistable-well systems. Ground truth: gray arrows and black lines. Panels show best seed by Jacobian similarity; check marks indicate correct qualitative recovery.

4.3 Predicting macaque dlPFC responses to micro-stimulation

We next show how iSSM can be used to predict neural responses to unseen micro-stimulation of known intervention sites. We applied linear and nonlinear iSSMs to electrophysiological recordings from macaque dlPFC (Nejatbakhsh et al., 2023), pooling data from 14 sessions of the same monkey comprising 686 control and 685 interventional trials across $N = 96$ electrodes and $M = 12$ stimulation sites ($D = 12$; Fig. 4a). Training details are in App. G.

We separate evaluation by trial type – control versus stimulation – to distinguish modelling of spontaneous dynamics from generalisation to unseen interventions. The nonlinear iSSM significantly outperforms the linear iSSM on reconstruction accuracy on both control and stimulation trials (Fig. 4c). On rollout accuracy, the largest gap appears under stimulation (Fig. 4d).

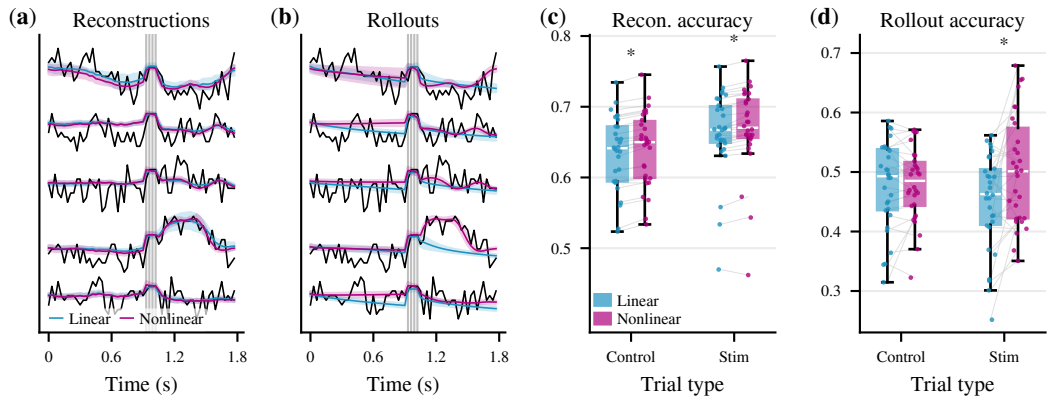

Figure 4: **Nonlinear iSSM outperforms linear iSSM on macaque prefrontal cortex recordings.** (a,b) Example reconstructions and rollouts: observations (black), linear iSSM (blue), nonlinear iSSM (magenta); mean ± s.d. across seeds. Gray lines mark interventions. (c,d) Reconstruction and rollout accuracy on held-out control and stimulation trials. Asterisks: $p < 0.05$.

These results suggest that dlPFC dynamics lie outside the linear iSSM hypothesis class. The rollout gap under stimulation, but not control, implies this matters most for predicting evoked responses.

4.4 Predicting human iEEG responses to single-pulse stimulation

We apply linear and nonlinear iSSMs to intracranial EEG from one epilepsy patient recorded during single-pulse electrical stimulation (Huang et al., 2023) ($N = 99$ channels, $M = 50$ sites, $K = 423$ trials, $D = 25$; Fig. 5a), extending the macaque analysis to a second species and recording modality. Test trials fall into two groups: 20 from sites seen during training; and 10 at a site, clinically labelled irritative zone, fully withheld from training. To enable predicting stimulation response at this site, never seen during training, we map each site's input column from electrode coordinates via a site-embedding network (App. H).

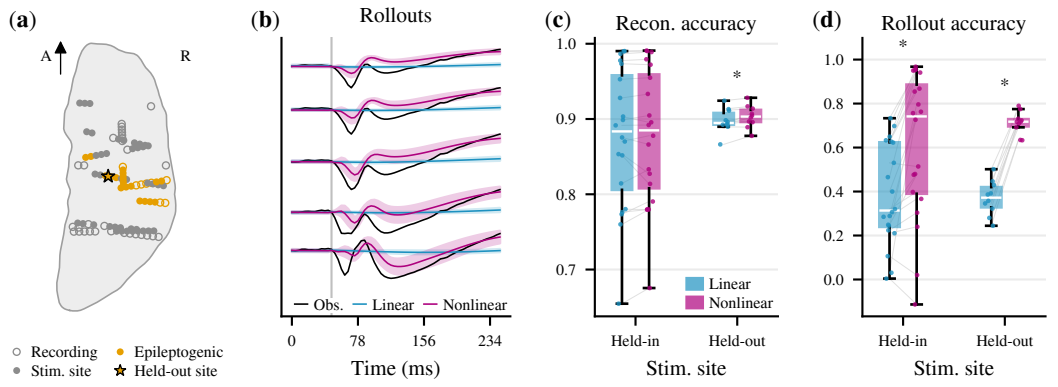

Figure 5: **Nonlinear iSSM predicts human iEEG stimulation responses more accurately, including at a held-out site.** (a) Axial view: recording contacts, stimulation sites, annotated epileptogenic contacts, and the held-out site. (b) Rollouts for the held-out site trial: observations (black), linear iSSM (blue), nonlinear iSSM (magenta); mean $\pm$ s.d. across seeds. Grey line: stimulation pulse. (c,d) Reconstruction and rollout accuracy on held-in and held-out site trials. Asterisks: $p < 0.05$.

On held-out trials from trained stimulation sites, both models reconstruct equally well, but the nonlinear iSSM predicts rollouts significantly better ($p = 0.0004$; Fig. 5b–d). At the held-out site, the nonlinear iSSM reaches rollout accuracy 0.71 ± 0.02 versus 0.37 ± 0.02 for the linear iSSM ($p = 0.002$; mean $\pm$ s.e.m. across held-out trials), with a small reconstruction advantage.

Note that our theory constrains the input map only on observed stimulation sites (Remark 2); thus, the held-out-site evaluation tests coordinate-based interpolation and is included here only as a proof of concept. What makes this empirical generalization test possible is that the held-out site lies 3.5 mm from the nearest held-in site, and recorded responses provide ground truth for validation.

4.5 Designing interventions and evaluating latent recovery

Our theory offers practical guidance on how future stimulation experiments can be designed to maximize identifiability: the theorems are population statements requiring the observation law to be matched over the latent support, so identification holds only where training trajectories visit. Thus, latent accuracy should improve with intervention number and diversity. We test this on the bistable well system (details in App. I), and replicate on two other systems (App. I.2).

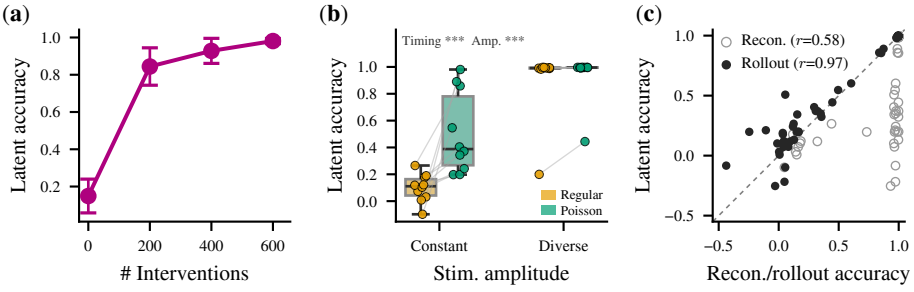

Figure 6: **Numerous and diverse interventions improve latent recovery.** (a) Latent accuracy improves with more training interventions. (b) Diverse amplitudes and irregular timing yield higher latent accuracy; boxes: 10 seeds, $^{***}p < 0.001$ main effect. (c) Latent accuracy versus rollout (black) and reconstruction (gray) accuracy; gray dashed line: identity.

First, we sweep the Poisson stimulation rate over $\lambda \in \{0, 0.25, 0.5, 0.75\}\,\mathrm{s}^{-1}$, generating 0 to $\sim$600 interventions for training (Fig. 6a). Training nonlinear iSSM on each condition, shows latent accuracy increases monotonically with the number of interventions from near chance to near ceiling. Next, we compare stimulation protocols, crossing (constant or diverse) amplitude with (regular or

Poisson) timing at a matched total number of pulses. Amplitude diversity has the larger main effect on latent accuracy (Cohen's $d = 1.86$, $p = 5.7 \times 10^{-6}$), although irregular timing also helps ($d = 0.71$, $p = 1.4 \times 10^{-5}$; Fig. 6b). The fixed-amplitude, regular-timing protocol performs worst.

As ground-truth latents are typically unavailable in real data, we ask which observable metric best reflects latent recovery. Pooling all seeds and conditions from Fig. 6a–b, rollout accuracy tracks latent accuracy closely (points scatter around the identity line), whereas the more commonly used reconstruction accuracy stays near ceiling even when latent recovery fails (Fig. 6c).

5 Discussion

We generalize iSSM identifiability guarantees to injective emissions and nonlinear latent dynamics. We demonstrate two applications – adjudicating between decision-making models, and predicting stimulus response in real neural recordings – which speak to central neuroscience objectives.

Implications for neuroscience. A core objective in systems neuroscience is to identify dynamic regimes of cognition: e.g., prior work used dynamical inference on neural recordings to characterize evidence accumulation for decision-making (Latimer et al., 2015; Luo et al., 2025). Yet, conclusions remain contested (Zoltowski et al., 2019; Steinemann et al., 2024); in part because equally good fits can imply different dynamics, which led Genkin & Engel (2020) to propose an empirical model selection approach that prioritizes consistency of inferred dynamics. Our framework offers a complementary route that combines formal identifiability guarantees with considerable expressiveness.

A second objective, with direct clinical relevance, is predicting neural response to stimulation – e.g., to localize seizure onset zones in epilepsy or to optimize deep brain stimulation targets. Whole-brain biophysical models pursue this via virtual stimulations (Hashemi et al., 2023; Deco & Kringelbach, 2014). However, they are strongly hypothesis-driven, commonly fitting only a few parameters to data (Deco et al., 2008; 2018), and current state-of-the-art models achieve comparatively poor fits (Herzog et al., 2024). iSSM offers a data-driven alternative, and we show that it predicts stimulus response in both macaque and human intracranial recordings, and enable response prediction at unseen stimulation sites. Notably, an exploratory analysis suggests that virtual stimulations, including at several untrained sites, may discriminate healthy from epileptogenic tissue (App. H.4); though this extends beyond our identifiability guarantees (see Limitations).

Limitations. Theorem 3 assumes hard interventions. If microstimulation is not perfectly hard, latent and rollout accuracy may degrade due to misspecification. Nevertheless, toy-model simulations show partial robustness as interventions soften (App. J; Fig. 12); and our theory requires only stochastically hard interventions rather than a deterministic clamp to a known value. Established techniques such as dynamic clamps (Prinz et al., 2004) or optogenetic silencing (Wiegert et al., 2017) plausibly operate within this regime. Mapping experimental techniques onto hard/soft regimes is an important avenue for future work.

For stimulation-response prediction, identifiability holds only over the observed support of latent inputs (Remark 2). Predicting response at unseen sites is therefore extrapolation that our theory does not guarantee. Additionally, we analysed data from only one patient; and the 99 shaft contacts are unlikely to sufficiently sample the full hemisphere. Reliable prediction at distant sites likely requires denser coverage and validation across multiple patients.

Stimulation artefacts, which corrupt signal at stimulated channels, pose a further practical challenge. In the human intracranial recordings, we mitigate this by interpolating and masking stimulated recording sites (App. H.1), and we extend our identifiability theory to cover fitting under such observation masking (App. A.6).

