PRIME: Poincaré return induced measure for learning partially observed dynamical systems

Yiting Duan[1,*] **Longyan Tan**[2] **Hao Wu**[1]
Mehdi Tavakol[1,*] **Yi Guo**[1,*]

[1]Western Sydney University, Australia
[2]University of Cambridge, United Kingdom

Abstract

Learning dynamics from sparse partial observations presents a fundamental dilemma: the pointwise mean squared error (MSE) training objective may induce local contraction and empirically lead to long-term rollout collapse, whereas invariant-measure matching preserves global geometry but lacks long-term temporal ordering constraints. To resolve this, we introduce the **P**oincaré **R**eturn **I**nduced **M**easur**e** (**PRIME**) framework. PRIME bridges this gap by imposing long-term temporal constraints through recurrence-level temporal organization. Instead of relying on unstable pointwise matching, PRIME refines the learned dynamics by leveraging recurrence-level temporal information, specifically by aligning the empirical recurrence graph measures constructed from consecutive recurrence events. We evaluate PRIME on the partially observed Lorenz63 system, sensor-observed Kuramoto-Sivashinsky equation, and flow past a cylinder. Across all benchmarks, PRIME effectively mitigates long-term rollout collapse and improves temporal consistency while maintaining or enhancing long-term geometric and spectral fidelity. Our source code is available at `https://github.com/CrispDyt/2026-NeurIPS-PRIME.git`.

1 Introduction

Learning dynamical systems from data is a central problem in science and engineering, especially for analyzing complex phenomena in chaotic systems, fluid dynamics [48], weather forecasting [2], and neuroscience [46, 23]. The primary goal is to learn a model that can autoregressively unroll trajectories from an initial state. Previous learning frameworks [9, 56, 59, 29, 12, 55] have primarily considered cases where the system state is fully observed, thus the observations are approximately Markovian. However, the full state of many real-world systems is not visible, and one may only have access to low-dimensional projections. In contrast to the ideal case, such observable processes are generally non-Markovian: the same observation may correspond to multiple latent states and multiple admissible futures. Consequently, methods adopting state or derivative matching based on the mean squared error (MSE) may lead to variance-shrunk or contracted rollouts rather than preserving faithful long-term dynamics [49, 25, 47, 54].

Fortunately, many of these chaotic nonlinear systems are dissipative with attractors; trajectories converge to an attractor characterized by an invariant measure [17]. Thus, a natural alternative is to shift the learning paradigm toward matching the invariant measures of the attractor, since models are capable of capturing the long-term statistics while preserving the global geometric shape of the system [22, 4, 3]. Recently, significant efforts have been directed toward this field, primarily focusing on the estimation and optimization of invariant measures [26, 13, 7].

*Correspondence to: Yiting Duan (`22173342@student.westernsydney.edu.au`); Mehdi Tavakol (`M.Tavakol@westernsydney.edu.au`); Yi Guo (`y.guo@westernsydney.edu.au`).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, a critical limitation of invariant-measure matching is that it constrains only the long-time distribution of trajectories, without preserving their temporal ordering. In practice, existing approaches often address this limitation by treating the measure-preserving term as a geometric regularizer and combining it with a pointwise MSE loss to incorporate temporal information, with the goal of achieving both short and long-term accuracy [49, 13]. However, when the system is partially observed and sparsely sampled, the first-order pointwise information becomes insufficient to serve as an effective temporal constraint (as demonstrated in Figure 1). We focus on learning an autonomous predictor in the partially observed space, where the predictor is evaluated by free rollout from an initial state. This setting differs from forecasting models that employ a finite history-window during inference to mitigate the non-Markovian ambiguity introduced by partial observation.

Thus, our core research question is: *How can long-term temporal structure be incorporated into a measure-based learning framework for modeling dynamical systems, without reverting to unreliable pointwise matching under sparsely sampled and partially observed conditions?*

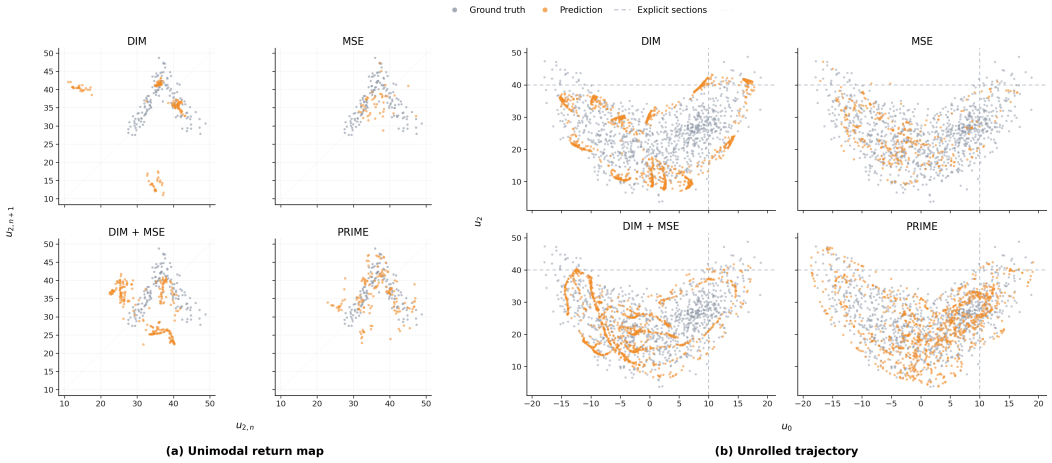

Figure 1: Unrolled trajectories of a three-layer MLP backbone trained with four different objectives ($\mathcal{L}^{\mathrm{DIM}}$, $\mathcal{L}^{\mathrm{MSE}}$, $\mathcal{L}^{\mathrm{PRIME}} + \mathcal{L}^{\mathrm{DIM}}$ and $\mathcal{L}^{\mathrm{DIM}} + \mathcal{L}^{\mathrm{MSE}}$) for the partially observed Lorenz63 system, i.e., $h(\boldsymbol{u}) = (u_0, u_2)$ with added Gaussian noise. (a) Unimodal return maps of unrolled trajectories for all models. (b) Unrolled trajectory from the same initial condition across all models.

The answer to this question is embedded in our contributions. First, we identify a long-term rollout failure mode caused by pointwise MSE training objectives under sparse and partial observations. We demonstrate that when a system is partially observed, an MSE-based predictor may act as a conditional-mean operator that contracts autoregressive rollouts toward spurious low-dimensional attractors. Second, we point out that the measure-matching training objective lacks temporal constraints, and the pointwise matching is not the optimal way to add them under partial observability. Third, we propose a novel training objective, PRIME (Poincaré Return Induced Measure), which incorporates recurrence-level temporal information to refine temporal structure consistency without relying on pointwise matching.

2 Related work

Learning dynamical systems from observational data is an extensively studied topic. Neural operators aim to learn integro-differential operators directly using neural network architectures [12, 34, 32, 21]. These methods leverage the universal approximation property of neural networks and predominantly rely on optimizing the pointwise MSE loss. Some of these approaches have been extended to partially observed systems and dissipative systems by incorporating dissipative terms or regularizing with other invariants, such as the Lyapunov spectrum [31, 45]. Another popular paradigm is learning dynamical systems based on the Koopman operator, where the evolution of the dynamics is approximated in a higher-dimensional latent space, constructed using polynomial basis functions or an autoencoder [37, 56, 59].

Table 1: Comparison of temporal constraints and learning objectives.

Method	Source of Temporal Constraint	Learning Objective
PRIME	recurrence graph measure (long-term) + n-step pushforward of DIM (short-term)	$\mathcal{L}^{\mathrm{DIM}} + \mathcal{L}^{\mathrm{PRIME}}$
MSE	pointwise MSE	$\mathcal{L}^{\mathrm{MSE}}$
DIM	n-step pushforward of DIM (short-term)	$\mathcal{L}^{\mathrm{DIM}}$
DIM + MSE	pointwise MSE + n-step pushforward of DIM (short-term)	$\mathcal{L}^{\mathrm{DIM}} + \mathcal{L}^{\mathrm{MSE}}$

Recently, growing efforts have been directed toward the data-driven learning of invariant measures which characterize the long-term behavior of the dynamical systems. Existing system identification methods incorporate an invariant-measure-preserving loss term in addition to a pointwise matching term to better capture long-term dynamics, such as DySLIM [49, 13, 57]. Moreover, Botvinick-Greenhouse et al. [4] proposed an Eulerian approach to learn dynamics based on the state invariant measure of low-dimensional ODEs using PDE-constrained optimization. Furthermore, it has been theoretically proven that a measure-theoretic embedding exists between such lifted manifolds [6].

Although matching invariant measures can increase the long-term rollout stability of the chaotic attractor, it is insufficient for dynamical system identification. Consequently, it is usually applied with a pointwise MSE loss for unique system identification. However, the invariant measure on the time-delay coordinate [5] can be used for unique system identification under certain conditions. However, optimizing only the DIM loss, i.e., $\mathcal{L}^{\mathrm{DIM}}$, constrains the generated rollouts to match the same stationary invariant distribution and finite local n-step temporal statistics by aligning the original delay invariant measure with its n-step pushforward measures. In other words, the geometry of the attractor may be correctly recovered while the long-term temporal ordering remains incorrect. Instead of imposing temporal constraints through pointwise MSE matching, PRIME leverages recurrence information, including recurrence events and their associated recurrence times, as a mechanism for long-term temporal supervision. Specifically, the learning objective combines recurrence transition measure matching and marginal-preserving terms with the DIM loss, where the recurrence transition measure captures long-term temporal dependencies while the DIM loss provides short-term temporal constraints. The comparison among different classes of methods is summarized in Table 1.

3 Background

We focus on a nonlinear autonomous dynamical system $\partial_t \boldsymbol{u} = \mathcal{F}[\boldsymbol{u}(t)]$, where $\boldsymbol{u}(t) \in \mathcal{M}$ is the state of the system at time t, and $\mathcal{M}$ is a d-dimensional compact manifold. For a given fixed time step Δt, the system is discretized as: $\boldsymbol{u}_{k+1} = \Phi^{\Delta t}(\boldsymbol{u}_k)$, with the flow $\Phi^{\Delta t} : \mathcal{M} \to \mathcal{M}$ obtained by integrating the equation over the period Δt. For notational simplicity, we drop the superscript Δt. Thus, given an initial state $\boldsymbol{u}_0$, we can use the flow operator to unroll the trajectory as $\boldsymbol{u}_k = \Phi(\boldsymbol{u}_{k-1}) = \Phi^2(\boldsymbol{u}_{k-2}) = \cdots = \Phi^k \boldsymbol{u}_0$. In practice, the dimensionality of the dynamical system is often high, while the choice of the measurement function is limited by physical constraints. Consequently, we typically obtain only partial observations: $\boldsymbol{y}_k \in \mathcal{N} \subseteq \mathbb{R}^\ell$, where $\ell < d$.

Let $\mathcal{M}$ be a measurable space equipped with a Borel σ-algebra $\mathcal{A}$ and a probability measure ν. If the flow Φ is a measurable map, it induces a pushforward measure $\Phi\#\nu : \mathcal{A} \to [0, 1]$, defined as $\Phi\#\nu(A) = \nu(\Phi^{-1}(A))$ for all $A \in \mathcal{A}$. The measure ν is called an *invariant measure* if it is preserved under the flow, i.e., $\Phi\#\nu = \nu$. An attractor is a subset of the state space that captures the system's long-term behavior. A defining characteristic of an attractor is that trajectories starting from its basin of attraction rapidly converge to this set, which serves as the support of the invariant measure [58, 17]. In this work, we consider trajectories approaching an attractor and assume that their long-time empirical distributions converge to an invariant probability measure supported on the attractor. While the invariant measure successfully captures the long-term statistical behavior of trajectories, an invariant measure alone does not uniquely identify the dynamics. For example, two distinct, non-diffeomorphic systems can share the same invariant measure. Consequently, invariant measure matching is typically employed as a geometric regularizer in conjunction with a pointwise loss function [49].

Recent theoretical advances [5] investigate the properties of the invariant measures on the reconstructed delay-coordinate system constructed via the time-delay map $\Psi^m_{y_i,\Phi} : \mathcal{M} \to \mathcal{X} \subseteq \mathbb{R}^m$,

defined as $\Psi^m_{y_i,\Phi}(\boldsymbol{u}) = (y_i(\boldsymbol{u}), y_i(\Phi^{-\tau}\boldsymbol{u}), ..., y_i(\Phi^{-(m-1)\tau}(\boldsymbol{u})))^\top$, where y_i is the i-th scalar component of the measurements $\boldsymbol{y}$, and τ denotes a fixed time-delay lag measured in observation steps. They prove that if two dynamical systems share the same delay invariant measure (DIM) on the same delay coordinate, they are topologically equivalent on their support. Moreover, they propose the following measure-preserving training objective for the predictor Φ_θ:

$$\mathcal{L}^{\mathrm{DIM}}(\theta) = w_{\mathrm{delay}}\mathcal{L}^{\mathrm{delay}} + w_{\mathrm{state}}\mathcal{L}^{\mathrm{measure}} = w_{\mathrm{delay}} \sum_{i=1}^{\ell} \mathcal{D}(\Psi^m_{y_i,\Phi_\theta}\#\nu, \nu_{y_i}) + w_{\mathrm{state}}\mathcal{D}(\Phi_\theta\#\nu, \nu) \quad (1)$$

Here, ν and ν_{y_i} denote the invariant measures in the observation and delay-coordinate spaces, respectively, and w_{state} and w_{delay} are hyperparameters. Given n observations, these measures are estimated empirically as $\hat{\nu} = \frac{1}{n}\sum_{k=1}^{n} \delta_{\boldsymbol{y}_k}$ and $\hat{\nu}_{y_i} = \frac{1}{n-(m-1)\tau}\sum_{k=(m-1)\tau+1}^{n} \delta_{\boldsymbol{z}_k^{(i)}}$, where δ is the Dirac measure and $\boldsymbol{z}_k^{(i)}$ is the delay vector constructed via $\Psi^m_{y_i,\Phi}$. The corresponding model-induced distributions are approximated using rollouts initialized from observed states, and $\mathcal{D}$ quantifies discrepancies between probability measures. The performance of $\mathcal{L}^{\mathrm{DIM}}$ also depends on the quality of delay-coordinate reconstruction [16], as discussed in Appendix A.

