OBSERVATIONALLY EQUIVALENT, INTERVENTIONALLY DIFFERENT: INTERVENTION IDENTIFIABILITY IN MULTISCALE WORLD MODELS

Anonymous authors

ABSTRACT

A world model can predict every observed coarse trajectory correctly while giving the wrong answer to an intervention on a hidden fast mechanism. We formalize this gap through observational fibers: an intervention target is identifiable exactly when it is constant over all models with the same observational law. This yields a two-point minimax lower bound and a general gauge certificate for non-identifiability. For a slow diffusion driven by a fast Ornstein–Uhlenbeck mechanism, we construct a continuum of microscopic models with the same complete averaged slow-path law but different responses to replacement of the fast conditional mean. We then characterize the missing cross-scale information. Frozen conditional means or a calibrated stationary variance identify the intervention response; one positive-lag covariance together with the variance separates the fast mixing rate from its diffusion scale. A perturbation bound quantifies the instability of intervention transport when the fast mean response is weak. The results distinguish observational prediction, microscopic parameter recovery, and target-specific intervention identification. They also show why architectural clock factorization alone cannot resolve an informational ambiguity in a coarse world model.

1 INTRODUCTION

World models are commonly trained to reproduce observational trajectories and then used to answer questions about distribution shift, control, or intervention. In multiscale systems these two uses need not coincide. A coarse trajectory may determine an effective slow dynamics while leaving the fast mechanism that generated it unidentified. Replacing that hidden mechanism can then produce different coarse responses in world models that are indistinguishable from observational data.

This paper asks a population-level question: which intervention targets are determined by the observational law of a coarse world model, and what additional cross-scale information makes them determined? We treat a world model as a member θ of a model class, write $O(\theta)$ for its observational law, and write $T_u(\theta)$ for the post-intervention target under intervention u. The observational fiber $[\theta]_O$ contains all models that fit the same observational law. Intervention transport is possible exactly when T_u is constant on this fiber.

We use the standard factorization criterion for identifiability and its two-point minimax consequence. It separates failure of estimation from failure of identification: unlimited observational data do not resolve a target that varies inside an observational fiber. A group action provides a convenient certificate. Any gauge that preserves O but changes T_u proves non-identifiability.

We then specialize the framework to a slow variable driven by a fast Ornstein–Uhlenbeck mechanism. The averaged observational law identifies only four composite coefficients. We construct a continuum of fast mechanisms with the same complete reduced slow-path law but different responses to a replacement of the fast conditional mean. This strengthens a parameter-counting ambiguity into an intervention separation.

The negative result also identifies the missing information. Conditional stationary fast snapshots at two slow states identify the mean-response coefficient and are sufficient to identify the coarse effect of mean-mechanism replacements. One positive-lag covariance additionally identifies the fast mixing rate and diffusion scale separately. We characterize target sufficiency within the specified

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.

measurement protocols and give a perturbation bound showing how anchor error is amplified near a weak mean-response coefficient.

The contribution is not a claim that multiple clocks are more expressive. Scheduled multirate updates can be flattened into an augmented single-clock transition, and a sampled Markov process has a one-clock kernel. The obstruction studied here is informational: observational prediction at a coarse scale need not identify how a hidden mechanism changes under intervention. This distinction is relevant to latent dynamics, causal representation learning, model-based control, and multiscale forecasting.

In summary, we make three contributions.

First, we formulate intervention identifiability through observational fibers: a target is identifiable exactly when it is constant over all models inducing the same observational law. This yields a general two-point minimax lower bound and a gauge-based certificate for non-identifiability.

Second, we instantiate this framework in a slow–fast stochastic system and construct a continuum of microscopic models that are observationally indistinguishable at the averaged level but respond differently to the same intervention on the hidden fast mechanism.

Third, we characterize what additional cross-scale information is sufficient to identify the intervention target without fully recovering the microscopic generator. For the OU specialization, conditional fast means identify the mean-replacement target; a calibrated stationary variance supplies an alternative target anchor; and variance plus one lag covariance separate the fast rate from its diffusion scale. We further quantify the stability of target transport under approximate anchor recovery and evaluate the resulting finite-sample behavior empirically.

2 Related Work

Multiscale dynamics and learned time-steppers. Stochastic averaging and homogenization characterize reduced dynamics under explicit mixing, scaling, regularity, horizon, and convergence assumptions (Pavliotis & Stuart, 2008). Data-driven reduction and multiscale drift estimation study which coarse variables and effective coefficients can be recovered (Papavasiliou et al., 2009; Dsilva et al., 2016). Hierarchical neural time-steppers learn flow maps at several horizons (Liu et al., 2022), while MTS3 instantiates multiple temporal scales inside a probabilistic world model (Shaj et al., 2023). Stride-conditioned operators likewise learn a family of time-advancement maps; a multi-stepsize mixture-of-experts operator uses the requested stride to route temporal experts (Pan et al., 2026). If the training information consists only of the same coarse observational trajectories, fitting several strides constrains prediction without selecting a microscopic member of an exact observational fiber. Finer observed strides can change the information protocol for finite-scale models, so comparisons must state which trajectories are supplied. These works primarily ask how multiscale structure can be exploited for prediction or simulation. Our question is different: once a coarse observational law is fixed, does it determine the response to a specified intervention on a hidden fast mechanism?

Sampling and dynamical identifiability. A sampled transition law need not determine a unique continuous-time generator. Koopman-based sampling theory makes the corresponding aliasing conditions explicit (Zeng et al., 2024). We likewise distinguish a coarse transition law from its microscopic generators, but our target is not recovery of the unintervened generator. Instead, we ask whether a specified intervention functional is invariant over all microscopic models compatible with the same observational law.

Structural identifiability in partially observed stochastic systems. Structural identifiability studies which parameters, or combinations of parameters, are uniquely determined by an ideal observation protocol. Recent work extends this perspective to partially observed stochastic differential equations, using observable moment dynamics to characterize identifiable combinations in linear and near-linear SDEs, including Ornstein–Uhlenbeck systems (Browning et al., 2025). Their differential-algebra construction eliminates unobserved moments from recurrence relations and makes the initialization protocol part of the identification question. Here the frozen generator supplies explicit invariant moments, and the slow generator supplies the observable coefficient com-

binations. This is closely related to our slow–fast specialization, but the object of interest differs. We do not require the complete microscopic parameter vector to be identifiable. Instead, we ask whether a particular intervention target is constant over the remaining observational equivalence class. Consequently, a target may be identifiable even when parts of the microscopic generator are not, and target-sufficient cross-scale information may be weaker than that required for full system identification.

Causal representation learning under interventions. Causal representation learning also studies how additional environments or interventions reduce latent ambiguity. Weakly supervised intervention pairs can render latent causal representations identifiable (Brehmer et al., 2022), while stronger interventional assumptions yield identifiability up to restricted transformations or, under suitable conditions, recovery of latent causal structure (Ahuja et al., 2023; Varici et al., 2024). Our setting is complementary. We do not assume interventional training environments are available to identify the latent representation. Instead, we start from a coarse observational dynamical law and ask whether a *specified future intervention query* is already determined, and if not, which cross-scale measurements are sufficient to make that target identifiable. Perfect interventions in Ahuja et al. (2023) identify latent factors up to permutation and scaling under their assumptions, whereas imperfect interventions can leave block-affine ambiguity. A numerical replacement value in a dynamical intervention must therefore be specified in a calibrated coordinate; invariance of an observational representation does not by itself fix the meaning of that replacement.

Causal abstraction and intervention consistency. Exact and approximate causal abstractions require agreement under specified interventions, not only under observational distributions (Rubenstein et al., 2017; Beckers et al., 2020). Our observational-fiber criterion gives a target-specific formulation of this distinction for stochastic dynamical models. The OU specialization makes the equivalence class induced by temporal aggregation explicit, derives a non-vanishing intervention separation, and identifies cross-scale measurements that are sufficient for selected intervention targets.

Identifiability up to symmetries. Latent-variable and causal representation models are often identifiable only up to transformations that leave the relevant observed distributions unchanged (Brehmer et al., 2022; Ahuja et al., 2023). We adopt the same quotient perspective, but the object of interest is not necessarily the latent representation or microscopic parameter itself. The relevant question is whether an intervention functional is invariant on the observational quotient. This target-specific view permits cross-scale anchors that identify the requested response without requiring complete identification of the underlying microscopic model.

3 Problem Definition

Let Θ be a class of stochastic dynamical models. An observation protocol is represented by a map $O : \Theta \to \mathcal{P}(\mathcal{Y})$, where $O(\theta)$ denotes the population law of the observed trajectory under model θ.