Progressing towards improved identifiability in real data. Accurate identification of something as complex as neural dynamics needs adequate data. Typical interventional datasets, however, include only a few channels, fixed amplitudes, regularly timed pulses, and limited repetitions. Our toy-model simulations clearly show that more diverse and numerous stimulations substantially improve latent recovery. We hope that these results motivate richer interventional protocols to help advance towards identifying causal mechanisms of neural dynamics.

AI USE STATEMENT

In this work, we used generative AI tools (Claude Code and ChatGPT) to assist in the writing of proofs (with presentation and proof checking; see below), provide feedback on research methodology and experiments, interpret results, and via coding assistance: implement methods, generate synthetic data, clean and reformat datasets. We have not used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, help develop theoretical models or conceptual frameworks, or propose or refine hypotheses, and translation and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, summarize existing literature, draft parts of the paper (such as this section) and edit it for readability. We have reviewed all AI-assisted work. All theoretical results and proofs were developed by the authors; AI was used to discuss proofs in related work, check our proofs for errors, and improve their clarity in the writeup. Experiments and methods were designed by the authors; AI gave feedback and assisted with coding, including implementing the preprocessing pipeline for the human iEEG data according to the pipeline published by the original authors (Huang et al., 2023). All AI-assisted code was reviewed and tested by the authors. Results were interpreted by the authors, with AI used for discussion. AI drafted the ethics, reproducibility and AI use statements; the main text was written by the authors, with AI used for polishing, streamlining and typo elimination. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only previously published neural recordings; no new data were collected. The human iEEG data are the publicly available, de-identified OpenNeuro dataset `ds004457` (Huang et al., 2023), recorded under Mayo Clinic Institutional Review Board approval with written informed consent from each patient, as reported by the original authors. The macaque recordings were previously described by Nejatbakhsh et al. (2023). The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Anonymized code for all experiments is available at `https://anonymous.4open.science/r/issm-code-E235/`. Assumptions for our identifiability results are stated in Sec. 3.2, with complete proofs in App. A. For each experiment, the appendix details data generation or preprocessing, model architectures, training procedures and computational resources (App. D–I); statistical tests are described in App. C. The human iEEG data are publicly available on OpenNeuro (`ds004457`).

REFERENCES

Hillel Adesnik and Lamiae Abdeladim. Probing neural codes with two-photon holographic optogenetics. *Nature neuroscience*, 24(10):1356–1366, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pp. 372–407. PMLR, 2023.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *International Conference on Machine Learning*, 2024. URL `https://arxiv.org/abs/2305.15925`.

Carles Balsells-Rodas, Toshiko Matsui, Pedro AM Mediano, Yixin Wang, and Yingzhen Li. On the identifiability of regime-switching models with multi-lag dependencies. *arXiv preprint arXiv:2601.03325*, 2026.

Jürgen Bolz, Charles D Gilbert, and Torsten N Wiesel. Pharmacological analysis of cortical circuitry. *Trends in Neurosciences*, 12(8):292–296, 1989.

Marlene R Cohen and William T Newsome. What electrical microstimulation has revealed about the neural basis of cognition. *Current opinion in neurobiology*, 14(2):169–177, 2004.

Kayvon Daie, Lorenzo Fontolan, Shaul Druckmann, and Karel Svoboda. Feedforward amplification in recurrent networks underlies paradoxical neural coding. *bioRxiv*, 2023. URL https://doi.org/10.1101/2023.08.04.552026.

Gustavo Deco and Morten L. Kringelbach. Great expectations: Using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron*, 84:892–905, 2014. URL https://doi.org/10.1016/j.neuron.2014.08.034.

Gustavo Deco, Viktor Jirsa, Peter A. Robinson, Michael Breakspear, and Karl J. Friston. The dynamic brain: From spiking neurons to neural masses and cortical fields. *PLoS Computational Biology*, 4, 2008. URL https://doi.org/10.1371/journal.pcbi.1000092.

Gustavo Deco, Josephine Cruzat, Joana Cabral, Gitte Moos Knudsen, Robin L. Carhart-Harris, Peter C. Whybrow, Nikos K. Logothetis, and Morten L. Kringelbach. Whole-brain multimodal neuroimaging model using serotonin receptor maps explains non-linear functional effects of lsd. *Current biology : CB*, 28 19: 3065–3074.e6, 2018. URL https://doi.org/10.1016/j.cub.2018.07.083.

Lief Fenno, Ofer Yizhar, and Karl Deisseroth. The development and application of optogenetics. *Annual review of neuroscience*, 34:389–412, 2011. URL https://doi.org/10.1146/annurev-neuro-061010-113817.

Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1:445–66, 1961. URL https://doi.org/10.1016/S0006-3495%2861%2986902-6.

Aniruddh R Galgali, Maneesh Sahani, and Valerio Mante. Residual dynamics resolves recurrent contributions to neural computation. *Nature Neuroscience*, 26(2):326–338, 2023.

Mikhail Genkin and Tatiana A. Engel. Moving beyond generalization to accurate interpretation of flexible models. *Nature machine intelligence*, 2:674–683, 2020. URL https://doi.org/10.1038/s42256-020-00242-6.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, 2020.

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. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021.

Meysam Hashemi, Anirudh Nihalani Vattikonda, Jayant Jha, Viktor Sip, Marmaduke M. Woodman, Fabrice Bartolomei, and Viktor Jirsa. Amortized bayesian inference on generative dynamical network models of epilepsy using deep neural density estimators. *Neural networks : the official journal of the International Neural Network Society*, 163:178–194, 2023. URL https://doi.org/10.1016/j.neunet.2023.03.040.

Rubén Herzog, Pedro A. M. Mediano, Fernando E. Rosas, Andrea I. Luppi, Yonatan Sanz-Perl, Enzo Tagliazucchi, Morten L. Kringelbach, Rodrigo Cofré, and Gustavo Deco. Neural mass modeling for the masses: Democratizing access to whole-brain biophysical modeling with fastdmf. *Network Neuroscience*, 8:1590–1612, 2024. URL https://doi.org/10.1162/netn_a_00410.

Mark H Histed, Amy M Ni, and John HR Maunsell. Insights into cortical mechanisms of behavior from microstimulation experiments. *Progress in neurobiology*, 103:115–130, 2013.

Harvey Huang, Nicholas M. Gregg, Gabriela Ojeda Valencia, Benjamin H. Brinkmann, Brian Nils Lundstrom, Gregory A. Worrell, Kai J. Miller, and Dora Hermes. Electrical stimulation of temporal and limbic circuitry produces distinct responses in human ventral temporal cortex. *The Journal of Neuroscience*, 43:4434 – 4447, 2023. URL https://doi.org/10.1523/JNEUROSCI.1325-22.2023.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Jin ichi Nagumo, Suguru Arimoto, and Shuji Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50:2061–2070, 1962. URL https://doi.org/10.1109/JRPROC.1962.288235.

Hidehiko K Inagaki, Lorenzo Fontolan, Sandro Romani, and Karel Svoboda. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, 566(7743):212–217, 2019.

Mitra Javadzadeh, Marine Schimel, Sonja B. Hofer, Yashar Ahmadian, and Guillaume Hennequin. Dynamic consensus-building between neocortical areas via long-range connections. *bioRxiv*, 2024. URL https://doi.org/10.1101/2024.11.27.625691.

Andrew Kyle Lampinen, Stephanie CY Chan, Yuxuan Li, and Katherine Hermann. Representation biases: variance is not always a good proxy for importance. *eneuro*, 13(3):ENEURO–0461, 2026.

Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pp. 13557–13603. PMLR, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *International Conference on Learning Representations*, 2023.

David Lipshutz, Amin Nejatbakhsh, and Alex H Williams. Disentangling recurrent neural dynamics with stochastic representational geometry. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Thomas Zhihao Luo, Timothy Doyeon Kim, Diksha Gupta, Adrian G Bondy, Charles D Kopec, Verity A Elliott, Brian DePasquale, and Carlos D Brody. Transitions in dynamical regime and neural mode during perceptual decisions. *Nature*, 646(8087):1156–1166, 2025.

Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 32, 2019.

Asma Motiwala, Joana Soldado-Magraner, Aaron P. Batista, Matthew A. Smith, and Byron M. Yu. Brain-computer interfaces as a causal probe for scientific inquiry. *Trends in cognitive sciences*, 30(1):40 – 53, 2026. URL https://doi.org/10.1016/j.tics.2025.06.017.

Amin Nejatbakhsh and Yixin Wang. Identifying neural dynamics using interventional state space models. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=n7qKt6gjl9.

Amin Nejatbakhsh, Francesco Fumarola, Saleh Esteki, Taro Toyoizumi, Roozbeh Kiani, and Luca Mazzucato. Predicting the effect of micro-stimulation on macaque prefrontal activity based on spontaneous circuit dynamics. *Physical review research*, 5, 2023. URL https://doi.org/10.1103/physrevresearch.5.043211.

Astrid A. Prinz, L. F. Abbott, and Eve Marder. The dynamic clamp comes of age. *Trends in neurosciences*, 27 (4):218–24, 2004. URL https://doi.org/10.1016/J.TINS.2004.02.004.

William Qian, Jacob A Zavatone-Veth, Benjamin S Ruben, and Cengiz Pehlevan. Partial observation can induce mechanistic mismatches in data-constrained models of neural dynamics. *Advances in Neural Information Processing Systems*, 37:67467–67510, 2024.

Andrew R. Sedler, Chris VerSteeg, and Chethan Pandarinath. Expressive architectures enhance interpretability of dynamics-based neural population models. *Neurons, behavior, data analysis, and theory*, 2023. URL https://arxiv.org/abs/2212.03771.

Natalie A. Steinemann, Gabriel M. Stine, Eric M. Trautmann, Ariel Zylberberg, Daniel M. Wolpert, and Michael N. Shadlen. Direct observation of the neural computations underlying a single decision. *eLife*, 12, 2024. URL https://doi.org/10.7554/eLife.90859.3.

Christopher Versteeg, Jonathan D McCart, Mitchell Ostrow, David M Zoltowski, Clayton B Washington, Laura Driscoll, Olivier Codol, Jonathan A Michaels, Scott W Linderman, David Sussillo, et al. Computation-through-dynamics benchmark: Simulated datasets and quality metrics for dynamical models of neural activity. *bioRxiv*, 2025.

J. Simon Wiegert, Mathias Mahn, Matthias Prigge, Yoav Printz, and Ofer Yizhar. Silencing neurons: Tools, applications, and experimental constraints. *Neuron*, 95:504 – 529, 2017. URL https://doi.org/10.1016/j.neuron.2017.06.050.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

Congxi Zhang and Yongchun Xie. Identifiable representation and model learning for latent dynamic systems. *Space: Science & Technology*, 5:0267, 2025.

David M. Zoltowski, Kenneth W. Latimer, Jacob L. Yates, Alexander C. Huk, and Jonathan W. Pillow. Discrete stepping and nonlinear ramping dynamics underlie spiking responses of lip neurons during decision-making. *Neuron*, 102, 2019. URL http://dx.doi.org/10.1016/j.neuron.2019.04.031.

A Identifiability Proofs

A.1 Setup

We consider the interventional state space model

$$y_t \sim P(y_t \mid f_\theta(x_t)), \tag{7}$$

$$x_{t+1} = m_t \odot g(x_t) + Bu_t + \varepsilon_t, \tag{8}$$

where $x_t \in \mathbb{R}^D$, $y_t \in \mathbb{R}^N$, $m_t := \mathbf{1}\{Bu_t = 0\} \in \{0,1\}^D$ is the intervention mask, and $\odot$ denotes elementwise multiplication. The inputs u_t are exogenous and observed, i.e., independent of the latent states and of the process noise.