4 Topological collapse of MSE under partial observations

The goal is to learn a parametric operator Φ_θ operating within the partially observed space $\mathcal{N}$, and to evaluate its performance via autoregressive rollouts $\hat{\boldsymbol{y}}_{t+k} = \Phi^k_\theta(\boldsymbol{y}_t)$. Classical methods train the model Φ_θ by minimizing the ℓ-step MSE loss between the predicted and partially observed trajectories, formulated as [8, 27]:

$$\mathcal{L}_{\ell\text{-step}}(\theta) = \mathbb{E}_t \sum_{k=1}^{\ell} \left\| \Phi^k_\theta(\boldsymbol{y}_t) - \boldsymbol{y}_{t+k} \right\|_2^2 . \quad (2)$$

where $\|\cdot\|_2$ denotes the Euclidean norm in the observation space, and $\mathbb{E}_t$ represents the empirical average along the trajectory. However, optimizing $\mathcal{L}_{\ell\text{-step}}$ introduces several computational and numerical difficulties. When ℓ is large, standard backpropagation through time requires storing all intermediate evaluations due to the successive integration steps. As the rollout horizon grows, this frequently leads to vanishing or exploding gradients [41]. This challenge motivates the multiple shooting strategy [39], which partitions the long trajectory into shorter sequential segments, with their respective initial states processed in parallel, instead of training the dynamics end-to-end over the entire horizon. Furthermore, techniques such as multi-step objectives [48, 49] or pushforward tricks [8] can be applied to improve learning stability further.

The population minimizer of a one-step predictor is:

$$\Phi^\star_\theta(\boldsymbol{y}_k) = \mathbb{E}\big[\boldsymbol{y}_{k+1} \mid \boldsymbol{y}_k\big]. \quad (3)$$

By the law of total covariance, we have:

$$\mathrm{Cov}(\boldsymbol{y}_{k+1}) - \mathrm{Cov}(\Phi^\star_\theta(\boldsymbol{y}_k)) = \mathbb{E}[\mathrm{Cov}(\boldsymbol{y}_{k+1} \mid \boldsymbol{y}_k)]. \quad (4)$$

When the autonomous system is fully observed, $\boldsymbol{y}_{k+1}$ is uniquely determined by $\boldsymbol{y}_k$, thus $\mathbb{E}[\mathrm{Cov}(\boldsymbol{y}_{k+1} \mid \boldsymbol{y}_k)] = 0$. However, when the system is partially observed, the observed low-dimensional dynamics are generally not dynamically closed, rendering the observed evolution non-Markovian. For example, any two coordinates of the Lorenz63 system fail to define a dynamically closed two-dimensional subsystem, as the next-time evolution depends on the third unobserved coordinate. As a result, when $\mathbb{E}[\mathrm{Cov}(\boldsymbol{y}_{k+1} \mid \boldsymbol{y}_k)]$ is nonzero, the predictor may shrink the covariance of the trajectory and generate rollouts toward low-variance spurious attractors.

Figure 2 illustrates that under partial observability, the unrolled trajectories of the Lorenz63 system spiral into a limit cycle or a fixed point with covariance contraction. Additional results in Appendix C demonstrate that this topological collapse may also appear across multi-step MSE variants, various heterogeneous baselines, and other systems, which confirms that the failure is a pervasive issue under partial observation, rather than merely an artifact of one-step training on the Lorenz63 system.

This collapse of MSE-based rollout predictions is a pervasive failure mode in learning temporal dynamics [40, 1, 30]. In this work, we focus on partially observed chaotic systems. Inspired by the fundamental mechanism of chaos, the stretching and folding of phase space driven by

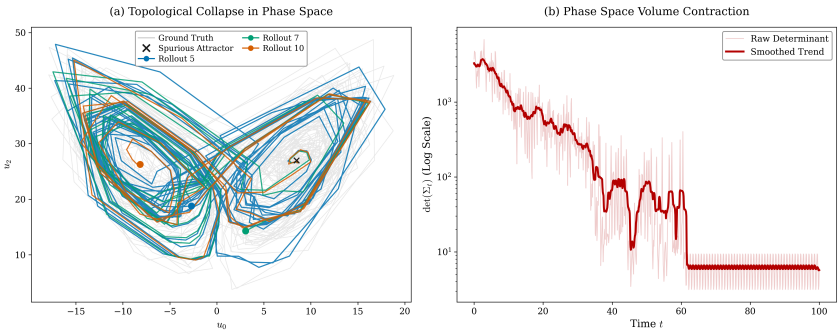

Figure 2: Topological collapse in the partially observed Lorenz63 dynamics ($h(\boldsymbol{u}) = (u_0, u_2)$) learned via a one-step autoregressive MSE objective with a three-layer MLP. Out of 30 randomly initialized rollouts, the trajectories degenerate into spurious low-dimensional attractors (e.g., 26 converge to a fixed point, while others fall into limit cycles). (a) Unrolled trajectories from three selected initial conditions all converge to a fixed point (black cross). (b) The determinant of the ensemble covariance over time, demonstrating macroscopic phase space volume contraction.

nonlinearities [51, 50], we posit that the capacity of an unrolled trajectory to remain chaotic is *governed by a competition between the system's inherent chaotic expansion (covariance amplification) and the information loss induced by partial observability*. The following informal theorem 1 describes a local condition for fixed-point collapse under partial observability. A formal version is provided in Appendix B.

Theorem 1. *Consider a one-step predictor Φ_θ trained with mean squared error using partial observations from a chaotic system. For observations in the local region $\boldsymbol{y}_t \in \overline{B_{\delta_0}}(\boldsymbol{y}_0)$, define Σ_{in} as the local covariance of $\boldsymbol{y}_t$, and $\Sigma_{\text{out}} := \text{Cov}_{\nu_{\delta_0}}(\boldsymbol{y}_{t+1})$ is the local output covariance. Define $\kappa = \lambda_{\max}(\Sigma_{\text{in}}^{-1/2} \text{Cov}_{\nu_{\delta_0}}(\boldsymbol{y}_{t+1})\Sigma_{\text{in}}^{-1/2})$ and $\eta = \lambda_{\min}(\Sigma_{\text{in}}^{-1/2}\mathbb{E}[\text{Cov}(\boldsymbol{y_{t+1}} \mid \boldsymbol{y}_t)]\Sigma_{\text{in}}^{-1/2})$. Under local conditions in Appendix B, if $\kappa - \eta$ is sufficiently small, then Φ_θ is locally contractive. Its rollout collapses to a fixed point.*

While Theorem 1 characterizes the local condition under which trajectories collapse into a fixed point, partial observability may also lead to other pathological behaviors, e.g., limit cycles, as reported in Appendix C. A promising solution to alleviate topological collapse under sparse partial observations is the use of measure-matching training objectives. However, a pure measure-matching objective is permutation-invariant, which can result in physically invalid rollouts. At the same time, a straightforward alternative could be a linear combination of the DIM term with a pointwise MSE loss for adding temporal constraints. However, under these severe conditions, pointwise temporal information could be unreliable, which may lead to suboptimal predictions, as shown in Tables 2 and 3. Consequently, a new paradigm that incorporates temporal structure instead of pointwise matching is imperative for modeling these ill-posed systems.

5 Methods

In this section, we introduce our training objective PRIME, which is the second-stage regularizer built on the DIM warm-up and inspired by classical suspension-flow theory of dynamical systems [24, 15]. Specifically, PRIME aligns the graph measure associated with the Poincaré return process, providing a long-term temporal constraint by jointly capturing recurrence statistics and recurrence transitions.

Let Φ^t denote the flow generated by the vector field $\mathcal{F}$. A Poincaré section Σ is a codimension-one hypersurface transverse to the flow, locally defined by $\Sigma = \{\boldsymbol{u} : h(\boldsymbol{u}) = 0\}$, with $\nabla h(\boldsymbol{u}) \cdot \mathcal{F}(\boldsymbol{u}) \neq 0$ for $\boldsymbol{u} \in \Sigma$. For points admitting a finite first return, the return time is defined as $\tau_\Sigma(\boldsymbol{u}) := \inf\{t > 0 : \Phi^t(\boldsymbol{u}) \in \Sigma\}$. The associated Poincaré return map is $P(\boldsymbol{u}) := \Phi^{\tau_\Sigma(\boldsymbol{u})}(\boldsymbol{u})$, which governs the transitions between successive crossings of Σ. Assuming that P admits an ergodic invariant probability measure ν_Σ, this induced measure characterizes the long-run distribution of return states on Σ for ν_Σ-almost every initial return state.

In our partially observed setting, recurrence events are extracted from observed trajectories without requiring exact intersections with a classical Poincaré section. Let $t_1 < t_2 < \cdots < t_N$ denote the observation indices of the detected recurrence events, with inter-event intervals $K_i = t_{i+1} - t_i$. For each event, we construct the recurrence coordinate $\boldsymbol{r}_i = [\boldsymbol{y}_{t_i}, \boldsymbol{z}_{t_i}]$, where $\boldsymbol{z}_{t_i} = (\boldsymbol{y}_{t_i-\tau}, \boldsymbol{y}_{t_i-2\tau}, \ldots, \boldsymbol{y}_{t_i-(m-1)\tau})$ denotes the time-delay component. For clarity, $\boldsymbol{r}_i^d$ and $\boldsymbol{r}_i^\theta$ denote the recurrence coordinates constructed from the observed data and model-generated trajectories, respectively.

The population-level target of PRIME is the stationary recurrence graph measure $\pi_\star := \mathcal{L}(\boldsymbol{r}_i, \boldsymbol{r}_{i+1})$ associated with the observed system. Let $\bar{\nu}_{\mathrm{rec}} := \mathcal{L}(\boldsymbol{r}_i)$ denote its marginal measure. The recurrence graph measure admits the disintegration $\pi_\star(dr, dr') = \bar{\nu}_{\mathrm{rec}}(dr) Q_\star(dr' \mid r)$, where $Q_\star(B \mid r) = \mathbb{P}(\boldsymbol{r}_{i+1} \in B \mid \boldsymbol{r}_i = r)$. Similarly, the recurrence graph measure associated with trajectories generated by the model is defined as $\pi_\theta := \mathcal{L}(\boldsymbol{r}_i^\theta, \boldsymbol{r}_{i+1}^\theta)$, which admits the same disintegration. The population-level PRIME objective is then defined as $\mathcal{L}^{\mathrm{PRIME}}(\theta) = D(\pi_\theta, \pi_\star)$, where D denotes a discrepancy between probability measures. The following proposition 2 establishes the identification of graph measure. All proofs in this section are provided in Appendix B.

Proposition 2 (Graph Measure Identification)**.** *If $\pi_\theta = \pi_\star$, then $\bar{\nu}_\theta = \bar{\nu}_{\mathrm{rec}}$ and $Q_\theta(\cdot|r) = Q_\star(\cdot|r)$ for $\bar{\nu}_{\mathrm{rec}}$-a.e. r.*

Our algorithm consists of two stages. In the first stage, we use the $\mathcal{L}^{\mathrm{DIM}}$ loss as a warm-up to capture the global attractor geometry via measure matching. In the second stage, we optimize the combined objective $\mathcal{L}^{\mathrm{PRIME}} + \mathcal{L}^{\mathrm{DIM}}$ to refine the long-term temporal organization by aligning empirical graph measures over consecutive transitions between recurrence events.

Two practical considerations motivate our approach. First, classical suspension flow theory characterizes the temporal structure of a flow through the induced return map, its invariant measure, and the associated roof-time function. A direct distributional approach to capturing the complete recurrence-level statistics would involve matching the joint law of the full return process, $\mathcal{L}(\ldots, \boldsymbol{r}_{i-1}, K_{i-1}, \boldsymbol{r}_i, K_i, \boldsymbol{r}_{i+1}, \ldots)$, which is computationally prohibitive. Second, constructing an explicit Poincaré section with transverse intersections and well-defined finite return times can be challenging, particularly under sparse and partial observations. Near-tangential crossings and ill-defined returns can make recurrence-event detection unreliable and compromise the numerical stability of subsequent model training.

To address these challenges, we construct a data-driven recurrence section as a union of local neighborhoods. Our goal is not to recover an exact global Poincaré section, but to extract stable recurrence events and their associated inter-event intervals from observed trajectories. We first augment the observations with time-delay coordinates, rank points by their estimated local velocity, and select active candidates to avoid nearly stationary or degenerate recurrence events. We then apply farthest-point sampling (FPS) to select spatially distributed landmarks, whose local neighborhoods constitute the data-driven recurrence section. Finally, adaptive thresholds are employed to ensure robust recurrence-event detection. Further details are provided in Appendix D.

Instead of matching the full return process, PRIME aligns empirical recurrence graph measures constructed from consecutive recurrence-coordinate pairs over recurrence blocks. For a recurrence block of L consecutive recurrence events, $\mathcal{B}_k = \{t_k, t_{k+1}, \ldots, t_{k+L-1}\}$, the associated recurrence coordinates are $\{\boldsymbol{r}_k, \boldsymbol{r}_{k+1}, \cdots, \boldsymbol{r}_{k+L-1}\}$. During training, the recurrence schedule is derived from the recurrence times extracted from the data. Specifically, starting from the current observation contained in $\boldsymbol{r}_i^\theta$, the predictor is unrolled for K_i observation steps to construct the next recurrence coordinate $\boldsymbol{r}_{i+1}^\theta$ along the generated trajectory. Thus, K_i serves as a recurrence-conditioned multi-shooting horizon, aligning the generated recurrence transitions with the empirical recurrence schedule.