The exact identifiability results use the complete observational law rather than a finite sample. This gives the observer the strongest information available under the specified observation protocol and therefore separates non-identifiability from estimation error. In particular, if two models satisfy $O(\theta) = O(\theta')$, then no amount of additional data drawn from this same protocol can distinguish them. Any common, parameter-independent Markov observation kernel, such as discrete subsampling followed by independent measurement noise, preserves equality of laws. Such coarsening cannot resolve an ambiguity already present in the complete law.

An intervention $u \in \mathcal{U}$ has an associated target $T_u : \Theta \to (\mathcal{Z}, d_Z)$, where the target may be a scalar response coefficient, a post-intervention transition law, a trajectory distribution, or another functional of the intervened model. The metric d_Z is chosen to match the target. For scalar targets such as a post-intervention drift coefficient, the metric is absolute distance; for distribution-valued targets, one may instead use a probability metric such as total variation or Wasserstein distance.

Definition 1 (Observational fiber)**.** *The observational fiber through θ is* $[\theta]_O := \{\theta' \in \Theta : O(\theta') = O(\theta)\}$.

Definition 2 (Intervention identifiability)**.** *The target T_u is identifiable from O on Θ if there exists a map ψ_u on $O(\Theta)$ such that $T_u = \psi_u \circ O$.*

Identifiability here is set-theoretic factorization; no continuity, statistical consistency, or measurable selector is asserted without additional assumptions.

This definition distinguishes three questions: prediction of the observational law, identification of microscopic parameters, and identification of an intervention target. A microscopic parameter may remain unidentified while a particular intervention target is identifiable. Conversely, a model may recover the observational law exactly while failing to identify a mechanism intervention.

An auxiliary cross-scale measurement is a map $A : \Theta \to \mathcal{A}$. The augmented observation protocol is (O, A). The measurement A is a *target-sufficient anchor* for intervention u when T_u is identifiable from this joint map. Target sufficiency is weaker than microscopic identifiability: the joint fiber may contain multiple microscopic models provided that they all agree on the requested target.

Exact identifiability depends only on equality of observational laws and does not require a particular metric on $\mathcal{P}(\mathcal{Y})$. For approximate statements, observational laws are equipped with a metric d_O. Typical choices include total variation, Wasserstein distance, or another probability metric appropriate to the observation space.

The target modulus is defined as $\omega_u(\delta) := \sup_{\theta,\theta' :\, d_O(O(\theta),O(\theta'))\le\delta} d_Z(T_u(\theta), T_u(\theta'))$.

The value $\omega_u(0)$ is the largest intervention separation that can remain inside an exact observational fiber. More generally, $\omega_u(\delta)$ quantifies how different intervention responses can remain among models whose observational laws are only δ apart. This approximate form will be useful when relating exact averaged-model identifiability to finite-ε slow–fast systems.

4 Identifiability on Observational Fibers

Theorem 1 (Fiber characterization)**.** *For a fixed intervention u, the following are equivalent: (i) T_u is identifiable from O; (ii) T_u is constant on every observational fiber; and (iii) $\omega_u(0) = 0$. The same statements hold for an augmented protocol after replacing O by (O, A).*

The theorem turns intervention identification into a property of the observation quotient, rather than of an estimator. Its main force comes from explicit fibers and from anchors that collapse them.

Proposition 1 (Two-point lower bound)**.** *Suppose θ_0, θ_1 induce the same law for every sample available to an estimator $\widehat{T}$, and let $\Delta = d_Z(T_u(\theta_0), T_u(\theta_1))$. Then*

$$\inf_{\widehat{T}} \max_{i\in\{0,1\}} \mathbb{E}_{\theta_i} d_Z\big(\widehat{T}, T_u(\theta_i)\big) \ge \Delta/2. \tag{1}$$

The bound holds for every sample size, including an infinite observational trajectory.

Proposition 2 (Gauge certificate)**.** *Let a group G act on Θ and suppose $O(g\theta) = O(\theta)$ for all $g \in G$. If for some g, θ, u one has $T_u(g\theta) \neq T_u(\theta)$, then T_u is not identifiable from O. Conversely, an anchor is target-sufficient if T_u is constant on every joint fiber of (O, A), even if the joint fiber contains multiple microscopic models.*

These results also clarify what clock factorization cannot establish. A deterministic multirate schedule with commensurate periods can be represented by a base-clock transition after augmenting the state with schedule phase and aggregation buffers. The stochastic version uses the ordered composition of the scheduled Markov kernels. Hence clock count alone does not remove an observational fiber. Additional measurements or restrictions on the model class do.

5 A Slow–Fast Intervention Separation

Model, time scale, and intervention. Let $\theta = (a, b, C, \kappa, \sigma_S, \sigma_M)$, with $\kappa > 0$, $\sigma_S, \sigma_M > 0$, and let $\varepsilon > 0$ be a known time-scale parameter. The Brownian motions W and V are independent. An intervention acts on a specified microscopic model: u_c changes C to the prescribed value c and holds $(a, b, \kappa, \sigma_S, \sigma_M, \varepsilon)$ fixed. Across an observational fiber these coefficients may vary jointly; holding them fixed under an intervention does not mean that they are known from observational data.

More generally, a mean-function replacement u_g maps the frozen generator $\kappa(\mu_\theta(m) - s)\partial_s + (\sigma_S^2/2)\partial_{ss}$ to $\kappa(g(m)-s)\partial_s+(\sigma_S^2/2)\partial_{ss}$. It replaces the associated transition semigroup as a whole, preserving its relaxation rate and diffusion scale. It is a mechanism intervention, not conditioning on $S = g(m)$ or setting the fast noise to zero. The linear choice $g(m) = cm$ gives u_c.

Consider the fast conditional dynamics and slow equation

$$dS_t = \kappa(CM_t - S_t)\,dt + \sigma_S\,dW_t, \tag{2}$$

$$dM_t = \varepsilon(aM_t + bS_t^2)\,dt + \sqrt{\varepsilon}\,\sigma_M\,dV_t. \tag{3}$$

For a controlled experiment that holds M fixed at m, write $\pi_{\theta,m}$ for the invariant fast law. Throughout, expressions such as $\mathbb{E}[S \mid M = m]$ and $\mathrm{Var}(S \mid M = m)$ refer to this *frozen-process* law, not the conditional law of a freely evolving finite-ε joint process. The stationary fast law is

$$S \mid M = m \sim \mathcal{N}(Cm, r), \qquad r := \frac{\sigma_S^2}{2\kappa}. \tag{4}$$

The averaged slow model is

$$d\bar{M}_t = \left(\alpha\bar{M}_t + \beta_0 + \beta_2\bar{M}_t^2\right)dt + \gamma\,dV_t, \tag{5}$$

where

$$\alpha = \varepsilon a, \qquad \beta_0 = \varepsilon br, \qquad \beta_2 = \varepsilon bC^2, \qquad \gamma = \sqrt{\varepsilon}\sigma_M. \tag{6}$$

On slow time $u = \varepsilon t$, the quadratic coefficients are $B_2 = \beta_2/\varepsilon = bC^2$ observationally and $Q_c = q_c/\varepsilon = bc^2$ after replacement. Comparisons as $\varepsilon \to 0$ use a fixed slow-time horizon and fixed microscopic parameters. A fixed separation in Q_c corresponds to an ε-scaled separation in the physical-time coefficient q_c.

To avoid irrelevant explosion questions from the quadratic example, fix a compact interval D with nonempty interior and observe the process stopped at its first exit from D. Assume $\gamma > 0$ and that the initial distribution has positive density on D. Our observational protocol is this complete stopped-path law on a fixed finite horizon, with a common known initial slow law.

Lemma 1 (Reduced-coefficient identification)**.** *The stopped-path law of equation* (5) *uniquely determines* $(\alpha, \beta_0, \beta_2, \gamma^2)$.

Thus the issue below is not failure to learn the reduced dynamics, but what those coefficients determine about the hidden fast mechanism.

The intervention changes the averaged quadratic coefficient to $q_c = \varepsilon bc^2$, while preserving α, β_0, and γ within each microscopic model. The target metric is $d_Z(q, q') = |q - q'|$.

Theorem 2 (Prediction-equivalent, intervention-different)**.** *Fix reduced coefficients* $(\alpha, \beta_0, \beta_2, \gamma)$ *with* $\beta_0, \beta_2 > 0$*. There exists a continuum of microscopic parameterizations that induce the same complete stopped-path law of the averaged process.*

For the mean-response replacement intervention u_c *with* $c \neq 0$*, the target* q_c *is not identifiable from this observational law. In particular, for any two admissible parameterizations with* $C_1^2 \neq C_2^2$*, the following relation holds:* $q_c(\theta_i) = \frac{\beta_2 c^2}{C_i^2}$, $\quad i \in \{1, 2\}$*, and hence* $q_c(\theta_1) \neq q_c(\theta_2)$*. Consequently, the two-point argument yields the minimax lower bound* $\inf_{\widehat{T}} \max_{i\in\{1,2\}} \mathbb{E}_{\theta_i} d_Z\left(\widehat{T}, T_{u_c}(\theta_i)\right) \geq \frac{|\beta_2|c^2}{2}\left|C_1^{-2} - C_2^{-2}\right|$.

