BEYOND RECONSTRUCTION: DYNAMICAL ADMISSIBILITY IN PHYSICS-BASED TEMPORAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning representations from observations supports physics-based modeling of evolving systems. Accurate reconstruction and small mechanistic residuals on sampled trajectories do not, by themselves, establish whether a learned representation supports closed latent evolution. We study the dynamical admissibility of time-varying representations, connecting temporal causal representation learning with dynamical systems. Using semiconjugacy and exact lumpability, we distinguish latent closure from compatibility with a prescribed mechanistic flow. Invertible reparameterizations preserve closure through conjugate dynamics, whereas retaining the prescribed flow requires them to commute with that flow. In approximate settings, we separate representation-induced nonclosure from transition-model error, and derive finite-horizon bounds relating nonclosure to reconstruction that identify when accurate reconstruction controls nonclosure and when unresolved future-state information imposes a tradeoff between future reconstruction accuracy and latent closure. Experiments on controlled synthetic systems and real-world cerebral perfusion data illustrate these relationships and the complementary roles of reconstruction accuracy, latent-transition consistency, and mechanistic consistency.

1 INTRODUCTION

Learning representations from high-dimensional observations provides a means of modeling evolving physical systems. In physics-based models, these representations can serve as internal states governed by mechanistic equations. In biomedical applications, such states are intended to describe physiological processes underlying clinical measurements and support forecasts of their evolution. Physics-informed neural networks (PINNs) combine observation fitting with constraints derived from governing differential equations, encouraging mechanistic consistency of inferred physical quantities, including unobserved states (Raissi et al., 2019). However, satisfying these constraints on sampled trajectories does not by itself establish that a learned latent representation retains sufficient information to determine its *own future evolution*.

A complementary perspective is offered by causal representation learning (CRL), which seeks to recover causally meaningful latent variables from observations (Schölkopf et al., 2021). A central focus of CRL is identifiability: whether latent variables can be recovered up to a specified equivalence class under assumptions on the data-generating process and available supervision. Temporal CRL further exploits temporal dependencies, transition mechanisms, or interventions to identify latent causal processes from sequential observations (Yao et al., 2021; 2022).

However, many studies rely on synthetic or rendered visual benchmarks whose latent variables and causal dependencies are specified *by construction* (Lippe et al., 2022a;c). Their applicability to evolving physiological states inferred from real biomedical observations remains less explored. Recent work connects CRL to scientific dynamical systems by treating trajectory-specific physical parameters as latent variables (Yao et al., 2024). Crucially, these latent quantities are time-invariant: they characterize an entire trajectory without themselves evolving along it. In many real-world biomedical applications, however, the quantities of interest are physiological states that change within a trajectory (Laubenbacher et al., 2024; Sun et al., 2023). Recovering such states requires learning variables whose own evolution admits a well-defined dynamical description.

Our focus is the *dynamical admissibility* of learned time-varying representations. Given an observed-to-latent representation $P : \mathcal{X} \to \mathcal{Z}$ with $z_t = P(x_t)$, we ask whether z_t is merely an encoding of

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.

x_t, or whether it retains sufficient information to determine *its own future evolution*. In the exact case, observed states mapped to the same latent state must remain indistinguishable in latent space under forward evolution; departures from this condition motivate a quantitative notion of approximate dynamical admissibility. Even when a representation supports closed evolution, compatibility with a prescribed mechanism remains a separate question. Accurate reconstruction and small mechanistic residuals on sampled trajectories do not, by themselves, establish these properties.

These considerations raise two challenges: characterizing latent closure and compatibility with prescribed mechanisms, and distinguishing representation-induced nonclosure from transition-model error when both representations and transitions are learned from data:

(C1) Dynamical Validity of Evolving Latent Representations. Existing CRL-dynamical systems connections primarily consider trajectory-level latent parameters that remain fixed along a trajectory (Yao et al., 2024). When the latent variable itself evolves, it is no longer sufficient to identify a representation at each time point: the current latent state must also determine a well-defined future latent state. This question is further complicated by the reparameterization ambiguity inherent to CRL, which generally identifies latent variables only up to an equivalence class determined by the available supervision and structural assumptions (Khemakhem et al., 2020; Brehmer et al., 2022; Lippe et al., 2022c; von Kügelgen et al., 2023). In physics-based settings, the latent coordinates may further need to obey prescribed mechanistic dynamics. A central question is therefore which reparameterizations permitted by CRL remain compatible with a prescribed mechanistic flow.

(C2) From State-Space Characterizations to Learned Representations. Prior works have established several state-space characterizations of when a lower-dimensional representation admits consistent dynamics, through notions such as semiconjugacy and lumpability (Li & Rabitz, 1990; Horstmeyer & Atay, 2016; Cardelli et al., 2023). However, these analyses typically consider a specified dynamical system and a prescribed reduction, whereas in representation learning both the encoding map and its latent transition are inferred from finite trajectories. Thus, a finite-horizon discrepancy may arise from information lost by the representation or from approximation error in the learned transition. A principled analysis is needed to distinguish representation-induced nonclosure from transition-model error and clarify how the former relates to reconstruction and latent prediction.

Contributions. To address these challenges, we analyze exact and approximate dynamical admissibility of learned temporal representations. Our contributions are threefold:

- **Exact closure and mechanistic compatibility.** With classical semiconjugacy and exact lumpability, we formulate when a *time-varying* representation defines a closed latent dynamical state. As shown in Figure 1, we distinguish preservation of closure from preservation of a prescribed flow: invertible reparameterizations preserve closure through conjugate dynamics, whereas retaining the original flow is equivalent to commuting with that flow. This characterizes the transformations within a CRL identifiability class that remain compatible with the prescribed mechanism.
- **Finite-horizon bounds for approximate admissibility.** We separate prediction error into irreducible nonclosure induced by the representation and error of the learned transition. We derive finite-horizon bounds linking nonclosure to reconstruction, which establish when accurate reconstruction controls nonclosure of latent dynamics, and when unresolved future-state information imposes a tradeoff between accurate future reconstruction and consistent latent evolution.
- **Evaluation on synthetic systems and real cerebral perfusion data.** We examine these relationships using controlled synthetic systems and real-world cerebral perfusion data. Synthetic experiments illustrate how reconstruction and latent-transition accuracy can diverge under limited representation capacity. Real-world cerebral perfusion experiments on latent hemodynamics modeling further evaluate the trajectory estimation, latent-transition consistency, and mechanistic consistency across complementary modeling regimes, demonstrating the complementary information provided by reconstruction fidelity, latent-transition consistency, and mechanistic consistency.

2 Dynamical Admissibility of Time-Varying Representations

In this section, we study when temporal representations induce a well-defined latent flow. In Section 2.1, we formulate this property using the notion of semiconjugacy and classical criterion of exact lumpability. In Section 2.2, we further examine how latent coordinate transformations affect dynamical admissibility and identify those that remain consistent with the original latent flow map.

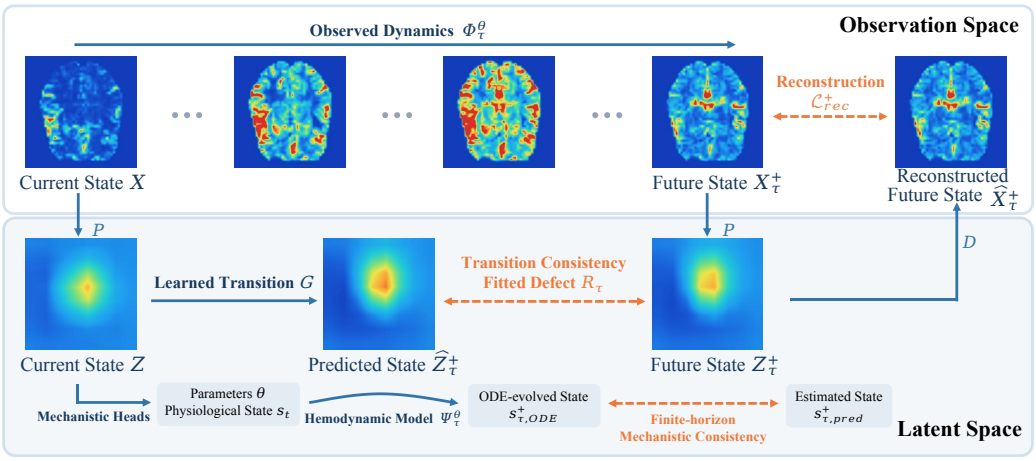

Figure 1: Overview of the PI-TR (Physics-Informed Temporal Representation) framework for time-varying representations in dynamical models, with cerebral perfusion hemodynamics (Section 4.2) as an example. PI-TR separately assesses reconstruction fidelity, agreement between predicted and encoded future latent states, and finite-horizon consistency with the prescribed hemodynamic model.

Dynamical System Setup. Consider a continuous-time dynamical system:

$$\dot{x}(t) = f_\theta(x(t)), \quad x(t) \in \mathcal{X} \quad \text{and} \quad \theta \in \Theta, \tag{2.1}$$

where $\mathcal{X}$ denotes the state space, Θ denotes the parameter space of time-invariant physical parameters, and f_θ induces a flow $\Phi_t^\theta : \mathcal{X} \to \mathcal{X}$. Here, we are interested in a time-varying latent representation

$$z(t) = P(x(t)), \quad P : \mathcal{X} \to \mathcal{Z}. \tag{2.2}$$

Here, $x(t)$ denotes the high-dimensional observed dynamical data measured at time t. Further, $P(\cdot)$ plays the role of a learned representation in the context of temporal CRL, which maps $x(t)$ at each time point to a lower-dimensional latent variable. As a concrete example, $z(t)$ may represent time-varying physiological quantities such as hemodynamic states in blood flow modeling (Lee & Liu, 2026; Kuang et al., 2024), which serve as the latent state representation of a patient-specific dynamical model. We instantiate this setting in cerebral hemodynamics in Section 4.

For $z(t)$ to serve as a dynamical state, its current value must contain sufficient information to determine its future evolution. $P(\cdot)$ produces a latent trajectory, but does not guarantee that the resulting $z(t)$ is a *valid* state of a dynamical system. This motivates the following structural question:

Question 2.1. *When does a time-varying representation define a valid latent dynamical state?*

2.1 Preliminaries: Exact Latent Closure in Temporal CRL

We first consider the exact setting: representation $P(\cdot)$ induces a well-defined latent flow $\Psi_t^\theta : \mathcal{Z} \to \mathcal{Z}$. The consistency between the original and latent evolutions is expressed by the classical notion of semiconjugacy (Brin & Stuck, 2002; Hasselblatt & Katok, 2003; Katok & Hasselblatt, 1995):

Definition 2.2 (Semiconjugacy of the latent dynamics)**.** We say that the representation P semi-conjugates the original dynamics Φ_t^θ to the latent dynamics Ψ_t^θ if the following relation holds:

$$P \circ \Phi_t^\theta = \Psi_t^\theta \circ P \quad \text{for all} \quad t \geq 0. \tag{2.3}$$

The commuting relation (2.3) has a direct implication: any two states mapped to the same latent state must have identical latent images under the original flow at every future time. In dynamical systems, this condition is captured by the notion of exact lumpability, leading to the following proposition:

Proposition 2.3 (Exact Lumpability, (Horstmeyer & Atay, 2016; Tomlin et al., 1997; Li & Rabitz, 1989))**.** Consider (2.1) and (2.2). Suppose that $P : \mathcal{X} \to \mathcal{Z}$ is surjective. Then, there exists a

unique latent evolution $\Psi_t^\theta : \mathcal{Z} \to \mathcal{Z}$ such that P semiconjugates the original dynamic Φ_t^θ to Ψ_t^θ, if and only if, for any $x, x' \in \mathcal{X}$ satisfying $P(x) = P(x')$, the following holds for all $t \geq 0$:

$$P\big(\Phi_\theta^t(x)\big) = P\big(\Phi_\theta^t(x')\big). \tag{2.4}$$

Further, when (2.4) holds, the induced latent evolution is uniquely defined by

$$\Psi_t^\theta\big(P(x)\big) := P\big(\Phi_t^\theta(x)\big).$$

In the temporal CRL setting, Proposition 2.3 implies that a representation admits *exact latent closure* precisely when the distinctions removed by $P(\cdot)$ remain irrelevant to its future latent evolution. This condition is referred to as *exact lumpability*, and we use the two terms interchangeably throughout.