For each recurrence block $\mathcal{B}_k$, the empirical recurrence graph measures constructed from the observed data and model-generated trajectories are defined using the Dirac measure δ as follows:

$$\hat{\pi}_d^{(k)} = \frac{1}{L-1} \sum_{j=k}^{k+L-2} \delta_{(\boldsymbol{r}_j^d, \boldsymbol{r}_{j+1}^d)}, \quad \hat{\pi}_\theta^{(k)} = \frac{1}{L-1} \sum_{j=k}^{k+L-2} \delta_{(\boldsymbol{r}_j^\theta, \boldsymbol{r}_{j+1}^\theta)}. \tag{5}$$

The corresponding source and target marginals are:

$$\hat{\mu}_{v,c}^{(k)} = \frac{1}{L-1} \sum_{j=k}^{k+L-2} \delta_{\boldsymbol{r}_{j+c}^v}, \quad \text{for } v \in \{d, \theta\},\ c \in \{0, 1\}. \tag{6}$$

The PRIME loss is then defined as:

$$\mathcal{L}^{\mathrm{PRIME}} = \mathbb{E}_{\mathcal{B}_k}\left[\frac{1}{L}\left[\mathcal{D}\left(\hat{\pi}_\theta^{(k)}, \hat{\pi}_d^{(k)}\right) + w_0\mathcal{D}\left(\hat{\mu}_{\theta,0}^{(k)}, \hat{\mu}_{d,0}^{(k)}\right) + w_1\mathcal{D}\left(\hat{\mu}_{\theta,1}^{(k)}, \hat{\mu}_{d,1}^{(k)}\right)\right]\right]. \tag{7}$$

Here, $\mathcal{D}$ denotes the distance between two probability distributions, which is the Sinkhorn divergence with a squared Euclidean cost and an entropy regularization parameter (controlled by a blur parameter $\varepsilon > 0$). The first term aligns the empirical graph measure over consecutive recurrence transitions, while the two marginal terms stabilize the source and target distributions within each recurrence block. The empirical recurrence schedule is used only during training to define the recurrence-conditioned multi-shooting horizons. During autonomous evaluation, no empirical recurrence times or inter-event intervals are provided to the model. The trained model is rolled out autoregressively from an initial state, and the same recurrence section extractor is applied post hoc to this generated trajectory. The full implementation details are presented in Appendix D Algorithm 2. The following proposition 3 establishes the Wasserstein consistency of the empirical recurrence graph measure estimator. However, the consistency result concerns the empirical recurrence transition measure and does not guarantee recovery of the learned flow or its invariant measure.

Proposition 3 (Empirical Graph Measure Consistency)**.** *Let $\mathcal{R} \subseteq \mathbb{R}^d$ be closed, and let r_i, $i \geq 0$ be a stationary and ergodic process with stationary marginal $\bar{\nu}_{\mathrm{rec}} = \mathcal{L}(r_i)$. Define the empirical estimator of $\pi_\star$ by $\widehat{\pi}_N = \frac{1}{N}\sum_{i=0}^{N-1}\delta_{(r_i, r_{i+1})}$. Fix $p \geq 1$, and equip $\mathcal{R}\times\mathcal{R}$ with the product metric $d_{\times,p}((r_1, r_1'), (r_2, r_2')) = \left(\|r_1 - r_2\|^p + \|r_1' - r_2'\|^p\right)^{1/p}$. If $\int_{\mathcal{R}}\|r\|^p d\bar{\nu}_{\mathrm{rec}}(r) < \infty$, then $W_p(\widehat{\pi}_N, \pi_\star) \to 0$ almost surely as $N \to \infty$.*

The following corollary 4 establishes a positive Wasserstein separation bound when the model-generated recurrence states collapse onto an incorrect closed set.

Corollary 4 (Collapsed-Support Separation Bound)**.** *Under the assumptions of Proposition 3, let $\widehat{\pi}_{\theta,N} = \frac{1}{N}\sum_{i=0}^{N-1}\delta_{(\boldsymbol{r}_i^\theta, \boldsymbol{r}_{i+1}^\theta)}$ denote the empirical recurrence transition measure generated by the model. Let $C \subseteq \mathcal{R}$ be nonempty and closed, and suppose that $\boldsymbol{r}_{i+1}^\theta \in C$ for all $i \geq 0$. Defining $d(\boldsymbol{r}, C) = \inf_{\boldsymbol{c}\in C}\|\boldsymbol{r} - \boldsymbol{c}\|$, we have*

$$W_p^p(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N) \geq \frac{1}{N}\sum_{i=0}^{N-1} d(\boldsymbol{r}_{i+1}, C)^p. \tag{8}$$

Consequently, $\liminf_{N\to\infty} W_p^p(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N) \geq \int_{\mathcal{R}} d(\boldsymbol{r}, C)^p\, d\bar{\nu}_{\mathrm{rec}}(\boldsymbol{r})$ almost surely. If $\bar{\nu}_{\mathrm{rec}}(C^c) > 0$, the right-hand side is strictly positive.

6 Numerical experiments

Benchmarks and baselines Our experiments are designed as objective-controlled comparisons. PRIME is a training objective for autonomous predictors on the partially observed state space, where delay coordinates and recurrence events are only used for training, rather than providing history windows during rollout. Therefore, in the main experiments, all methods share the same backbone architecture and the same training data, and differ only in the training objective $\mathcal{L}^{\mathrm{obj}} \in \{\mathrm{DIM}, \mathrm{MSE}, \mathrm{MSE} + \mathrm{DIM}, \mathrm{PRIME}\}$ across three benchmarks: the Lorenz63 system, KS system, and flow past a cylinder. Relevant theoretical results and numerical evidence concerning invariant measures for the systems underlying these benchmarks are provided in [53, 14, 35, 11, 10]. For a fair comparison, all models within a given benchmark share the same backbone architecture and the same dataset. We also report the performance of history-window and heterogeneous forecasting baselines separately as a reference in Table 4. Experiments are repeated with five different random seeds using PyTorch [43].

Evaluation We evaluate the models based on their ability to maintain temporal consistency and preserve long-term statistics and spectral fidelity. For temporal accuracy, we compute the Normalized Mean Squared Error (**NMSE**) over a finite prediction window, normalized by the mean squared magnitude of the reference trajectory, the Distance of Recurrence Time Distributions (**DRTD**), defined as the 1D empirical Wasserstein-1 distance between the ground-truth and predicted Poincaré recurrence time, and the Time Correlation Metric (**TCM**) [49]. To evaluate long-term physical fidelity, we use the Sinkhorn Divergence (**SD**) between the empirical distributions of the attractors to quantify

global geometric overlap. Additionally, we also quantify spectral consistency using a log-power spectral density (**Log-PSD**) distance via Welch's method. Although PRIME uses recurrence-level information for training, no empirical recurrence times or inter-event intervals are provided for generating rollouts. For DRTD, the explicit Poincaré sections used for testing are distinct from the recurrence section used for training. For further details and precise definitions of the evaluation criteria, see Appendix E.

6.1 Results and comparison

Lorenz63. The first system we examine is the Lorenz63 system [33], a simplified representation of nonlinear atmospheric convection characterized by symmetric properties. For a fair comparison, both our method and the baselines employ a three-layer MLP architecture trained on 500 observations via one-step prediction. Figure 1 shows the unrolled trajectories generated by models trained under four different objectives. For our analysis, we select two explicit Poincaré sections, $\Sigma_1 := \{u_0 = 10\}$, $\Sigma_2 := \{u_2 = 40\}$, alongside a third section (Σ_3) induced by the unimodal return map, which is defined as the successive transition map of the u_2 coordinate. As detailed in Tables 2 and 3, PRIME provides the best overall trade-off among attractor geometry, temporal consistency, and spectral properties, while maintaining the non-collapsed rollout behavior. Although the MSE training objective achieves the lowest long-horizon NMSE, this metric is misleading here, as the generated trajectories collapse to a spurious low-dimensional attractor.

Table 2: Temporal consistency diagnosis across all benchmarks, where Σ_i refers to the explicit section for DRTD calculation. TCM is evaluated over 32 rollout initial conditions for each of the five random seeds. DRTD uses a common rollout initial condition and is summarized over seeds with sufficient return events. The complete results are provided in the Appendix F.

System	Model	TCM(↓)	Poincaré DRTD (↓)		
			$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	$\boldsymbol{\Sigma_3}$
Lorenz63	PRIME	$\mathbf{0.28 \pm 0.51}$	$\mathbf{0.48 \pm 0.14}$	1.60 ± 1.54	$\mathbf{0.080 \pm 0.032}$
	MSE	1.40 ± 1.79	1.72 ± 0.90	$5.00 \pm 0.00^*$	0.30 ± 0.19
	DIM	0.61 ± 1.27	0.64 ± 0.22	$\mathbf{0.86 \pm 0.32}$	0.13 ± 0.062
	DIM+MSE	0.30 ± 0.40	0.84 ± 0.64	3.22 ± 1.37	0.40 ± 0.060
KS	PRIME	$\mathbf{10.18 \pm 18.78}$	$\mathbf{7.30 \pm 1.58}$	$\mathbf{6.49 \pm 3.12}$	12.33 ± 7.70
	MSE	88.03 ± 114.19	11.40 ± 4.74	11.16 ± 3.09	$\mathbf{9.80 \pm 3.57}$
	DIM	91.67 ± 120.64	10.39 ± 3.66	11.18 ± 4.34	11.33 ± 7.24
	DIM+MSE	84.65 ± 103.30	10.70 ± 7.64	7.58 ± 2.12	16.57 ± 5.18
CFD	PRIME	$\mathbf{0.011 \pm 0.0058}$	$\mathbf{(2.66 \pm 1.24) \times 10^{-3}}$	$\mathbf{(2.35 \pm 1.82) \times 10^{-3}}$	$\mathbf{(5.42 \pm 1.25) \times 10^{-3}}$
	MSE	0.09 ± 0.07	$(2.86 \pm 0.86) \times 10^{-3}$	$(4.87 \pm 2.47) \times 10^{-3}$	$(6.46 \pm 2.72) \times 10^{-3}$
	DIM	0.14 ± 0.08	$(4.64 \pm 3.87) \times 10^{-3}$	$(4.63 \pm 3.56) \times 10^{-3}$	$(7.53 \pm 2.57) \times 10^{-3}$
	DIM+MSE	0.10 ± 0.04	$(4.55 \pm 1.36) \times 10^{-3}$	$(3.47 \pm 0.93) \times 10^{-3}$	$(7.86 \pm 0.53) \times 10^{-3}$

* Only one of the five models produced enough section crossings to compute return intervals.

Kuramoto-Sivashinsky. Next, we evaluate our approach on a more challenging setting: the Kuramoto-Sivashinsky (KS) equation, a high-order, one-dimensional chaotic PDE [42]. For this task, we employ a one-step three-layer MLP across all training objectives, and the system is initialized with a sine wave perturbed by Gaussian noise. The discretized KS system is highly chaotic [44], featuring a state dimension of 200 over a spatial domain of size 100. Based on empirical formulations [38, 19], its global attractor possesses a dimension between 27 and 35. In our experimental setting, 500 observations are collected from only five sensors, a severely ill-posed, partially observed scenario.

Tables 2 and 3 report the rollout performance of the trained models. For the DRTD metric evaluation, we use three explicit Poincaré sections, $\Sigma_1 := \{u_0 = 0\}$, $\Sigma_2 := \{u_1 = 0\}$, and $\Sigma_3 := \{u_2 = 0\}$. The MSE-based model exhibits severe long-horizon error accumulation and diverges in these rollouts. In contrast, DIM and DIM+MSE-based models alleviate this error growth, but still accumulate substantial long-horizon error. PRIME achieves the lowest NMSE for both short and long horizons, while also preserving the attractor geometry and recurrence-level temporal structure more accurately, which indicates that PRIME improves rollout stability and reduces the occurrence of collapse or divergence.

Flow past a cylinder. Finally, we study 2D fluid flow past a cylinder at a Reynolds number $Re = 70$, governed by the Navier-Stokes equations [28]. This system exhibits classical vortex shedding and

Table 3: Performance comparison across the Lorenz63, Kuramoto-Sivashinsky (KS), and flow past a cylinder benchmarks. All metrics are evaluated over 160 runs (32 trajectories for each of the 5 random seeds).

System	Model	Short- & long-term NMSE (↓)			Structure Metrics (↓)	
		NMSE 20	NMSE 100	NMSE 400	SD	Log-PSD
Lorenz63	PRIME	$\mathbf{0.20 \pm 0.13}$	0.29 ± 0.062	0.32 ± 0.039	$\mathbf{8.11 \pm 10.43}$	$\mathbf{0.74 \pm 0.29}$
	MSE	0.20 ± 0.13	$\mathbf{0.29 \pm 0.060}$	$\mathbf{0.28 \pm 0.040}$	53.06 ± 31.61	3.85 ± 3.38
	DIM	0.26 ± 0.16	0.35 ± 0.083	0.37 ± 0.066	21.37 ± 24.10	1.81 ± 2.092
	DIM+MSE	0.21 ± 0.12	0.28 ± 0.068	0.29 ± 0.054	31.91 ± 26.83	1.91 ± 2.12
KS	PRIME	$\mathbf{1.30 \pm 0.52}$	$\mathbf{1.46 \pm 0.23}$	$\mathbf{1.47 \pm 0.12}$	$\mathbf{1.99 \pm 0.78}$	1.66 ± 1.24
	MSE	1.92 ± 0.87	7.21 ± 30.14	$(2.57 \pm 9.88) \times 10^2$	$(1.19 \pm 3.92) \times 10^4$	1.59 ± 0.25
	DIM	1.72 ± 0.61	2.22 ± 0.93	16.03 ± 46.36	$(1.20 \pm 2.73) \times 10^3$	$\mathbf{0.96 \pm 0.63}$
	DIM+MSE	1.68 ± 0.78	2.52 ± 1.77	11.70 ± 31.21	$(3.36 \pm 9.46) \times 10^2$	1.48 ± 0.85
CFD	PRIME	$\mathbf{(6.39 \pm 2.91) \times 10^{-5}}$	$\mathbf{(6.91 \pm 3.04) \times 10^{-5}}$	$\mathbf{(1.69 \pm 1.17) \times 10^{-4}}$	$\mathbf{(3.91 \pm 0.39) \times 10^{-3}}$	$\mathbf{0.74 \pm 0.12}$
	MSE	$(1.12 \pm 0.62) \times 10^{-4}$	$(3.80 \pm 3.85) \times 10^{-4}$	$(4.37 \pm 5.02) \times 10^{-3}$	$(5.81 \pm 0.72) \times 10^{-3}$	1.33 ± 0.60
	DIM	$(1.32 \pm 1.16) \times 10^{-4}$	$(2.06 \pm 2.73) \times 10^{-3}$	0.02 ± 0.02	$(4.34 \pm 0.45) \times 10^{-3}$	1.38 ± 0.43
	DIM+MSE	$(8.60 \pm 4.14) \times 10^{-5}$	$(9.92 \pm 5.36) \times 10^{-4}$	0.01 ± 0.01	$(4.02 \pm 0.20) \times 10^{-3}$	1.32 ± 0.34

Table 4: Heterogeneous forecasting baselines under the 2D partially observed Lorenz63 benchmark. All models are evaluated from the same 30 rollout initial conditions. DRTD is averaged over $\Sigma_1 := \{u_0 = 10\}$, $\Sigma_2 := \{u_2 = 40\}$ and the recurrence section induced by the unimodal map.