Exact scaling and finite time-scale separation. For $\lambda > 0$, the transformation $(C, b, \sigma_S, S_0) \mapsto (\lambda C, b/\lambda^2, \lambda\sigma_S, \lambda S_0)$ with (a, κ, σ_M) fixed preserves the full finite-ε slow path under the coupling $S'_t = \lambda S_t$, $M'_t = M_t$. It therefore defines an exact observational subfiber, provided the hidden initialization transforms with the model. Applying the same numerical replacement c gives different $Q_c = bc^2$ on this subfiber. This symmetry acts on a calibrated physical fast variable: a pure change of units would also transform c and would describe a different query.

The larger averaged fiber also permits changes in κ that need not preserve the finite-ε law. For uniform approximation, restrict the microscopic coefficients to a compact set with $\kappa \geq \kappa_* > 0$, uniformly bounded eighth initial fast moments, and a common slow initial law. A smooth bounded

continuation of the slow coordinate outside D gives a nonexplosive localized model. On a fixed slow-time interval, uniform frozen mixing and the polynomial OU Poisson solution give $\sup_\theta W_1(O_\varepsilon^{\mathrm{loc}}(\theta), \bar{O}^{\mathrm{loc}}(\theta)) \le K_T\sqrt{\varepsilon}$, where W_1 uses the supremum path norm. Two models on the same averaged fiber consequently have localized observational distance at most $2K_T\sqrt{\varepsilon}$, while their slow-time targets can remain separated. The localization, moment estimates, and coupling argument are specified with the proofs. This statement concerns the continued processes; stopping each process at its own exit requires a separate exit-time argument.

Small Wasserstein distance quantifies approximation of Lipschitz path functionals. It does not imply indistinguishability of arbitrary tests or transfer an exact minimax bound to unlimited finite-ε data. For sample laws P_0, P_1, an approximate two-point bound instead involves $\Delta(1 - \mathrm{TV}(P_0, P_1))/2$.

Let a population anchor contain the conditional mean at two distinct slow states $m_1 \neq m_2$ and the conditional variance at either state:

$$A_{\mathrm{stat}}(\theta) = \Big(\mathbb{E}[S \mid M = m_1], \mathbb{E}[S \mid M = m_2], \mathrm{Var}(S \mid M = m_1)\Big). \tag{7}$$

Theorem 3 (Stationary anchor for mean replacement)**.** *Assume $C \neq 0$ and $m_1 \neq m_2$. Under (O, A_{stat}), the replacement target is identifiable:*

$$C = \frac{\mathbb{E}[S \mid M = m_2] - \mathbb{E}[S \mid M = m_1]}{m_2 - m_1}, \qquad q_c = \frac{\beta_2 c^2}{C^2}. \tag{8}$$

The variance additionally identifies $r = \frac{\sigma_S^2}{2\kappa}$. For $r > 0$ and $C \neq 0$, the identified quantities satisfy the consistency relation

$$\frac{\beta_0}{r} = \frac{\beta_2}{C^2} = \varepsilon b. \tag{9}$$

Stationary snapshots alone do not identify κ and σ_S separately.

Target sufficiency versus anchor minimality. For $c \neq 0$ and $\beta_2 \neq 0$, the mean-response replacement target satisfies $q_c = \frac{\beta_2 c^2}{C^2}$. Hence the target is identifiable whenever the augmented observation protocol identifies C^2. Conversely, if C^2 remains free within a joint observational fiber, then q_c remains non-identifiable.

Thus identification of C^2 is necessary and sufficient on the augmented fiber. In the zero-intercept family, one frozen mean at any known $m \neq 0$ already gives $C = \mathbb{E}[S \mid M = m]/m$; a mean at $m = 0$ contains no information about C. Two distinct means are sufficient and also eliminate a common unknown additive sensor offset. They are minimal within that offset-contaminated mean-only protocol, since one equation $\mathbb{E}[Y \mid m] = Cm + d$ cannot separate C and d.

A calibrated variance is an alternative target-sufficient statistic: $q_c = \beta_0 c^2/r$ when $\beta_0, r > 0$. It identifies $C^2 = \beta_2 r/\beta_0$ without identifying the sign of C. Thus neither two means nor variance is universally required; the choice depends on what can be measured reliably.

Corollary 1 (One lag identifies fast rate and scale)**.** *If the anchor additionally contains a stationary lag covariance at a known $\tau > 0$, then, for $r > 0$,*

$$\kappa = -\tau^{-1} \log \frac{\mathrm{Cov}(S_t, S_{t+\tau} \mid M = m)}{r}, \qquad \sigma_S = \sqrt{2\kappa r}. \tag{10}$$

Thus one positive lag breaks the remaining OU rate–diffusion gauge.

Theorem 4 (Robust transport from an approximate anchor)**.** *Let $|C| \ge C_{\min} > 0$, $|\widehat{C} - C| \le C_{\min}/2$, and $|\widehat{\beta}_2 - \beta_2| \le e_\beta$. For*

$$\widehat{q}_c = \frac{\widehat{\beta}_2 c^2}{\widehat{C}^2}, \tag{11}$$

the following relation holds:

$$|\widehat{q}_c - q_c| \le c^2 \left(\frac{4 e_\beta}{C_{\min}^2} + \frac{16|\beta_2|}{C_{\min}^3} |\widehat{C} - C| \right). \tag{12}$$

Consequently the anchor is stable away from $C = 0$, while intervention transport is ill-conditioned as the observational mean response vanishes.

Discrete and noisy anchors. For frozen observations $Y_j = S_{j\tau} + \xi_j$, independent zero-mean sensor noise with known variance ν^2 preserves the mean and gives $\mathrm{Var}(Y) = r + \nu^2$ and $K_\ell = \mathrm{Cov}(Y_j, Y_{j+\ell}) = re^{-\kappa\ell\tau}$ for $\ell \geq 1$. Subtracting ν^2 therefore preserves the variance and one-lag identification formulas. If ν^2 is unknown, two positive lags identify $\kappa = -\tau^{-1}\log(K_2/K_1)$ and $r = K_1^2/K_2$ when $0 < K_2 < K_1$. Temporally correlated noise requires a calibrated covariance correction; noisy labels for the frozen slow state require a separate measurement-error model. For slow-only data, a common sampling/noise kernel preserves equality of observational laws, but equality alone does not ensure that sampled data identify the reduced coefficients.

Drift approximation and target bias. Suppose the fitted quadratic coefficient has estimation error e_{est} and model approximation bias δ_β. The transport bound holds with $e_\beta = e_{\mathrm{est}} + |\delta_\beta|$. More concretely, if a physical-time reduced drift is $\alpha m + \beta_0 + \beta_2 m^2 + R(m)$ with $\|R\|_\infty \leq \rho$, and fitting uses a design measure ν, put $\phi(m) = (1, m, m^2)^\top$ and $G = \int \phi\phi^\top d\nu$. If G is nonsingular, the projection bias is bounded by $\rho\|e_3^\top G^{-1}\| \int \|\phi\| d\nu$. This bounds coefficient-induced error for the specified OU intervention; an additional discrepancy b_u between the true intervention target and $\beta_2 c^2/C^2$ adds $|b_u|$ to the error bound. Observational drift fit alone cannot bound that mechanism discrepancy.

Invariant targets and non-Gaussian mechanisms. The observational coefficients $(\alpha, \beta_0, \beta_2, \gamma^2)$ and bounded functionals of the averaged path are identifiable without anchors. A proportional mean change $C \leftarrow \lambda C$ also has identifiable quadratic response $\lambda^2\beta_2$, unlike replacement by a fixed c. For a centered location-scale fast invariant family $S = Cm + sZ$, with fixed centered non-Gaussian law for Z and finite variance v_Z, the same quadratic coupling depends only on C^2 and $s^2 v_Z$. Mean replacement that preserves the residual law therefore has the same target-sufficiency formulas. The exponential lag-covariance identity is specific to OU dynamics. A nonlinear diffusion construction and the role of higher moments for other couplings are given with the proofs.

6 Finite-Scale Dynamics and Sampled Anchors

The numerical study separates dynamical approximation from estimation of a target-sufficient statistic. All measurements are generated synthetically; reported uncertainty is the standard error across independent random seeds.

Finite-scale observational and interventional behavior. The microscopic parameters are fixed at $(a, b, C, \kappa, \sigma_S) = (-0.5, 1, 0.3, 1, 0.5)$ and $(-0.5, 0.25, 0.6, 2, \sqrt{2})$, with $\sigma_M = 0.3$ in both models. Their slow-time reduced coefficients are $(a, br, bC^2, \sigma_M) = (-0.5, 0.125, 0.09, 0.3)$. Simulations cover $u = \varepsilon t \in [0, 1]$, starting at $M_0 = 0.5$ with stationary frozen fast initialization. Terminal slow variables are compared under shared Brownian increments. The numerical integrator uses a frozen-OU fast update and an explicit slow update with step $\varepsilon/40$. Each of 30 seeds supplies 128 paths. The interval $[-2, 2]$ is absorbing at the discretized boundary; no path reaches that boundary in this sweep.