2.2 Temporal CRL under Latent Reparameterization

Building on Section 2.1, we now consider the coordinate ambiguity inherent to CRL, where identifiability is generally established only up to an equivalence class rather than a unique latent coordinate system. Depending on the supervision and structural assumptions, existing results identify causal variables up to permutations and component-wise reparameterizations (Brehmer et al., 2022), or up to invertible affine transformations and more restricted subclasses (Ahuja et al., 2023). When such representations are interpreted as dynamical states, this identifiability ambiguity must also be reconciled with the exact closure criterion (Proposition 2.3). This requires determining which CRL-admissible reparameterizations preserve exact latent closure, and which remain consistent with a prescribed latent dynamics. We therefore consider an invertible latent reparameterization $h : \mathcal{Z} \to \mathcal{Z}$ within such an identifiability class and characterize when these dynamical properties are preserved.

Theorem 2.4 (Dynamical Equivalence under Latent Reparameterization)**.** Let $h : \mathcal{Z} \to \mathcal{Z}$ be an invertible map. Then, if $P(\cdot)$ is surjective and semiconjugates Φ_t^θ to Ψ_t^θ, the transformed representation $\hat{P} := h \circ P$ semiconjugates Φ_t^θ to $\hat{\Psi}_t^\theta := h \circ \Psi_t^\theta \circ h^{-1}$. Furthermore, the relation

$$\hat{P} \circ \Phi_t^\theta = \Psi_t^\theta \circ \hat{P} \quad \text{for all} \quad t \geq 0$$

holds, if and only if the invertible map $h(\cdot)$ commutes with the original latent dynamics Ψ_t^θ, i.e.,

$$h \circ \Psi_t^\theta = \Psi_t^\theta \circ h \quad \text{for all} \quad t \geq 0.$$

Theorem 2.4 separates preservation of latent closure from preservation of a prescribed mechanistic flow. Any invertible change of latent coordinates preserves closure when the flow is transformed accordingly. Additional restrictions arise when the mechanistic flow is prescribed, and held fixed: admissible reparameterizations must then commute with that flow. The remaining coordinate ambiguity is therefore given by the intersection of the CRL equivalence class with the symmetries of the prescribed dynamics. (Please see Appendix C.1 for a detailed proof.)

Remark 2.5 (CRL Equivalence Class $\leftrightarrow$ Dynamical Symmetry)**.** Let $\mathcal{H}_{\text{CRL}}$ denote the class of invertible latent reparameterizations $h : \mathcal{Z} \to \mathcal{Z}$ left unresolved by some CRL identifiability results, such as permutations (Brehmer et al., 2022) or affine maps (Ahuja et al., 2023). We further introduce the dynamical symmetry class as the set of invertible maps that commute with the latent flow:

$$\mathcal{H}_{\text{dyn}}(\Psi^\theta) := \left\{ h \mid h \circ \Psi_t^\theta = \Psi_t^\theta \circ h,\ \forall t \geq 0 \right\}.$$

Then, requiring the transformed representation to follow the same prescribed flow restricts the CRL equivalence class to $\mathcal{H}_{\text{CRL}} \cap \mathcal{H}_{\text{dyn}}(\Psi^\theta)$, excluding reparameterizations that alter the latent evolution.

3 Approximate Dynamical Admissibility from Finite Trajectories

Building on the exact latent-closure criterion in Section 2, we analyze approximate representations and latent transitions learned from sequential observations. We consider a representation P, a decoder $D : \mathcal{Z} \to \mathcal{X}$, and a transition map $G : \mathcal{Z} \to \mathcal{Z}$, which may be trained jointly to preserve observation-level information and model temporal evolution (Lippe et al., 2022c)[1].

[1]Throughout this section, P, D, G are measurable, all required second moments and losses are finite, and conditional means of future latent states given current ones lie in $\mathcal{Z}$ almost surely for every horizon considered.

Finite-Horizon Lumpability and Reconstruction. Joint training does not by itself establish exact latent closure in the sense of Proposition 2.3. At a given prediction horizon, latent prediction discrepancies can arise because *(i)* P discards information needed to determine the future latent state, or *(ii)* G fails to attain the best prediction supported by P. We first separate representation-induced nonclosure from transition-model error, then examine how reconstruction constrains the former.

We use uppercase X, Z to denote random and latent states sampled along trajectories. We quantify prediction discrepancies by comparing the represented future with a candidate latent transition:

Definition 3.1 (Prediction Horizon and Semiconjugacy Defect)**.** Given a reference time $t \geq 0$, a prediction horizon $\tau > 0$ specifies how far forward in time the flow advances the state $x(t)$:

$$X_\tau^+ := X(t+\tau) = \Phi_\tau^\theta(X(t))$$

for some fixed ODE parameters $\theta \in \Theta$. Then, for a given representation $P : \mathcal{X} \to \mathcal{Z}$, a map $G : \mathcal{Z} \to \mathcal{Z}$ over τ, and a probability measure μ on $\mathcal{X}$, we define the semiconjugacy defect as:

$$\mathcal{R}_\tau(P, G; \mu) := \int_{\mathcal{X}} \|P(X_\tau^+) - G \circ P(X)\|^2 \, \mathrm{d}\mu(X). \tag{3.1}$$

Further, the irreducible finite-horizon lumpability error takes the form:

$$\mathcal{E}_\tau(P; \mu) := \inf_G \mathcal{R}_\tau(P, G; \mu),$$

where the infimum ranges over all measurable latent estimators G with θ held fixed.

To separate these error sources, let $X \sim \mu$, $Z = P(X)$, $Z_\tau^+ = P(X_\tau^+)$. The conditional mean $\mathbb{E}[Z_\tau^+ \mid Z]$ minimizes squared-error prediction of future latent state from the current representation. Consequently, the semiconjugacy defect admits the decomposition (see Appendix C.2 for derivation):

$$\mathcal{R}_\tau(P, G; \mu) = \underbrace{\mathcal{E}_\tau(P; \mu)}_{\text{representation-induced nonclosure}} + \underbrace{\mathbb{E}\left[\left\|\mathbb{E}[Z_\tau^+ \mid Z] - G(Z)\right\|^2\right]}_{\text{transition-model error}}. \tag{3.2}$$

The first term measures variation in the future latent state that remains unresolved by the current representation and *cannot* be reduced by improving G while holding P fixed. The second term measures the additional error of the fitted transition relative to the optimal predictor supported by P. Therefore, a nonzero fitted defect does not by itself imply representation-induced nonclosure.

With separated representation-induced nonclosure and transition-model error, we now relate the former to representation-retained information by measuring current and future state reconstruction[2]:

Definition 3.2 (Reconstruction Loss)**.** For a given decoder $D : \mathcal{Z} \to \mathcal{X}$, we define the reconstruction losses at t and $t + \tau$:

$$\mathcal{L}_{\mathrm{rec}}(P, D; \mu) := \mathop{\mathbb{E}}_{X \sim \mu} \|X - D(Z)\|^2, \quad \mathcal{L}_{\mathrm{rec}}^+(P, D; \mu, \tau) := \mathop{\mathbb{E}}_{X \sim \mu} \|X_\tau^+ - D(Z_\tau^+)\|^2. \tag{3.3}$$

Definition 3.1 and Definition 3.2 lead to the following theorem which relates reconstruction error to the irreducible latent nonclosure $\mathcal{E}_\tau(P; \mu)$. (Please see Appendices C.2 and C.3 for a detailed proof.)

Theorem 3.3 (Reconstruction–Lumpability Bounds)**.** Assume that $P \circ \Phi_\tau^\theta$ is L_τ-Lipschitz on the support of μ. Then, for any representation P and any decoder D, it holds:

$$\mathcal{E}_\tau(P; \mu) \leq L_\tau^2 \mathcal{L}_{\mathrm{rec}}(P, D; \mu). \tag{3.4}$$

Furthermore, suppose that the decoder D is L_D-Lipschitz. Then, the following holds:

$$\sqrt{\mathcal{L}_{\mathrm{rec}}^+(P, D; \mu, \tau)} + L_D \sqrt{\mathcal{E}_\tau(P; \mu)} \geq \Delta_\tau(P; \mu), \tag{3.5}$$

for any representation P and any decoder D, where

$$\Delta_\tau(P; \mu) := \Big(\mathop{\mathbb{E}}_{X \sim \mu} \|X_\tau^+ - \mathbb{E}[X_\tau^+ \mid Z]\|^2 \Big)^{1/2}. \tag{3.6}$$

[2]Future reconstruction uses the encoded future state through $D(Z_\tau^+)$; forecasting instead uses $D(G(Z))$.

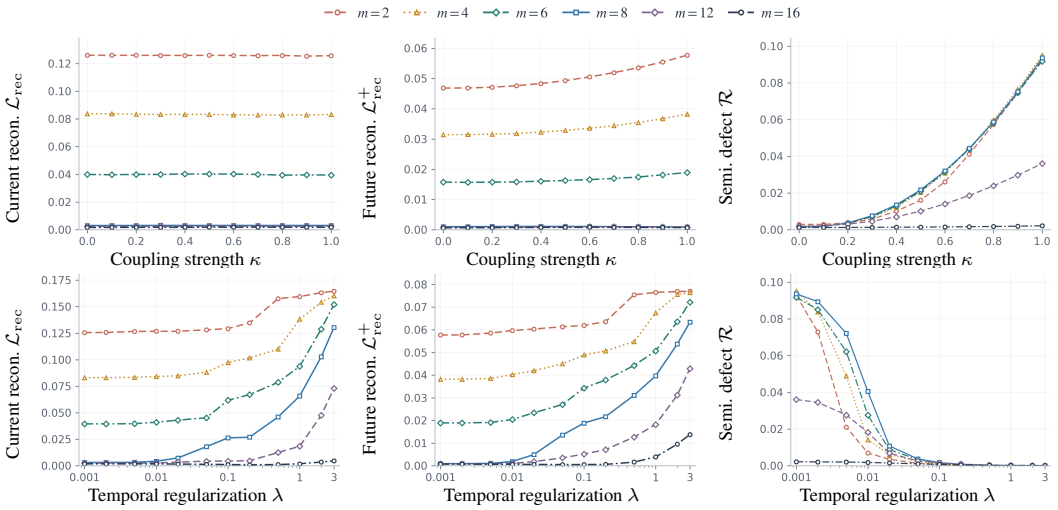

Figure 2: Synthetic coupling κ (top) and temporal regularization λ (bottom) sweeps. Columns show $\mathcal{L}_{\text{rec}}$, $\mathcal{L}^{+}_{\text{rec}}$, and $\mathcal{R}_{\tau}$. Colors denote latent dimension m, and curves average five seeds at $\tau = 0.5$.

Remark 3.4 (Reconstruction vs Lumpability)**.** (3.4) shows that accurate current-state reconstruction controls irreducible latent nonclosure under the stated regularity assumption. Accurate reconstruction is not necessary for closure, however, since discarded information may be irrelevant to latent evolution. (3.5) establishes the tradeoff between reconstruction and latent closure: if the unresolved future-state variation $\Delta_{\tau}(P;\mu)$ remains bounded away from zero and decoder sensitivity remains bounded, future reconstruction error and latent nonclosure cannot both approach zero.

Remark 3.5 (Regression Interpretation of $\Delta_{\tau}(P;\mu)$)**.** Under the squared-error loss, the conditional mean $\mathbb{E}[X^{+}_{\tau} \mid Z]$ is the optimal estimator of the future state X^{+}_{τ} from the current representation Z. Therefore, the lower bound given in (3.5) quantifies the future-state variation that cannot be estimated from the current representation, independently of any particular decoder or latent transition.

4 Experiments

In this section, we examine the dynamical properties of learned temporal representations in two complementary settings. In Section 4.1, we use a synthetic system to isolate how dynamical coupling, representation capacity, and temporal regularization affect reconstruction and the fitted latent-transition defect. In Section 4.2, we consider a cerebral perfusion digital twin to examine these properties in a physiologically grounded setting with high-dimensional observations and mechanistic constraints.