Model	NMSE 100 ↓	NMSE 400 ↓	SD ↓	Log-PSD ↓	DRTD ↓
NGRC-NVAR [20]	468.30 ± 1551.00	503.20 ± 1650.00	1.93×10^5	4.22	0.39
Reservoir-ESN [36]	$1.10 \times 10^4 \pm 2430.00$	$1.13 \times 10^4 \pm 1036.00$	4.20×10^6	4.26	0.80
LSTM [54]	0.26 ± 0.06	0.31 ± 0.03	1.78	0.64	0.65
Koopman-AE [37]	0.27 ± 0.05	0.31 ± 0.03	1.14	0.71	1.95
DySLIM-style [49]	0.22 ± 0.04	0.21 ± 0.02	63.24	9.13	0.32
PRIME	0.31 ± 0.05	0.32 ± 0.03	3.61	0.62	0.43

presents a challenging partially observed scenario. We collect 200 observations from nine probes in the Kármán vortex street, corrupted by Gaussian noise, as shown in Figure 3. We apply a three-layer MLP coupled with a 20-step training strategy across all training objectives.

Figure 3 illustrates the experimental setup and the rollout behavior from a fixed initialization. Models that incorporate measure matching and pointwise matching successfully capture the general oscillatory structure, but all of them suffer from progressive phase shifting during extended rollouts. In contrast, from Tables 2 and 3, PRIME achieves the best prediction accuracy, attractor geometry, spectral fidelity, and recurrence-level temporal consistency, where the three Poincaré sections are $\Sigma_1 := \{u_5 = 4\}, \Sigma_2 := \{u_4 = 4\}$ and $\Sigma_3 := \{u_6 = 4\}$. The variability in the Lorenz63 ablation results is largely driven by one outlying seed. Details are reported in Table 7 in Appendix F.

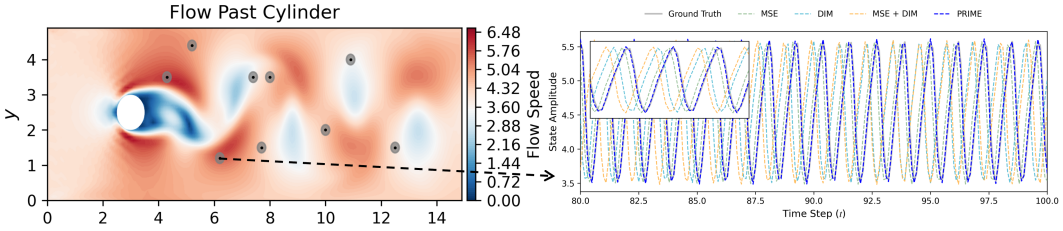

Figure 3: Forecasting the dynamics of noisy observations in a Kármán vortex street using four training objectives. Left: Snapshot of the flow and illustration of the partially observed velocity probes. Right: Trajectory forecasts for the selected probe.

6.2 Ablation study

In this section, we revisit the Lorenz63 system and conduct ablation studies to evaluate the effects of the recurrence schedule, the recurrence section, and loss terms in PRIME. We conclude that the

Table 5: Mechanism verification on the Lorenz63 system. Left: recurrence schedule ablations varying the recurrence block timing schedule K. Right: Recurrence section ablations varying the selection of recurrence events. NMSE 400 is reported as mean $\pm$ std over 30 rollout initializations.

(a) Recurrence schedule

Variant	NMSE 400 ↓	SD ↓	Log-PSD ↓	DRTD ↓
DIM-only	0.310 ± 0.037	57.054	0.849	0.430
PRIME-fixed-K	0.281 ± 0.026	75.286	3.933	0.601
PRIME-shuffled-K	$\mathbf{0.267 \pm 0.015}$	41.068	3.178	0.657
PRIME-correct-K	0.341 ± 0.028	**2.332**	**0.434**	**0.262**

(b) Recurrence section

Variant	NMSE 400 ↓	SD ↓	Log-PSD ↓	DRTD ↓
DIM-only	0.314 ± 0.015	71.05	4.133	0.436
Random section	0.304 ± 0.035	50.98	3.674	0.441
Data-driven	0.325 ± 0.025	**1.447**	**0.423**	**0.330**
Explicit Z-band	$\mathbf{0.253 \pm 0.012}$	40.42	1.417	0.788

recurrence schedule and the recurrence section are both necessary. We also ablate the data-driven recurrence section, with detailed results provided in Appendix F. The comparison of computational cost and inference time is shown in Appendix G.

Recurrence schedule. The fixed-K variant replaces all inter-event intervals with their median value. In contrast, the shuffled-K variant preserves the marginal distribution of the intervals but destroys their state-dependent pairing with the recurrence coordinates. In contrast to PRIME, as illustrated in Table 5 (a), the fixed-K variant collapses to a fixed point, while the shuffled-K variant collapses to a spurious limit cycle. These pathological behaviors confirm the critical role of maintaining the state-dependent temporal structure within the recurrence block.

Recurrence section and recurrence events. We next ablate the construction of the recurrence section. The random section baseline uses the identical number of recurrence events as the data-driven recurrence section, but samples them randomly. Similarly, we restrict the explicit Z-band section to achieve the same number of recurrence events as the data-driven recurrence section. Table 5 (b) demonstrates that the efficacy of PRIME stems from capturing dynamically meaningful recurrence events, rather than merely benefiting from temporal information from an additional, arbitrary recurrence block, since all other methods suffer topological collapse. Additional ablations in Appendix F support the importance of recurrence events for constructing informative long-term transition measures.

Loss term. We further ablate the PRIME loss terms and the role of $\mathcal{L}^{\mathrm{DIM}}$ in Appendix F, which supports retaining the DIM loss during the second stage. The unrolled trajectories from the marginal-only variant collapse to a point. Adding the graph-measure term of the induced return map alleviates the topological collapse. At the same time, the full PRIME objective achieves the best geometry, which confirms that the critical mechanism to avoid topological collapse is the graph measure matching, whereas marginal matching alone is insufficient.

7 Conclusion

In this work, we propose PRIME, a training objective that integrates the recurrence-level temporal information into the measure-matching training framework for learning dynamics under sparse and partial observations. Specifically, by aligning the empirical graph measure of the induced return map, we demonstrated that PRIME reduces topological collapse and improves the recurrence-level temporal consistency, long-term attractor geometry, and spectral fidelity. These improvements were consistently validated across a range of challenging benchmarks, including the Lorenz63 system, the KS equation, and fluid flow past a cylinder. We hope that the learning paradigm of the recurrence-level temporal alignment with measure-matching can be generalized to real-world partially observed dynamical systems, such as complex fluid dynamics forecasting from sparse sensor arrays.

Limitations

For numerical implementation, PRIME employs the empirical inter-event interval K_i as a recurrence-conditioned multi-shooting horizon during training. The model is evaluated via autonomous rollout without access to empirical recurrence times, but the model does not generate the roof time itself. A natural extension is a generative suspension model that jointly learns the return map transition, the roof time distribution, and the trajectory evolution within the block. Furthermore, PRIME introduces a non-negligible computational overhead during training. Compared with DIM, each training iteration is approximately 10-16x more expensive because of the recurrence graph measure matching.

Acknowledgements

Yiting Duan acknowledges the financial support provided by the WentWest Primary Health Network in the form of a PhD scholarship.

References

[1] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

[2] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[3] Jonah Botvinick-Greenhouse and Yunan Yang. On the unique recovery of transport maps and vector fields from finite measure-valued data. *arXiv preprint arXiv:2604.07671*, 2026.

[4] Jonah Botvinick-Greenhouse, Robert Martin, and Yunan Yang. Learning dynamics on invariant measures using pde-constrained optimization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(6), 2023.

[5] Jonah Botvinick-Greenhouse, Robert Martin, and Yunan Yang. Invariant measures in time-delay coordinates for unique dynamical system identification. *Physical Review Letters*, 135(16): 167202, 2025.

[6] Jonah Botvinick-Greenhouse, Maria Oprea, Romit Maulik, and Yunan Yang. Measure-theoretic time-delay embedding. *Journal of Statistical Physics*, 192(12):171, 2025.

[7] Jason J Bramburger and Giovanni Fantuzzi. Data-driven discovery of invariant measures. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 480 (2286), 2024.

[8] Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. *arXiv preprint arXiv:2202.03376*, 2022.

[9] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[10] Mattia Cenedese, Joar Axås, Bastian Bäuerlein, Kerstin Avila, and George Haller. Data-driven modeling and prediction of non-linearizable dynamics via spectral submanifolds. *Nature communications*, 13(1):872, 2022.

[11] Mickaël D Chekroun and Nathan E Glatt-Holtz. Invariant measures for dissipative dynamical systems: abstract results and applications. *Communications in Mathematical Physics*, 316(3): 723–761, 2012.

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

[13] Xiaoyuan Cheng, Yi He, Yiming Yang, Xiao Xue, Sibo Cheng, Daniel Giles, Xiaohang Tang, and Yukun Hu. Learning chaos in a linear way. *arXiv preprint arXiv:2503.14702*, 2025.

[14] Pierre Collet, Jean-Pierre Eckmann, Henri Epstein, and Joachim Stubbe. A global attracting set for the kuramoto-sivashinsky equation. *Communications in Mathematical Physics*, 152: 203–214, 1993.

[15] Isaac P Cornfeld, Sergej V Fomin, and Yakov Grigorevich Sinai. *Ergodic theory*. Springer Science & Business Media, 2012.

[16] Yiting Duan, Zhikun Zhang, and Yi Guo. Wasserstein geometry of information loss in nonlinear dynamical systems. *arXiv preprint arXiv:2601.22814*, 2026.

[17] J-P Eckmann and David Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of modern physics*, 57(3):617, 1985.

[18] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 2681–2690. PMLR, 2019.

[19] Ciprian Foias, Basil Nicolaenko, George R Sell, and Roger Temam. Inertial manifolds for the kuramoto-sivashinsky equation and an estimate of their lowest dimension. *Journal de Mathématiques Pures et Appliquées*, 67(3):197–226, 1988.

[20] Daniel J Gauthier, Erik Bollt, Aaron Griffith, and Wendson AS Barbosa. Next generation reservoir computing. *Nature communications*, 12(1):5564, 2021.

[21] Jingren Hou, Hong Wang, Pengyu Xu, Chang Gao, Huafeng Liu, and Liping Jing. Learning neural operators from partial observations via latent autoregressive modeling. *arXiv preprint arXiv:2601.15547*, 2026.

[22] Ruoxi Jiang, Peter Y Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors. *Advances in Neural Information Processing Systems*, 36:27645–27669, 2023.

[23] Justin Jude, Matthew G Perich, Lee E Miller, and Matthias H Hennig. Robust alignment of cross-session recordings of neural population activity by behaviour via unsupervised domain adaptation. *arXiv preprint arXiv:2202.06159*, 2022.

[24] Anatole Katok, AB Katok, and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*. Number 54. Cambridge university press, 1995.

[25] Ryan Keisler. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*, 2022.

[26] Aaron Kirtland, Jonah Botvinick-Greenhouse, Marianne DeBrito, Megan Osborne, Casey Johnson, Robert S Martin, Samuel J Araki, and Daniel Q Eckhardt. An unstructured mesh approach to nonlinear noise reduction for coupled systems. *SIAM Journal on Applied Dynamical Systems*, 22(4):2927–2944, 2023.

[27] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

[28] Felix Koehler. Machine learning and simulation. `https://github.com/ceyron/machine-learning-and-simulation`, 2024. Source code repository.

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

[30] Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[31] Jiayi Li, Zhaonan Wang, and Flora D Salim. Sgno: Spectral generator neural operators for stable long horizon pde rollouts. *arXiv preprint arXiv:2602.18801*, 2026.

[32] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

[33] Edward N Lorenz. Deterministic nonperiodic flow 1. In *Universality in Chaos, 2nd edition*, pages 367–378. Routledge, 2017.

[34] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.

[35] Grzegorz Lukaszewicz. Pullback attractors and statistical solutions for 2-d navier-stokes equations. *Discrete Cont. Dyn. Syst.-B*, 9:643–659, 2008.

[36] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer science review*, 3(3):127–149, 2009.

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

[38] Paul Manneville. Liapounov exponents for the kuramoto-sivashinsky model. In *Macroscopic Modelling of Turbulent Flows: Proceedings of a Workshop Held at INRIA, Sophia-Antipolis, France, December 10–14, 1984*, pages 319–326. Springer, 2005.

[39] Stefano Massaroli, Michael Poli, Sho Sonoda, Taiji Suzuki, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Differentiable multiple shooting layers. *Advances in Neural Information Processing Systems*, 34:16532–16544, 2021.