The coupled observational RMS decreases from 0.038 at $\varepsilon = 0.5$ to 0.010 at $\varepsilon = 0.02$, and the discrepancy from the common averaged process decreases from 0.112 to 0.026. Under $C \leftarrow 0.45$, the slow-time coefficient gap is 0.151875, while the terminal mean gap remains about 0.031–0.035. Each intervention run starts from its own frozen equilibrium; these are post-equilibration responses, not instantaneous switches of a shared hidden state. A same-rate scaling pair has identical observational slow paths to machine precision. The different-rate results measure approach to averaging, whereas the same-rate construction measures an exact symmetry.

Which fast statistics identify the target? At each known frozen state $m = -1, +1$, $N/4$ independent stationary pairs are collected at lag 1: exactly N scalar measurements per run. All feature ablations use these same samples and independent sensor noise. The mean estimate uses $\widehat{q}_c = \beta_2 c^2/\widehat{C}^2$; the variance alternative uses $\widehat{q}_c = \beta_0 c^2/\widehat{r}$, subtracting the known noise variance. Both receive the same population reduced coefficients, isolating anchor estimation from slow-drift estimation. The lag statistic is used only to separate rate and diffusion scale.

With sensor standard deviation 0.1, increasing N from 100 to 2000 reduces mean-anchor target error from 0.210 to 0.051 and variance-anchor error from 0.148 to 0.032. Variance is therefore an alternative route to this target, rather than a mandatory addition to the mean anchor. At $N = 500$, the rate error is 0.134, compared with target errors 0.098 and 0.059. Separate sweeps use contiguous fast samples and noisy slow observations. At fixed fast sample count, closely spaced observations produce higher error because of dependence. When the slow quadratic coefficient is also estimated from short noisy bursts, relative target error is 0.499 without slow sensor noise and 0.662 at sensor standard deviation 0.05. Population target sufficiency therefore does not remove drift-estimation error or poor finite-sample conditioning.

Table 1: Finite-scale slow dynamics on $u \in [0, 1]$ with fixed microscopic coefficients. Entries are mean $\pm$ standard error across 30 seeds of 128 paths. RMS discrepancies use a specified synchronous coupling at the terminal time; they are not estimates of total variation or of a full-path distribution metric. The mean gap compares the two equilibrated post-intervention systems.

ε	Coupled obs. RMS	Averaging RMS	Post-intervention mean gap
0.5	0.0381 ± 0.0008	0.1119 ± 0.0029	0.0330 ± 0.0014
0.2	0.0301 ± 0.0008	0.0829 ± 0.0020	0.0346 ± 0.0010
0.1	0.0214 ± 0.0005	0.0580 ± 0.0011	0.0326 ± 0.0006
0.05	0.0150 ± 0.0002	0.0411 ± 0.0006	0.0315 ± 0.0005
0.02	0.0097 ± 0.0002	0.0263 ± 0.0004	0.0308 ± 0.0005

Table 2: Anchor feature ablation using exactly the same N scalar fast measurements, with sensor standard deviation 0.1 and known reduced coefficients. Entries are mean relative error $\pm$ standard error across 200 seeds. The rate estimate is valid in 199/200 runs at $N = 100$ and 200/200 otherwise; reported rate errors condition on validity. All target estimates are valid.

N	Mean anchor: q_c	Variance anchor: q_c	Variance: r	Lag: κ
100	0.210 ± 0.013	0.148 ± 0.009	0.134 ± 0.007	0.409 ± 0.030
500	0.098 ± 0.006	0.059 ± 0.003	0.058 ± 0.003	0.134 ± 0.009
2000	0.051 ± 0.003	0.032 ± 0.002	0.032 ± 0.002	0.066 ± 0.004

7 Implications for World Models

The theory yields three implications for learned world models.

Evaluate targets, not parameter names. An unidentified microscopic parameter does not by itself prevent a target from being identified. Conversely, excellent observational likelihood does not certify an intervention target that varies along the observational fiber. Evaluation should therefore specify both the intervention target and the information protocol before reporting parameter or mechanism recovery.

Match information before comparing architectures. An architecture supplied with fast stationary moments or mechanism-level information is not information-equivalent to one trained only on slow trajectories. Such a comparison measures the value of the additional cross-scale information jointly with the architectural inductive bias. In the OU model, identifying C^2 makes the mean-response replacement target constant on the augmented observational fiber.

Target identification may require less information than full system identification. A specific intervention target need not require recovery of the entire fast generator. In the OU example, conditional means of the fast variable identify the mean-response coefficient C, which is sufficient to determine the mean-replacement target q_c. The stationary variance additionally identifies the composite scale

$$r = \frac{\sigma_S^2}{2\kappa},$$

but does not separate κ from σ_S. Recovering these two generator parameters separately requires temporal fast information, such as one positive-lag covariance. Thus the information required depends on the target being queried.

For illustration, the parameter sets

$$(a, b, C, \kappa, \sigma_S) = (-0.5, 1, 0.3, 1, 0.5) \tag{13}$$

and

$$(-0.5, 0.25, 0.6, 2, \sqrt{2}) \tag{14}$$

have identical reduced coefficients

$$(br, a, bC^2) = (0.125, -0.5, 0.09). \tag{15}$$

After the common replacement $C \leftarrow 0.45$, their quadratic coefficients are 0.2025 and 0.050625 before multiplication by ε. This deterministic witness is reproduced by the accompanying audit script; it illustrates observational equivalence with intervention separation, not empirical evidence for a general architectural advantage.

8 Discussion and Conclusion

We characterized when a coarse observational world model determines the response to an intervention on a hidden fast mechanism. Exact observational prediction is sufficient for a specified intervention target only when that target is constant on the corresponding observational fiber. In the OU-driven slow–fast model, the averaged slow law leaves a continuum of fast mechanisms observationally equivalent, while a mean-response replacement of the fast mechanism produces different coarse responses across that equivalence class. Target-sufficient cross-scale information can collapse the relevant ambiguity without requiring complete recovery of the microscopic generator.

The explicit OU calculations separate two mechanisms of observational ambiguity. A fast-state scaling preserves full finite-scale slow paths under compatible initialization, whereas changing the fast relaxation rate generally preserves only the averaged law. Uniform approximation over compact, mixing parameter families controls Lipschitz path observables after smooth localization; it is distinct from a statistical indistinguishability guarantee for arbitrary finite-scale data. The non-Gaussian location-scale construction retains the moment-based target identification, while the OU rate formula relies on exponential covariance decay.

The numerical study measures these distinctions in a controlled setting. Frozen anchors provide calibrated cross-scale information, and finite-sample errors depend on correlation, sensor calibration, and drift estimation. The experiments do not establish identification from uncontrolled noisy slow-only trajectories or finite-sample learning rates for neural estimators. The architecture control measures observational prediction under a common input protocol and does not test intervention recovery.

Our population-level observation protocol also separates identifiability from statistical estimation. The complete stopped-path law is intentionally stronger than the discrete, noisy, and partially observed trajectories available in practice. Such observation schemes cannot remove an ambiguity that already exists at the population-law level, but they introduce additional statistical questions about finite-sample recovery of the proposed anchors and intervention targets.

The practical message is therefore target-specific. Good observational fit does not by itself validate an intervention on a hidden mechanism. A valid intervention claim must either establish that the requested target is invariant over the observational fiber or specify additional cross-scale information that makes the target identifiable. This distinction separates observational prediction, microscopic system identification, and intervention validity in multiscale world models.

AI use statement

Generative AI tools were used to assist with language editing, clarity, and structural refinement of the manuscript. All mathematical arguments, theoretical claims, experimental design, numerical results, and cited sources were reviewed and verified by the authors. The authors take full responsibility for the content of the paper.

Reproducibility statement

The principal results of this work are analytical and can be reproduced directly from the definitions, constructions, and proofs given in the paper and appendices. The multiscale observation maps, intervention classes, notions of observational and interventional equivalence, and identifiability criteria are defined explicitly. Each non-identifiability result is accompanied by a constructive pair or family of world models satisfying the stated observational constraints while differing under the specified interventions. The appendices provide the intermediate lemmas and proof details needed to verify the main theorems. Numerical illustrations are generated from the stated model equations and parameter settings and are intended as direct demonstrations of the theoretical constructions rather than as evidence dependent on large-scale training. Consequently, the central claims can be independently checked without access to proprietary data or models.