4.1 Controlled Synthetic Experiments

Controlled Coupling under Compression. We design a synthetic system to examine how reconstruction and latent dynamical consistency respond to the dynamical relevance of compressed state components. We include two comparisons: (i) varying physical coupling at a fixed latent dimension changes the dependence of future states on these components, and (ii) varying latent dimension at a fixed coupling changes the capacity available to represent them. We consider $a, b \in \mathbb{R}^8$ governed by:

$$\dot{a}_i = 0.2 \cdot (a_{i-1} - 2a_i + a_{i+1}) - a_i \quad \text{and} \quad \dot{b} = \kappa \cdot \left(\frac{3a_i}{1 + a_i}\right) - b_i,$$

where $\kappa \geq 0$ controls the strength of the nonlinear coupling from a_i to b_i, and $i \in \{1, \ldots, 8\}$, with periodic boundary conditions $a_0(t) := a_8(t)$ and $a_9(t) := a_1(t)$. The initial coordinates are sampled independently from $\text{Uniform}[0, 2]$, and the observed state is $x(t) = (0.1a(t), b(t)) \in \mathbb{R}^{16}$. Notably, the observation scaling makes errors in a relatively inexpensive under squared reconstruction loss, while the b-dynamics are closed on their own only at $\kappa = 0$. In particular, increasing κ strengthens their dependence on a without increasing the reconstruction penalty assigned to errors in a.

Regime Comparison

Regime	Image Encoding	Learned Latent Transition	Mechanistic Dynamics
PINN	✗	✗	✓
Mech-ID	✓	✗	✓
CITRIS	✓	✓	✗
PI-TR	✓	✓	✓

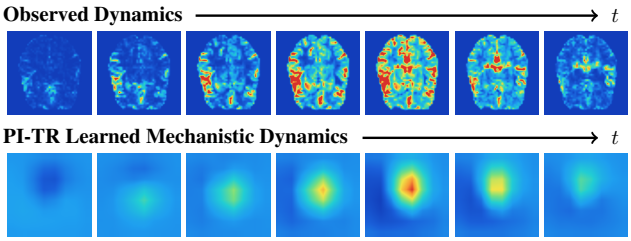

Figure 3: Architectural and mechanistic comparison of the four regimes. We illustrate PI-TR's learned mechanistic dynamics with spatially mapped scores of their first principal components.

Setup. We jointly train a Gaussian encoder, a decoder, and a Gaussian latent transition on observation pairs $(x(0), x(\tau))$ with $\tau = 0.5$. Specifically, the encoder maps each observation to an m-dimensional Gaussian posterior, the decoder reconstructs the full 16-dimensional state, and the transition models latent evolution over τ. The encoder and decoder each use two hidden layers of width 128, while the transition uses one hidden layer of width 128 with a residual mean update. All hidden activations are $\mathrm{Tanh}(\cdot)$. Further, our training objective follows the reconstruction-transition formulation given in (4.2), with $\lambda \geq 0$ weighting the temporal KL term. We investigate the roles of three experimental parameters: *(i)* the coupling strength κ controls the dynamical coupling from a to b, *(ii)* the latent dimension m sets the degree of dimensional compression, and *(iii)* the temporal regularization λ weights the KL divergence between the future posterior and the transition prior relative to reconstruction. For each $m \in \{2, 4, 6, 8, 12, 16\}$, we conduct two sweeps: eleven coupling strengths $\kappa \in \{0, 0.1, \ldots, 1\}$ at fixed $\lambda = 10^{-3}$, and twelve weights $\lambda \in [10^{-3}, 3]$ at fixed $\kappa = 1$. Initial states and data splits are shared across settings, and each configuration is trained with five seeds. The complete objective, network constraints, and training details are provided in Appendix D.

Evaluation and Results We report the current and future reconstruction loss $\mathcal{L}_{\mathrm{rec}}$, $\mathcal{L}^{+}_{\mathrm{rec}}$, and fitted semiconjugacy defect $\mathcal{R}$ under different scenarios in Figure 2. For $L_D > 0$, Theorem 3.3 yields

$$\mathcal{R}_\tau \geq \left(\Delta_\tau - \sqrt{\mathcal{L}^{+}_{\mathrm{rec}}}\right)^2_{+} / L_D^2, \quad \text{where} \quad (\cdot)_+ := \max\{\cdot, 0\} \tag{4.1}$$

Hence, the defect has a positive lower bound whenever Δ_τ exceeds $\sqrt{\mathcal{L}^{+}_{\mathrm{rec}}}$. We examine this constraint by separately varying κ and λ at each fixed latent dimension m and all other hyperparameters.

- Dynamical coupling κ. Because increasing κ strengthens the influence of a on future b without increasing its reconstruction weight, information that is inexpensive to discard can become increasingly important for future evolution. The top row of Figure 2 exhibits the corresponding separation, with reconstruction remaining largely stable while the fitted defect increases, particularly under compression. The inequality (4.1) provides a conditional explanation: if insufficient retention of a leads to larger Δ_τ, while reconstruction remains accurate, the defect cannot be arbitrarily small.

- Temporal regularization λ. With κ and m fixed, the λ sweep keeps the dynamics and representation capacity unchanged, varying only the relative weight of temporal consistency during joint training. Increasing λ places greater emphasis on agreement between the transition prior and the future encoder posterior. Although this penalty acts on latent distributions, the bottom row of Figure 2 shows a corresponding trade-off in the evaluated errors, with stronger regularization reducing the fitted defect at a reconstruction cost under compression. This behavior is consistent with (4.1), which requires sufficiently small Δ_τ to accommodate both accurate future reconstruction and a small defect.

4.2 Real-World Cerebral Perfusion with Latent-State Hemodynamics

In cerebral perfusion imaging, the measured high-dimensional imaging time sequences (observed dynamics in Figure 3) arise from patients' evolving cerebral blood flow (Liu et al., 2021). Beyond estimating time-invariant, trajectory-specific ODE parameters, we examine learned representations of the changing physiological state and their ability to support subsequent temporal prediction. We compare trajectory-level mechanistic identification with learned temporal representations, with and without constraints derived from a prescribed hemodynamic model (Lee & Liu, 2026). These comparisons assess reconstruction fidelity, latent-transition accuracy, and mechanistic consistency. Due to space limitations, we defer the clinical background and implementation details to Appendix D.

Table 1: Trajectory estimation and temporal consistency. SH and LH denote shorter and longer evaluation horizons, respectively. Values are mean ± sample standard deviation over 20 seeds.

Trajectory Estimation Error				
Metric	PINN (ODE)	Mech-ID	CITRIS	**PI-TR**
RMSE (↓)	0.260 ± 0.0357	0.026 ± 0.0019	0.032 ± 0.0027	$\mathbf{0.025 \pm 0.0003}$

Latent Transition Defect (3.1) $\times 10^3$ (↓)						
Method	Early-SH	Early-LH	Peak-SH	Peak-LH	Late-SH	Late-LH
CITRIS	0.330 ± 0.087	0.294 ± 0.090	0.247 ± 0.132	0.256 ± 0.153	0.234 ± 0.138	0.238 ± 0.156
PI-TR	0.526 ± 0.137	0.482 ± 0.141	$\mathbf{0.235 \pm 0.141}$	$\mathbf{0.232 \pm 0.156}$	$\mathbf{0.220 \pm 0.137}$	$\mathbf{0.208 \pm 0.149}$

Reconstruction Error (3.3) $\times 10^1$ (↓)						
Method	Early-SH	Early-LH	Peak-SH	Peak-LH	Late-SH	Late-LH
CITRIS	0.238 ± 0.0037	0.220 ± 0.0037	0.085 ± 0.0004	0.064 ± 0.0002	0.072 ± 0.0003	0.057 ± 0.0002
PI-TR	$\mathbf{0.237 \pm 0.0035}$	$\mathbf{0.220 \pm 0.0037}$	0.086 ± 0.0005	0.065 ± 0.0003	0.072 ± 0.0004	0.057 ± 0.0003

Compared Models (Figure 3). We consider four regimes that differ in the quantities inferred from an observed trajectory and in whether physics-based temporal evolution are explicitly incorporated.

- PINN (Physics-Informed Neural Network (Raissi et al., 2019)). PINN takes the observed data sequences and directly predicts two trajectories: an observation-space trajectory reconstructing the measured data, and a lower-dimensional latent-state trajectory representing the evolving physiology. The two trajectories are constrained by their respective prescribed physical models, governing the observed data and the latent hemodynamics. PINN provides a physics-based baseline that models both trajectories, but does *not* learn an observation-to-latent representation or a latent-space transition.

- Mech-ID (Mechanistic Identification (Yao et al., 2024)). Mech-ID treats trajectory-specific ODE parameters as latent variables. An observed trajectory is mapped to a single set of latent mechanistic parameters, and an ODE solver generates the corresponding trajectory under prescribed governing equations. Therefore, Mech-ID learns the latent dynamics underlying an observed sequence, but assumes that the inferred latent quantities are *time-invariant* throughout the observed trajectory.

- CITRIS (Causal Identifiability from Temporal Intervened Sequences (Lippe et al., 2022c)). CITRIS[3] learns a time-varying representation and explicitly models its temporal evolution. The formulation contains an observation-to-latent encoder $q_\phi(z_t \mid x_t)$, a reconstruction decoder $p_\psi(x_t \mid z_t)$, and a transition prior $p_\gamma(z_{t+1} \mid z_t)$ that models the latent-state evolution, with ϕ, ψ, and γ denoting the respective network parameters. CITRIS jointly learns all parameters by optimizing

$$\mathcal{L}_{\text{CITRIS}} := \Big[-\mathbb{E}\left[\log p_\psi(x_{t+1}|z_{t+1})\right] + \lambda\, \mathbb{E}\left[D_{\text{KL}}\left(q_\phi(\cdot|x_{t+1}) \,\|\, p_\gamma(\cdot|z_t)\right)\right] \Big], \tag{4.2}$$

where $\lambda \geq 0$ controls the strength of temporal regularization, and the expectation in the first and the second term is taken with respect to $z_{t+1} \sim q_\phi(\cdot \mid x_{t+1})$ and $z_t \sim q_\phi(\cdot \mid x_t)$, respectively. With the transition prior $p_\gamma(z_{t+1} \mid z_t)$ serving as the variational prior for the next latent state, CITRIS learns latent temporal dynamics, yet does so *without* incorporating the governing hemodynamic equations.

- PI-TR (Physics-Informed Temporal Representation, Figure 1). PI-TR retains CITRIS' encoder, decoder, and latent transition while *explicitly connecting* the learned representations to cerebral hemodynamics. Mechanistic heads infer physiological states and patient-specific ODE parameters from latent encodings. Physical residuals are evaluated on quantities inferred from both encoded and transition-predicted representations, and trajectories generated by the model are fitted to observations. PI-TR therefore couples learned latent temporal evolution with the known hemodynamics mechanism.

Reconstruction Fidelity and Latent Temporal Consistency. We examine how accurately the four regimes reproduce the observed trajectories, and whether reconstruction fidelity is accompanied by consistent latent evolution. Table 1 reports the root mean squared error (RMSE) of the observed-state trajectories, computed over valid time points and spatial regions within each evaluation window and then averaged across validation windows. PI-TR achieves the lowest trajectory RMSE, closely followed by Mech-ID, while PINN and CITRIS exhibit larger errors. For CITRIS and PI-TR, their shared encoder-decoder and learned-transition structure additionally permits evaluation of the latent

[3]Our implementation follows the learning principles of CITRIS, with adaptations detailed in Section D.2.2.

transition defect (3.1) and future-endpoint reconstruction error (3.3). We report selected scenarios from an exploratory start-horizon sweep. Peak starts at the time index where the spatially averaged observed signal reaches its maximum. Early starts halfway between the sequence beginning and this index (rounded down), and Late starts one time step after it. Shorter Horizon (SH) and Longer Horizon (LH) correspond to horizons of 9 and 11 time steps, respectively. PI-TR reconstructs slightly better but exhibits larger transition defects at the Early start, while the ordering reverses at the Peak and Late starts. These contrasting rankings illustrate that reconstruction fidelity and consistency under a learned transition capture different aspects of a temporal representation.

Verification of Mechanistic Consistency. Beyond observation fitting and latent temporal consistency, we further verify *finite-horizon mechanistic consistency*, which measures how far a model-estimated state trajectory departs from the one generated by the governing ODE under identical initial conditions and inputs[4]. More precisely, it reflects the propagation of local ODE residuals over the evaluation horizon, together with numerical integration error.