[40] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[41] Jonas Mikhaeil, Zahra Monfared, and Daniel Durstewitz. On the difficulty of learning chaotic dynamics with rnns. *Advances in neural information processing systems*, 35:11297–11312, 2022.

[42] AM Obukhov. Kolmogorov flow and laboratory simulation of it. *Russ. Math. Surv*, 38(4): 113–126, 1983.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[44] Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical review letters*, 120(2):024102, 2018.

[45] Jason A Platt, Stephen G Penny, Timothy A Smith, Tse-Chun Chen, and Henry DI Abarbanel. Constraining chaos: Enforcing dynamical invariants in the training of recurrent neural networks. *arXiv preprint arXiv:2304.12865*, 2023.

[46] Mikhail I Rabinovich, Pablo Varona, Allen I Selverston, and Henry DI Abarbanel. Dynamical principles in neuroscience. *Reviews of modern physics*, 78(4):1213–1265, 2006.

[47] 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.

[48] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

[49] Yair Schiff, Zhong Yi Wan, Jeffrey B Parker, Stephan Hoyer, Volodymyr Kuleshov, Fei Sha, and Leonardo Zepeda-Núñez. Dyslim: Dynamics stable learning by invariant measure for chaotic systems. *arXiv preprint arXiv:2402.04467*, 2024.

[50] Stephen Smale. Differentiable dynamical systems. *Bulletin of the American mathematical Society*, 73(6):747–817, 1967.

[51] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (studies in nonlinearity)*, volume 1. Westview press, 2001.

[52] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, pages 366–381. Springer, 2006.

[53] Warwick Tucker. A rigorous ode solver and smale's 14th problem. *Foundations of computational mathematics*, 2(1):53–117, 2002.

[54] Pantelis R Vlachas, Wonmin Byeon, Zhong Y Wan, Themistoklis P Sapsis, and Petros Koumoutsakos. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213), 2018.

[55] Benjamin Walker, Andrew D McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The lie brackets make a difference. *arXiv preprint arXiv:2402.18512*, 2024.

[56] Yuanchao Xu, Kaidi Shao, Nikos Logothetis, and Zhongwei Shen. Reskoopnet: Learning koopman representations for complex dynamics with spectral residuals. *arXiv preprint arXiv:2501.00701*, 2025.

[57] Yunan Yang, Levon Nurbekyan, Elisa Negrini, Robert Martin, and Mirjeta Pasha. Optimal transport for parameter identification of chaotic dynamics via invariant measures. *SIAM Journal on Applied Dynamical Systems*, 22(1):269–310, 2023.

[58] Lai-Sang Young. What are srb measures, and which dynamical systems have them? *Journal of statistical physics*, 108(5):733–754, 2002.

[59] Shimin Zhang, Ziyuan Ye, Yinsong Yan, Zeyang Song, Yujie Wu, and Jibin Wu. Koopstd: Reliable similarity analysis between dynamical systems via approximating koopman spectrum with timescale decoupling. In *Forty-second International Conference on Machine Learning*, 2025.

A Time-delay reconstruction and delay invariant measure matching

The invariant measure ν is a probability measure supported on the attractor of the dynamical system. For a discrete dynamical system Φ, the invariant property guarantees:

$$\Phi\#\nu = \nu,$$

and, more generally, $(\Phi^q)\#\nu = \nu$ for any integer pushforward horizon $q \geq 1$. For a learned predictor Φ_θ, this naturally introduces a distributional constraint:

$$\Phi_\theta^q\#\nu = \nu,$$

where ν is estimated from observed trajectories on the attractor. This invariant measure matching helps preserve the long-term geometry of the attractor. However, matching the invariant measure in the observation space alone is not sufficient for unique system identification, since distinct dynamical systems may share the same invariant measure while having different temporal organizations.

The DIM method strengthens invariant measure matching by comparing invariant measures on the reconstructed delay coordinates. For a scalar observable y_i and parameters for time-delay reconstruction (m, τ), the delay-coordinate map is defined as

$$\Psi^m_{y_i,\Phi}(\boldsymbol{u}) = \big(y_i(\boldsymbol{u}), y_i(\Phi^{-\tau}\boldsymbol{u}), \ldots, y_i(\Phi^{-(m-1)\tau}\boldsymbol{u})\big).$$

The corresponding delay invariant measure is the pushforward measure $\Psi^m_{y_i,\Phi}\#\nu$. The theoretical motivation of DIM is that, under suitable genericity, embedding, and periodic-orbit regularity assumptions, equality of delay-coordinate invariant measures implies topological equivalence of the underlying dynamics on the supports of their invariant measures [5]. Thus, compared with matching the invariant measure only in the observation space, DIM imposes a stronger geometric constraint by matching the invariant measure in a reconstructed delay-coordinate space. Based on this idea, DIM formulates system identification as an optimization problem that combines state-space invariant measure matching with delay-coordinate invariant measure matching:

However, the effectiveness of DIM depends on the quality of the time-delay reconstruction, and in particular on whether the delay map is sufficiently informative. Under mild regularity conditions, Takens' embedding theorem [52] guarantees that a generic observation function yields an embedding

Algorithm 1 Delay invariant measure matching.

Input. Partial observations $\{\boldsymbol{y}_t\}_{t=1}^{T} \subseteq \mathbb{R}^{\ell}$.
(i) Select parameters m, τ for time-delay reconstruction and a pushforward horizon q.
(ii) Construct the empirical attractor measure ν and the empirical delay-coordinate measures $\hat{\mu}^{m}_{y_i,\Phi}$ for each observed coordinate y_i, $i = 1, \ldots, \ell$.
(iii) Solve

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \left[w_{\text{state}} \mathcal{D}\big((\Phi_\theta^q)\#\nu, \nu\big) + w_{\text{delay}} \sum_{i=1}^{\ell} \mathcal{D}\left(\Psi^{m}_{y_i,\Phi_\theta}\#\nu, \hat{\mu}^{m}_{y_i,\Phi}\right) \right].$$

Output. Trained predictor $\Phi_{\tilde{\theta}}$.

(i.e., continuous injective function) from the underlying state space to the reconstructed space. However, this guarantee is asymptotic and existential. For a fixed measurement function and finite data, the delay map may be poorly conditioned or even non-injective on the relevant parts of the attractor. When injectivity fails, different latent states can share nearly identical delay coordinates while having different future evolutions. In this case, the non-injectivity of the time-delay reconstruction does not admit a reliable single-valued deterministic closure, which may degrade the injectivity of the time-delay reconstruction and the downstream forecasting performance.

We use the Rössler system

$$\begin{aligned} \dot{u}_0 &= -u_1 - u_2 \\ \dot{u}_1 &= u_0 + 0.2u_1 \\ \dot{u}_2 &= 0.2 + u_2(u_0 - 5.7) \end{aligned}$$

to illustrate the DIM pipeline and to demonstrate how the quality of the time-delay reconstruction affects downstream forecasting. We compare two scalar observations, y_0 and y_2, for time-delay reconstruction. For each observation, we train a three-layer MLP with 256 hidden units and $\tanh(\cdot)$ activations using the DIM objective. The model is optimized with Adam for 3000 epochs using a learning rate of 10^{-3}. We fix the delay $\tau = 30$, the embedding dimension $m = 5$, and the pushforward horizon $q = 20$. We adopt the unbiased empirical estimate of the energy distance, which is connected to Maximum Mean Discrepancy and can be efficiently accelerated via GPU-based optimal transport solvers [18]. For validation, 30 initial conditions are randomly selected from the attractor, and rollout prediction errors are evaluated within the Lyapunov-time regime.

Figure 4 reports the cumulative MSE over the prediction horizon with 95% confidence intervals. The forecasting error based on the u_2 delay reconstruction grows faster than that based on the u_0 reconstruction. This indicates that the two scalar measurements induce delay-coordinate representations of unequal predictive quality. Detailed analysis and quantification of this non-injectivity effect are provided in [16]. This example highlights a practical limitation of DIM: although DIM provides a powerful geometry-preserving objective, its forecasting performance still depends on the informativeness of the available measurement function.

B Proofs of theoretical results

B.1 Proof of local contraction under partial observability

The following theorem describes the mechanism by which a one-step MSE-trained predictor can lead to local collapse to a fixed point under partial observability.

Theorem 5. *Consider a one-step autoregressive predictor Φ_θ trained via the mean squared error on a discrete, partially observed dynamical system. Let ν_{δ_0} denote the joint law of $(\boldsymbol{y}_t, \boldsymbol{y}_{t+1})$ conditioned on $\boldsymbol{y}_t \in \overline{B_{\delta_0}}(\boldsymbol{y}_0)$, and let $\bar{\boldsymbol{y}}_{\delta_0} := \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_t]$. Assume the input covariance $\Sigma_{in} := \mathrm{Cov}_{\nu_{\delta_0}}(\boldsymbol{y}_t)$ is positive definite. Define the output covariance matrix as $\Sigma_{out} := \mathrm{Cov}_{\nu_{\delta_0}}(\boldsymbol{y}_{t+1})$, $M_{out} = \Sigma_{in}^{-1/2}\Sigma_{out}\Sigma_{in}^{-1/2}$, and $M_{info} = \Sigma_{in}^{-1/2}\mathbb{E}_{\nu_{\delta_0}}[\mathrm{Cov}(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t)]\Sigma_{in}^{-1/2}$. Define $\kappa = \lambda_{\max}(M_{out})$ and $\eta = \lambda_{\min}(M_{info})$. Let $\Phi^\star(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t = \boldsymbol{y}]$ denote the population MSE minimizer on ν_{δ_0}. Assume that $\Phi^\star$ has Hessian norm bounded by L, and its Jacobian is L_Σ-Lipschitz. Define constants $a = dL\sqrt{C_\Sigma}$ and $b = dL^2 C_\Sigma$, where C_Σ is the condition number of Σ_{in}, and $\alpha = L\sqrt{C_\Sigma}$. Define*

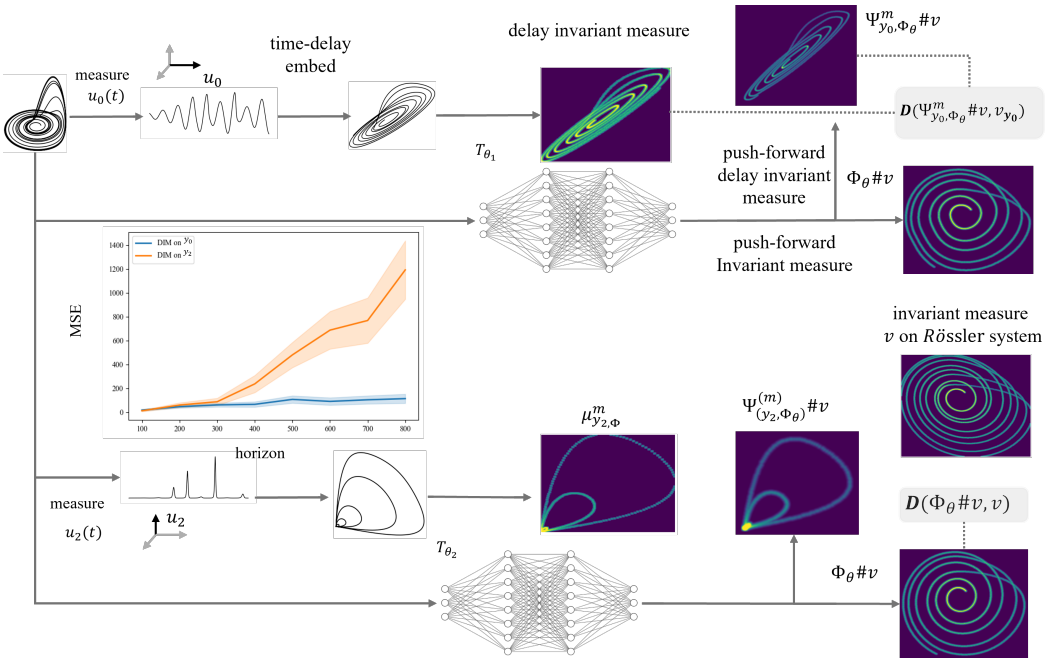

Figure 4: DIM on the Rössler system using scalar observations y_0 and y_2 for time-delay reconstruction. The figure reports cumulative rollout MSE with 95% confidence intervals, illustrating that the forecasting performance of DIM depends on the quality of the delay-coordinate reconstruction induced by the measurement function.

$s_{\max} := a\delta_0 + \sqrt{a^2\delta_0^2 + \kappa - \eta + b\delta_0^2}$ *and* $m_{drift} := \|\mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1}] - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}}$, *where* $\|\cdot\|_{\Sigma_{in}}$ *is the Mahalanobis norm. For vectors, define* $\|\boldsymbol{v}\|_{\Sigma_{in}} := \|\Sigma_{in}^{-1/2}\boldsymbol{v}\|_2$. *For matrices, use the induced norm* $\|A\|_{\Sigma_{in}} := \|\Sigma_{in}^{-1/2}A\Sigma_{in}^{1/2}\|_2$. *If the following conditions are satisfied:*

- $\kappa - \eta < 1 - 2a\delta_0 - b\delta_0^2, \quad a\delta_0 < 1/2,$
- *Suppose there exists* δ' *and a local sub-ball* $\overline{\mathcal{B}^{\Sigma}_{\delta'}(\bar{\boldsymbol{y}}_{\delta_0})}$, *such that there exist* C^0 *and* C^1 *error bounds* $\sup_{\boldsymbol{y}\in\overline{\mathcal{B}^{\Sigma}_{\delta'}(\bar{\boldsymbol{y}}_{\delta_0})}} \|\Phi_\theta(\boldsymbol{y}) - \Phi^\star(\boldsymbol{y})\|_{\Sigma_{in}} \le \varepsilon_0$ *and* $\sup_{\boldsymbol{y}\in\overline{\mathcal{B}^{\Sigma}_{\delta'}(\bar{\boldsymbol{y}}_{\delta_0})}} \|\nabla\Phi_\theta(\boldsymbol{y}) - \nabla\Phi^\star(\boldsymbol{y})\|_{\Sigma_{in}} \le \varepsilon_1$, *and* $q := s_{\max} + L_\Sigma\delta' + \varepsilon_1$ *satisfies* $q < 1$ *and* $m_{drift} + \sqrt{b}\delta_0 + \varepsilon_0 \le (1-q)\delta'$.