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407, 2023.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 606–615, 2020. URL https://proceedings.mlr.press/v115/beckers20a.html.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alexander P. Browning, Michael J. Chappell, Hamid Rahkooy, Torkel E. Loman, and Ruth E. Baker. Exact identifiability analysis for a class of partially observed near-linear stochastic differential equation models. *arXiv preprint arXiv:2503.19241*, 2025.

Carmeline J. Dsilva, Ronen Talmon, C. William Gear, Ronald R. Coifman, and Ioannis G. Kevrekidis. Data-driven reduction for a class of multiscale fast-slow stochastic dynamical systems. *SIAM Journal on Applied Dynamical Systems*, 15(3):1327–1351, 2016. doi: 10.1137/151004896. URL https://doi.org/10.1137/151004896.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Yuying Liu, J. Nathan Kutz, and Steven L. Brunton. Hierarchical deep learning of multiscale differential equation time-steppers. *Philosophical Transactions of the Royal Society A*, 380(2229):20210200, 2022. doi: 10.1098/rsta.2021.0200. URL https://doi.org/10.1098/rsta.2021.0200.

Guanyu Pan, Huiyu Yang, Yunpeng Wang, Zikun Xu, Jianchun Wang, and Nianyu Yi. Stable fine-time-step long-horizon turbulence prediction with a multi-stepsize mixture-of-experts neural operator. *arXiv preprint arXiv:2604.12794*, 2026. URL https://arxiv.org/abs/2604.12794.

Anastasia Papavasiliou, Grigorios A. Pavliotis, and Andrew M. Stuart. Maximum likelihood drift estimation for multiscale diffusions. *Stochastic Processes and their Applications*, 119(10):3173–3210, 2009. doi: 10.1016/j.spa.2009.05.003. URL https://doi.org/10.1016/j.spa.2009.05.003.

Grigorios A. Pavliotis and Andrew M. Stuart. *Multiscale Methods: Averaging and Homogenization*, volume 53 of *Texts in Applied Mathematics*. Springer, 2008. doi: 10.1007/978-0-387-73829-1. URL https://doi.org/10.1007/978-0-387-73829-1.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf.

Vaisakh Shaj, Saleh Gholam Zadeh, Ozan Demir, Luiz Ricardo Douat, and Gerhard Neumann. Multi time scale world models. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/54d8aab579b5a9ed3395764c7341ebec-Abstract-Conference.html.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2314–2322, 2024.

Zhexuan Zeng, Zuogong Yue, Alexandre Mauroy, Jorge Goncalves, and Ye Yuan. A sampling theorem for exact identification of continuous-time nonlinear dynamical systems. *IEEE Transactions on Automatic Control*, 69, 2024. doi: 10.1109/TAC.2024.3409639. URL https://doi.org/10.1109/TAC.2024.3409639.

A Proofs

A.1 Proof of Theorem 1

If $T_u = \psi_u \circ O$, then equal observational laws have equal targets, so the target is constant on every observational fiber.

Conversely, suppose that T_u is constant on every observational fiber. Define

$$\psi_u(P) = T_u(\theta)$$

for any θ satisfying $O(\theta) = P$. Fiber constancy makes this definition independent of the representative, and therefore

$$T_u = \psi_u \circ O.$$

Finally, $\omega_u(0) = 0$ exactly when every pair θ, θ' satisfying

$$O(\theta) = O(\theta')$$

also satisfies

$$d_{\mathcal{Z}}\left(T_u(\theta), T_u(\theta')\right) = 0.$$

Hence the three statements are equivalent.

The augmented claim is identical after replacing the observation map O by the joint map (O, A).

A.2 Proof of Proposition 1

For every realization of $\widehat{T}$, the triangle inequality gives

$$d_{\mathcal{Z}}\left(\widehat{T}, T_u(\theta_0)\right) + d_{\mathcal{Z}}\left(\widehat{T}, T_u(\theta_1)\right) \geq \Delta, \tag{16}$$

where

$$\Delta = d_{\mathcal{Z}}\left(T_u(\theta_0), T_u(\theta_1)\right).$$

By assumption, the data distribution is identical under θ_0 and θ_1, and hence the distribution of $\widehat{T}$ is also identical under the two models. Taking expectation under this common distribution yields

$$\mathbb{E}_{\theta_0} d_{\mathcal{Z}}\left(\widehat{T}, T_u(\theta_0)\right) + \mathbb{E}_{\theta_1} d_{\mathcal{Z}}\left(\widehat{T}, T_u(\theta_1)\right) \geq \Delta.$$

Therefore at least one of the two risks is at least $\Delta/2$, giving

$$\inf_{\widehat{T}} \max_{i \in \{0,1\}} \mathbb{E}_{\theta_i} d_{\mathcal{Z}}\left(\widehat{T}, T_u(\theta_i)\right) \geq \frac{\Delta}{2}.$$

A.3 Proof of Proposition 2

Suppose there exist $g \in G$ and $\theta \in \Theta$ such that

$$O(g\theta) = O(\theta)$$

but

$$T_u(g\theta) \neq T_u(\theta).$$

Then θ and $g\theta$ lie in the same observational fiber while having distinct intervention targets. By Theorem 1, T_u is therefore not identifiable from O.

For the anchor statement, apply Theorem 1 to the augmented observation map

$$\theta \mapsto \big(O(\theta), A(\theta)\big).$$

The anchor is target-sufficient exactly when T_u is constant on every joint fiber of this augmented map, even if multiple microscopic models remain in the same joint fiber.

A.4 Proof of Lemma 1

For every interior state m, the stopped-path law of the averaged process determines the infinitesimal conditional first and second moments. In particular,

$$\lim_{h\downarrow 0} h^{-1}\mathbb{E}\left[\bar{M}_{t+h} - m \mid \bar{M}_t = m\right] = \beta_0 + \alpha m + \beta_2 m^2 \tag{17}$$

and

$$\lim_{h\downarrow 0} h^{-1}\mathbb{E}\left[(\bar{M}_{t+h} - m)^2 \mid \bar{M}_t = m\right] = \gamma^2. \tag{18}$$

Positive initial density and nondegenerate diffusion make these quantities available throughout the interior of the stopped domain. Equality of the first expression on an interval identifies the coefficients of the polynomial drift,

$$(\alpha, \beta_0, \beta_2),$$

while the second expression identifies γ^2.

Hence the complete stopped-path law uniquely determines

$$(\alpha, \beta_0, \beta_2, \gamma^2).$$

A.5 Proof of Theorem 2

Fix reduced coefficients

$$(\alpha, \beta_0, \beta_2, \gamma), \qquad \beta_0, \beta_2 > 0.$$

Choose any

$$C \neq 0$$

and any

$$\varepsilon > 0,$$

and define

$$b = \frac{\beta_2}{\varepsilon C^2}, \qquad r = \frac{\beta_0 C^2}{\beta_2}, \qquad a = \frac{\alpha}{\varepsilon}, \qquad \sigma_M = \frac{\gamma}{\sqrt{\varepsilon}}. \tag{19}$$

For any $\kappa > 0$, set

$$\sigma_S = \sqrt{2\kappa r}. \tag{20}$$

Then

$$r = \frac{\sigma_S^2}{2\kappa},$$

and the induced averaged coefficients are

$$\varepsilon a = \alpha, \qquad \varepsilon b r = \beta_0, \qquad \varepsilon b C^2 = \beta_2, \qquad \sqrt{\varepsilon}\sigma_M = \gamma.$$

Thus all such microscopic parameterizations induce the same averaged SDE and therefore the same complete stopped-path law. Varying C, ε, or κ produces a continuum of microscopic models within the same observational fiber.

Under the mean-response replacement intervention u_c, the microscopic coefficient C is replaced by the common value c while the remaining microscopic coefficients are held fixed. The resulting post-intervention quadratic coefficient is

$$q_c(\theta) = \varepsilon b c^2 = \frac{\beta_2 c^2}{C^2}. \tag{21}$$

Hence for two admissible parameterizations satisfying

$$C_1^2 \neq C_2^2,$$

we obtain

$$q_c(\theta_1) = \frac{\beta_2 c^2}{C_1^2} \neq \frac{\beta_2 c^2}{C_2^2} = q_c(\theta_2).$$

Therefore q_c is not identifiable from the averaged observational law.

Applying Proposition 1 to the two models gives

$$\inf_{\widehat{T}} \max_{i\in\{1,2\}} \mathbb{E}_{\theta_i} d_{\mathcal{Z}}\left(\widehat{T}, T_{u_c}(\theta_i)\right) \geq \frac{|\beta_2| c^2}{2} \left| C_1^{-2} - C_2^{-2} \right|.$$

A.6 Proof of Theorem 3

For a frozen slow state $M = m$, the stationary fast OU law satisfies

$$\mathbb{E}[S \mid M = m] = Cm$$

and

$$\mathrm{Var}(S \mid M = m) = r = \frac{\sigma_S^2}{2\kappa}.$$

Therefore, for two distinct slow states $m_1 \neq m_2$,

$$C = \frac{\mathbb{E}[S \mid M = m_2] - \mathbb{E}[S \mid M = m_1]}{m_2 - m_1}. \tag{22}$$

The reduced coefficient β_2 is already identified from the slow observational law by Lemma 1. Hence

$$q_c = \frac{\beta_2 c^2}{C^2} \tag{23}$$

is identified from the augmented observation protocol (O, A_{stat}).