Throughout, we compare two parameterizations that induce the same distribution over observations, and denote the second one by hats:

$$y_t \sim P(y_t \mid f_{\hat{\theta}}(\hat{x}_t)), \tag{9}$$

$$\hat{x}_{t+1} = \hat{m}_t \odot \hat{g}(\hat{x}_t) + \hat{B}u_t + \hat{\varepsilon}_t. \tag{10}$$

A.2 Step 1: Identification of the noiseless representation

This step follows the same strategy as Step 1 in Nejatbakhsh & Wang (2025) and is briefly restated for completeness.

Under the bounded completeness assumption on the observation model (Assumption 1), the distribution of the noiseless signal

$$\bar{y}_t := f_\theta(x_t) \tag{11}$$

is identifiable from the observed distribution of y_t. The argument follows directly from Nejatbakhsh & Wang (2025): the observed marginal satisfies

$$P(y_t) = \int P(y_t \mid \bar{y})\, dP_{\bar{y}_t}(\bar{y}), \tag{12}$$

where $P_{\bar{y}_t}$ denotes the law of $\bar{y}_t$. If two laws $P_{\bar{y}_t}$ and $P'_{\bar{y}_t}$ induce the same $P(y_t)$, their difference is a finite signed measure satisfying the premise of Assumption 1, so $P_{\bar{y}_t} = P'_{\bar{y}_t}$.

Reduction to the noiseless setting. Since $\bar{y}_t = f_\theta(x_t)$ is identifiable from the observed data, we may, without loss of generality, redefine $y_t := \bar{y}_t$ and work with the noiseless representation in the remainder of the proof. All subsequent arguments are therefore carried out under the equivalent model

$$y_t = f_\theta(x_t). \tag{13}$$

A.3 Step 2: Affine identifiability with smooth injective decoders

We now provide an alternative route to affine identifiability that replaces the piecewise linearity assumption of Nejatbakhsh & Wang (2025) by smoothness, injectivity, and Gaussian latent transitions.

Assumptions. We adopt the assumptions of Theorem 2. In particular:

(i) The decoders are injective and continuously differentiable, with continuously differentiable inverses,

(ii) the latent dynamics follow a Gaussian first-order Markov process with positive definite covariance, and

(iii) there exists a nonempty open set $U \subseteq \mathrm{supp}(x_{t-1})$ on which $\det J_g(x) \neq 0$.

Proof. Because the two parameterizations induce the same distribution on the noiseless variables $(\bar{y}_1, \ldots, \bar{y}_T)$, they also induce the same joint distribution on each pair $(\bar{y}_{t-1}, \bar{y}_t)$, and thus, the same conditional distribution on $\bar{y}_t \mid \bar{y}_{t-1}$ for any $t > 1$:

$$p_\theta(\bar{y}_t \mid \bar{y}_{t-1}) = p_{\hat{\theta}}(\bar{y}_t \mid \bar{y}_{t-1}). \tag{14}$$

Given injectivity of the decoder, we may pull back this equality to latent coordinates using $x_t = f_\theta^{-1}(y_t)$ and $\hat{x}_t = f_{\hat{\theta}}^{-1}(y_t)$:

$$\mathrm{vol}\, J_{f_\theta^{-1}}(y_t)\, p_\theta(x_t \mid x_{t-1}) = \mathrm{vol}\, J_{f_{\hat{\theta}}^{-1}}(y_t)\, p_{\hat{\theta}}(\hat{x}_t \mid \hat{x}_{t-1}). \tag{15}$$

Invertibility of the decoders ensures that a mapping $\psi(x_t) = \hat{x}_t$ exists, where $\psi = f_{\hat{\theta}}^{-1} \circ f_\theta$. Using this definition and taking logarithms yields

$$\log \mathrm{vol}\, J_{f_\theta^{-1}}(y_t) + \log p_\theta(x_t \mid x_{t-1}) = \log \mathrm{vol}\, J_{f_{\hat{\theta}}^{-1}}(y_t) + \log p_{\hat{\theta}}(\psi(x_t) \mid \psi(x_{t-1})). \tag{16}$$

Now apply the mixed derivative $\nabla_{x_t} \nabla_{x_{t-1}}$. The Jacobian-volume terms vanish because they depend on x_t but not on x_{t-1}. Hence

$$\nabla_{x_t} \nabla_{x_{t-1}} \log p_\theta(x_t \mid x_{t-1}) = \nabla_{x_t} \nabla_{x_{t-1}} \log p_{\hat{\theta}}(\psi(x_t) \mid \psi(x_{t-1})). \tag{17}$$

For Gaussian latent transitions,

$$\log p_\theta(x_t \mid x_{t-1}) = -\frac{1}{2}(x_t - g_\phi(x_{t-1}))^\top Q^{-1}(x_t - g_\phi(x_{t-1})) + C, \tag{18}$$

where C does not depend jointly on (x_t, x_{t-1}). Therefore,

$$\nabla_{x_t} \nabla_{x_{t-1}} \log p_\theta(x_t \mid x_{t-1}) = Q^{-1} J_{g_\phi}(x_{t-1}). \tag{19}$$

Analogously, for the hatted parameterization, the chain rule gives

$$\nabla_{x_t} \nabla_{x_{t-1}} \log p_{\hat{\theta}}(\psi(x_t) \mid \psi(x_{t-1})) = J_\psi(x_t)^\top \hat{Q}^{-1} J_{g_{\hat{\phi}}}(\psi(x_{t-1})) J_\psi(x_{t-1}). \tag{20}$$

Substituting into (17) yields

$$Q^{-1} J_{g_\phi}(x_{t-1}) = J_\psi(x_t)^\top \hat{Q}^{-1} J_{g_{\hat{\phi}}}(\psi(x_{t-1})) J_\psi(x_{t-1}). \tag{21}$$

By Assumption (iii), there exists $x^* \in \mathbb{R}^D$ such that $J_{g_\phi}(x^*)$ is full rank on a neighborhood of x^*. By continuity of g_ϕ (Assumption (i)), there exists an nonempty set $\mathcal{B}_{x^*} \subset \mathbb{R}^D$ such that $\det J_{g_\phi}(x) \neq 0$ for any $x \in \mathcal{B}_{x^*}$. This implies $J_{g_\phi}(x)$ is invertible on a nonempty set. Since all matrices in (21) are of size $D \times D$, this implies that each factor on the right-hand side must also be full rank on that set. In particular, $J_\psi(x_t)$, $\hat{Q}$, $J_{g_{\hat{\phi}}}(\psi(x_{t-1}))$, and $J_\psi(x_{t-1})$ are invertible for any $x_{t-1} \in \mathcal{B}_{x^*}$, and we may therefore rearrange the expression as

$$J_\psi(x_t)^\top = Q^{-1} J_{g_\phi}(x_{t-1}) J_\psi(x_{t-1})^{-1} J_{g_{\hat{\phi}}}(\psi(x_{t-1}))^{-1} \hat{Q}. \tag{22}$$

By Assumption (ii) and for any $x_{t-1} \in \mathcal{B}_{x^*}$, the conditional distribution of $x_t | x_{t-1}$ has full support for $t > 2$. Therefore, the above equation holds for any $x_t \in \mathbb{R}^D$. Since the right-hand side depends on x_{t-1} but not on x_t, this implies $J_\psi(x_t)$ is constant with respect to x_t, and therefore, ψ is affine:

$$\hat{x}_t = \psi(x_t) = M x_t + c, \tag{23}$$

which proves affine identifiability. □

Remark 1. Step 3 below only requires affine identifiability. Thus it can be combined either with Theorem 2 or, alternatively, with Theorem 3.5 of Balsells-Rodas et al. (2024) as in the original iSSM proof (Nejatbakhsh & Wang, 2025).

A.4 Step 3: Refinement under interventions

We now show how hard interventions refine affine identifiability (Theorem 2) to block-wise permutation-scaling-shift identifiability.

Notation. We will use the following notation. For an index $j \in \{1, \dots, D\}$, we denote by $-j$ the complement of j, i.e., the set of all latent indices except j. Accordingly, for a vector x_t, we write $x_{t,-j}$ for the subvector obtained by removing the j-th coordinate, and for a matrix M, we write $M_{j',-j}$ for the entries of the j'-th row excluding the j-th column.

Assumptions. We assume affine identifiability (Theorem 2), the faithfulness (Assumption (i)), and independent process noise with non-zero characteristic function (Assumption (ii)).

Statement. Given the assumptions, we show that hard interventions constrain the structure of M. In particular, in the single-intervention case, if coordinate j is intervened on, then there exists a coordinate j' in the estimated latent space such that

$$M_{j',-j} = 0. \tag{24}$$

This implies that the j'-th estimated latent depends only on the true j-th latent (up to scaling and shift). Hence, the intervened latent dimension is identifiable up to permutation, scaling, and shift.

Proof. Following Nejatbakhsh & Wang (2025), we first treat the single-intervention case. Since the inputs are exogenous and observed, we condition throughout on a fixed intervention value $u_t = u$; all independence statements below are understood under this conditioning. Suppose coordinate j is intervened on. Then the structural equation for that coordinate becomes

$$x_{t+1,j} = (Bu)_j + \varepsilon_{t,j}, \tag{25}$$

so that $x_{t+1,j}$ is the process noise up to a constant shift. By construction, $x_{t+1,j}$ is independent of all past latents:

$$x_{t+1,j} \perp x_t. \tag{26}$$

Assume further that in the estimated latent space there exists an index j' such that

$$\hat{x}_{t+1,j'} \perp \hat{x}_t. \tag{27}$$

Using affine identifiability $\hat{x}_{t+1} = Mx_{t+1} + c$, we expand

$$\hat{x}_{t+1,j'} = M_{j',-j}^\top x_{t+1,-j} + M_{j',j} x_{t+1,j} + c_{j'}. \tag{28}$$

Since constants do not affect independence, Eq. (27) implies

$$M_{j',-j}^\top x_{t+1,-j} + M_{j',j} x_{t+1,j} \perp Mx_t. \tag{29}$$

By Eq. (26), the term $M_{j',j} x_{t+1,j}$ is already independent of Mx_t. We now show that this forces the remaining term $M_{j',-j}^\top x_{t+1,-j}$ to be independent of Mx_t as well.

Define

$$U := M_{j',-j}^\top x_{t+1,-j}, \qquad V := M_{j',j} x_{t+1,j}, \qquad Z := Mx_t. \tag{30}$$

Then Eq. (29) is $U + V \perp Z$. Moreover, by Eq. (26),

$$V \perp Z. \tag{31}$$

By the independent noise Assumption (ii) and Eq. (25), the intervened coordinate depends only on the intervention and its own noise, so

$$V \perp U. \tag{32}$$

Finally, by Eq. (25) V is the process noise up to scaling and a constant shift, so by Assumption (ii) the characteristic function $\varphi_V(t)$ is nonzero for all $t \in \mathbb{R}$.