Then every trajectory initialized in $\overline{\mathcal{B}^{\Sigma}_{\delta'}(\bar{\boldsymbol{y}}_{\delta_0})}$ *remains in this sub-ball and converges to a unique fixed point.*

Proof. The population minimizer of the mean squared error risk is the conditional expectation:

$$\Phi^\star(\boldsymbol{y}_t) = \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t].$$

Therefore, by the law of total covariance under the conditional measure ν_{δ_0}, we have:

$$\mathrm{Cov}_{\nu_{\delta_0}}(\Phi^\star(\boldsymbol{y}_t)) = \Sigma_{out} - \mathbb{E}_{\nu_{\delta_0}}[\mathrm{Cov}(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t)] \succeq 0. \tag{9}$$

Define $\gamma_{\mathrm{dir}} := \lambda_{\max}(M_{\mathrm{out}} - M_{\mathrm{info}})$. By Weyl's inequality:

$$\gamma_{\mathrm{dir}} = \lambda_{\max}\Big(M_{\mathrm{out}} - M_{\mathrm{info}}\Big) \le \kappa - \eta. \tag{10}$$

Let $\boldsymbol{z}_t = \boldsymbol{y}_t - \bar{\boldsymbol{y}}_{\delta_0}$. Applying the Taylor expansion of $\Phi^\star$ around $\bar{\boldsymbol{y}}_{\delta_0}$ yields:

$$\Phi^\star(\boldsymbol{y}_t) = \Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) + J^\star\boldsymbol{z}_t + \boldsymbol{r}(\boldsymbol{z}_t). \tag{11}$$

Since both $\boldsymbol{y}_t$ and $\bar{\boldsymbol{y}}_{\delta_0}$ reside within $\overline{B_{\delta_0}}(\boldsymbol{y}_0)$, it follows that $\|\boldsymbol{z}_t\|_2 \le 2\delta_0$. The Hessian bound implies $\|\boldsymbol{r}(\boldsymbol{z}_t)\|_2 \le \frac{L}{2}\|\boldsymbol{z}_t\|_2^2 \le L\delta_0\|\boldsymbol{z}_t\|_2$.

Define $\tilde{\boldsymbol{z}}_t = \Sigma_{in}^{-1/2}\boldsymbol{z}_t$, $B^\star = \Sigma_{in}^{-1/2}J^\star\Sigma_{in}^{1/2}$, $s := \|B^\star\|_2 = \|J^\star\|_{\Sigma_{in}}$. Also let $\tilde{\boldsymbol{r}} = \Sigma_{in}^{-1/2}\boldsymbol{r}(\boldsymbol{z}_t)$. Then $\tilde{\boldsymbol{r}}$ is bounded by:

$$\begin{aligned}\|\tilde{\boldsymbol{r}}\|_2 = \|\Sigma_{in}^{-1/2}\boldsymbol{r}(\boldsymbol{z}_t)\|_2 \le \|\Sigma_{in}^{-1/2}\|_2\|\boldsymbol{r}(\boldsymbol{z}_t)\|_2 &\le L\delta_0\|\Sigma_{in}^{-1/2}\|_2\|\Sigma_{in}^{1/2}\|_2\|\tilde{z}_t\|_2 \\ &\le (L\sqrt{C_\Sigma})\delta_0\|\tilde{\boldsymbol{z}}_t\|_2\end{aligned} \tag{12}$$

Since $\mathrm{Cov}_{\nu_{\delta_0}}(\tilde{\boldsymbol{z}}_t) = I_d$, we have $\mathbb{E}_{\nu_{\delta_0}}\|\tilde{\boldsymbol{z}}_t\|_2^2 = d$, and consequently, the spectral norms of the cross-covariance and auto-covariance of the remainder are bounded by:

$$\begin{aligned}\|\,\mathrm{Cov}_{\nu_{\delta_0}}(B^\star\tilde{\boldsymbol{z}}_t, \tilde{\boldsymbol{r}})\|_2 = \|\mathbb{E}_{\nu_{\delta_0}}[B^\star\tilde{\boldsymbol{z}}_t\tilde{\boldsymbol{r}}^\top]\|_2 \le s\,\mathbb{E}_{\nu_{\delta_0}}[\|\tilde{\boldsymbol{z}}_t\|_2\|\tilde{\boldsymbol{r}}\|_2] \le (d\alpha)\delta_0 s = a\delta_0 s, \\ \|\,\mathrm{Cov}_{\nu_{\delta_0}}(\tilde{\boldsymbol{r}})\|_2 \le \mathbb{E}_{\nu_{\delta_0}}[\|\tilde{\boldsymbol{r}}\|_2^2] \le (d\alpha^2)\delta_0^2 = b\delta_0^2.\end{aligned}$$

By the triangle inequality, the norm of the composite remainder covariance matrix $\tilde{R} := \mathrm{Cov}_{\nu_{\delta_0}}(B^\star\tilde{\boldsymbol{z}}_t, \tilde{\boldsymbol{r}}) + \mathrm{Cov}_{\nu_{\delta_0}}(\tilde{\boldsymbol{r}}, B^\star\tilde{\boldsymbol{z}}_t) + \mathrm{Cov}_{\nu_{\delta_0}}(\tilde{\boldsymbol{r}})$ is bounded by $\|\tilde{R}\|_2 \le 2a\delta_0 s + b\delta_0^2$, and the norm of the expected remainder is bounded by $\|\mathbb{E}_{\nu_{\delta_0}}[\tilde{\boldsymbol{r}}]\|_2 \le \sqrt{\mathbb{E}_{\nu_{\delta_0}}[\|\tilde{\boldsymbol{r}}\|_2^2]} \le \sqrt{b}\delta_0$ via Jensen's inequality.

Since

$$M_{\text{out}} - M_{\text{info}} = \mathrm{Cov}_{\nu_{\delta_0}}(\Sigma_{in}^{-1/2}\Phi^\star(\boldsymbol{y}_t)) = B^\star(B^\star)^\top + \tilde{R}$$

where $\|\tilde{R}\|_2 \le 2a\delta_0 s + b\delta_0^2$, we have:

$$s^2 = \|B^\star(B^\star)^\top\| \le \|M_{\text{out}} - M_{\text{info}}\|_2 + \|\tilde{R}\|_2 \le \gamma_{\text{dir}} + 2a\delta_0 s + b\delta_0^2. \tag{13}$$

Solving this quadratic inequality for s yields an upper bound for the positive root:

$$s \le a\delta_0 + \sqrt{a^2\delta_0^2 + \gamma_{\text{dir}} + b\delta_0^2} \le a\delta_0 + \sqrt{a^2\delta_0^2 + \kappa - \eta + b\delta_0^2} = s_{\max}. \tag{14}$$

Substituting $\gamma_{\text{dir}} \le \kappa - \eta$ and applying $\kappa - \eta + b\delta_0^2 < 1 - 2a\delta_0$, the term inside the square root is strictly less than $(1 - a\delta_0)^2$. Thus, $s_{\max} < 1$.

Under the second condition, there exists a sub-ball $\overline{\mathcal{B}_{\delta'}^{\Sigma}(\bar{\boldsymbol{y}}_{\delta_0})}$, such that the upper bound of the local Jacobian of the MSE-based predictor Φ_θ satisfies the condition $\sup_{\boldsymbol{y}\in\overline{\mathcal{B}_{\delta'}^{\Sigma}(\bar{\boldsymbol{y}}_{\delta_0})}}\|\nabla\Phi_\theta(\boldsymbol{y})\|_{\Sigma_{in}} \le s_{\max} + L_\Sigma\delta' + \varepsilon_1 = q < 1$. By the law of total expectation over ν_{δ_0}, we get:

$$\mathbb{E}_{\nu_{\delta_0}}[\Phi^\star(\boldsymbol{y}_t)] = \mathbb{E}_{\nu_{\delta_0}}[\mathbb{E}[\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t]] = \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1}], \tag{15}$$

then, taking expectations over Eq. (11) yields:

$$\mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1}] = \Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) + J^\star\mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{z}_t] + \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{r}(\boldsymbol{z}_t)] = \Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) + \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{r}(\boldsymbol{z}_t)], \tag{16}$$

and rearranging the term provides:

$$\|\Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) - \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1}]\|_{\Sigma_{in}} = \|\mathbb{E}_{\nu_{\delta_0}}[\tilde{\boldsymbol{r}}]\|_2 \le \sqrt{b}\delta_0. \tag{17}$$

Applying the triangle inequality, we get:

$$\begin{aligned}\|\Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} &\le \|\mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1}] - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} + \|\Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) - \mathbb{E}_{\nu_{\delta_0}}[\boldsymbol{y}_{t+1}]\|_{\Sigma_{in}} \\ &\le m_{\text{drift}} + \sqrt{b}\delta_0.\end{aligned} \tag{18}$$

Using the C^0 approximation bound, we obtain:

$$\begin{aligned}\|\Phi_\theta(\bar{\boldsymbol{y}}_{\delta_0}) - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} &\le \|\Phi_\theta(\bar{\boldsymbol{y}}_{\delta_0}) - \Phi^\star(\bar{\boldsymbol{y}}_{\delta_0})\|_{\Sigma_{in}} + \|\Phi^\star(\bar{\boldsymbol{y}}_{\delta_0}) - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} \\ &\le \varepsilon_0 + m_{\text{drift}} + \sqrt{b}\delta_0 \le (1-q)\delta'.\end{aligned} \tag{19}$$

Applying the mean value theorem, for any $\boldsymbol{y} \in \overline{\mathcal{B}_{\delta'}^{\Sigma}(\bar{\boldsymbol{y}}_{\delta_0})}$:

$$\begin{aligned}\|\Phi_\theta(\boldsymbol{y}) - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} &\le \|\Phi_\theta(\boldsymbol{y}) - \Phi_\theta(\bar{\boldsymbol{y}}_{\delta_0})\|_{\Sigma_{in}} + \|\Phi_\theta(\bar{\boldsymbol{y}}_{\delta_0}) - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} \\ &\le q\|\boldsymbol{y} - \bar{\boldsymbol{y}}_{\delta_0}\|_{\Sigma_{in}} + (1-q)\delta' \le \delta'.\end{aligned} \tag{20}$$

Let $D := \overline{\mathcal{B}_{\delta'}^{\Sigma}(\bar{\boldsymbol{y}}_{\delta_0})}$. Since $\sup_{\boldsymbol{y}\in D}\|\nabla\Phi_\theta(\boldsymbol{y})\|_{\Sigma_{in}} \le q < 1$, then Φ_θ is a contraction on D. Since D is complete, the Banach fixed-point theorem yields a unique fixed point in D, and every trajectory initialized in D remains in D and converges to that fixed point. □

Proposition 6 (Graph Measure Identification)**.** *If $\pi_\theta = \pi_\star$, then $\bar{\nu}_\theta = \bar{\nu}_{\mathrm{rec}}$ and $Q_\theta(\cdot|r) = Q_\star(\cdot|r)$ for $\bar{\nu}_{\mathrm{rec}}$-a.e. r.*

Proof. Taking the first marginal of $\pi_\theta = \pi_\star$ implies $\bar{\nu}_\theta = \bar{\nu}_{\mathrm{rec}}$. Thus, given the first coordinate of $\pi_\star$, the uniqueness of regular conditional distributions yields $Q_\theta(\cdot \mid r) = Q_\star(\cdot \mid r)$ for $\bar{\nu}_{\mathrm{rec}}$-a.e. r. □

Proposition 7. *Let $\mathcal{R} \subseteq \mathbb{R}^d$ be closed, and let r_i, $i \geq 0$ be a stationary and ergodic process with stationary marginal $\bar{\nu}_{\mathrm{rec}} = \mathcal{L}(r_i)$. Define the empirical estimator of $\pi_\star$ by $\widehat{\pi}_N = \frac{1}{N}\sum_{i=0}^{N-1} \delta_{(r_i, r_{i+1})}$. Fix $p \geq 1$, and equip $\mathcal{R} \times \mathcal{R}$ with the product metric $d_{\times,p}((r_1, r_1'), (r_2, r_2')) = (\|r_1 - r_2\|^p + \|r_1' - r_2'\|^p)^{1/p}$. If $\int_{\mathcal{R}} \|r\|^p d\bar{\nu}_{\mathrm{rec}}(r) < \infty$, then $W_p(\widehat{\pi}_N, \pi_\star) \to 0$ almost surely as $N \to \infty$.*

Proof. Let $z_i = (r_i, r_{i+1})$. Since $\{r_i\}$ is stationary and ergodic, the pair process $\{z_i\}$ is also stationary and ergodic with law $\pi_\star$. For every bounded continuous f, $\int f \, d\widehat{\pi}_N = \frac{1}{N}\sum_{i=0}^{N-1} f(z_i) \to \mathbb{E}[f(z_0)] = \int f \, d\pi_\star$ almost surely by Birkhoff's ergodic theorem. Since $\mathcal{R} \times \mathcal{R}$ is Polish, a standard countable convergence-determining-class argument gives $\widehat{\pi}_N \Rightarrow \pi_\star$ almost surely. For $z = (r, r')$, define $\|z\|_{\times,p}^p = \|r\|^p + \|r'\|^p$. Since both marginals of $\pi_\star$ equal $\bar{\nu}_{\mathrm{rec}}$,

$$\int_{\mathcal{R}\times\mathcal{R}} (\|r\|^p + \|r'\|^p) \, d\pi_\star(r, r') = 2 \int_{\mathcal{R}} \|r\|^p \, d\bar{\nu}_{\mathrm{rec}}(r) < \infty. \tag{21}$$