Conversely, for $c \neq 0$ and $\beta_2 \neq 0$, if C^2 remains free within a joint observational fiber, then the construction in Theorem 2 preserves the slow observational law while changing

$$q_c = \frac{\beta_2 c^2}{C^2}.$$

Thus identification of C^2 is necessary and sufficient for this target under the stated nondegeneracy conditions.

The conditional stationary variance additionally identifies

$$r = \frac{\sigma_S^2}{2\kappa},$$

but does not separate κ from σ_S.

A.7 Proof of Corollary 1

For a stationary Ornstein–Uhlenbeck process with frozen slow state $M = m$,

$$\mathrm{Cov}\left(S_t, S_{t+\tau} \mid M = m\right) = re^{-\kappa\tau}. \tag{24}$$

Since r is already identified from the stationary variance, one positive-lag covariance at known $\tau > 0$ gives

$$\kappa = -\frac{1}{\tau}\log\left(\frac{\mathrm{Cov}(S_t, S_{t+\tau} \mid M = m)}{r}\right). \tag{25}$$

Once κ is known,

$$\sigma_S = \sqrt{2\kappa r}. \tag{26}$$

Thus one positive-lag covariance separates the fast mixing rate from the diffusion scale.

A.8 Proof of Theorem 4

Recall

$$q_c = \frac{\beta_2 c^2}{C^2}, \qquad \widehat{q}_c = \frac{\widehat{\beta}_2 c^2}{\widehat{C}^2}.$$

Add and subtract

$$\frac{\beta_2 c^2}{\widehat{C}^2}.$$

Then

$$|\widehat{q}_c - q_c| \le c^2 \frac{|\widehat{\beta}_2 - \beta_2|}{|\widehat{C}|^2} + c^2|\beta_2|\left|\widehat{C}^{-2} - C^{-2}\right| \tag{27}$$

$$\le c^2 \frac{e_\beta}{|\widehat{C}|^2} + c^2|\beta_2|\left|\widehat{C}^{-2} - C^{-2}\right|. \tag{28}$$

Because

$$|C| \ge C_{\min}$$

and

$$|\widehat{C} - C| \le \frac{C_{\min}}{2},$$

we have

$$|\widehat{C}| \ge \frac{C_{\min}}{2}.$$

Therefore

$$c^2 \frac{e_\beta}{|\widehat{C}|^2} \le \frac{4c^2 e_\beta}{C_{\min}^2}.$$

The derivative of

$$x \mapsto x^{-2}$$

is

$$-2x^{-3}.$$

Along the interval between C and $\widehat{C}$, the absolute value of x is at least $C_{\min}/2$. Hence

$$\left|\frac{d}{dx}x^{-2}\right| \le \frac{16}{C_{\min}^3}.$$

By the mean-value theorem,

$$\left|\widehat{C}^{-2} - C^{-2}\right| \le \frac{16}{C_{\min}^3}|\widehat{C} - C|.$$

Combining the two bounds yields

$$|\widehat{q}_c - q_c| \le c^2\left(\frac{4e_\beta}{C_{\min}^2} + \frac{16|\beta_2|}{C_{\min}^3}|\widehat{C} - C|\right). \tag{29}$$

Thus intervention transport is stable away from $C = 0$ and becomes ill-conditioned as the observational mean-response coefficient approaches zero.

A.9 Uniform averaging on a localized parameter family

Let π be a fixed smooth bounded function with bounded first two derivatives that equals the identity on the compact interval D. In slow time, consider

$$dS_u^\varepsilon = \varepsilon^{-1}\kappa(C\pi(M_u^\varepsilon) - S_u^\varepsilon)du + \varepsilon^{-1/2}\sigma_S dW_u, \qquad dM_u^\varepsilon = (a\pi(M_u^\varepsilon) + b(S_u^\varepsilon)^2)du + \sigma_M dV_u.$$

Its averaged drift is $\bar{F}(m) = a\pi(m) + b(C^2\pi(m)^2 + r)$, which is globally Lipschitz. The original and localized dynamics agree up to exit from D. The coefficients range over a compact set with $\kappa \geq \kappa_* > 0$; the initial slow variable has a common law and the initial fast variables have uniformly bounded eighth moments. Brownian motions are independent.

For $z = s - C\pi(m)$, the frozen Poisson equation $\mathcal{L}_m\chi = s^2 - C^2\pi(m)^2 - r$ has the explicit solution

$$\chi(m,s) = -\frac{z^2}{2\kappa} - \frac{2C\pi(m)z}{\kappa}.$$

Its coefficients and the required derivatives have uniform polynomial bounds. Frozen dissipativity and the bounded forcing $C\pi(m)$ give $\sup_{u\leq T,\theta,\varepsilon\leq 1}\mathbb{E}|S_u^\varepsilon|^8 < \infty$. Applying Itô's formula to the fourth moment and the martingale maximal inequality also gives $\mathbb{E}\sup_{u\leq T}|S_u^\varepsilon|^4 \leq K_T/\varepsilon$. The latter bound is deliberately coarse and suffices for the boundary term.

Write $R_u = \int_0^u [(S_v^\varepsilon)^2 - C^2\pi(M_v^\varepsilon)^2 - r]dv$. Itô's formula yields

$$R_u = \varepsilon(\chi_u - \chi_0) - \varepsilon\int_0^u \mathcal{L}_{\text{slow}}\chi_v\, dv - \sqrt{\varepsilon}\int_0^u \sigma_S\partial_s\chi_v\, dW_v - \varepsilon\int_0^u \sigma_M\partial_m\chi_v\, dV_v.$$

Here $\mathcal{L}_{\text{slow}} = F\partial_m + (\sigma_M^2/2)\partial_{mm}$. The preceding moment bounds control the boundary term and the cubic growth in $\mathcal{L}_{\text{slow}}\chi$. The martingale maximal inequality then gives $\sup_\theta \mathbb{E}\sup_{u\leq T}|R_u|^2 \leq K_T\varepsilon$. Couple the averaged and full slow processes with the same V and initial slow state. Their difference equals the integral of the difference in $\bar{F}$ plus bR_u. Gronwall's inequality gives

$$\sup_\theta \mathbb{E}\sup_{u\leq T}|M_u^\varepsilon - \bar{M}_u|^2 \leq K_T\varepsilon.$$

This coupling implies the stated supremum-path Wasserstein bound. Its constant depends on the compact parameter set, T, π, and the initial moment bound. It is not uniform as $\kappa_* \downarrow 0$ or over unbounded coefficients, and it does not establish total-variation convergence.

A.10 Exact scaling at finite separation

Take $\lambda > 0$, $C' = \lambda C$, $b' = b/\lambda^2$, $\sigma_S' = \lambda\sigma_S$, and $S_0' = \lambda S_0$, with common $a, \kappa, \sigma_M, \varepsilon, M_0, W, V$. Multiplying the fast SDE by λ shows that $S_t' = \lambda S_t$ solves the transformed fast equation when $M_t' = M_t$. Since $b'(S_t')^2 = bS_t^2$, the slow equations agree. Pathwise uniqueness up to the common exit time proves equality of the stopped slow paths. Any parameter-independent sampling and sensor kernel preserves this equality. The intervention $C \leftarrow c$ is not equivariant under this transformation: bc^2 changes to bc^2/λ^2. This is a statement about a fixed physical replacement value; changing units would also rescale c.

A.11 Alternative anchors and noisy moments

At a known nonzero frozen state the zero-intercept mean Cm identifies C. With unknown common additive sensor offset d, two distinct means determine both C and d, whereas one mean permits arbitrary C compensated by d. From the observational coefficients, $\beta_0 = \varepsilon br$ and $\beta_2 = \varepsilon bC^2$ imply $q_c = \beta_0 c^2/r$ and $C^2 = \beta_2 r/\beta_0$ when $\beta_0, r > 0$. Thus variance alone can be sufficient in this family.

Independent centered measurement noise contributes only to lag-zero covariance. Consequently $K_1 = re^{-\kappa\tau}$ and $K_2 = re^{-2\kappa\tau}$ identify $r = K_1^2/K_2$ and $\kappa = -\tau^{-1}\log(K_2/K_1)$. Unknown temporal noise covariance or an unknown state-dependent sensor mean invalidates these cancellations.