We evaluate from a starting index 80% of the way from the sequence beginning to the observed peak, rounded down, following an exploratory start-time sweep. In Table 2, $\Delta \mathbf{s} = (\Delta s_1, \Delta s_2)$ denotes the difference between the ODE-integrated state and the model-estimated state at the evaluation horizon; s_1 is the observation-side state and s_2 the hidden twin state. Relative error (Rel. Error) normalizes the state discrepancy by the magnitude of the network-estimated future state. Results are averaged across cases and summarized as mean $\pm$ sample standard deviation over 20 seeds. PI-TR exhibits lower joint and relative discrepancies across all reported horizons, primarily through smaller errors in s_1. Results for s_2 are mixed, indicating that the reduction in aggregate discrepancy is not perfectly uniform across the coupled state components.

Table 2: Finite-horizon mechanistic consistency verification.

Horizon	Method	$\|\Delta \mathbf{s}\|_2$ $\times 10^3$ (↓)	$\|\Delta s_1\|$ $\times 10^3$ (↓)	$\|\Delta s_2\|$ $\times 10^3$ (↓)	Rel. Error $\times 10^3$ (↓)
$\tau = 1$	PINN	134.560 ± 13.483	134.487 ± 13.479	1.664 ± 0.415	1280.864 ± 207.390
	PI-TR	**32.057 ± 4.920**	**31.871 ± 5.032**	**1.598 ± 0.768**	**182.802 ± 32.557**
$\tau = 2$	PINN	204.158 ± 36.277	204.136 ± 36.276	0.905 ± 0.330	1076.838 ± 189.013
	PI-TR	**48.417 ± 6.928**	**48.294 ± 6.992**	1.756 ± 0.989	**274.320 ± 43.045**
$\tau = 3$	PINN	333.019 ± 59.776	333.000 ± 59.775	1.606 ± 0.561	2330.372 ± 434.803
	PI-TR	**70.164 ± 11.976**	**70.072 ± 12.001**	1.793 ± 1.065	**393.795 ± 73.378**
$\tau = 4$	PINN	444.636 ± 64.640	444.610 ± 64.636	1.206 ± 0.476	1919.314 ± 287.872
	PI-TR	**85.528 ± 17.615**	**85.456 ± 17.635**	1.803 ± 1.096	**474.683 ± 110.654**
$\tau = 5$	PINN	457.504 ± 72.330	457.493 ± 72.326	1.434 ± 0.625	2712.386 ± 441.367
	PI-TR	**90.898 ± 23.533**	**90.825 ± 23.561**	1.811 ± 1.111	**500.494 ± 149.587**
$\tau = 6$	PINN	502.921 ± 68.174	502.901 ± 68.173	1.080 ± 0.493	2458.499 ± 344.974
	PI-TR	**87.305 ± 27.813**	**87.217 ± 27.856**	1.821 ± 1.120	**478.816 ± 176.591**
$\tau = 7$	PINN	454.324 ± 64.655	454.316 ± 64.653	1.186 ± 0.554	2930.913 ± 445.814
	PI-TR	**79.490 ± 27.925**	**79.386 ± 27.970**	1.834 ± 1.126	**434.170 ± 179.508**

Remark 4.1 (Limits of latent consistency verification)**.** Though PINN achieves lower mechanistic discrepancies in s_2 at most horizons, its direct trajectory parameterization provides neither an encoder-decoder representation nor a learned latent transition, making the diagnostics in (3.3) and (3.1) undefined for this regime. The s_2 advantage assesses one component of mechanistic consistency without establishing the additional representation-level properties examined for CITRIS and PI-TR.

5 Conclusion

We address a crucial gap in learning time-varying representations for physics-based models by extending prior formulations based on time-invariant, trajectory-specific physical parameters to evolving and dynamically admissible latent states. Building on classical closure criteria, we distinguish preservation of latent closure from preservation of a prescribed mechanistic flow, clarifying the restrictions on CRL reparameterizations. Under approximate settings, we separate representation-induced nonclosure from transition-model error, and derive finite-horizon bounds that establish when accurate reconstruction controls nonclosure and when unresolved future-state information imposes a tradeoff between future reconstruction accuracy and latent closure. Synthetic and real-world cerebral perfusion experiments illustrate the complementary roles of reconstruction accuracy, learned-transition consistency, and mechanistic consistency, paving the way for broader frameworks that integrate data-driven representations with mechanistic models while explicitly assessing dynamical admissibility.

[4]Only PINN and PI-TR estimate both state trajectories and ODE parameters, hence enable such verification (see regime comparison in Figure 3). CITRIS lacks the required mechanistic state and parameter readouts, while Mech-ID produces trajectories directly through numerical integration, with no learned trajectory to verify.

REFERENCES

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

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Michael Brin and Garrett Stuck. *Introduction to dynamical systems*, volume 1. Cambridge university press The Edinburgh Building, Cambridge CB2 2RU, UK, 2002.

Steven L Brunton, Bingni W Brunton, Joshua L Proctor, and J Nathan Kutz. Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PloS one*, 11(2):e0150171, 2016.

Luca Cardelli, Giuseppe Squillace, Mirco Tribastone, Max Tschaikowski, and Andrea Vandin. Formal lumping of polynomial differential equations through approximate equivalences. *Journal of Logical and Algebraic Methods in Programming*, 134:100876, 2023.

Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45): 22445–22451, 2019.

Jiale Chen, Dingling Yao, Adeel Pervez, Dan Alistarh, and Francesco Locatello. Scalable mechanistic neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 10018–10039, 2025.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Rodero, Filip Loncaric, Yingjing Feng, Andrew Gilbert, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, et al. The 'digital twin' to enable the vision of precision cardiology. *European heart journal*, 41(48):4556–4564, 2020.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pp. 586–596, 1981.

Andreas Fieselmann, Markus Kowarschik, Arundhuti Ganguly, Joachim Hornegger, and Rebecca Fahrig. Deconvolution-based CT and MR brain perfusion measurement: theoretical model revisited and practical implementation details. *International Journal of Biomedical Imaging*, 2011(1): 467563, 2011.

Juan L Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. *arXiv preprint arXiv:2502.20099*, 2025.

Boris Hasselblatt and Anatole Katok. A first course in dynamics. *Cambridge UP, Cambridge*, 343: 344, 2003.

Leonhard Horstmeyer and Fatihcan M Atay. Characterization of exact lumpability for vector fields on smooth manifolds. *Differential Geometry and its Applications*, 48:46–60, 2016.

Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*. Cambridge university press, 1995.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Keying Kuang, Frances Dean, Jack B Jedlicki, David Ouyang, Anthony Philippakis, David Sontag, and Ahmed Alaa. Med-real2sim: Non-invasive medical digital twins using physics-informed self-supervised learning. *Advances in Neural Information Processing Systems*, 37:5757–5788, 2024.

Pranamya Kulkarni, Puranjay Datta, Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Ropes: Robotic pose estimation via score-based causal representation learning. *arXiv preprint arXiv:2510.20884*, 2025.

Reinhard Laubenbacher, Borna Mehrad, Ilya Shmulevich, and Natalia Trayanova. Digital twins in medicine. *Nature computational science*, 4(3):184–191, 2024.

Yi-Chen Lee and Peirong Liu. HemoPIC: A physics-informed cerebral hemodynamics digital twin for brain perfusion. *arXiv preprint arXiv:2607.08799*, 2026.

Genyuan Li and Herschel Rabitz. A general analysis of exact lumping in chemical kinetics. *Chemical engineering science*, 44(6):1413–1430, 1989.

Genyuan Li and Herschel Rabitz. A general analysis of approximate lumping in chemical kinetics. *Chemical engineering science*, 45(4):977–1002, 1990.

Qianxiao Li, Felix Dietrich, Erik M Bollt, and Ioannis G Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10), 2017.

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. *arXiv preprint arXiv:2206.06169*, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Intervention design for causal representation learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022b.

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, 2022c.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. In *Uncertainty in Artificial Intelligence*, pp. 1263–1273. PMLR, 2023.

Peirong Liu, Yueh Z Lee, Stephen R Aylward, and Marc Niethammer. Perfusion imaging: an advection diffusion approach. *IEEE transactions on medical imaging*, 40(12):3424–3435, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

Samuel E Otto and Clarence W Rowley. Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593, 2019.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Tianze Sun, Xiwang He, and Zhonghai Li. Digital twin in healthcare: Recent updates and challenges. *Digital health*, 9:20552076221149651, 2023.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. *Advances in neural information processing systems*, 30, 2017.

Alison S Tomlin, Genyuan Li, Herschel Rabitz, and János Tóth. The effect of lumping and expanding on kinetic differential equations. *SIAM Journal on Applied Mathematics*, 57(6):1531–1556, 1997.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

Stefan Winzeck, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, et al. ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Frontiers in neurology*, 9:679, 2018.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems*, 37: 71705–71736, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, volume 2025, pp. 53847–53890, 2025.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

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

Appendix

A Related Work, Limitations and Future Directions **13**

B Notation Table **15**

C Proof of Main Text **15**
C.1 Proof of Theorem 2.4 16
C.2 Proof of Theorem 3.3: Part (I) 17
C.3 Proof of Theorem 3.3: Part (II) 19

D Experimental Details **20**
D.1 Controlled Synthetic Dynamical System 20
D.1.1 Data Generation and Preprocessing 20
D.1.2 Training and Evaluation Protocol 20
D.2 Cerebral Perfusion Hemodynamics Digital Twin 21
D.2.1 Physiological background and governing dynamics 21
D.2.2 Data and Model Implementation 23
D.2.3 Additional Evaluations 24

AI USE STATEMENT

We used large language models (LLMs) to assist with writing and editing, including improving clarity, grammar, and conciseness. We also used LLMs for literature retrieval and discovery, such as identifying potentially missing related work. All technical content and results are original contributions of the authors, and the authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is methodological and involves no prospective human-subject recruitment or clinical intervention. The cerebral perfusion experiments use publicly available imaging data for research purposes. The proposed models are not intended for clinical diagnosis or treatment decision-making

REPRODUCIBILITY STATEMENT

Complete proofs and assumptions are provided in Appendix C. Experimental details including architectures, hyperparameters, data processing, and evaluation procedures are detailed in Appendix D. The source code for reproducing the reported experiments is submitted together with the manuscript.

A RELATED WORK, LIMITATIONS AND FUTURE DIRECTIONS

Causal Representation Learning in Scientific and Dynamical Systems. Several recent works have explored CRL in scientific and dynamical settings. Yao et al. (2024) demonstrate this connection on global atmospheric dynamics and real-world climate data, using causal latent variables to recover physically meaningful factors from temporal observations and linking representation learning with mechanistic dynamical models. Their formulation provides a concrete bridge between CRL and scientific system identification. Subsequent works have broadened this direction toward scalable mechanistic modeling of long temporal sequences (Chen et al., 2025) and empirical evaluation of CRL on controlled physical systems (Gamella et al., 2025). Related studies further connect CRL with invariance principles in ecological applications (Yao et al., 2025) and interventional representation learning for robotic state estimation (Kulkarni et al., 2025). In contrast, our work focuses instead on time-varying latent states, where the learned representation itself participates in the system evolution.

Temporal Causal Representation Learning. Temporal CRL has developed several ways of exploiting sequential structure beyond static observations. Subsequent work on temporal intervened sequences extends identifiability to instantaneous causal effects and joint graph recovery (Lippe et al., 2022a), studies how intervention experiments should be designed to identify latent causal variables (Lippe et al., 2022b), and relaxes known intervention targets by learning causal variables from unknown binary interactions in embodied environments (Lippe et al., 2023). These formulations use temporal dependencies, interventions, or actions to constrain causal representations, with the temporal structure represented primarily through conditional transitions rather than an explicitly specified governing flow. This distinction matters for our setting: when latent variables are additionally used as states of a mechanistic dynamical model, their temporal evolution must satisfy structural requirements beyond causal identifiability alone, such as the consistency with the governing state dynamics.

Dynamical Systems and Model Reduction. Beyond CRL, a separate line of work studies when lower-dimensional coordinates inherit consistent dynamics from a higher-dimensional system. Semiconjugacy and lumpability provide classical characterizations of this property, with exact conditions developed for differential equations and smooth dynamical systems (Li & Rabitz, 1989; Tomlin et al., 1997; Horstmeyer & Atay, 2016). Approximate variants have also been studied for nonlinear kinetic systems and polynomial differential equations (Li & Rabitz, 1990; Cardelli et al., 2023). These formulations typically assume a given dynamical system together with a prescribed reduction. Our setting instead concerns representations and latent transitions learned from finite trajectories, requiring the effects of representation loss, transition approximation, and reconstruction to be considered jointly.