We will show that these facts imply $U \perp Z$. Specifically, V is a function only of the intervened noise $\varepsilon_{t,j}$ and exogenous intervention value, while (U, Z) is a function only of x_t, $\varepsilon_{t,-j}$, and non-intervened exogenous variables, mutual independence of the process noise and exogenous interventions implies

$$V \perp (U, Z).$$

In more detail, since $U + V \perp Z$,

$$\varphi_{U+V,Z}(t, s) = \varphi_{U+V}(t)\varphi_Z(s). \tag{33}$$

Using $U \perp V$ gives

$$\varphi_{U+V}(t) = \varphi_U(t)\varphi_V(t), \tag{34}$$

and therefore

$$\varphi_{U+V,Z}(t,s) = \varphi_U(t)\varphi_V(t)\varphi_Z(s). \tag{35}$$

On the other hand,

$$\varphi_{U+V,Z}(t,s) = \mathbb{E}\Big[e^{it(U+V)}e^{isZ}\Big] \tag{36}$$

$$= \mathbb{E}\big[e^{itU}e^{isZ}\,\mathbb{E}\big[e^{itV} \mid U,Z\big]\big]. \tag{37}$$

Since $V \perp U$ and $V \perp Z$, we have $V \perp (U,Z)$, so

$$\mathbb{E}\big[e^{itV} \mid U,Z\big] = \varphi_V(t), \tag{38}$$

hence

$$\varphi_{U+V,Z}(t,s) = \varphi_{U,Z}(t,s)\varphi_V(t). \tag{39}$$

Combining (35) and (39) and dividing by $\varphi_V(t) \neq 0$ yields

$$\varphi_{U,Z}(t,s) = \varphi_U(t)\varphi_Z(s). \tag{40}$$

Thus $U \perp Z$, i.e.

$$M_{j',-j}^{\top} x_{t+1,-j} \perp Mx_t. \tag{41}$$

Since j is the only intervened coordinate, $x_{t+1,-j} = g(x_t)_{-j} + \varepsilon_{t,-j}$ does not depend on u. The joint law of (U,Z) is therefore the same for every intervention value targeting j, and Eq. (41) continues to hold after marginalising over these values.

The remaining proof follows the same argument as in Nejatbakhsh & Wang (2025) and is included here for completeness. Specifically, define vectors $v, w \in \mathbb{R}^D$ by

$$w := M_{j'}, \qquad v_j := 0, \qquad v_{-j} := M_{j',-j}, \tag{42}$$

where $M_{j'}$ denotes the j'-th row of M. Then (41) can be written as

$$v^{\top} x_{t+1} \perp w^{\top} x_t. \tag{43}$$

By invertibility of M, we have $w \neq 0$. Therefore Assumption (i) implies $v = 0$. Hence

$$M_{j',-j} = 0. \tag{44}$$

So the coordinate j' in the estimated latent space depends only on the true intervened coordinate j, up to scaling and shift. This proves our statement for the single-intervention case.

For multiple simultaneous interventions, the same argument applies blockwise. Let $\mathcal{I}$ denote the set of intervened coordinates. Then $x_{t+1,\mathcal{I}} \perp x_t$ conditional on $u_t = u$, and repeating the characteristic-function argument shows that the corresponding rows of M cannot mix coordinates in $\mathcal{I}$ with coordinates in $-\mathcal{I}$. Therefore the affine map must be block diagonal up to permutation, with one block for intervened coordinates and one block for unintervened coordinates. This yields block identifiability up to permutation of blocks and an invertible affine transformation within each block. □

From block to full identifiability. Under the unordered pairs condition (Assumption 2), the above argument applies to every pair of latent coordinates, forcing all off-diagonal entries of M to vanish. Hence $M = \Lambda\Pi$, where Π is a permutation matrix and Λ is diagonal and invertible, yielding identifiability up to permutation, scaling, and shift as shown by Nejatbakhsh & Wang (2025).

A.5 Nonlinear input maps

Corollary 4 also holds when the linear input map Bu_t is replaced by an arbitrary deterministic map $h(u_t)$, provided there exists an intervention-free regime on the support of u.

Proposition 5 (Nonlinear input maps)**.** *Replace Bu_t in Eq. (2) by an arbitrary deterministic map $h : \mathbb{R}^M \to \mathbb{R}^D$, and let $m_t := \mathbf{1}\{h(u_t) = 0\}$. Assume the conditions of Corollary 4, with Assumption 2 applied to the coordinates that receive nonzero input. If there exists $u^0 \in \mathrm{supp}(u)$ such that $h(u^0) = 0$, then the latent variables remain identifiable up to permutation, scaling, and shift.*

Proof of Proposition 5. Let $s_t = h(u_t)$, so $m_{t,j} = \mathbf{1}\{s_{t,j} = 0\}$. We verify that the three-step proof of Corollary 4 remains valid when replacing Bu by $h(u)$ with only one additional assumption on the input map h, needed for *step 2*.

Step 1 is unchanged because it does not involve the input map.

Step 2 requires an intervention-free regime because the transition Jacobian $\mathrm{diag}(m_t)J_g(x_t)$ is singular during active interventions. Thus, we need to assume that there exists $u^0 \in \mathrm{supp}(u)$ with $h(u^0) = 0$. At $u_t = u^0$, we then have $m_t = \mathbf{1}$ and $x_{t+1} = g(x_t) + \varepsilon_t$, so Theorem 2 applies. The resulting affine map $\psi(x) = Mx + c$ is independent of the intervention environment and therefore holds both at u^0 and under active interventions.

Step 3 follows the same argument as in App. A.4: conditioning on $u_t = u$ fixes both $h(u)$ and the intervened coordinates, and an intervened coordinate satisfies Eq. (25) with $(Bu)_j$ replaced by $h_j(u)$. Applying Assumption 2 to the coordinates receiving nonzero input yields $M = \Lambda\Pi$. □

Remark 2. Note that the data constrains h only on $\mathrm{supp}(u)$. If the observed interventions comprise finitely many channel-amplitude combinations, one-hot encoding these combinations allows a linear input map with one column per combination to represent any h exactly on the observed support.

A.6 Observation masking

In the human iEEG study, we mask out observation channels that correspond to active stimulation sites on each trial to exclude saturation artefacts. We show here that the identifiability guarantees above remain valid in this setting with modest additional conditions.

Definition 1 (Exogenous observation masking)**.** Let $\boldsymbol{o} \in \{0,1\}^N$ denote a binary mask indicating which observation coordinates are retained. On trial k, a mask $\boldsymbol{o}_k$ applies. The mask is known and determined only by observed exogenous variables; the observation noise is independent of these variables and the latent states taken jointly. Write $\mathcal{O}(\boldsymbol{o}) := \{n : o_n = 1\}$ for the indices of the observed coordinates and let $P_{\boldsymbol{o}} : \mathbb{R}^N \to \mathbb{R}^{|\mathcal{O}(\boldsymbol{o})|}$ denote projection onto those coordinates. Finally, define the masked noiseless signal as $\bar{y}_{t,\mathcal{O}(\boldsymbol{o})} := P_{\boldsymbol{o}} f_\theta(x_t)$, and masked observations as $y_{t,\mathcal{O}(\boldsymbol{o})} := P_{\boldsymbol{o}} y_t$.

We verify that the three-step proof of Corollary 4 remains valid under Definition 1. For each step, we state what the masking changes, give sufficient conditions under which the step still holds, and discuss plausibility in the human iEEG analysis.

Step 1: bounded completeness under observation masking. Step 1 (App. A.2) requires bounded completeness (Assumption 1) for the observation model. Under masking, the analogous requirement is that $P(y_{t,\mathcal{O}(\boldsymbol{o})} \mid \bar{y}_{t,\mathcal{O}(\boldsymbol{o})})$ is bounded complete in the sense of Assumption 1. Below we provide a sufficient condition that applies to the human iEEG analysis.

Proposition 6 (Bounded completeness under observation masking)**.** *Suppose that:*

- *(i) observations are generated by additive noise $y_t = \bar{y}_t + \eta_t$, where $\bar{y}_t$ is the noiseless signal and $\eta_t \perp \bar{y}_t$;*
- *(ii) the noise η_t admits a density p_η on $\mathbb{R}^N$; and*
- *(iii) the characteristic function φ_η of η_t satisfies $\varphi_\eta(s) \neq 0$ for all $s \in \mathbb{R}^N$.*

Then the conditional distribution $P(y_t \mid \bar{y}_t)$ satisfies Assumption 1. The same conclusion holds under any mask satisfying Definition 1.

Proof. We first prove the claim for the unmasked case, and later we show that the argument also holds under masking.

By (i) and (ii), the conditional density of y_t given $\bar{y}_t$ is $p_\eta(y_t - \bar{y}_t)$. To verify Assumption 1, we must show that for every finite signed measure μ on $\mathbb{R}^N$,

$$\int p_\eta(y - r)\, d\mu(r) = 0 \quad \text{for all } y \in \mathbb{R}^N \tag{45}$$

implies $\mu = 0$. The left-hand side is the convolution $(p_\eta * \mu)(y)$, an integrable function of y since $p_\eta \in L^1$ and μ is finite. Taking Fourier transforms in y, with $\widehat{\mu}(s) := \int e^{is^\top r}\, d\mu(r)$, gives

$$\widehat{\mu}(s)\, \varphi_\eta(s) = 0 \quad \text{for all } s. \tag{46}$$

By (iii), $\varphi_\eta(s) \neq 0$ for all s, thus $\widehat{\mu}(s) = 0$ for all s. Uniqueness of the Fourier–Stieltjes transform of finite signed measures then implies $\mu = 0$, as required.

Now we show that the conditions are preserved under masking. By Definition 1, the mask is exogenous and fixed, so the additive noise model and independence of (i) are preserved:

$$y_{t,\mathcal{O}(\boldsymbol{o})} = P_{\boldsymbol{o}}\bar{y}_t + P_{\boldsymbol{o}}\eta_t \quad \text{and} \quad \eta_{t,\mathcal{O}(\boldsymbol{o})} \perp \bar{y}_{t,\mathcal{O}(\boldsymbol{o})}. \tag{47}$$

Condition (ii) is preserved because the marginal of a density is a density, i.e., integrating p_η over the masked-out coordinates yields a density $p_{\eta_{\boldsymbol{o}}}$ of $P_{\boldsymbol{o}}\eta_t$ on $\mathbb{R}^{|\mathcal{O}(\boldsymbol{o})|}$.

Finally, the characteristic function of the marginal noise satisfies

$$\varphi_{\eta_{\boldsymbol{o}}}(s_{\boldsymbol{o}}) = \varphi_\eta\big(P_{\boldsymbol{o}}^\top s_{\boldsymbol{o}}\big)\,, \tag{48}$$

so a non-vanishing φ_η implies $\varphi_{\eta_{\boldsymbol{o}}}(s_{\boldsymbol{o}}) \neq 0$ for all $s_{\boldsymbol{o}}$. Hence, the above convolution argument shown for the unmasked case also applies to the masked case. □

Remark 3. In the human iEEG analysis, the mask is fully determined by whichever stimulation site is active during that trial, so Definition 1 applies. Additionally, we assume Gaussian observation noise, $y_t = \bar{y}_t + \eta_t$ with $\eta_t \sim \mathcal{N}(\mu, \Sigma_\eta)$ and Σ_η positive definite. Gaussian noise is additive and its characteristic function never vanishes. The non-singular covariance ensures that η_t admits a density on $\mathbb{R}^N$. Thus, Proposition 6 implies that Step 1 holds under masking in this application.

Step 2: affine identifiability under observation masking. Observation masking may break Assumption (i) of Step 2 (App. A.3), which requires the emission function to be injective and continuously differentiable, with continuously differentiable inverse on its image: projecting to $\mathcal{O}(\boldsymbol{o})$ can break injectivity even when the full decoder f_θ is injective. Below we give sufficient conditions under which affine identifiability is preserved, then justify their plausibility in our setting.