A.12 Projection bias under reduced-drift misspecification

For a design distribution ν and feature vector $\phi = (1, m, m^2)^\top$, the population normal equations give the coefficient perturbation $G^{-1} \int \phi R \, d\nu$, where $G = \int \phi\phi^\top d\nu$. Taking the quadratic coordinate and applying the norm inequality yields $|\delta_\beta| \le \rho \|e_3^\top G^{-1}\| \int \|\phi\| d\nu$. The triangle inequality separates coefficient estimation, approximation, and any discrepancy in the intervention functional itself. The last term requires a mechanism assumption and cannot be inferred from the observed reduced drift alone.

A.13 A non-Gaussian frozen fast family

Let $U(z) = z^2/2 + \eta z^4/4$ with $\eta > 0$, and let $s_0 > 0$ denote a residual scale. For a frozen slow state, consider the nonlinear diffusion

$$dS_t = -\kappa s_0 U'((S_t - Cm)/s_0)dt + s_0\sqrt{2\kappa}\, dW_t.$$

Its invariant law is $S = Cm + s_0 Z$, where Z has density proportional to $e^{-U(z)}$. This symmetric, non-Gaussian law has finite variance v_Z, so the averaged quadratic coupling is $b(C^2m^2 + s_0^2 v_Z)$. Holding η fixed and scaling (C, s_0, b) to $(\lambda C, \lambda s_0, b/\lambda^2)$ preserves these coefficients. Replacing Cm by cm in the drift while retaining s_0, κ, η produces quadratic response bc^2. Frozen mean and variance anchors therefore retain their target-sufficiency roles in this family. For a polynomial slow coupling of degree p, the averaged drift depends on conditional moments up to degree p; those moments are natural anchor candidates when the intervention specifies how their underlying law changes. Neither a universal finite-moment sufficiency claim nor the OU single-lag rate formula follows for an unrestricted nonlinear fast generator.

B Architecture-Only Control under Slow-Only Observations

This experiment examines whether explicit multirate recurrent structure provides a predictive advantage when both architectures receive the same slow-only observations. It is intended as an architectural control rather than as a direct test of intervention identifiability: the models are evaluated on observational prediction, not on recovery of the hidden intervention target.

B.1 Experimental Setup

System and protocol. We use the `NonlinearMixing` slow–fast SDE

$$dS_t = \kappa(CM_t - S_t)\, dt + \sigma_S\, dW_t, \tag{30}$$

$$dM_t = \varepsilon(aM_t + bS_t^2)\, dt + \sqrt{\varepsilon}\, \sigma_M\, dV_t, \tag{31}$$

with default parameters

$$(\kappa, C, \sigma_S, a, b, \sigma_M) = (1.0, 0.3, 0.5, -0.5, 1.0, 0.3)$$

and

$$\varepsilon = 0.1.$$

Training trajectories cover

$$T_{\text{train}} = 80.0,$$

corresponding to approximately 8Δ for $\Delta = 10$ at $\varepsilon = 0.1$.

Test trajectories cover

$$T_{\text{test}} = 40.0,$$

with

$$N_{\text{test}} = 8$$

initial conditions and

$$R = 16$$

stochastic realizations for each initial condition.

Observations are subsampled at

$$h_{\text{obs}} = 1.0.$$

Models. We compare two learned recurrent model families.

- **SingleRateGRU.** A standard `nn.GRU` receives the slow observation M_t at each model step. We evaluate one- and two-layer variants with hidden dimension

  $$h = 24.$$

  The one-layer model has 1994 trainable parameters and the two-layer model has 5594 trainable parameters. A linear output head predicts

  $$(\mu_t, \log \sigma_t) = \mathrm{Linear}(z_t).$$

- **LearnedMultiRateGRU.** The multirate model contains a fast GRU updated at every fine step and a slow GRU updated every K fine steps. Slow-to-fast conditioning uses

  $$f_{t+\delta} = \mathrm{GRU}_f\left(f_t, [M_t, s_t]\right),$$

  while fast-to-slow communication aggregates fast states through their empirical mean and standard deviation.

  We use fast and slow hidden dimensions of 16 and evaluate

  $$K \in \{1, 5, 20\}.$$

  The resulting model has 4162 trainable parameters.

The parameter counts are reported explicitly because the architectures are not exactly parameter matched in this depth sweep. The experiment therefore tests the robustness of the observed predictive pattern across architecture depth and clock ratio rather than establishing a strictly capacity-matched comparison.

Training protocol. Both model families are trained with Adam using learning rate

$$10^{-2},$$

gradient clipping at 1.0, Gaussian negative log-likelihood, 30 epochs, and batch size 16.

For every experimental cell, the models use the same training trajectories and random seeds. Results are averaged over three seeds.

Evaluation protocol. At test time, each model is rolled out autoregressively for 40 steps at

$$h = 1.0,$$

feeding the previous prediction as the next model input.

Predictive quality is measured using Gaussian CRPS (Hersbach, 2000), averaged over

$$N_{\mathrm{test}} \times R$$

test trajectories.

B.2 Results

Across all six evaluated configurations, the SingleRateGRU obtains lower CRPS than the LearnedMultiRateGRU. With one recurrent layer, single-rate CRPS ranges from 0.269 to 0.282, compared with 0.296 to 0.309 for the multirate model. With two layers in the single-rate baseline, the corresponding ranges are 0.227–0.230 and 0.267–0.299.

As the clock ratio increases from $K = 1$ to $K = 20$, the relative CRPS gap decreases from 10.0% to 5.3% in the one-layer comparison and from 30.1% to 17.5% in the two-layer comparison. The single-rate model retains the lower score in every tested configuration. The table reports all six configurations and their parameter counts.

n_{layers}	K	Single CRPS	Multi CRPS	ΔCRPS (%)	$(P_{\text{single}}, P_{\text{multi}})$
1	1	0.269	0.296	+10.0	(1994, 4162)
1	5	0.282	0.309	+9.7	(1994, 4162)
1	20	0.282	0.297	+5.3	(1994, 4162)
2	1	0.230	0.299	+30.1	(5594, 4162)
2	5	0.227	0.280	+23.2	(5594, 4162)
2	20	0.227	0.267	+17.5	(5594, 4162)

Table 3: Architecture-depth and clock-ratio sweep on `NonlinearMixing` with $\varepsilon = 0.1$, $h_{\text{obs}} = 1.0$, and three random seeds. ΔCRPS reports the relative increase of the multirate CRPS over the single-rate CRPS, so positive values indicate lower CRPS for the single-rate model. Parameter counts are shown explicitly because this experiment is not a strict parameter-matched comparison.

B.3 Interpretation

This experiment is consistent with, but does not by itself prove, the information-theoretic distinction developed in the main paper. Both recurrent architectures receive the same slow-only observational protocol; changing the recurrent update structure does not introduce additional observations of the hidden fast mechanism.

The result should therefore be interpreted narrowly. In the evaluated regime, explicit multirate recurrence does not produce a predictive advantage over the single-rate baseline. More importantly for the theory developed in this paper, architectural factorization should not be confused with additional cross-scale information. Whether an intervention target is identifiable is determined by the observational fiber and the available anchors, not by the number of clocks used to parameterize the predictor.

A direct empirical test of intervention identifiability would additionally require evaluating recovery of the target q_c across observationally equivalent microscopic systems; the present experiment evaluates observational prediction only.

C Clock Flattening Boundary

Consider a deterministic multirate system whose component update periods are integer multiples of a common base step. A single-clock representation can be constructed by augmenting the state with a phase variable taking values modulo the least common multiple of all update periods, together with any aggregation buffers needed by the original schedule.

At each base tick, the augmented transition executes, in the original prescribed order, every component update scheduled at that phase and then updates the phase and aggregation buffers.

Induction over base ticks gives equality of trajectory laws between the multirate system and the augmented single-clock representation.

The same idea extends to stochastic transitions by composing the scheduled Markov kernels in their prescribed order at each phase. This is ordered kernel composition; it does not require distinct component kernels to satisfy a common semigroup identity.

Therefore the number of explicit clocks is not, by itself, a source of additional observational information. Clock factorization can change inductive bias, execution, or parameterization, but it cannot collapse an observational fiber without additional measurements or restrictions on the model class.

D Deterministic Witness and Reproducibility

D.1 Deterministic Witness

Consider the two microscopic parameter sets

$$(a, b, C, \kappa, \sigma_S) = (-0.5, 1, 0.3, 1, 0.5) \tag{32}$$

and

$$(a, b, C, \kappa, \sigma_S) = (-0.5, 0.25, 0.6, 2, \sqrt{2}). \tag{33}$$

For the first parameterization,

$$r = \frac{\sigma_S^2}{2\kappa} = \frac{0.25}{2} = 0.125,$$

and therefore

$$br = 0.125, \qquad bC^2 = 0.09.$$

For the second parameterization,

$$r = \frac{2}{4} = 0.5,$$

and therefore

$$br = 0.125, \qquad bC^2 = 0.25(0.6)^2 = 0.09.$$

Thus the two microscopic systems have identical reduced combinations

$$(br, a, bC^2) = (0.125, -0.5, 0.09).$$

After the common mean-response replacement

$$C \leftarrow 0.45,$$

the quadratic coefficients before multiplication by ε are

$$1 \times 0.45^2 = 0.2025$$

and

$$0.25 \times 0.45^2 = 0.050625,$$

respectively.