Koopman Representations and Learned Latent Dynamics. The Koopman-based representation learning is another closely related line of work which seeks simplified descriptions of nonlinear dynamics in transformed coordinates. Specifically, classical Koopman theory lifts a nonlinear flow Φ_t to the evolution of observable functions through the linear operator $\mathcal{K}_t g = g \circ \Phi_t$ (Koopman, 1931; Mezić, 2005), enabling spectral and modal analysis of nonlinear systems with linear-operator tools. Since this operator generally acts on an infinite-dimensional function space, data-driven methods such as dynamic mode decomposition (DMD) and extended dynamic mode decomposition (EDMD) approximate its action within finite collections of observables (Williams et al., 2015; Brunton et al., 2016). Their effectiveness, however, depends strongly on the choice of observables, motivating later approaches that learn these functions directly from data using neural networks and autoencoders (Li et al., 2017; Takeishi et al., 2017; Lusch et al., 2018; Otto & Rowley, 2019). Modern deep Koopman methods accordingly learn a finite-dimensional representation P together with a linear latent evolution K_t, encouraging the approximate commuting relation $P \circ \Phi_t \approx K_t \circ P$. In this sense, they seek an approximate semiconjugacy to a particular linear latent dynamical system. Other related approaches relax linearity by learning other forms of latent evolution, including sparse nonlinear governing equations through sparse identification of nonlinear dynamics (SINDy) with autoencoders (Champion et al., 2019) and neural ODEs in continuous-time latent-variable models (Chen et al., 2018; Rubanova et al., 2019). Differently, our setting is instead motivated by physics-based temporal models, where the latent dynamics may be nonlinear and may already be prescribed by mechanistic equations, making linearization neither an assumption nor the primary objective. Accordingly, we treat dynamical closure as a property of the representation itself, independently of any particular parameterization of the latent transition, and characterize when a learned P admits a well-defined latent evolution at all. In approximate settings, this perspective isolates irreducible representation-induced nonclosure from the error of a fitted transition, rather than merging both effects into the residual of a chosen latent dynamics model. Moreover, we relate the intrinsic nonclosure to finite-horizon reconstruction and characterize which latent reparameterizations preserve closure and the prescribed mechanistic flow, providing a unified view of dynamical admissibility in the learned temporal representations.

Limitations and Future Directions. Our analysis focuses on deterministic latent dynamics and finite prediction horizons, leaving stochastic systems and longer-horizon behavior for future study. The finite-horizon bounds are developed under squared-error objectives and regularity conditions, and extending them to broader losses and stochastic representations would further generalize the analysis. In addition, our current results characterize the dynamical requirements imposed on temporal representations, while a tighter integration with CRL identifiability theory under interventions or causal structural assumptions remains an important direction. Finally, the real-world evaluation focuses on cerebral perfusion, motivating validation across a broader range of scientific and biomedical systems.

B NOTATION TABLE

Table 3: Mathematical Notations and Symbols

Symbol	Description
$\mathcal{X}$	State space of the observed dynamical system (2.1)
Θ	Parameter space of time-invariant physical parameters (2.1)
$x(t)$	High-dimensional observed dynamical state at time t (2.1)
θ	Time-invariant physical parameters governing the system dynamics (2.1)
f_θ	Vector field governing the observed dynamical system (2.1)
Φ_t^θ	Flow map induced by the original dynamics (2.1)
$\mathcal{Z}$	State space of the learned latent representation (2.2)
$z(t)$	Learned latent state at time t (2.2)
$P : \mathcal{X} \to \mathcal{Z}$	Representation map from the observed state to the latent state (2.2)
Ψ_t^θ	Latent flow induced by P under exact latent closure (Definition 2.2)
$h : \mathcal{Z} \to \mathcal{Z}$	Invertible reparameterization of the latent coordinates (Theorem 2.4)
$X(t)$	Random reference state sampled at time t (Definition 3.1)
μ	Probability measure over reference states in $\mathcal{X}$ (Definition 3.1)
τ	Prediction horizon from time t to $t+\tau$ (Definition 3.1)
X_τ^+	Future state at time $t+\tau$ after evolution over horizon τ (Definition 3.1)
$G : \mathcal{Z} \to \mathcal{Z}$	Candidate latent transition over horizon τ (Definition 3.1)
$\mathcal{R}_\tau(P, G; \mu)$	Finite-horizon semiconjugacy defect for P and G (3.1)
$\mathcal{E}_\tau(P; \mu)$	Irreducible nonclosure after optimizing over G (Definition 3.1)
$Z(t)$	Random latent state corresponding to $X(t)$ (Definition 3.2)
Z_τ^+	Future latent state at time $t+\tau$ (Definition 3.2)
$D : \mathcal{Z} \to \mathcal{X}$	Decoder mapping latent states to the observed state space (Definition 3.2)
$\mathcal{L}_{\text{rec}}$	Reconstruction error of the reference state at time t (3.3)
$\mathcal{L}_{\text{rec}}^+$	Reconstruction error of the future state at time $t+\tau$ (3.3)
L_τ	Lipschitz constant of $P \circ \Phi_\tau^\theta$ (Theorem 3.3)
L_D	Lipschitz constant of the decoder D (Theorem 3.3)
$\Delta_\tau(P; \mu)$	Residual future-state variation given the current representation (3.6)

C PROOF OF MAIN TEXT

In this section, we provide proofs for the results stated in the main text. Appendix C.1 characterizes how latent reparameterizations affect the induced evolution and when the original latent flow is retained (Theorem 2.4). Appendix C.2 and Appendix C.3 derive the two finite-horizon bounds relating reconstruction to latent closure (Theorem 3.3), respectively. We first provide a helper lemma:

Lemma C.1 (Conditional Independent-Copy Variance Identity, Modified from (Boucheron et al., 2003; Efron & Stein, 1981))**.** Let $U, U' \in \mathbb{R}^m$ and $W \in \mathbb{R}^k$ be random vectors, with U and U' both having finite second moments. Suppose that U and U' are conditionally i.i.d. given W; i.e.,

$$U, U' \mid W \overset{\text{i.i.d.}}{\sim} \nu_W(\cdot)$$

for some conditional distribution $\nu_W(\cdot)$. Then, the following identity holds:

$$\mathbb{E}\left[\|U - U'\|^2\right] = 2 \cdot \mathbb{E}\left[U - \mathbb{E}[\|U \mid W]\|^2\right].$$

Proof. Since U and U' have the same conditional distribution given W, we have

$$\mathbb{E}[U' \mid W] = \mathbb{E}[U \mid W].$$

Define

$$R := U - \mathbb{E}[U \mid W], \qquad R' := \mathbb{E}[U \mid W] - U'.$$

Then $U - U' = R + R'$. Note that conditioned on W, R and R' are independent, and

$$\mathbb{E}[R \mid W] = 0, \qquad \mathbb{E}[R' \mid W] = 0.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\|U-U'\|^2 \mid W\right] \\
&= \mathbb{E}\left[\|R+R'\|^2 \mid W\right] \\
&= \mathbb{E}\left[\|R\|^2 \mid W\right] + \mathbb{E}\left[\|R'\|^2 \mid W\right] + 2\,\mathbb{E}\left[\langle R, R'\rangle \mid W\right] \\
&= \mathbb{E}\left[\|R\|^2 \mid W\right] + \mathbb{E}\left[\|R'\|^2 \mid W\right] + 2\sum_{i=1}^{m} \mathbb{E}[R_i R_i' \mid W] \\
&= \mathbb{E}\left[\|R\|^2 \mid W\right] + \mathbb{E}\left[\|R'\|^2 \mid W\right] + 2\sum_{i=1}^{m} \mathbb{E}[R_i \mid W]\,\mathbb{E}[R_i' \mid W] \\
&= \mathbb{E}\left[\|R\|^2 \mid W\right] + \mathbb{E}\left[\|R'\|^2 \mid W\right].
\end{aligned}$$

Since U and U' are identically distributed conditioned on W,

$$\mathbb{E}\left[\|R'\|^2 \mid W\right] = \mathbb{E}\left[\|U' - \mathbb{E}[U \mid W]\|^2 \mid W\right] = \mathbb{E}\left[\|U - \mathbb{E}[U \mid W]\|^2 \mid W\right] = \mathbb{E}\left[\|R\|^2 \mid W\right].$$

Hence,

$$\mathbb{E}\left[\|U-U'\|^2 \mid W\right] = 2\,\mathbb{E}\left[\left\|U - \mathbb{E}[U \mid W]\right\|^2 \mid W\right],$$

and therefore

$$\frac{1}{2}\,\mathbb{E}\left[\|U-U'\|^2 \mid W\right] = \mathbb{E}\left[\left\|U - \mathbb{E}[U \mid W]\right\|^2 \mid W\right].$$

Finally, taking expectations on both sides and applying the tower property gives

$$\frac{1}{2}\,\mathbb{E}\left[\|U-U'\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|U - \mathbb{E}[U \mid W]\|^2 \mid W\right]\right] = \mathbb{E}\left[\|U - \mathbb{E}[U \mid W]\|^2\right].$$

This completes the proof. □

C.1 Proof of Theorem 2.4

We now present the proof of Theorem 2.4.

Theorem C.2 (Theorem 2.4 Restated, Dynamical Equivalence under Latent Reparameterization)**.** Let $h : \mathcal{Z} \to \mathcal{Z}$ be an invertible map. Then, if $P(\cdot)$ is surjective and semiconjugates Φ_t^θ to Ψ_t^θ, the transformed representation $\hat{P} := h \circ P$ semiconjugates Φ_t^θ to $\hat{\Psi}_t^\theta := h \circ \Psi_t^\theta \circ h^{-1}$. Furthermore, the relation

$$\hat{P} \circ \Phi_t^\theta = \Psi_t^\theta \circ \hat{P} \quad \text{for all} \quad t \geq 0$$

holds, if and only if the invertible map $h(\cdot)$ commutes with the original latent dynamics Ψ_t^θ, i.e.,

$$h \circ \Psi_t^\theta = \Psi_t^\theta \circ h \quad \text{for all} \quad t \geq 0.$$

Proof. By Definition 2.2 and substituting $\hat{P} := h \circ P$, the first part of Theorem 2.4 follows directly:

$$h \circ P \circ \Phi_t^\theta = h \circ \Psi_t^\theta \circ P = (h \circ \Psi_t^\theta \circ h^{-1}) \circ h \circ P.$$

Then, we prove both directions of the second part.

- ($\Rightarrow$). Assume

$$\hat{P} \circ \Phi_t^\theta = \Psi_t^\theta \circ \hat{P}.$$

  Since $\hat{P} := h \circ P$,

$$h \circ P \circ \Phi_t^\theta = \Psi_t^\theta \circ h \circ P.$$

By the semiconjugacy of P, we have

$$h \circ \Psi_t^\theta \circ P = \Psi_t^\theta \circ h \circ P.$$

Hence, h must commute with the prescribed latent flow: $h \circ \Psi_t^\theta = \Psi_t^\theta \circ h$.

- ($\Leftarrow$). Assume

$$h \circ \Psi_t^\theta = \Psi_t^\theta \circ h.$$

Using the semiconjugacy of P again, we have:

$$h \circ \Psi_t^\theta \circ P = \Psi_t^\theta \circ h \circ P \Rightarrow h \circ P \circ \Phi_t^\theta = \Psi_t^\theta \circ h \circ P. \tag{C.1}$$

By $\hat{P} := h \circ P$, the implied equality in (C.1) becomes

$$\hat{P} \circ \Phi_t^\theta = \Psi_t^\theta \circ \hat{P}.$$

Therefore, $\hat{P}$ semiconjugates the original dynamics to the same prescribed latent flow.

This completes the proof. □

C.2 Proof of Theorem 3.3: Part (I)

We now present the proof for the first part of Theorem 3.3.

Theorem C.3 (Theorem 3.3 Part (I) Restated, Reconstruction–Lumpability Bounds)**.** Assume that $P \circ \Phi_\tau^\theta$ is L_τ-Lipschitz on the support of μ. Then, for any $P(\cdot)$ and any $D(\cdot)$, it holds:

$$\mathcal{E}_\tau(P; \mu) \leq L_\tau^2 \mathcal{L}_{\text{rec}}(P, D; \mu).$$

Proof. For simplicity, we shorthand $X(t)$ with X. We prove Theorem C.3 with three steps.