Proposition 7 (Affine identifiability under observation masking)**.** *Assume that all conditions of Theorem 2 hold, and that for every mask the data contain transitions without active interventions ($m_t = \mathbf{1}$). Assume additionally that:*

- *(i) for every mask $\boldsymbol{o}$, the restricted decoders $P_{\boldsymbol{o}} \circ f_\theta$ and $P_{\boldsymbol{o}} \circ f_{\hat{\theta}}$ are injective and continuously differentiable, with continuously differentiable inverses on their images;*
- *(ii) for every pair of masks $\boldsymbol{o}, \boldsymbol{o}'$, the restricted decoders $P_{\mathcal{O}(\boldsymbol{o})\cap\mathcal{O}(\boldsymbol{o}')} \circ f_\theta$ and $P_{\mathcal{O}(\boldsymbol{o})\cap\mathcal{O}(\boldsymbol{o}')} \circ f_{\hat{\theta}}$ are injective.*

Then affine identifiability is preserved under observation masking in the sense of Definition 1.

Proof. Fix two masks $\boldsymbol{o}$ and $\boldsymbol{o}'$. By (i), we can apply Step 2 separately under each mask, using its transitions without active interventions, yielding affine maps $\psi_{\boldsymbol{o}}(x) = M_{\boldsymbol{o}}x + c_{\boldsymbol{o}}$ and $\psi_{\boldsymbol{o}'}(x) = M_{\boldsymbol{o}'}x + c_{\boldsymbol{o}'}$ such that

$$P_{\boldsymbol{o}} f_\theta(x) = P_{\boldsymbol{o}} f_{\hat{\theta}}(\psi_{\boldsymbol{o}}(x)), \qquad P_{\boldsymbol{o}'} f_\theta(x) = P_{\boldsymbol{o}'} f_{\hat{\theta}}(\psi_{\boldsymbol{o}'}(x)). \tag{49}$$

Restricting both identities to $\mathcal{O}(\boldsymbol{o}) \cap \mathcal{O}(\boldsymbol{o}')$ gives

$$P_{\mathcal{O}(\boldsymbol{o})\cap\mathcal{O}(\boldsymbol{o}')} f_\theta(x) = P_{\mathcal{O}(\boldsymbol{o})\cap\mathcal{O}(\boldsymbol{o}')} f_{\hat{\theta}}(\psi_{\boldsymbol{o}}(x)) = P_{\mathcal{O}(\boldsymbol{o})\cap\mathcal{O}(\boldsymbol{o}')} f_{\hat{\theta}}(\psi_{\boldsymbol{o}'}(x)). \tag{50}$$

By (ii), injectivity of $P_{\mathcal{O}(\boldsymbol{o})\cap\mathcal{O}(\boldsymbol{o}')} \circ f_{\hat{\theta}}$ implies $\psi_{\boldsymbol{o}}(x) = \psi_{\boldsymbol{o}'}(x)$ for all $x \in \mathbb{R}^D$, which is the latent support under both masks by Assumption (ii). Hence $M_{\boldsymbol{o}} = M_{\boldsymbol{o}'}$ and $c_{\boldsymbol{o}} = c_{\boldsymbol{o}'}$. Since $\boldsymbol{o}$ and $\boldsymbol{o}'$ were arbitrary, a single affine map $\hat{x} = Mx + c$ holds across all mask patterns. □

Remark 4. In the human iEEG analysis, each mask removes only the two channels at the active stimulation site (at most four under pairwise overlap, out of $N = 99$ outputs). For $P_{\boldsymbol{o}} \circ f_\theta$ to fail injectivity, distinct latents $x \neq x'$ would have to agree on all unmasked channels while differing

only on masked ones. That is, the ~ 4 masked coordinates would have to contain unique information needed to distinguish these latents that is absent from the remaining ~ 95 channels. A similar argument applies to the continuously differentiable inverse in (i): it requires the Jacobian $P_{\boldsymbol{o}} J_{f_\theta}(x)$ to keep full rank D at every x, which fails only if, at some x, a small change of the latent state along some direction affects the masked channels alone. The masked channels are the ones closest to the stimulated tissue; it is therefore not unthinkable that they contain identifying information that other channels lack. However, other contacts on the same electrode shaft remain observed, and stimulation also evokes responses at distant channels. We therefore consider such unique information in the few masked channels unlikely and treat conditions (i) and (ii) as a mild robustness assumption. Finally, each trial contains a single stimulation pulse, so all other time steps are free of active interventions, as required by Proposition 7.

Step 3: refinement under interventions. Step 3 (App. A.4) concerns only the latent transition under hard interventions and does not involve the observation model. What is relevant in this context is that Step 3 combines constraints obtained from different intervention channels to restrict a single affine transformation M. The conclusion of Proposition 7 that the affine maps obtained under different observation masks must coincide, is therefore essential. However, once Steps 1 and 2 establish a common affine map $\hat{x} = Mx + c$ across masks, Step 3 remains valid and requires no additional assumptions.

B Variational Inference

We perform parameter learning and latent state inference using variational inference. Given observations $y_{1:T}$ and interventions $u_{1:T}$, our goal is to infer the posterior distribution over latent trajectories $p_\Theta(x_{1:T} \mid y_{1:T}, u_{1:T})$ while learning model parameters $\Theta = (\theta, \phi, B, Q)$.

Since exact inference is intractable due to nonlinearities in both the transition and emission models, we introduce an approximate posterior $q_\Phi(x_{1:T} \mid y_{1:T}, u_{1:T})$ and optimize the evidence lower bound (ELBO):

$$\mathcal{L}(\Phi, \Theta) = \mathbb{E}_{q_\Phi}\left[\log p_\Theta(y_{1:T}, x_{1:T} \mid u_{1:T}) - \log q_\Phi(x_{1:T} \mid y_{1:T}, u_{1:T})\right]. \tag{51}$$

We adopt amortized variational inference and parameterize the approximate posterior using a recurrent neural network. Specifically, we use an LSTM-based recognition model to produce the mean and variance of a Gaussian approximate posterior over latent states at each time step. Latent variables are reparameterized as $x_t = \mu_t + \sigma_t \odot \eta_t$, where $\eta_t \sim \mathcal{N}(0, I)$.

To match the interventional structure from Eq. (2), we apply the mask m_t to the approximate posterior mean during training:

$$\mu_t \leftarrow m_t \odot \mu_t + B u_t. \tag{52}$$

The model parameters and variational parameters are jointly optimized using stochastic gradient descent on the negative ELBO.

C Statistical Analysis

Unless stated otherwise, each model is trained with 10 random seeds. Reported means are accompanied by the standard error of the mean (s.e.m.). When reporting per-trial performance in real-data experiments, we first average predictions across seeds for each held-out trial, then report the mean and s.e.m. across trials within each evaluation tier (e.g., control vs. stimulation). Otherwise, reported accuracy metrics are mean $\pm$ s.e.m. across seeds. Model comparisons use a two-sided paired Wilcoxon signed-rank test on seed-level scores ($n = 10$). Where applicable, p-values are Benjamini-Hochberg FDR-corrected for multiple comparisons.

D FitzHugh-Nagumo Experiment Details

We provide full details for the simulation experiments in Sec. 4.1.

D.1 Data Generation

We simulate the two-dimensional FHN system

$$\dot{v} = v - \frac{v^3}{3} - w + I_{\text{ext}}, \qquad \dot{w} = \frac{v + a - bw}{\tau}, \tag{53}$$

with parameters a=0.7, b=0.8, τ=12.5, I_{ext}=0.5 (oscillatory regime; true latent dimension D=2). Trajectories have length T=50 s at Δt=0.1 s, preceded by a 5-second burn-in that is discarded before training. We use 250 training trajectories (batch size 25) and 10 validation trajectories. Initial conditions are sampled as $x_0 \sim \mathcal{N}(0, \sigma_0^2 I_D)$ with σ_0=0.05 before the burn-in. Latent dynamics include additive Gaussian process noise with σ_x=0.05; observations are noiseless (σ_y=0).

Observation model. Observations are obtained via $y_t = \tanh(Cx_t)$, where $C \in \mathbb{R}^{N \times D}$ is a fixed random matrix with orthonormal columns (N=4 observations).

Interventions. Interventions are delivered on M=4 input channels via random (Poisson) pulse stimulation: at each time step a single channel is activated with probability r_{stim}=0.20, with pulse duration 0.25 s, amplitude drawn uniformly from $[-2, 2]$, and additive Gaussian noise σ=0.25. Of the training trajectories, 10% are non-interventional (control trials) with no stimulation applied.

D.2 Model Architecture

Linear and nonlinear variants share the same emission and recognition architectures and differ only in the dynamics parameterization.

Dynamics. The nonlinear model uses $g_\phi(x_t) = \text{MLP}(x_t)$, implemented as a two-hidden-layer network with 8 units per layer and GeLU activations. The linear model uses $g_\phi(x_t) = Ax_t$, with $A \in \mathbb{R}^{D \times D}$ learned. Both share a learned input matrix $B \in \mathbb{R}^{D \times M}$, initialized to the identity (zero-padded as needed) and subject to the B-sparsity mechanisms below.

B-sparsity mechanisms. Two mechanisms enforce sparsity on B. First, at each forward pass we treat entries with $|B_{ij}| < \tau_B$ as inactive when computing the hard-intervention mask (τ_B=10^{-3} fixed for all experiments), so tiny numerical residuals do not trigger spurious interventions. Second, during training we hard-threshold B elementwise after each optimizer step (outside backpropagation): entries with $|B_{ij}| < \tau_t$ are set to zero, where τ_t increases linearly from 0 to $\tau_{\max}$=0.10 over the first p_τ=0.10 fraction of the maximum number of epochs and is fixed thereafter. We use hard thresholding rather than an ℓ_1 penalty so that surviving entries are not shrunk toward zero while small residual entries are set to exact zero; dense support on B also violates the unordered-pairs condition needed for Corollary 4. The warmup avoids premature pruning early in training. On FHN (nonlinear iSSM, 10 seeds), disabling the schedule ($\tau_{\max}$=0) reduced latent accuracy from 0.97 ± 0.02 to 0.74 ± 0.04 and degraded B recovery (paired Wilcoxon signed-rank test: both p=0.004); although a milder threshold $\tau_{\max}$=0.05 already matched the default model's performance.

Emission. The emission model factorises into a linear readout and a nonlinear residual correction:

$$f_\theta(x_t) = W_{\text{skip}}\, x_t + \text{MLP}_\theta(x_t), \tag{54}$$

where $W_{\text{skip}} \in \mathbb{R}^{N \times D}$ is a learned linear readout and the nonlinear branch projects x_t to H_{em}=32 dimensions (linear + GeLU), passes it through n_{blocks}=2 pre-norm residual blocks (LayerNorm $\to$ Linear $\to$ GeLU $\to$ Linear with skip connection), and projects back to N. The nonlinear branch is initialised with near-zero output weights so the model starts approximately linear.

Recognition. The approximate posterior $q_\Phi(x_{1:T} \mid y_{1:T}, u_{1:T})$ is parameterized by a shared-backbone LSTM: inputs (y_t, u_t) are linearly projected to H_{rec}=32 dimensions (LayerNorm + GeLU), processed by a single-layer LSTM, and decoded by two separate heads for the posterior mean and log-variance.

D.3 Training

Models are trained to maximize the ELBO (standard β=1 VAE objective) using Adam with learning rate 3×10^{-3}, gradient clipping (max norm 1.0), and ℓ_2 weight decay 10^{-3}. The KL term is linearly warmed up over the first p_β=0.20 fraction of epochs; the training-time B-sparsity warmup follows the schedule above. Early stopping monitors rollout MAE on the validation set with patience 50 epochs, up to a maximum of 1000 epochs.

D.4 Evaluation Protocol

All toy-model results in Secs. 4.1, 4.2, and 4.5 use the same evaluation protocol. After training, we score each seed on 100 newly simulated test trajectories from the true dynamical system. Using the training Δt and burn-in T_{burn}, we simulate $T_{\mathrm{burn}} + T_{\mathrm{encode}} + T$ seconds per trial (T_{encode}=5 s, T=100 s) and discard the burn-in.