Moreover, $\int \|z\|_{\times,p}^p \, d\widehat{\pi}_N(z) = \frac{1}{N}\sum_{i=0}^{N-1} \big(\|r_i\|^p + \|r_{i+1}\|^p\big)$. Applying Birkhoff's theorem to $r \mapsto \|r\|^p$ gives $\int \|z\|_{\times,p}^p d\widehat{\pi}_N(z) \to \int \|z\|_{\times,p}^p d\pi_\star(z)$ almost surely. Weak convergence together with convergence of the p-th moments therefore implies $W_p(\widehat{\pi}_N, \pi_\star) \to 0$ almost surely. □

Corollary 8 (Collapsed-Support Separation Bound)**.** *Under the assumptions of Proposition 3, let $\widehat{\pi}_{\theta,N} = \frac{1}{N}\sum_{i=0}^{N-1} \delta_{(\boldsymbol{r}_i^\theta, \boldsymbol{r}_{i+1}^\theta)}$ denote the empirical recurrence transition measure generated by the model. Let $C \subseteq \mathcal{R}$ be nonempty and closed, and suppose that $\boldsymbol{r}_{i+1}^\theta \in C$ for all $i \geq 0$. Defining $d(\boldsymbol{r}, C) = \inf_{\boldsymbol{c}\in C} \|\boldsymbol{r} - \boldsymbol{c}\|$, we have*

$$W_p^p(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N) \geq \frac{1}{N}\sum_{i=0}^{N-1} d(\boldsymbol{r}_{i+1}, C)^p. \tag{22}$$

Consequently, $\liminf_{N\to\infty} W_p^p(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N) \geq \int_{\mathcal{R}} d(\boldsymbol{r}, C)^p \, d\bar{\nu}_{\mathrm{rec}}(\boldsymbol{r})$ *almost surely. If $\bar{\nu}_{\mathrm{rec}}(C^c) > 0$, the right-hand side is strictly positive.*

Proof. Let $\gamma \in \Gamma\big(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N\big)$ be an arbitrary coupling and denote $z^\theta = (r^\theta, r^{\theta\prime})$. Since the first marginal of γ is $\widehat{\pi}_{\theta,N}$ and $r_{i+1}^\theta \in C$ for every i, thus $r^{\theta\prime} \in C$ for γ-almost every (z^θ, z) and $d_{\times,p}(z^\theta, z)^p = \|r^\theta - r\|^p + \|r^{\theta\prime} - r'\|^p \geq \|r^{\theta\prime} - r'\|^p \geq d(r', C)^p$. Integrating with respect to γ yields

$$\begin{aligned} \int d_{\times,p}(z^\theta, z)^p \, d\gamma(z^\theta, z) &\geq \int d(r', C)^p \, d\gamma(z^\theta, z) \\ &= \int_{\mathcal{R}\times\mathcal{R}} d(r', C)^p \, d\widehat{\pi}_N(r, r') \\ &= \frac{1}{N}\sum_{i=0}^{N-1} d(r_{i+1}, C)^p. \end{aligned} \tag{23}$$

Taking the infimum over all such couplings, we have

$$W_p^p(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N) \geq \frac{1}{N}\sum_{i=0}^{N-1} d(r_{i+1}, C)^p. \tag{24}$$

For any $c_0 \in C$, then $d(r, C)^p \leq \|r - c_0\|^p \leq 2^{p-1}\big(\|r\|^p + \|c_0\|^p\big)$. Thus, under the assumptions, the function $r \mapsto d(r, C)^p$ is integrable with respect to $\bar{\nu}_{\mathrm{rec}}$. Since $\{r_{i+1}\}_{i\geq 0}$ is stationary and ergodic, Birkhoff's ergodic theorem gives

$$\frac{1}{N}\sum_{i=0}^{N-1} d(r_{i+1}, C)^p \longrightarrow \int_{\mathcal{R}} d(r, C)^p \, d\bar{\nu}_{\mathrm{rec}}(r) \tag{25}$$

almost surely. Combining this with the finite-sample bound gives

$$\liminf_{N\to\infty} W_p^p(\widehat{\pi}_{\theta,N}, \widehat{\pi}_N) \geq \int_{\mathcal{R}} d(r, C)^p \, d\bar{\nu}_{\mathrm{rec}}(r) \tag{26}$$

almost surely. Since C is closed, $d(r, C) = 0 \Leftrightarrow r \in C$. If $\bar{\nu}_{\mathrm{rec}}(C^c) > 0$, then $d(r, C)^p > 0$ on a positive $\bar{\nu}_{\mathrm{rec}}$. Therefore,

$$\int_{\mathcal{R}} d(r, C)^p \, d\bar{\nu}_{\mathrm{rec}}(r) > 0. \tag{27}$$

□

C MSE-induced topological collapse and heterogeneous baselines

C.1 Visualizations of MSE-induced topological collapse

We visualize the collapse behavior of autonomous predictors trained with pointwise MSE-based objectives. These examples are intended to illustrate the failure modes discussed in Section 4: under sparse partial observations, an MSE-trained deterministic predictor can generate rollouts that contract toward spurious fixed points, limit cycles, or unstable divergent trajectories.

Lorenz63

We first consider the partially observed Lorenz63 system. Figure 5 shows 30 autonomous rollouts generated by a one-step MSE-trained predictor from distinct initial conditions. Among these rollouts, 26 collapse to a fixed point, while the remaining four converge to spurious limit cycles. Figure 6 shows the corresponding behavior for a predictor trained with a 20-step MSE objective. In this case, 6 out of 30 rollouts collapse to a fixed point, 23 converge to spurious limit cycles, and one rollout diverges as the generated trajectory leaves the learned basin of attraction. These results show that the multi-step MSE training strategy does not eliminate the collapse phenomenon; rather, it changes the dominant failure mode from fixed-point collapse to spurious limit cycle behavior.

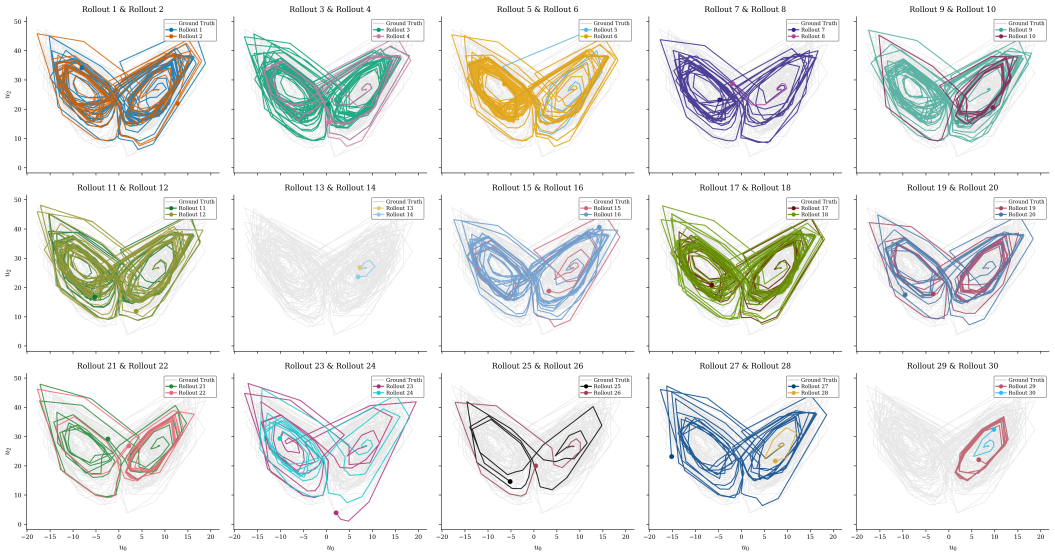

Figure 5: Autonomous rollouts of a one-step MSE-trained predictor from 30 distinct initial conditions in the partially observed Lorenz63 system. Most trajectories collapse to a spurious fixed point, while a small number converge to artificial limit cycles.

Kuramoto-Sivashinsky

We also visualize this phenomenon on the partially observed Kuramoto-Sivashinsky system. Figures 7 and 8 show autonomous rollouts from one-step and 20-step MSE-trained predictors, respectively.

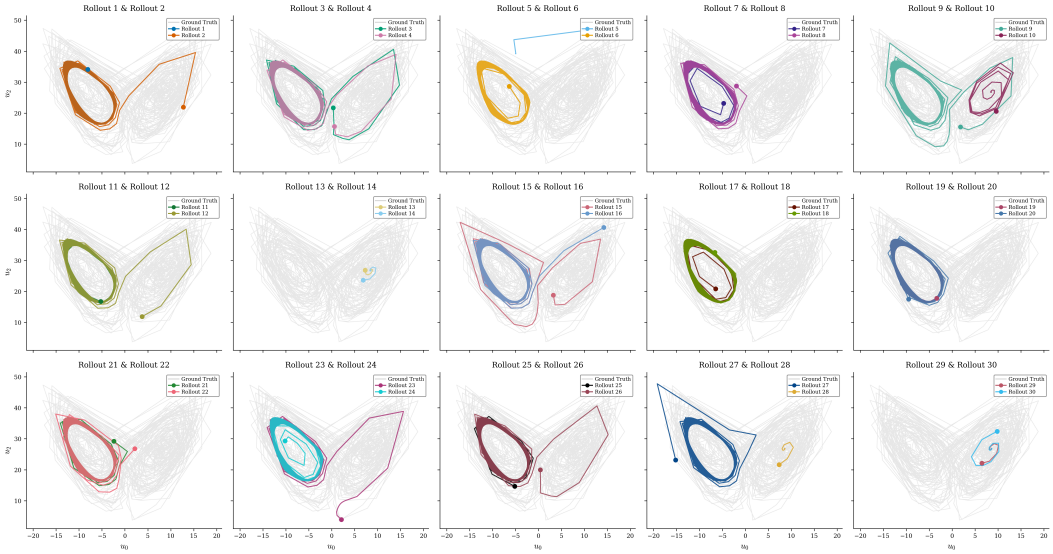

Figure 6: Autonomous rollouts of a 20-step MSE-trained predictor from 30 distinct initial conditions in the partially observed Lorenz63 system. Compared with the one-step model, the multi-step objective reduces fixed-point collapse but produces many spurious limit-cycle rollouts and one divergent trajectory.

For the one-step model, 5 out of 30 rollouts collapse to spurious limit cycles and 14 diverge, while the remaining rollouts preserve chaotic behavior over the plotted horizon. While the 20-step model alleviates model divergence, the generated trajectories collapse to spurious low-dimensional attractors. This behavior is consistent with the Lorenz63 results: although multi-step MSE training may stabilize short-horizon optimization, it does not resolve the covariance-contraction mechanism induced by pointwise prediction under partial observability.

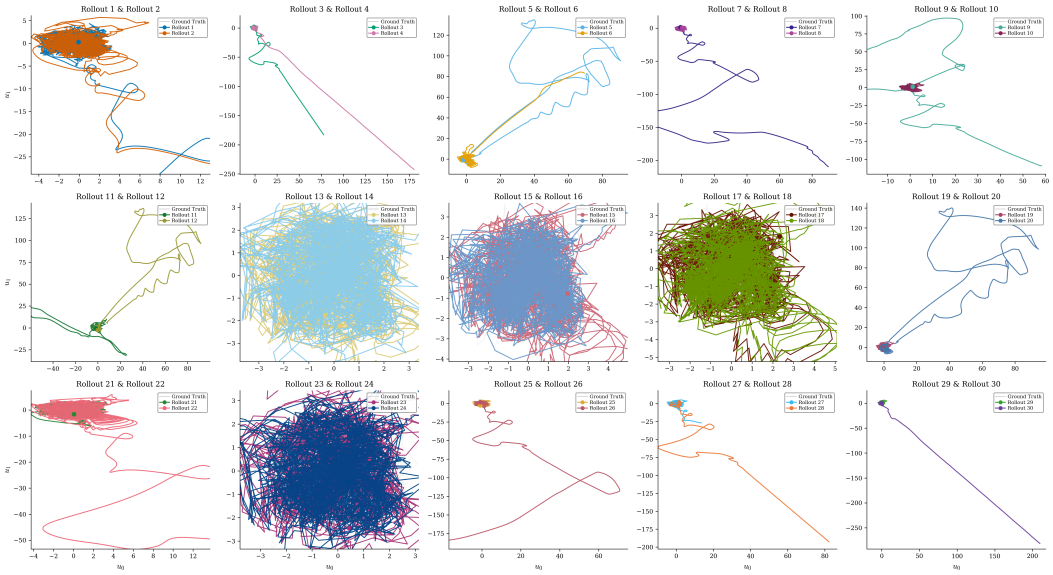

Figure 7: Autonomous rollouts of a one-step MSE-trained predictor from 30 distinct initial conditions in the partially observed Kuramoto-Sivashinsky system. The rollouts exhibit mixed failure modes, including divergence and spurious recurrent behavior.

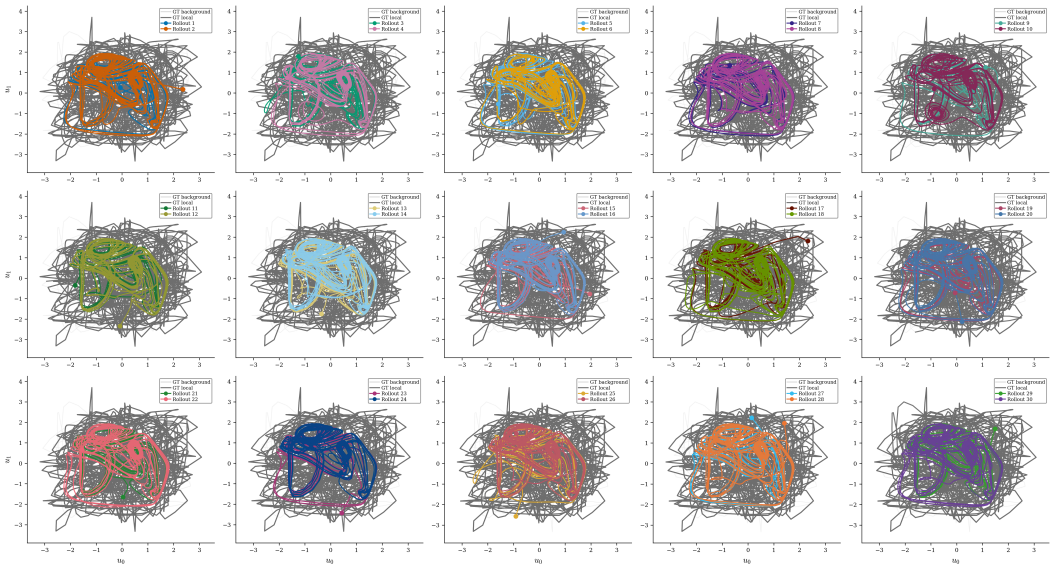

Figure 8: Autonomous rollouts of a 20-step MSE-trained predictor from 30 distinct initial conditions in the partially observed Kuramoto-Sivashinsky system. The multi-step objective alleviates the divergence but drives the generated trajectories toward spurious low-dimensional attractors.