- **Step 1: Construct an auxiliary state.** Recall that $X_\tau^+ = \Phi_\tau^\theta(X(t))$, $Z = P(X)$ and $Z_\tau^+ = P(X_\tau^+)$. Let $\mu_Z(\cdot)$ denote the conditional distribution of X on Z, and $\nu_Z(\cdot)$ denote the conditional distribution of Z_τ^+ on Z. That is, for every $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Z}$, we have:

$$\mu_Z(A) := \mathbb{P}(X \in A \mid Z) \quad \text{and} \quad \nu_Z(B) := \mathbb{P}(Z_\tau^+ \in B \mid Z).$$

Conditioned on the state of Z, let X' be an independent copy of X. Then, we have:

$$P(X') = P(X) = Z \quad \text{a.s.} \quad \text{and} \quad \mathbb{P}(X' \in A \mid Z) = \mathbb{P}(X \in A \mid Z) \tag{C.2}$$

for every $A \subseteq \mathcal{X}$. Further, by construction, the auxiliary state X' has the same marginal distribution as X, since $\mathbb{E}[\mathbb{1}\{X \in A\} \mid Z] = \mathbb{P}(X \in A \mid Z)$ and

$$\mathbb{P}(X' \in A) = \mathbb{E}[\mathbb{P}(X' \in A \mid Z)] = \mathbb{E}[\mathbb{P}(X \in A \mid Z)] = \mathbb{P}(X \in A),$$

where we apply the tower property for the first and last equality, and use (C.2) for the second equality. Therefore, by the construction of X',

$$X, X' \mid Z \overset{\text{i.i.d.}}{\sim} \mu_Z(\cdot). \tag{C.3}$$

Correspondingly, X' also induces its future state and representation:

$$X_\tau^{'+} := \Phi_\tau^\theta(X') \quad \text{and} \quad Z_\tau^{'+} := P(X_\tau^{'+}).$$

Applying deterministic map $P(\cdot) \circ \Phi_\tau^\theta(\cdot)$, (C.3) implies

$$Z_\tau^+, Z_\tau^{'+} \mid Z \overset{\text{i.i.d.}}{\sim} \nu_Z(\cdot). \tag{C.4}$$

- **Step 2: Identify the best estimator.** Recall Definition 3.1 and Definition 3.2. The irreducible finite-horizon lumpability takes the form:

$$\mathcal{E}_\tau(P;\mu) = \inf_G \mathcal{R}_\tau(P, G; \mu) = \inf_G \mathbb{E}\left[\|Z_\tau^+ - G(Z)\|^2\right].$$

Essentially, $\mathcal{E}_\tau$ is the minimum mean squared error for estimating the future representation Z^+ using only the current representation Z. This is a standard regression problem under square error loss, for which the optimal predictor is the conditional mean $G^* = \mathbb{E}[Z_\tau^+ \mid Z]$.

For completeness, here we show this directly. For any predictor G, we have that

$$\begin{aligned}
&\mathbb{E}\left[\|Z_\tau^+ - G(Z)\|^2\right] \\
&= \mathbb{E}\left[\| \underbrace{Z_\tau^+ - \mathbb{E}[Z_\tau^+ \mid Z]}_{:=(\mathrm{T}_1)} \|^2\right] + \mathbb{E}\left[\| \underbrace{\mathbb{E}[Z_\tau^+ \mid Z] - G(Z)}_{:=(\mathrm{T}_2)} \|^2\right] + 2\,\mathbb{E}\left[\langle (\mathrm{T}_1), (\mathrm{T}_2)\rangle\right] \\
&= \mathbb{E}\left[\|Z_\tau^+ - \mathbb{E}[Z_\tau^+ \mid Z]\|^2\right] + \mathbb{E}\left[\|\,\mathbb{E}\left[Z_\tau^+ \mid Z\right] - G(Z)\|^2\right] + 2\,\mathbb{E}\left[\mathbb{E}[\langle (\mathrm{T}_1), (\mathrm{T}_2)\rangle \mid Z]\right]. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By the tower property)}
\end{aligned}$$

Since (T_2) is determined by Z, we have

$$\mathbb{E}\left[\mathbb{E}[\langle (\mathrm{T}_1), (\mathrm{T}_2)\rangle \mid Z]\right] = \mathbb{E}\left[\langle \mathbb{E}[(\mathrm{T}_1) \mid Z], (\mathrm{T}_2)\rangle\right] = 0,$$

where we invoke:

$$\mathbb{E}[(\mathrm{T}_1) \mid Z] = \mathbb{E}[Z_\tau^+ - \mathbb{E}[Z_\tau^+ \mid Z] \mid Z] = 0$$

in the last equality. Therefore, the cross term vanishes and we have:

$$\mathbb{E}\left[\|Z_\tau^+ - G(Z)\|^2\right] = \mathbb{E}\left[\|Z_\tau^+ - \mathbb{E}[Z_\tau^+ \mid Z]\|^2\right] + \mathbb{E}\left[\|\,\mathbb{E}\left[Z_\tau^+ \mid Z\right] - G(Z)\|^2\right],$$

which implies

$$\mathcal{E}_\tau(P;\mu) = \inf_G \mathbb{E}\left[\|Z_\tau^+ - G(Z)\|^2\right] = \mathbb{E}\left[\|Z_\tau^+ - \mathbb{E}[Z_\tau^+ \mid Z]\|^2\right]. \tag{C.5}$$

- **Step 3: Express the optimal error through the auxiliary pair.** Finally, we show the upper bounds by connecting Step 1 and Step 2. Specifically, on the one hand, we have:

$$\mathbb{E}\left[\, Z_\tau^+ - Z_\tau^{+'}\|^2\right] = 2\,\mathbb{E}\left[\|Z_\tau^+ - \mathbb{E}[Z_\tau^+|Z]\|^2\right] = 2\mathcal{E}_\tau(P;\mu),$$

where we use (C.4) and Lemma C.1 for the first equality; (C.5) for the second equality.

Since $P \circ \Phi_\tau^\theta$ is L_τ-Lipschitz on the support of μ, and $X, X' \mid Z \overset{\text{i.i.d.}}{\sim} \mu_Z(\cdot)$, we have

$$2\mathcal{E}_\tau(P;\mu) = \mathbb{E}\left[\|Z_\tau^+ - Z_\tau^{+'}\|^2\right] \le L_\tau^2\,\mathbb{E}\left[\|X - X'\|^2\right] = 2L_\tau^2\,\mathbb{E}\left[\|X - \mathbb{E}[X \mid Z]\|^2\right],$$

where the inequality follows from the Lipschitz continuity of $P \circ \Phi_\tau^\theta$, and the last equality follows again from (C.3) and Lemma C.1. Therefore,

$$\mathcal{E}_\tau(P;\mu) \le L_\tau^2\,\mathbb{E}\left[\|X - \mathbb{E}[X \mid Z]\|^2\right]. \tag{C.6}$$

It remains to relate the right-hand side to the reconstruction loss. By Definition 3.2,

$$\begin{aligned}
&\mathcal{L}_{\mathrm{rec}}(P, D; \mu) \\
&= \mathbb{E}\left[\|X - D(Z)\|^2\right] \\
&= \mathbb{E}\left[\| \underbrace{X - \mathbb{E}[X \mid Z]}_{:=U} + \underbrace{\mathbb{E}[X \mid Z] - D(Z)}_{:=V} \|^2\right] \\
&= \mathbb{E}\left[\|U\|^2\right] + \mathbb{E}\left[\|V\|^2\right] + 2\,\mathbb{E}\left[\langle U, V\rangle\right].
\end{aligned}$$

Since V is determined by Z, the cross term satisfies

$$\mathbb{E}\left[\langle U, V\rangle\right] = \mathbb{E}\left[\,\mathbb{E}\left[\langle U, V\rangle \mid Z\right]\right] = \mathbb{E}\left[\left\langle\, \mathbb{E}[U \mid Z], V\right\rangle\right] = 0,$$

where

$$\mathbb{E}[U \mid Z] = \mathbb{E}\left[X - \mathbb{E}[X \mid Z] \mid Z\right] = 0.$$

Hence,

$$\mathcal{L}_{\text{rec}} = \mathbb{E}\left[\|X - \mathbb{E}[X \mid Z]\|^2\right] + \mathbb{E}\left[\|\,\mathbb{E}[X \mid Z] - D(Z)\|^2\right] \geq \mathbb{E}\left[\|X - \mathbb{E}[X \mid Z]\|^2\right]. \tag{C.7}$$

Applying (C.6) to (C.7), we have

$$\mathcal{E}_\tau(P;\mu) \leq L_\tau^2 \mathcal{L}_{\text{rec}},$$

This completes the proof for the first part of Theorem 3.3. □

C.3 Proof of Theorem 3.3: Part (II)

Here, we prove the second part of Theorem 3.3.

Theorem C.4 (Theorem 3.3 Part (II) Restated, Reconstruction–Lumpability Bounds)**.** Suppose that the decoder $D(\cdot)$ is L_D-Lipschitz. Then, the following holds:

$$\sqrt{\mathcal{L}^+_{\text{rec}}(P, D; \mu, \tau)} + L_D\sqrt{\mathcal{E}_\tau(P;\mu)} \geq \Delta_\tau(P;\mu),$$

for any representation $P(\cdot)$ and any decoder $D(\cdot)$, where

$$\Delta_\tau(P;\mu) := \big(\mathop{\mathbb{E}}_{X\sim\mu} \|X^+_\tau - \mathbb{E}[X^+_\tau \mid Z]\|^2\big)^{1/2}.$$

Proof. We prove the second part with the following two steps.

- **Step 1: Relate Δ_τ to the decoded optimal latent estimator.** Recall (C.5). The optimal predictor of Z^+_τ from Z, denoted by G^*, under the squared loss is the conditional mean:

$$G^*(Z) = \mathbb{E}[Z^+_\tau \mid Z] \quad \text{and} \quad \mathcal{E}_\tau(P;\mu) = \mathbb{E}\left[\|Z^+_\tau - G^*(Z)\|^2\right].$$

  Since $D(G^*(Z))$ is a function of Z, the conditional mean $\mathbb{E}[X^+_\tau|Z]$ minimizes the mean-squared error for predicting X^+_τ from Z. In particular,

$$\begin{aligned}
&\mathbb{E}\left[\|X^+_\tau - D(G^*(Z))\|^2\right] \\
&= \mathbb{E}\left[\|X^+_\tau - \mathbb{E}[X^+_\tau \mid Z]\|^2\right] + \mathbb{E}\left[\|\,\mathbb{E}[X^+_\tau \mid Z] - D(G^*(Z))\|^2\right] \\
&\geq \mathbb{E}\left[\|X^+_\tau - \mathbb{E}[X^+_\tau \mid Z]\|^2\right] = \Delta^2_\tau(P;\mu),
\end{aligned}$$

  where the cross term vanishes since

$$\mathbb{E}\left[X^+_\tau - \mathbb{E}[X^+_\tau \mid Z] \mid Z\right] = 0$$

  and $\mathbb{E}[X^+_\tau \mid Z] - D(G^*(Z))$ is determined by Z. Therefore,

$$\Delta_\tau(P;\mu) \leq \Big(\mathbb{E}\left[\|X^+_\tau - D(G^*(Z))\|^2\right]\Big)^{1/2}. \tag{C.8}$$

- **Step 2: Decompose the decoded prediction error.** We introduce a term $D(Z^+_\tau)$ and get

$$\begin{aligned}
&\Big(\mathbb{E}\left[\|X^+_\tau - D(G^*(Z))\|^2\right]\Big)^{1/2} \qquad \text{(C.9)} \\
&\leq \underbrace{\Big(\mathbb{E}\left[\|X^+_\tau - D(Z^+_\tau)\|^2\right]\Big)^{1/2}}_{\sqrt{\mathcal{L}^+_{\text{rec}}}} + \Big(\mathbb{E}\left[\|D(Z^+_\tau) - D(G^*(Z))\|^2\right]\Big)^{1/2}
\end{aligned}$$

  by Definition 3.2 and the triangle inequality. Then, since $D(\cdot)$ is L_D-Lipschitz, we have

$$\Big(\mathbb{E}\left[\|D(Z^+_\tau) - D(G^*(Z))\|^2\right]\Big)^{1/2} \leq L_D\Big(\mathbb{E}\left[\|Z^+_\tau - G^*(Z)\|^2\right]\Big)^{1/2} = L_D\sqrt{\mathcal{E}_\tau(P;\mu)}. \tag{C.10}$$

  Finally, applying (C.9) and (C.10) to (C.8), we obtain

$$\Delta_\tau(P;\mu) \leq \sqrt{\mathcal{L}^+_{\text{rec}}} + L_D\sqrt{\mathcal{E}_\tau(P;\mu)}.$$

This completes the proof for the second part of Theorem 3.3. □

D EXPERIMENTAL DETAILS

In this section, we provide additional details for the two experimental settings shown in Section 4. Appendix D.1 describes the controlled synthetic system, including data generation, model parameterization, training, and evaluation. Appendix D.2 details the cerebral perfusion digital twin, including the governing physiology, data processing, model implementations, and additional evaluations.