Test interventions. Every test trial receives Poisson pulse stimulation. Pulse duration, input noise, and the amplitude interval are inherited from the training stimulator, with the amplitude range scaled by 1.2 to probe interventions stronger than those seen during training. The 100 test trials are distributed approximately evenly over 7 stimulation rates uniformly spaced in $[0.2, 0.8]\,\mathrm{s}^{-1}$.

Trial structure and scoring. Reconstruction, rollout, and latent accuracy (Sec. 3.3) are computed on the 100 s evaluation segment (after the encode window). We compute one Pearson correlation per trial then average over trials within each seed. Correlations use only non-interventional timesteps where all input channels are zero. Trials on which the model produces non-finite outputs (e.g. a diverged rollout) score zero.

Alignment. For latent accuracy we permutation-align inferred latents by Hungarian matching on the inferred and true input matrices B, then fit a per-dimension affine map (scale and shift) from the driven full-trial posterior at non-interventional timesteps.

D.5 Computational Resources

All FHN experiments were run on CPU with 2 GB RAM. Per seed, training takes approximately 30-45 min (variation due to early stopping).

E Van Der Pol Oscillator Experiments

The Van der Pol (VdP) oscillator with its square-like relaxation limit cycle has qualitatively different attractor geometry compared to FHN, but it also lies outside the linear iSSM hypothesis class.

We simulated VdP dynamics

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = \mu(1 - x_1^2)\, x_2 - x_1, \tag{55}$$

with $\mu = 2$ ($D = 2$ latent dimensions). Model architecture and training follow Apps. D.2 and D.3. Data generation follows the same protocol as App. D.1, with the following deviations: Δt=0.05 s; process noise σ_x=0.002; observation nonlinearity arsinh instead of $\tanh$; stimulation probability r_{stim}=0.45; recognition hidden dimension H_{rec}=64; learning rate 2×10^{-3}; maximum epochs 2000; and B-sparsity warmup fraction p_τ=0.05.

Observation nonlinearity. We use arsinh rather than $\tanh$ as the pointwise observation nonlinearity. The VdP oscillator at $\mu = 2$ produces oscillations with a substantially larger amplitude than FHN, and the random injective projection C can amplify this further. At such value ranges, $\tanh$ saturates and produces near-zero gradients at the extremes, creating an optimization challenge that is unrelated to identifiability. The arsinh nonlinearity is unbounded (growing as $\log(2|x|)$ for large $|x|$), preserves injectivity, and avoids gradient saturation.

The VdP experiments replicate the FHN trends: The nonlinear iSSM again outperforms its linear counterpart (Fig. 7), achieving rollout accuracy of $r = 0.94 \pm 0.01$ s.e.m. versus 0.58 ± 0.04

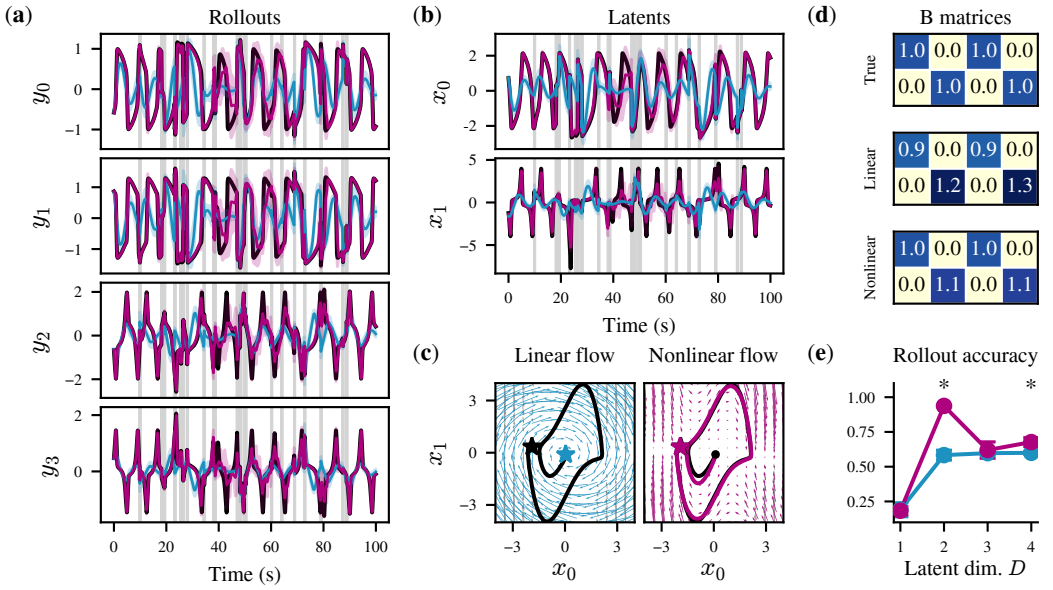

Figure 7: **Nonlinear iSSM identifies Van Der Pol oscillator latent dynamics.** (a) Example open-loop rollouts for one test trial (rollout accuracy): true data (black), linear iSSM (blue), nonlinear iSSM (magenta). Curves show mean across 10 seeds, shaded area indicates s.e.m; gray lines mark interventions. (b) Example aligned latent trajectories (latent accuracy; seed mean $\pm$ s.e.m). Nonlinear iSSM recovers the true latents; linear iSSM does not. (c) Latent flow fields from the seed with highest Jacobian cosine similarity to ground truth. Linear iSSM yields a spiral attractor and fails to recover the example latent trajectory (black); nonlinear iSSM recovers the relaxation-type limit cycle and the example trajectory. (d) Median inferred input map $\boldsymbol{B}$ after permutation-scale alignment. (e) Latent-dimension sweep over $D \in \{1, 2, 3, 4\}$ shows that model selection picks out the true dimension ($D = 2$).

and latent accuracy of $r = 0.93 \pm 0.01$ versus 0.55 ± 0.03 (Fig. 7a,b; Wilcoxon signed-rank test: both $p = 0.0020$). The best nonlinear model recovers the characteristic square-like VdP attractor, whereas the best linear model produces a spiral attractor (Fig. 7c). Nonlinear iSSM also more faithfully recovers $\boldsymbol{B}$ (Fig. 7d). Model selection picks out the true dimension ($D = 2$) for both models (Fig. 7e).

F Decision-Making Model Experiment Details

We provide full details for the simulation experiments in Sec. 4.2.

F.1 Dynamical Systems

We simulate three two-dimensional systems ($D = 2$) as follows.

Bistable well (nonlinear integration).

$$\dot{x}_1 = ax_1^3 + bx_1 + c_{\text{oh}}, \qquad \dot{x}_2 = cx_2 + c_{\text{oh}}, \tag{56}$$

with $a = -0.6$, b=2.0, $c = -1.0$, and c_{oh}=0.1. Evidence is integrated via an unstable saddle separating two stable choice states.

Line attractor (perfect integration).

$$\dot{\mathbf{x}} = A\mathbf{x}, \tag{57}$$

where $A = M(\theta)\, R\Lambda R^{-1}$ has one marginally stable direction (eigenvalue 0) and one stable direction ($\lambda_1 < 0$). Here $\Lambda = \text{diag}(0, \lambda_1)$, $R = [\mathbf{r}_0; \mathbf{l}_0]^\top$ contains the right and left eigenvectors, and

$M(\theta)$ is a 2×2 rotation matrix. We use $\mathbf{l}_0 = [1, 1]^\top$, $\mathbf{r}_0 = [1, 0]^\top$ (each normalized), $\lambda_1 = -1.0$, and $\theta = 45°$.

Point attractor (input-driven dynamics).

$$\dot{\mathbf{x}} = \mathrm{diag}(a_1, a_2)\, \mathbf{x}, \tag{58}$$

with $a_1 = -0.5$ and $a_2 = -1.0$. Both latent dimensions decay to a fixed point at the origin, so decision-related activity is inherited from inputs rather than generated locally.

F.2 Data Generation

The three systems share the same observation and intervention protocol. Trajectories have length $T = 20$ s at $\Delta t = 0.1$ s with no burn-in. We generate 40 training, 10 validation, and 100 test trajectories (batch size 20). All systems use $D = 2$ latent dimensions, $N = 10$ observations, and $M = 2$ input channels. Latent dynamics include additive Gaussian process noise with $\sigma_x = 0.1$; observations are noiseless ($\sigma_y = 0$).

Observation model. Observations are obtained via $y_t = \tanh(Cx_t)$, where $C \in \mathbb{R}^{N \times D}$ is a fixed random matrix with orthonormal columns.

Interventions. Interventions are delivered on $M = 2$ input channels via random pulse stimulation: at each time step a single channel is activated with probability $r_{\mathrm{stim}} = 0.75$, with pulse duration 0.25 s, amplitude drawn uniformly from $[-3, 3]$, and additive Gaussian noise $\sigma = 0.25$. All training trajectories contain interventions.

F.3 Training

We fit linear and nonlinear iSSMs to interventional data from each system. Model architecture and training follow Apps. D.2 and D.3, with learning rate 5×10^{-3} and batch size 20. Linear and nonlinear variants differ only in the dynamics parameterization. We train 10 random seeds per model class and evaluate on held-out trajectories with novel interventions.

F.4 Computational Resources

All decision-making experiments were run on CPU with 2 GB RAM. Per seed, training takes approximately 15–30 minutes (variation due to early stopping).

G Macaque dlPFC Dataset Experiment Details

G.1 Data

We use spike-count recordings from macaque prefrontal cortex (area 8Ar) collected during resting-state micro-stimulation experiments (Nejatbakhsh et al., 2023). Data were recorded at $f_s = 30$ Hz across $N = 96$ channels. Trials are extracted by windowing ± 1 s around each stimulation event (stimulation trials) or around randomly sampled time points matched in count (control trials), giving $T = 60$ bins per trial at $\Delta t = 33$ ms. Both conditions are used during training.

Input construction. The stimulation input $u_t \in \mathbb{R}^M$ is a binary indicator that is one at the stimulated electrode's channel during stimulation and zero elsewhere. We compress u to only the $M = 12$ electrodes that were ever stimulated across all sessions, so that each entry of u has a direct causal interpretation. Control trials have $u_t = 0$ throughout. Train, validation, and test sets are split with stratification by trial type (stimulation vs. control) to ensure both conditions are represented in each split; 10 trials are reserved for validation and 60 for testing.

G.2 Model Architecture

The linear and nonlinear models use the same implementation as the simulation experiments (App. D.2), with the dataset-specific modifications below. They differ only in their dynamics parameterization.

Likelihood. Because observations are spike counts, we use a Poisson likelihood: $p(y_t \mid x_t) = \mathrm{Poisson}(\exp(f_\theta(x_t)))$ applied independently per channel, where $f_\theta(x_t)$ outputs log-rates.

Dynamics. The nonlinear model uses $g_\phi(x_t) = \mathrm{MLP}(x_t)$, implemented as a three-hidden-layer network with 64 units per layer and GeLU activations. The linear model uses $g_\phi(x_t) = Ax_t$, with $A \in \mathbb{R}^{D\times D}$ learned. Both share a learned input matrix $B \in \mathbb{R}^{D\times M}$, initialized to an identity-like matrix with small perturbations and subject to the B-sparsity mechanisms described in App. D.2. We sweep the latent dimension over $D \in \{5, 12, 15, 20\}$ and report D=12 in the main text. The rollout advantage of the nonlinear iSSM holds at every D (Fig. 8).