Hence the two systems agree on the reduced observational coefficients but disagree on the response to the same hidden-mechanism intervention. This is a deterministic illustration of Theorem 2 rather than statistical evidence for an architectural claim.

D.2 Experimental Reproducibility

The deterministic checks are implemented in `validate_t3_r1.py`, which verifies the sampled-semigroup and averaging-gauge constructions to numerical precision. These checks illustrate the algebra and are not used as evidence for a statistical or architectural claim.

The architecture-control measurements are archived in `multiratefair/results_gru_depth.json`. The archive contains three-seed aggregate scores rather than seed-level dispersions, so the architecture comparison is descriptive and supplies no significance test.

All architecture-control configurations use the same data-generation, training, and evaluation pipeline. The compared models receive the same slow-only observations and are evaluated with the same trajectory realizations and random seeds.

The finite-scale simulations and sampled-anchor experiments are implemented in `experiments/run_identifiability.py`, with numerical results in `results/identifiability.json`. Their controlled initialization, sampling, noise calibration, and estimator validity conditions are specified with the experimental protocols.

E Sampling Protocols and Estimation Details

E.1 Finite-scale simulation

On slow time $u = \varepsilon t$, the fast generator is multiplied by $1/\varepsilon$ and the slow equation is $dM_u = (aM_u + bS_u^2)du + \sigma_M dV_u$. All microscopic coefficients stay fixed as ε changes. At each step, the fast update is the exact OU transition with the slow state frozen at its left endpoint; the slow drift also uses left-endpoint values. Brownian increments and standardized initial fast deviations are shared across models within each seed. Independent seeds give the reported standard errors. The observational RMS is $\{128^{-1}\sum_j (M_1^{(1,j)} - M_1^{(2,j)})^2\}^{1/2}$ per seed; the averaging RMS compares the first microscopic model with its averaged counterpart under the same slow noise. Such coupling costs bound terminal Wasserstein discrepancies but do not estimate their optimal transport cost. The post-intervention statistic is the mean of $M_1^{(u_c,1,j)} - M_1^{(u_c,2,j)}$ within a seed.

At $\varepsilon = 0.05$, halving the step from 0.00125 to 0.000625 changes the observational RMS from 0.01501 ± 0.00025 to 0.01521 ± 0.00025, the averaging RMS from 0.04114 ± 0.00060 to 0.04164 ± 0.00063, and the intervention mean gap from 0.03149 ± 0.00047 to 0.03135 ± 0.00048. This resolution check supports the displayed trends without supplying a uniform discretization error guarantee.

E.2 Stationary pairs, equal sample budgets, and sensor calibration

Anchor experiments use $(C, r, \kappa) = (0.3, 0.125, 1)$, $(\beta_0, \beta_2) = (0.0125, 0.009)$, and $c = 0.45$. For each frozen state $m = -1, +1$, independent pairs are generated as

$$S_0 = Cm + \sqrt{r}Z_0, \qquad S_1 = Cm + e^{-1}(S_0 - Cm) + \sqrt{r(1 - e^{-2})}Z_1,$$

with independent standard normal Z_0, Z_1. Independent Gaussian sensor noise of known standard deviation ν is added to both endpoints. The two state means define $\widehat{C}$. The variance estimate pools squared within-state residuals with denominator $N-2$ and subtracts ν^2. Pair dependence makes this finite-sample variance correction only approximate; the estimator is consistent across independent pairs. The lag estimate averages the paired residual products and gives $\widehat{\kappa} = -\log(\widehat{K}/\widehat{r})$.

The mean-target inverse is declared invalid if $|\widehat{C}| < 0.05$; the variance inverse requires $\widehat{r} > 0$, and the rate inverse requires $0 < \widehat{K} < \widehat{r}$. Invalid estimates are counted, and means and standard errors condition on the reported valid runs. No true parameter is substituted for a failed estimate. There are no invalid target estimates in the reported sweep. Rate inversions fail in 21/200 runs at $N = 100$, $\nu = 0.3$, illustrating that a population formula need not be numerically reliable at a small noisy sample size.

The means-only, variance-only, and lag calculations consume the same measurements. Collecting independent equilibrium pairs at controlled states is part of the experimental protocol; this is a sample-budget comparison, not a claim of equal physical preparation cost.

E.3 Contiguous samples and noisy reduced dynamics

The contiguous-sampling experiment uses exact stationary OU transitions at each frozen state with spacing $h \in \{0.1, 0.5, 1, 2\}$, 250 observations per state, and noise standard deviation 0 or 0.1. The duration per state is $249h$, so increasing h changes both temporal dependence and observation duration. Mean-anchor error at noise 0.1 decreases from 0.496 to 0.104 across this range.

To include slow-drift estimation, we simulate the averaged slow-time SDE with drift $-0.5m + 0.125 + 0.09m^2$ and diffusion 0.3. Each seed supplies 256 independent bursts at each of nine known initial states equally spaced on $[-1, 1]$. Burst duration is 0.1 and the Euler step is 0.001. Independent sensor noise is added to both endpoints. Least squares regresses the average increment divided by duration on $(1, m, m^2)$ at the controlled initial states. A separate set of 500 independent frozen fast samples with sensor standard deviation 0.1 provides $\widehat{C}$, and the target estimate is $\widehat{Q}_c = \widehat{B}_2 c^2/\widehat{C}^2$. The true $Q_c = 0.2025$ is used only to evaluate error. Finite-duration drift bias, diffusion noise, and sensor noise all enter this calculation. It uses a known initial-state design, not noisy covariates from an uncontrolled trajectory.

E.4 Reproducibility

The executable experiment is `experiments/run_identifiability.py`; `experiments/render_results.py` produces the table source from `results/identifiability.json`. The result file records the source hash, software versions, seed counts, and effective sample counts. The computations use Python 3.8 and NumPy 1.24.4 on CPU, with 30 seeds for finite-scale paths, 200 for the fast-anchor sweeps, and 100 for noisy slow bursts. Random-seed rules and all simulation constants are explicit in the script. These experiments use no trained neural estimator; the recurrent architecture control measures a separate observational prediction task.

Table 4: Independent stationary-pair anchors at lag $\tau = 1$. Target and rate errors are relative errors; uncertainty is the standard error across seeds. Invalid covariance inversions are counted, with no substitution of true parameters. All mean and variance target inversions are valid in this sweep.

N	ν	Mean q_c error	Variance q_c error	κ error	Valid κ
100	0.0	0.205 ± 0.013	0.129 ± 0.008	0.353 ± 0.026	198/200
100	0.1	0.210 ± 0.013	0.148 ± 0.009	0.409 ± 0.030	199/200
100	0.3	0.272 ± 0.021	0.271 ± 0.020	0.555 ± 0.038	179/200
500	0.0	0.101 ± 0.005	0.054 ± 0.003	0.123 ± 0.008	200/200
500	0.1	0.098 ± 0.006	0.059 ± 0.003	0.134 ± 0.009	200/200
500	0.3	0.110 ± 0.007	0.087 ± 0.005	0.239 ± 0.018	200/200
2000	0.0	0.049 ± 0.003	0.030 ± 0.002	0.061 ± 0.004	200/200
2000	0.1	0.051 ± 0.003	0.032 ± 0.002	0.066 ± 0.004	200/200
2000	0.3	0.062 ± 0.003	0.046 ± 0.003	0.116 ± 0.007	200/200

Table 5: Contiguous frozen-fast sampling with $N = 500$ total measurements and the mean anchor. Entries are mean relative target error $\pm$ standard error across 200 seeds. Increasing h increases duration as well as reducing serial dependence; sample count, rather than duration, is fixed.

h	$\nu = 0$	$\nu = 0.1$
0.1	0.489 ± 0.046	0.496 ± 0.048
0.5	0.179 ± 0.011	0.184 ± 0.011
1.0	0.128 ± 0.007	0.132 ± 0.007
2.0	0.098 ± 0.005	0.104 ± 0.006

Table 6: Joint reduced-drift and anchor estimation on slow time. Each run uses 2304 independent reduced-process bursts of duration 0.1 and 500 independent fast-anchor samples. The slow initial states are controlled and known. Entries are mean $\pm$ standard error across 100 seeds.

Slow sensor SD	B_2 estimation bias	Relative Q_c error
0.0	-0.0012 ± 0.0057	0.499 ± 0.037
0.02	-0.0004 ± 0.0059	0.534 ± 0.038
0.05	0.0007 ± 0.0071	0.662 ± 0.041