D.1 CONTROLLED SYNTHETIC DYNAMICAL SYSTEM

In this section, we present the synthetic data generation and network parameterization in Section D.1.1, followed by the training objective, optimization procedure, and evaluation protocol in Section D.1.2.

D.1.1 DATA GENERATION AND PREPROCESSING

We generate data from the coupled system in Section 4.1 by sampling the coordinates of $a(0), b(0) \in \mathbb{R}^8$ independently from $\mathrm{Uniform}[0, 2]$. The observations are $x(t) = (0.1a(t), b(t)) \in \mathbb{R}^{16}$, with no additional measurement noise. We compute $a(t)$ analytically and obtain each coordinate of $b(t)$ from

$$b_i(t) = e^{-t} b_i(0) + \kappa \int_0^t e^{-(t-u)} \frac{3a_i(u)}{1 + a_i(u)} \, \mathrm{d}u, \qquad i \in \{1, \dots, 8\}. \tag{D.1}$$

We evaluate (D.1) using 24-point Gauss-Legendre quadrature training, and validation contain 16,384 and 4,096 independently sampled initial states, respectively, with observation pairs $(x(0), x(0.5))$. The test set contains another 4,096 independent pairs of initial states sharing $b(0)$ but with independently sampled $a(0)$, yielding 8,192 states evaluated at horizons 0.5, 1, and 2. All configurations share the same initial states and data splits. Varying κ changes only the resulting future observations.

Before encoding, each observation is standardized using the means and standard deviations of the initial observation distribution. Specifically, the encoder receives $\widetilde{x}$ with coordinates

$$\widetilde{x}_i = \begin{cases} 10\sqrt{3}(x_i - 0.1), & i \in \{1, \dots, 8\}, \\ \sqrt{3}(x_i - 1), & i \in \{9, \dots, 16\}. \end{cases}$$

The same transformation is applied at every time point, while reconstruction is evaluated in the original observation scale. This puts the initial input coordinates on a common scale for encoding while preserving the smaller reconstruction penalty assigned to errors in a. We next specify the training objective and network parameterization, followed by optimization and evaluation procedures.

We discourage representation collapse through mean and covariance regularization, and control network sensitivity through variance clipping and spectral-norm constraints; implementation details are provided in the code. Training uses Adam for 8,000 updates with batch size 1,024, a cosine learning rate from 10^{-3} to 10^{-4}, and gradient-norm clipping at 5. We retain the checkpoint minimizing the full validation objective, evaluated every 250 updates. Both sweeps use five seeds per configuration, with all other training settings fixed. Evaluation uses encoder and transition means without posterior sampling. Reconstruction losses in (3.3) are averaged over test states and the 16 observation coordinates. The defect in (3.1) uses one, two, or four applications of G for $\tau = 0.5, 1, 2$, respectively. We average latent discrepancies over coordinates and normalize by the average coordinate variance of encoded training initial states, computed once per model and fixed across horizons. The curves in Figure 2 report means over five seeds at $\tau = 0.5$; Δ_τ is not directly estimated.

D.1.2 TRAINING AND EVALUATION PROTOCOL

We follow the temporal variational formulation in (4.2). For a fixed observation pair $x = x(0)$ and $x^+ = x(\tau)$ with $\tau = 0.5$, the reconstruction-transition objective is

$$\mathcal{L}_{\mathrm{CITRIS}} = - \mathop{\mathbb{E}}_{z^+ \sim q_\phi(\cdot|x^+)} \left[\log p_\psi(x^+|z^+) \right] + \lambda \mathop{\mathbb{E}}_{z \sim q_\phi(\cdot|x)} \left[D_{\mathrm{KL}}\big(q_\phi(\cdot|x^+) \,\|\, p_\gamma(\cdot|z)\big) \right].$$

Here, q_ϕ, p_ψ, and p_γ denote the encoder, decoder, and the transition distributions, respectively. The first expectation samples the future latent variable from its encoder posterior, and the second samples the current latent variable used to condition the transition prior. This objective is averaged over the training observation pairs, with λ weighting the temporal consistency relative to the reconstruction.

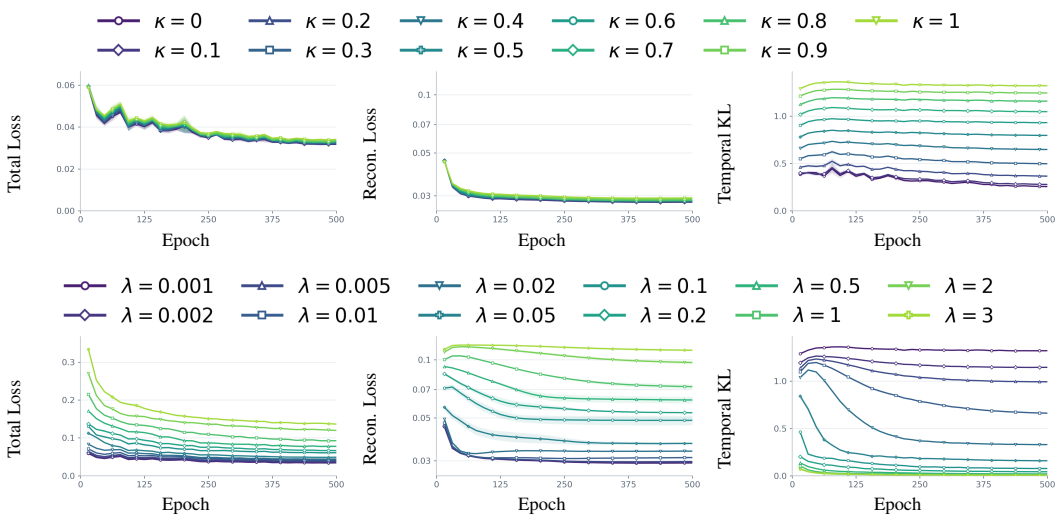

Figure 4: Validation losses for κ (top) and λ (bottom) sweeps at $m = 6$. Lines and bands show five-seed means and standard errors. Recon. uses a log scale. One epoch corresponds to 16 updates.

The encoder and transition parameterize diagonal Gaussian distributions,

$$q_\phi(z|x) = \mathcal{N}\big(P(x), \mathrm{diag}(\sigma_\phi^2(x))\big), \qquad p_\gamma(z^+|z) = \mathcal{N}\big(G(z), \mathrm{diag}(\sigma_\gamma^2(z))\big),$$

where $\sigma_\phi^2(x)$ and $\sigma_\gamma^2(z)$ denote coordinatewise variances. The encoder and decoder use two hidden layers of width 128, while the transition uses one hidden layer of width 128 with a residual mean update. Separate linear heads parameterize the Gaussian means and log variances, and all hidden layers use Tanh activations. The decoder outputs the reconstructed 16-dimensional observation under a fixed-variance Gaussian likelihood. In our analysis, P, D, and G correspond to the encoder mean, decoder output, and transition mean, respectively. For this synthetic setup, we average reconstruction over both endpoints and all 16 observation coordinates, and average the temporal KL divergence over the m latent coordinates. Each update uses one independently sampled latent variable per endpoint through reparameterization. We additionally include a standard-normal prior KL penalty, averaged over both endpoints and latent coordinates, with weight 10^{-4}. All three networks are trained jointly.

Training dynamics. Figure 4 complements the final reconstruction-defect evaluations given in Figure 2 by tracking the validation losses during joint training. At fixed $\lambda = 10^{-3}$, reconstruction converges similarly across coupling strengths, while stronger coupling yields larger temporal KL. At fixed $\kappa = 1$, increasing λ shifts the balance toward temporal consistency at a reconstruction cost, with this separation persisting late in training. These curves illustrate how the competing requirements are accommodated during optimization. In particular, the unweighted temporal KL measures distributional consistency between the future encoder posterior and the transition prior.

D.2 Cerebral Perfusion Hemodynamics Digital Twin

In this section, we provide supplementary details for the cerebral perfusion experiments in Section 4.2. Section D.2.1 describes the coupling between observed tracer dynamics and hidden hemodynamic states. Further, Section D.2.2 specifies the quantities inferred by each regime and their training procedures. Finally, Section D.2.3 reports additional error metrics and examines their sensitivity.

D.2.1 Physiological background and governing dynamics

In the broader biomedical context, an emerging direction is the development of medical digital twins, patient-specific computational models that connect clinical observations with the physiological processes underlying them (Laubenbacher et al., 2024; Corral-Acero et al., 2020). The long-term goal is to support virtual experiments on an individual patient's physiology, allowing disease progression and potential responses to alternative interventions to be explored before clinical intervention. Recent

Table 4: Quantities in the governing ODEs for each segmented brain region.

Symbol	Interpretations	Dynamics
$s_2(t)$	Hidden effective pressure relative to downstream pressure	Latent twin
θ_2	Vascular resistance controlling pressure-driven outflow	Shared
$s_1(t)$	Regional concentration proxy derived from imaging	Data
$u(t)$	Tracer inflow rate per unit effective volume	Data
θ_3	Effective tracer distribution volume	Data
θ_4	Constant blood inflow	Shared
θ_1	Vascular compliance	Latent twin

efforts in machine learning include physics-informed self-supervised approaches that estimate patient-specific cardiac model parameters from noninvasive echocardiography videos (Kuang et al., 2024).

Cerebral perfusion and the hemodynamic twin. We focus on cerebral perfusion, which characterizes how blood flows into and out of brain tissue over time. Essentially, perfusion MRI observes this process indirectly by recording signal changes as an injected contrast agent travels through the cerebral circulation. These measurements yield time-varying concentration proxies that describe contrast-agent transport (Fieselmann et al., 2011; Liu et al., 2021). From these observed transport dynamics, we seek to infer the underlying hemodynamic system. Following (Lee & Liu, 2026), we represent this system using a regional Windkessel model, which describes how blood inflow changes pressure within a tissue region and how that pressure drives blood outflow. The resulting twin consists of the inferred pressure trajectories and patient-specific parameters governing these dynamics. Below, we outline the governing equations and their connection to the observed transport process.

Observed tracer dynamics. The tracer is an injected contrast agent whose concentration provides an observable signal of transport through tissue. We divide the imaging domain into spatial regions, as detailed in Section D.2.2, and model the concentration state $s_{1,r}(t)$ in each region r through

$$\frac{\mathrm{d}s_{1,r}(t)}{\mathrm{d}t} = u_r(t) - \frac{s_{2,r}(t)s_{1,r}(t)}{\theta_{2,r}\theta_{3,r}}, \tag{D.2}$$

where $u_r(t)$ is the rate at which blood delivers tracer into the region per unit effective volume. The second term describes tracer removal, controlled by the hidden pressure state $s_{2,r}(t)$, vascular resistance $\theta_{2,r}$, and effective volume $\theta_{3,r}$. This regional mass balance assumes a well-mixed compartment with constant effective volume and negligible diffusive exchange. The input curve is estimated separately from the imaging data before training, with its scale fitted for each region.

Latent hemodynamic. The hidden state $s_{2,r}(t)$ represents effective pressure, with downstream pressure set to zero as the reference. Its evolution follows the Windkessel equation (Lee & Liu, 2026)

$$\theta_{1,r}\frac{\mathrm{d}s_{2,r}(t)}{\mathrm{d}t} = \theta_{4,r} - \frac{s_{2,r}(t)}{\theta_{2,r}}, \tag{D.3}$$

where $\theta_{1,r}$ is vascular compliance and $\theta_{4,r}$ is constant blood inflow. Pressure changes according to the imbalance between inflow and pressure-driven outflow. This same outflow determines tracer removal in (D.2), connecting the hidden and observed dynamics. Collecting states across the R regions gives $s_1(t) = (s_{1,r}(t))_{r=1}^R$ and $s_2(t) = (s_{2,r}(t))_{r=1}^R$, the observation-side and hidden twin states used in Table 2. The quantities are summarized in Table 4, with the regional subscript r omitted for brevity.