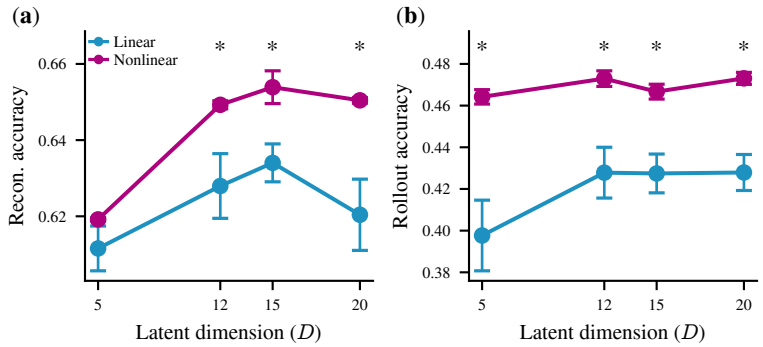

Figure 8: **Accuracy versus latent dimension D.** (a) Reconstruction and (b) rollout accuracy for linear (blue) and nonlinear (magenta) iSSMs; mean ± s.e.m. across 10 seeds. Asterisks mark significant differences ($p < 0.05$, FDR-corrected).

Emission. We use the residual-MLP emission described in App. D.2, with hidden dimension H_{em}=32, two residual blocks, and dropout probability 0.1.

Recognition. The approximate posterior $q_\Phi(x_{1:T} \mid y_{1:T}, u_{1:T})$ is parameterized by a *bidirectional* shared-backbone LSTM with hidden dimension H_{rec}=32, input dropout probability 0.1, and separate linear heads for the posterior mean and log-variance. Because the recognition network is applied offline to the full trial, bidirectionality is appropriate and improves posterior quality. During training, each neuron-level observation is independently masked with probability 0.1.

G.3 Training

We use the same training loop and ELBO objective as in App. D.3. We use Adam with learning rate 3×10^{-3}, batch size 25, ℓ_2 weight decay 10^{-2}, and gradient clipping at maximum norm 1.0. The KL term is warmed up linearly over the first p_β=0.25 fraction of epochs, while the training-time B-sparsity threshold increases to $\tau_{\max}$=0.05 over the first p_τ=0.50 fraction. Early stopping monitors rollout MAE on the validation set with patience 100 epochs, up to a maximum of 3000 epochs. These settings are shared by the linear and nonlinear models.

G.4 Evaluation

Reconstruction and rollout accuracy are computed on held-out test trials. Reconstruction encodes the full trial with the bidirectional recognition network (App. G.2). For rollouts, the encoder sees only pre-stimulation observations and inputs: a fixed window of 31 timesteps (≈ 1.0 s at 30 Hz), set to the earliest stimulation onset in the test set and shared across all trials including controls with

$u_t = 0$ throughout. $\boldsymbol{x}_0$ is taken from the start of that window and the dynamics are integrated open-loop with the known test inputs over the full trial. Rollout accuracy is the Pearson correlation between predicted and true observations, with the same channel masks as reconstruction.

G.5 Unordered pairs estimate

Permutation-scale-shift identifiability (Corollary 4) requires the unordered pairs condition (Assumption 2). To assess the degree to which the learned intervention design implements this condition, we read the τ_B-thresholded support of the trained B matrix (τ_B=10^{-3}; App. D.2). Let $Z_{dk} = \mathbf{1}[|B_{dk}| \geq \tau_B]$ indicate whether an intervention via channel k drives latent coordinate d. Channel k separates pair (i, j) iff it drives exactly one of the two coordinates, i.e. $Z_{i\cdot} \neq Z_{j\cdot}$. We report the fraction of separately intervenable pairs as:

$$\text{up_frac} = \frac{\#\{(i,j),\ i<j : Z_{i\cdot} \neq Z_{j\cdot}\}}{\binom{D}{2}}, \tag{59}$$

computed per seed and summarised as mean $\pm$ s.e.m. across seeds.

For the main nonlinear macaque model ($D = 12$), up_frac $= 93.3 \pm 1.3\%$ (mean $\pm$ s.e.m. across 10 seeds). Note that this estimate is not a validation of the unordered pairs condition, as we have no ground-truth latent interventions.

G.6 Computational Resources

All experiments were run on CPU with 8 GB RAM. Training a single model run takes approximately 3–6 hours depending on early stopping; both the linear and nonlinear iSSM were trained with the same budget.

H Human iEEG Dataset Experiment Details

H.1 Data

We use the human iEEG single-pulse electrical stimulation dataset of Huang et al. (2023) (OpenNeuro `ds004457`) for subject `sub-2`. Continuous SEEG was recorded at f_s=2048 Hz across 170 channels; biphasic single pulses were delivered to bipolar contact pairs at 4 or 6 mA. Following the event filter of Huang et al. (2023) (`status=good`), the preprocessed release contains 529 pulses at 60 bipolar stimulation sites.

Preprocessing. We replicate the signal-processing pipeline of Huang et al. (2023). For each pulse we extract a $[-1, +2]$ s window and apply, in order: (i) stimulation-artefact interpolation over $[-2, +15]$ ms (PCHIP spline fitted to surrounding clean samples), (ii) common-average reference over good non-stimulated channels, (iii) 60 Hz line-noise removal by spectrum fitting, and (iv) baseline correction (median over $[-0.5, -0.05]$ s).

Channel and trial selection. We drop clinically bad channels, channels without anatomical location, white-matter contacts, and white-matter stimulation sites (and trials at removed sites), leaving N=99 recording channels, M=50 stimulation sites, and 437 trials. Observations are decimated to 256 Hz with an anti-aliasing filter, cropped to $[-50, +200]$ ms around the pulse (T=64 bins at Δt=3.9 ms), and standardized per channel.

Input construction. The stimulation input $u_t \in \mathbb{R}^M$ is a single-sample impulse at pulse onset in the coordinate of the stimulated site, with amplitude equal to the delivered current (rescaled so the maximum maps to one) and zero elsewhere.

Masking and artefact trials. We exclude the two contacts of the stimulated pair from the likelihood for the whole trial (they saturate around the pulse). App. A.6 shows that identifiability is preserved under this masking, given modest additional conditions. The same whole-trial stim-contact masks are used when scoring validation and test trials; masked channels are excluded from

per-channel standardization and zero-filled before the recognition network. Additionally, 14 trials flagged by a joint variance and template-correlation criterion as having residual onset artefacts on any kept channel beyond the active stimulation pair are dropped, leaving 423 trials.

Splits. Every trial is a single pulse at one site, so we stratify by stimulation site. We hold out 10 trials for validation and 30 for testing, and train on the remaining 383. The test set has two tiers: a held-out-site tier with all 10 trials stimulating the right-hippocampal pair RC3-RC4 (clinically labelled irritative zone), excluded from training and validation; and a held-in tier with the remaining 20 test trials from sites seen during training.

H.2 Model Architecture

The models follow the macaque setup (App. G.2) with the modifications below; linear and nonlinear variants differ only in their dynamics parameterization.

Likelihood. Observations are continuous voltages, so we use a multivariate Gaussian likelihood $p(y_t \mid x_t) = \mathcal{N}(f_\theta(x_t), LL^\top)$ with a learned full covariance: L is lower triangular with per-channel scales on the diagonal and learned off-diagonal entries below.

Dynamics. As in App. G.2, except for the B parameterization (see below); we sweep $D \in \{15, 20, 25, 30\}$ and report D=25 in the main text. The rollout advantage of the nonlinear iSSM holds at every D (Fig. 9).

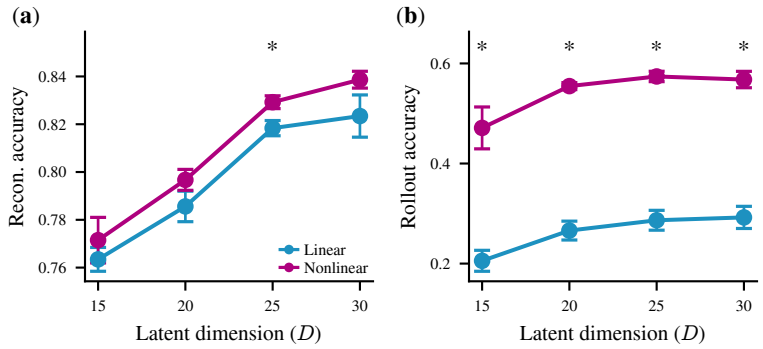

Figure 9: **Accuracy versus latent dimension D.** (a) Reconstruction and (b) rollout accuracy for linear (blue) and nonlinear (magenta) iSSMs; mean $\pm$ s.e.m. across 10 seeds. Asterisks mark significant differences ($p < 0.05$, FDR-corrected).

Site-embedding intervention matrix. Unlike the macaque models, B is not a free parameter: columns are predicted from electrode coordinates via a site-embedding network. Let $s_m \in \mathbb{R}^3$ be the midpoint of bipolar site m in the electrode coordinates provided with the dataset (`electrodes.tsv`). We center and isotropically scale the coordinates (root-mean-square deviation over all sites and axes) and map them through a shared MLP:

$$b_m = \mathrm{MLP}_\psi\left(\frac{s_m - \bar{s}}{\sigma}\right) \in \mathbb{R}^D, \qquad B = [\, b_1, \ldots, b_M \,] \in \mathbb{R}^{D \times M}. \tag{60}$$

MLP_ψ has one hidden layer of 32 units with GeLU activation; the output layer is initialized with Gaussian weights scaled by 0.1. This parameterization of B by coordinates lets us predict responses at sites that were never stimulated during training, and, more speculatively, at sites that were never stimulated at all (App. H.4). We note that this goes beyond our identifiability results, which constrain the input map only on the support of the observed inputs (Remark 2). Generalisation to unseen sites instead relies on the smoothness of the coordinate parameterisation, which we evaluate empirically.

Emission. We use the residual-MLP emission of App. D.2 with hidden dimension H_{em}=128, three residual blocks, and dropout probability 0.1.

Recognition. As in App. G.2, with H_{rec}=64 and input dropout probability 0.2.

H.3 Training

Training follows App. G.3, with learning rate 3×10^{-3}, batch size 32, and no training-time B-sparsity warmup (since B is computed from coordinates; the forward-pass threshold τ_B still applies). We train 10 random seeds per latent dimension. For the nonlinear model reported in the main text ($D = 25$), the unordered-pairs estimate (App. G.5) is up_frac $= 44.3 \pm 3.3\%$ (mean $\pm$ s.e.m. across 10 seeds). Evaluation follows App. G.4, with a fixed encoder window of 12 pre-stimulation timesteps (≈ 47 ms at Δt=3.9 ms).

H.4 Virtual Stimulation of Never-Stimulated Sites

We synthesize 77 candidate bipolar pairs from adjacent recorded contacts on the same lead (33 never stimulated in the data), annotate each with the most severe clinical label of its two contacts, and group them into epileptogenic (n=26; seizure-onset, early-propagation, or irritative zone) and healthy (n=51) pairs. Labels enter only at analysis time. For each pair we obtain a virtual column b from its midpoint coordinates and run two rollouts from each of 100 initial states (sampled from encoded latents at pulse onset across trials): one with a unit impulse along b at the first step, one autonomous. Then we difference the two to remove shared baseline drift. Both trajectories are decoded and we take $\|\Delta y(t)\|$, averaged over initial states over a 50-step (195 ms) window. From this curve we report the peak (response amplitude) and the mean from 100 ms to window end (response persistence). We repeat this analysis for each of the ten random seeds from model training and aggregate the resulting readouts before testing. Epileptogenic and healthy pairs are compared with Mann-Whitney U tests, with Benjamini-Hochberg correction over the two readouts and two models.

Under the nonlinear iSSM, epileptogenic pairs produce responses that are both larger (Cohen's $d = 1.08$, $p < 0.0001$) and more sustained ($d = 0.75$, $p = 0.0060$) than healthy pairs, matching the clinical expectation that seizure-onset tissue is hyperexcitable and hyperconnected; the linear iSSM shows neither effect (Fig. 10).

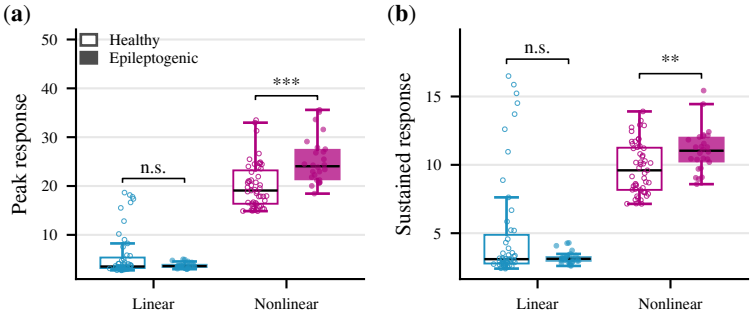

Figure 10: **Virtual stimulation of clinically annotated sites.** (a) Peak and (b) sustained network-wide response $\|\Delta \boldsymbol{y}\|$ to virtual stimulation of healthy ($n = 51$) and epileptogenic ($n = 26$) bipolar pairs, under the linear and the nonlinear iSSM. Mann-Whitney U tests, FDR-corrected over the two readouts and two models; $***$: $p < 0.001$.

Please note that we treat this as a pilot only. Input columns for many virtually stimulated sites are far from any sites actually stimulated in the dataset (Fig. 5a), and outside the scope of our identifiability results (Remark 2). Also, the present analysis uses only one subject and without independent replication.

H.5 Computational Resources

All experiments were run on CPU, with 4 cores and 6 GB RAM per run for 2–4 hours.

I Stimulation Protocol Experiment Details

We provide full details for the stimulation protocol design experiments in Sec. 4.5. Both analyses use the bistable well of App. F.1 and vary one property of the intervention set at a time, holding the model class fixed.

I.1 Diversity of Interventions

Data generation. Data generation follows App. F.2; only stimulator timing and amplitude parameters are varied, in a 2×2 factorial crossing pulse timing (regular or Poisson) with pulse amplitude (constant or diverse). Under Poisson timing, pulses arrive as a Poisson process at rate $\lambda{=}0.75\,\mathrm{s}^{-1}$, each assigned to a uniformly random input channel. Under regular timing, pulses cycle round-robin over the input channels with a fixed rest interval $d_{\mathrm{rest}} = 1/\lambda - d_{\mathrm{stim}} = 1.083\,\mathrm{s}$, so the two timings deliver the same expected number of pulses and differ only in their placement. Under diverse amplitudes, each pulse amplitude is drawn independently from $\mathcal{U}[-3, 3]$. Under constant amplitudes, a single amplitude is drawn once per run, with magnitude $\mathcal{U}[1.5, 3]$ and random sign, and used for every pulse in the run; amplitude and sign only vary across seeds.

Training. We train nonlinear iSSM only. Model architecture and training follow Apps. D.2 and D.3, with the settings of App. F.3. We train 10 random seeds for each of the four cells.

Computational resources. All runs were performed on CPU (2 cores, 2 GB RAM).

I.2 Replication on Point and Line Attractors

We repeat the intervention-count sweep and the amplitude-timing factorial of Sec. 4.5 on the point and line attractor systems of App. F.1, with the same training protocol as above. The qualitative trends match the bistable well (Fig. 6): latent accuracy rises with more interventions, diverse amplitudes and irregular timing improve recovery, and rollout accuracy tracks latent accuracy more closely than reconstruction does (Fig. 11).

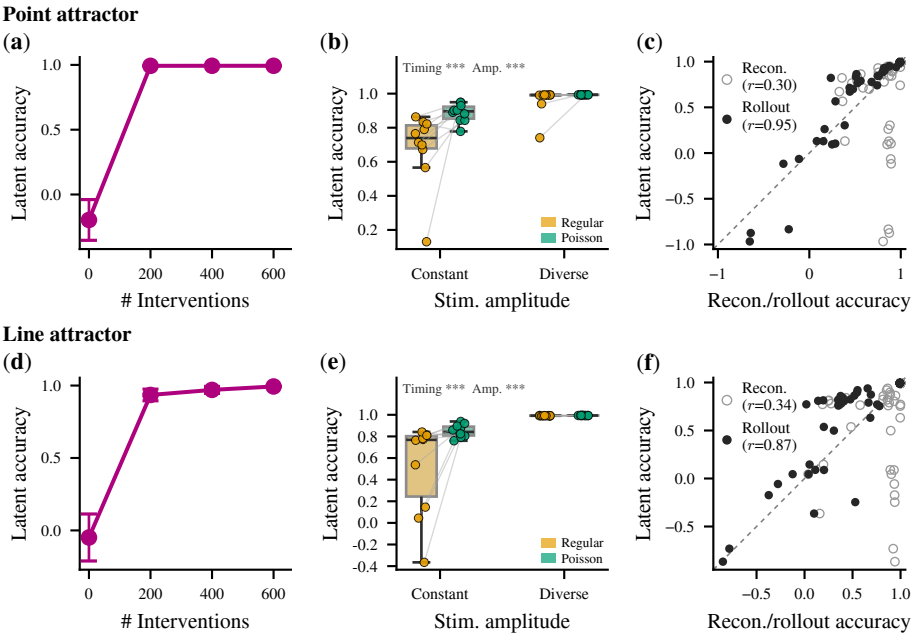

Figure 11: **Intervention-design trends replicate on point and line attractors.** Top row (a-c): point attractor; bottom row (d-f): line attractor. Panel layout matches Fig. 6: (a,d) latent accuracy versus intervention count; (b,e) amplitude-timing factorial (10 seeds); (c,f) latent accuracy versus rollout (black) and reconstruction (gray); dashed line: identity.

J INTERVENTION HARDNESS

Theorem 3 assumes hard interventions that decouple driven latents from the past state (Eq. (2)). We view intervention hardness as a gradual property that depends on the particular intervention technique. We therefore empirically test the robustness of latent identification to gradual violation of this assumption.

We sweep the data-generating intervention from a purely additive drive ($\alpha = 0$) to a perfectly hard clamp ($\alpha = 1$) in the bistable well system. We compared the iSSM against a standard SSM and an input-conditioned variant whose nonlinear drift takes $\boldsymbol{u}_t$ as an argument – the most expressive of the three, as its hypothesis class contains every α regime.

On the bistable well (Fig. 12a), latent accuracy reveals two patterns. First, iSSM identification improves gradually, not stepwise, as interventions harden; though it is poor under purely additive drives ($\alpha = 0$) when the hardness assumption is completely violated. Second, across all α, only iSSM ever recovers the true latents near-perfectly: at $\alpha = 1$ it reaches 0.98, significantly above either baseline measured in its own ideal regime ($p = 0.002$). We replicate this analysis on FitzHugh-Nagumo dynamics (Fig. 12b), finding qualitatively similar results.

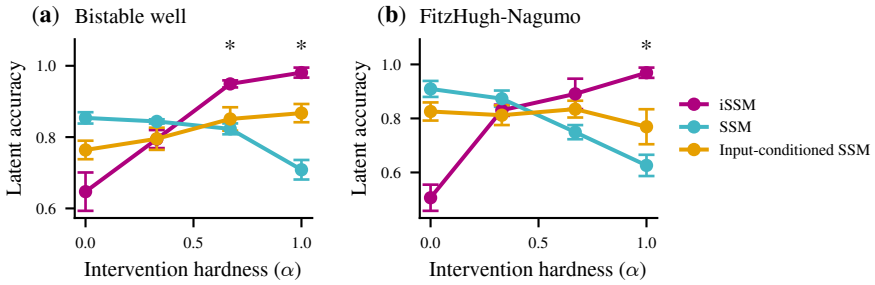

Figure 12: **Latent accuracy under softened interventions and model comparison.** Sweep over the intervention hardness parameter α of the data-generating process, from a purely additive drive ($\alpha = 0$) to a hard clamp ($\alpha = 1$), comparing nonlinear iSSM to two baselines. (a) Bistable well (App. J.1). (b) FitzHugh-Nagumo replication (App. J.2). Asterisks indicate α-regimes where iSSM significantly outperforms both baselines ($p < 0.05$).

J.1 BISTABLE WELL

Data generation. Data generation follows App. F.2, with 40 training and 10 validation trajectories (batch size 20), and with the hardness of the data-generating intervention swept. Let $x_{t+1}^{\text{free}} = g(x_t) + \epsilon_t$ denote the state that the system would have reached without stimulation, where g is the true one-step transition map. At every driven latent entry we blend the clamp with these free dynamics,

$$x_{t+1} = m_t \odot x_{t+1}^{\text{free}} + (\mathbf{1} - m_t) \odot \big(Bu_t + (1 - \alpha)\, x_{t+1}^{\text{free}}\big), \tag{61}$$

with the intervention mask $m_t = \mathbf{1}\{Bu_t = 0\}$ of Eq. (2); undriven entries are unaffected. At α=1 the driven entry is overwritten by Bu_t, recovering the hard intervention that Eq. (2) assumes; at α=0 the input is added on top of the free dynamics and the driven latent retains its dependence on the past state. We sweep $\alpha \in \{1, 0.67, 0.33, 0\}$. All other data properties are held at the values of App. F.2.

Model variants. The three models differ only in how the input u_t enters the latent transition; emission, recognition, likelihood and initial-condition modules are identical across arms. Writing MLP_ϕ for the drift network of App. D.2, the transitions are

$$\text{nonlinear iSSM:} \quad x_{t+1} = m_t \odot \text{MLP}_\phi(x_t) + Bu_t + \epsilon_t, \tag{62}$$

$$\text{standard SSM:} \quad x_{t+1} = \text{MLP}_\phi(x_t) + Bu_t + \epsilon_t, \tag{63}$$

$$\text{input-conditioned SSM:} \quad x_{t+1} = \text{MLP}_\phi([x_t, u_t]) + \epsilon_t, \tag{64}$$

with $m_t = \mathbf{1}\{Bu_t = 0\}$ as above.

Nonlinear iSSM is misspecified at $\alpha \neq 1$; standard SSM is correctly specified at $\alpha \neq 0$; the input-conditioned SSM is never misspecified.

Training. Model architecture and training follow Apps. D.2 and D.3, with the settings of App. F.3. We train 10 random seeds for each (model, α) condition.

Evaluation. Evaluation follows App. D.4, with one exception: since the input-conditioned SSM has no B, we index-align latents from latent trajectories rather than B columns for all models in this experiment.

Computational resources. All runs were performed on CPU (4 cores, 2 GB RAM for training and 4 GB for evaluation), with per-seed training cost comparable to App. F.4.

J.2 FitzHugh-Nagumo Replication

Data generation. Data generation follows App. D.1, with the intervention hardness blended as in Eq. (61) and swept over $\alpha \in \{1, 0.67, 0.33, 0\}$. All other data properties match App. D.1.

Model variants, training, and evaluation. We use the same three model classes (Eqs. (62)–(64)), training protocol (App. D.3), and evaluation protocol (App. D.4, with index alignment for the input-conditioned SSM) as in App. J.1. We train 10 random seeds for each (model, α) condition.

Computational resources. All runs were performed on CPU (4 cores, 2 GB RAM for training and 4 GB for evaluation), with per-seed training cost comparable to App. D.5.