Training details and ODE numerical integration. For CITRIS and PI-TR, we initialize the encoder, decoder, and latent transition from the same pretrained temporal VAE checkpoint. For Mech-ID, we reuse only its encoder. All three regimes are trained for 80 epochs on eight-observation windows with stride two and batch size two, using AdamW (Loshchilov & Hutter, 2017) with learning rates of 10^{-4} for pretrained components and 5×10^{-4} for mechanistic heads, weight decay of 10^{-4}, and gradient clipping at norm 1. For Mech-ID, we select the checkpoint with the lowest MSE between ODE-generated and observed concentration trajectories on the validation set. For CITRIS, selection minimizes the validation variational objective evaluated with the final KL weights.

For PI-TR, selection minimizes the validation physical objective after the 30-epoch warmup. For PINN, we train from scratch for 8,000 full-batch updates using AdamW with learning rate 10^{-3} and weight decay 10^{-5}. Its physical-loss weight increases to 1 over the first 2,000 updates, and we select the checkpoint with the lowest total loss on the validation set. For Mech-ID and PI-TR, the ODE integration uses explicit Euler steps at unit time-index intervals with nonnegativity clipping, initialized with observed concentration and inferred pressure. For all four regimes, we perform 20 seeded runs, with pretrained initialization held fixed across seeds where applicable.

D.2.2 Data and Model Implementation

Dataset and preprocessing. For the cerebral perfusion setting described above, here we use the MRI sequences from the Ischemic Stroke Lesion Segmentation (ISLES) 2017 challenge (Winzeck et al., 2018). We follow the MR signal-to-concentration conversion in (Liu et al., 2021) and adapt the anatomical segmentation and regional averaging procedure in (Lee & Liu, 2026). For completeness, we briefly restate these preprocessing steps at a high level. The raw MR signals are first converted into normalized concentration proxies. Then, SynthSeg (Billot et al., 2023), a brain MRI segmentation method, provides anatomical labels that are aligned with the perfusion images and combined into seven regional groups ($r \in \{1, 2, \ldots, 7\}$). Spatial averaging within each group yields the *observed concentration trajectory* $s_{1,r}(t)$ in (D.2). These trajectories, together with a separately estimated input curve, provide the data for fitting the coupled system in (D.2) and (D.3), whose pressure state $s_{2,r}(t)$ remains unobserved. The full image sequences are retained for representation learning. All four modeling regimes use the same fixed partition of 37 training cases and seven validation cases.

Implementation of PINN. PINN uses a feedforward network to directly estimate the regional state trajectories $s_{1,r}(t)$ and $s_{2,r}(t)$. The network takes normalized time, a learned embedding identifying the tissue region, and information from the precomputed input curve. This information includes the curve's current value, first difference, and normalized cumulative sum, together with its peak value, duration-normalized integral, normalized peak time, and baseline mean. Evaluating the network at successive time points produces the state trajectories. A separate head takes the region embedding and these four whole-curve summaries to estimate the regional mechanistic parameters and input scaling factors, which remain fixed over the sequence. Training fits the observed concentration trajectories and enforces the governing physics through squared finite-difference residual penalties for (D.2) and (D.3). Additional regularization penalizes successive pressure differences and log-parameter offsets from learned regional baselines. The PINN model has no image encoder, decoder, or latent transition.

Implementation of Mech-ID. Mech-ID adapts the mechanistic-identification view in (Yao et al., 2024) by mapping an initial image to parameters that generate a trajectory through integration. The first image of each window is processed by a three-dimensional convolutional encoder with the same architecture and pretrained initialization as those used in CITRIS and PI-TR detailed below. Its latent mean output is flattened and concatenated with the normalized time relative to the observed concentration peak. Two multilayer perceptron heads then estimate the initial hidden pressure state and the regional mechanistic parameters, including the input scaling factors. Together with the observed initial concentration and precomputed input curve, these estimates initialize the coupled ODEs in (D.2) and (D.3). Integrating these equations produces both state trajectories, with the inferred parameters held fixed throughout the window. The encoder and prediction heads are trained jointly to minimize the mean squared error between the ODE-generated and observed regional concentration trajectories. Physical dynamics are imposed through ODE integration, without a separate residual penalty. This implementation has no image decoder or independently learned latent transition.

Implementation of CITRIS. We implement the encoder $q_\phi(z_t|x_t)$ as a three-dimensional convolutional network that predicts a Gaussian latent distribution with diagonal covariance. The decoder $p_\psi(x_t|z_t)$ reconstructs the image through convolutional blocks and spatial upsampling. The transition prior $p_\gamma(z_{t+1}|z_t)$ is implemented by a convolutional residual network that predicts the next latent Gaussian distribution. The training objective extends (4.2) with an additional static KL term,

$$\begin{aligned}&\mathcal{L}_{\text{CITRIS}} \qquad\qquad\qquad\

Table 5: Across-seed variability of physiological parameter estimates.

Method	CV(θ_1) (↓)	CV(θ_2) (↓)	CV(θ_3) (↓)	CV(θ_4) (↓)
PINN	0.031	0.007	0.162	0.083
Mech-ID	0.773	0.350	0.295	0.398
CITRIS	N/A	N/A	N/A	N/A
PI-TR	**0.019**	0.051	**0.095**	**0.078**

inferred from the next observation, while the static KL term regularizes the posterior toward a standard Gaussian. In our implementation, reconstruction uses region-balanced MSE, supplemented by background MSE and spatial-gradient matching with weights 0.02 and 0.05, respectively. We set $\lambda = 1$ and increase β linearly from zero to 10^{-5} over the first 30 epochs. The transition network also receives the normalized time relative to the observed concentration peak. We remark that, this adaptation follows the temporal variational formulation of CITRIS (Lippe et al., 2022c) but omits intervention labels, intervention-specific latent partitions, and the intervention-target classifier. Moreover, we adopt the VAE formulation without the normalizing-flow extension of CITRIS-NF.

Implementation of PI-TR. While the adapted CITRIS baseline regularizes latent evolution through a learned transition prior, PI-TR additionally constrains the physiological dynamics represented by that evolution. Precisely, it retains the same encoder, decoder, latent transition, and reconstruction and KL terms (D.4). To evaluate physical residuals, we add two multilayer perceptron heads that map the learned latent representations to the physiological states and mechanistic parameters appearing in the governing equations. The first head is a state head that maps each latent state to regional pressure estimates $s_{2,r}(t)$, while the second head is a parameter head that averages latent states over the window and predicts regional mechanistic parameters and input scaling factors, held fixed within that window. Both heads are shared across the observation-encoded sequence and a latent sequence generated recursively by the transition network from the initial encoded state. For each sequence, finite-difference residuals of (D.2) and (D.3) are computed using the inferred pressures and parameters together with the observed regional concentration trajectories. An additional MSE term compares the observed concentration trajectories with those obtained by integrating the coupled ODEs, using the observed initial concentration and the initial pressure and parameters inferred from the transition-generated sequence. The physical objective combines residual penalties from both sequences, the ODE trajectory-fitting term, and regularization on pressure increments and parameter offsets. The two residual objectives and the ODE trajectory-fitting term receive equal weights. Within each residual objective, the pressure-increment and log-parameter-offset penalties have weights 0.01 and 0.001, respectively. The overall physical objective's weight increases linearly from zero to 0.05 over the first 30 epochs and remains fixed thereafter. All networks and heads are trained jointly.

D.2.3 Additional Evaluations

We assess the across-seed variability of the physiological parameters $\theta_1, \theta_2, \theta_3, \theta_4$. For each parameter, we first average its estimates over evaluation samples and regions within each seed. Its coefficient of variation (CV) is then the sample standard deviation of these seed-level averages divided by their absolute mean. CITRIS does not estimate physiological parameters, so its CVs are not applicable.

As shown in Table 5, PI-TR has the lowest CV for three parameters, while PINN has the lowest CV for θ_2. These statistics measure reproducibility across seeds, not parameter recovery accuracy.

Training dynamics. Figure 5 reports the training and validation objectives associated with the four implementations described in Section D.2.2. For PINN, the vertical axis measures the total objective combining observation fitting, residual penalties for the coupled governing equations, and pressure and parameter regularization. For Mech-ID, it measures the MSE between ODE-generated and observed concentration trajectories, since the governing dynamics are enforced through integration. For CITRIS, it measures the image reconstruction objective together with the temporal and static KL penalties. For PI-TR, it additionally includes the physical objective with weight 0.05, comprising residual penalties on both encoded and transition-generated sequences, ODE trajectory fitting, and mechanistic regularization. All curves use the final objective weights throughout, including during warmup, to separate changes in the loss components from changes in their scheduled weights.

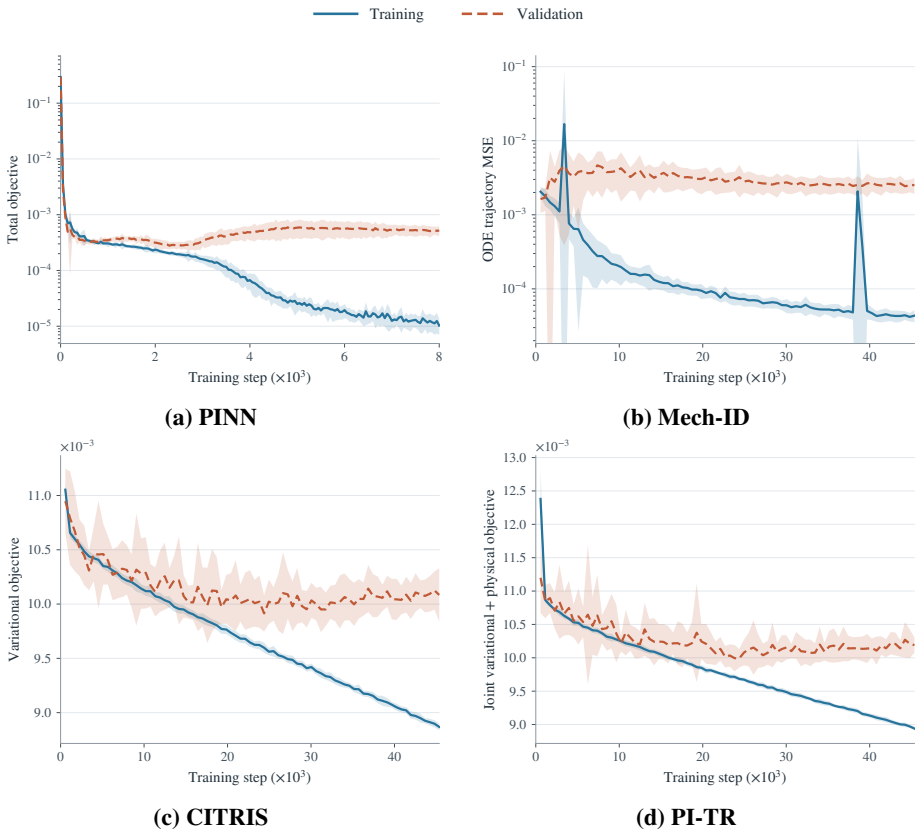

Figure 5: Training and validation objectives for the four cerebral perfusion regimes. Lines and shaded bands indicate the mean and sample standard deviation across 20 seeds. Objectives use the final loss weights throughout. For the mini-batch regimes, epoch-averaged training losses and end-of-epoch validation losses are plotted at cumulative training step counts. Objective definitions and computation per step differ across regimes. PI-TR checkpoint selection uses its physical validation objective.

The horizontal axis counts training steps, with one full-batch iteration per step for PINN and one mini-batch iteration for the other regimes. The latter retain their original recording frequency, with epoch-averaged training losses and end-of-epoch validation losses plotted at cumulative step counts. All four regimes reduce their training objectives, while validation behavior varies. For PINN, the validation objective decreases initially and subsequently rises as training continues. Mech-ID exhibits a pronounced separation between training and validation losses, with improvements in training fit providing limited validation gains. CITRIS and PI-TR show modest initial reductions in validation objectives, followed by a plateau with fluctuations despite continued decreases in the training loss.