C.2 Heterogeneous forecasting baselines

The main experiments in the paper are objective-controlled comparisons: all methods share the same autonomous predictor backbone and differ only in the training objective. To provide additional context, we further compare PRIME with several heterogeneous forecasting baselines under the same 2D partially observed Lorenz63 protocol. These baselines should not be interpreted as objective-controlled comparisons, since they differ not only in the training objective but also in the forecasting architecture and the information available to the predictor. In particular, NGRC-NVAR, Reservoir-ESN, LSTM, and Koopman-AE use history windows during autonomous forecasting to mitigate the ambiguity induced by non-Markovian partial observations. In contrast, PRIME serves as an autonomous predictor using only a three-layer MLP in the observed space for rollout generation. For PRIME, the recurrence-level and delay information are exclusively exploited through the training objective, rather than being fed as an additional history-window input during autonomous prediction. Table 4 therefore serves as a heterogeneous reference rather than a strictly fair comparison. Figure 9 shows autonomous rollouts from a common initial state and for history-based baselines, the required history window is taken from the same observed trajectory and ends at the same forecast origin; all plotted predictions start from the same observed initial point. The results show that history-window architectures such as LSTM and Koopman-AE achieve strong NMSE and attractor-overlap scores in this setting. However, PRIME remains competitive despite using only the current observation as the predictor state during rollout. In particular, PRIME achieves strong spectral fidelity and recurrence-level temporal consistency, while avoiding the spurious collapse observed in the DySLIM-style invariant-measure baseline. This supports the main claim that recurrence-level temporal alignment provides a useful alternative to injecting temporal information through history-window forecasting architectures.

D Algorithm for data-driven recurrence section construction

The data-driven recurrence section is designed to extract stable recurrence events from finite, partially observed trajectories, rather than to recover an exact global Poincaré section. The construction uses time-delay augmentation to make the recurrence coordinate more informative, speed filtering to avoid nearly stationary or degenerate events, and farthest point sampling to distribute local recurrence neighborhoods across the observed attractor. The entry and exit thresholds are chosen adaptively from

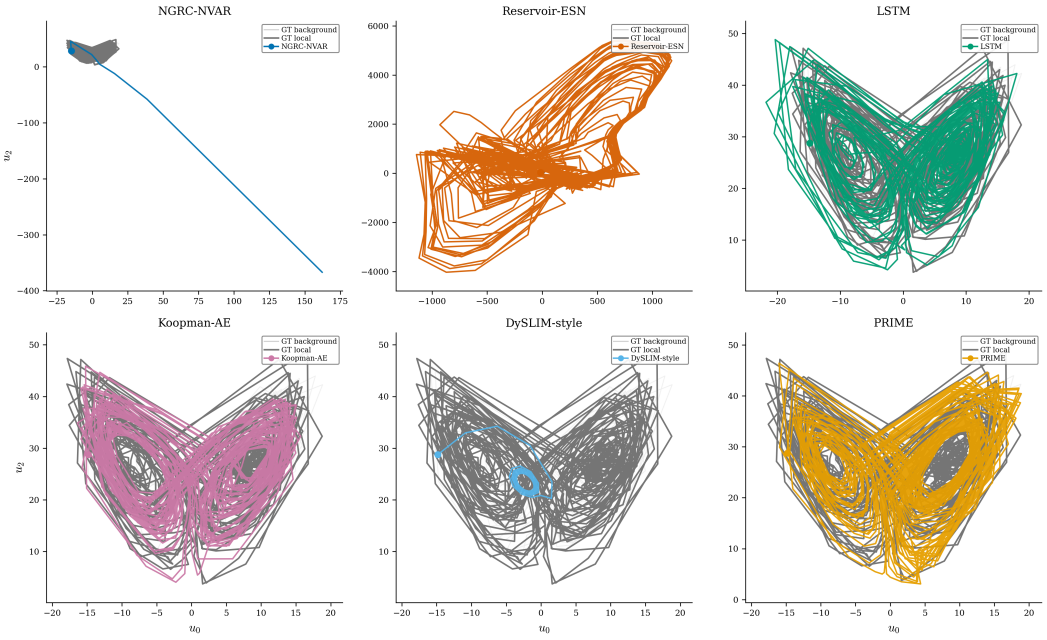

Figure 9: Autonomous rollouts of all trained models from a common forecast origin. For history-window baselines, the input history starts from the same observed trajectory and ends at the same forecast origin; all plotted predictions start from the same observed initial point.

landmark distances and incorporate hysteresis, which prevents repeated detections caused by grazing or short-time oscillations near the same neighborhood. The same recurrence section parameters are then fixed and reused during autonomous evaluation to extract recurrence events from generated trajectories. The full algorithm for constructing the data-driven recurrence section is summarized in Algorithm 2.

E Experimental setup and evaluation criteria

E.1 Experimental setup

Lorenz63

The governing equations of the Lorenz63 system are

$$\begin{aligned}\frac{du_0}{dt} &= \sigma(u_1 - u_0),\\ \frac{du_1}{dt} &= u_0(\rho - u_2) - u_1,\\ \frac{du_2}{dt} &= u_0 u_1 - \beta u_2,\end{aligned}$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. We integrate the system using the Euler method with an integration step of $\Delta t = 0.01$, and downsample by a factor of 10 to yield an observation interval of $\Delta t_{\text{obs}} = 0.1$. The observation function is $h(\boldsymbol{u}) = (u_0, u_2)$. Since the Lorenz63 system has the C_2 symmetry

$$(u_0, u_1, u_2) \mapsto (-u_0, -u_1, u_2),$$

a delay reconstruction based only on the u_2 coordinate cannot distinguish the two symmetric lobes. Therefore, we only use the u_0 coordinate as the delay channel for constructing recurrence events, with a delay lag $\tau = 1$ and an embedding dimension of $m = 5$.

Algorithm 2 Data-driven recurrence section construction.

Input. Partial observations $\{\boldsymbol{y}_t\}_{t=1}^{T} \subseteq \mathbb{R}^{\ell}$, m, τ, speed quantile q_{speed}, candidate fraction ρ_{cand}, number of landmarks M, hyperparameters $(\mu_{\text{in}}, \mu_{\text{out}})$, and minimum inter-event gap L_{sec}.
(i) Set $t_{\min} = (m-1)\tau$. For each $t \geq t_{\min}$, construct the delay coordinate

$$\boldsymbol{z}_t = \big(\boldsymbol{y}_{t-\tau}, \boldsymbol{y}_{t-2\tau}, \ldots, \boldsymbol{y}_{t-(m-1)\tau}\big),$$

and form the augmented recurrence coordinate

$$\boldsymbol{r}_t = [\boldsymbol{y}_t, \boldsymbol{z}_t].$$

(ii) Compute the local speed in the scaled observation space,

$$v_t = \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|_2,$$

and define the active index pool

$$\mathcal{I}_{\text{active}} = \{t \geq t_{\min} : v_t \geq \text{Quantile}(\{v_s : s \geq t_{\min}\}, q_{\text{speed}})\}.$$

Uniformly subsample $\mathcal{I}_{\text{active}}$ with fraction ρ_{cand} to obtain the candidate $\mathcal{I}_{\text{cand}}$.
(iii) Standardize the augmented recurrence coordinates:

$$\widetilde{\boldsymbol{r}}_t = \frac{\boldsymbol{r}_t - \boldsymbol{m}_{\text{cand}}}{\boldsymbol{s}_{\text{cand}}},$$

where $(\boldsymbol{m}_{\text{cand}}, \boldsymbol{s}_{\text{cand}})$ are computed from candidate coordinates $\{\boldsymbol{r}_t : t \in \mathcal{I}_{\text{cand}}\}$.
(iv) Apply Farthest Point Sampling to the standardized candidate set

$$\{\widetilde{\boldsymbol{r}}_t : t \in \mathcal{I}_{\text{cand}}\}$$

and obtain M landmarks

$$\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}.$$

(v) Let d_{med} be the median nearest-neighbor distance among landmarks in the standardized recurrence-coordinate space. Set the thresholds

$$\varepsilon_{\text{in}} = \mu_{\text{in}} d_{\text{med}}, \qquad \varepsilon_{\text{out}} = \mu_{\text{out}} \varepsilon_{\text{in}}.$$

(vi) Extract recurrence events by scanning the standardized trajectory $\{\widetilde{\boldsymbol{r}}_t\}_{t=t_{\min}}^{T}$. A time t is recorded as a recurrence event if

$$\min_{\boldsymbol{c} \in \mathcal{C}} \|\widetilde{\boldsymbol{r}}_t - \boldsymbol{c}\|_2 \leq \varepsilon_{\text{in}}$$

and $t - t_{\text{last}} \geq L_{\text{sec}}$. After recording an event, the detector remains inactive until the trajectory exits the corresponding landmark neighborhood, i.e.,

$$\|\widetilde{\boldsymbol{r}}_t - \boldsymbol{c}_{\text{active}}\|_2 \geq \varepsilon_{\text{out}}.$$

Output. Recurrence times $\{t_i\}_{i=1}^{N}$, return coordinates $\{\boldsymbol{r}_{t_i}\}_{i=1}^{N}$, and recurrence-section parameters

$$\mathcal{S}_{\text{rec}} = \big(\mathcal{C}, \boldsymbol{m}_{\text{cand}}, \boldsymbol{s}_{\text{cand}}, \varepsilon_{\text{in}}, \varepsilon_{\text{out}}, L_{\text{sec}}\big).$$

For the data-driven recurrence section, we use $M = 200$ landmarks, speed quantile $q_{\text{speed}} = 0.3$, candidate fraction $\rho_{\text{cand}} = 0.5$, a minimum inter-event gap of $L_{\text{sec}} = 5$, and the entry and exit multipliers are set to $\mu_{\text{in}} = 0.35$ and $\mu_{\text{out}} = 1.8$, respectively.

For training, we use a three-layer MLP backbone with a width of 100. The total number of training iterations is 4000. The first 3000 iterations are used as a warm-up stage to optimize the training objective $\mathcal{L}^{\text{DIM}}$. The pushforward lag for the invariant measure is set to $q = 2$, and the stage-1 batch size is 512. Both stage-1 and stage-2 use the Adam optimizer with a learning rate of 10^{-3}. After the warm-up stage, we alternate one stage-2 update and one stage-1 update. In the second stage, we optimize the joint objective

$$\mathcal{L}^{\text{DIM}} + w_2 \mathcal{L}^{\text{PRIME}}$$

with $w_2 = 0.5$. For the PRIME recurrence block loss, we use recurrence blocks of length $L = 32$ and sample 16 recurrence blocks per stage-2 update. The state and delay components are weighted

by $w_{\text{state}} = 1.0$ and $w_{\text{delay}} = 0.7$. The source and target marginal weights are set to $w_0 = 0.20$, and $w_1 = 0.20$. In this work, all distributional discrepancies are computed using the Sinkhorn divergence with a squared Euclidean cost. The Sinkhorn blur parameter is linearly annealed from 0.2 to 0.1 during training.

Kuramoto-Sivashinsky system

We next evaluate our method on the one-dimensional Kuramoto-Sivashinsky (KS) equation on a periodic domain:

$$\partial_t u = -u\,\partial_x u - \theta\,\partial_{xx} u - \theta\,\partial_{xxxx} u, \qquad x \in [0, L],$$

where $L = 100$ and $\theta = 1$. The spatial domain is discretized using $N = 200$ uniformly spaced grid points. The initial condition is

$$u(x, 0) = \sin\left(\frac{2\pi x}{L}\right) + 10^{-2}\xi(x),$$

where $\xi(x)$ denotes an i.i.d. Gaussian perturbation. We simulate the system using a pseudo-spectral exponential time-stepping scheme with an integration step of $\Delta t = 0.1$ and apply the standard $2/3$ de-aliasing rule to the nonlinear term. The simulated trajectory is downsampled by a factor of 10, yielding an observation interval of $\Delta t_{\text{obs}} = 1.0$. The partially observed trajectory is constructed by measuring the solution at five fixed spatial grid indices: $\{0, 50, 80, 100, 150\}$. Thus, each observation satisfies $\boldsymbol{y}_t \in \mathbb{R}^5$. We use the downsampled sensor trajectory for training and evaluation. For the noisy training observations, we add Gaussian noise $\mathcal{N} \sim (0, 0.25)$ to each sensor channel.

To construct the data-driven recurrence section, we employ a multivariate time-delay map defined as:

$$\Psi^3_{\boldsymbol{y},\Phi}(\boldsymbol{u}) = \left(\Psi^3_{y_1,\Phi}(\boldsymbol{u})^\top, \Psi^3_{y_2,\Phi}(\boldsymbol{u})^\top, \ldots, \Psi^3_{y_5,\Phi}(\boldsymbol{u})^\top\right)^\top \in \mathbb{R}^{15},$$

by concatenating the scalar time-delay reconstructions from all five sensor channels, where the delay lag is set to $\tau = 10$ and the embedding dimension is set to $m = 3$. We use $M = 128$ landmarks, speed quantile $q_{\text{speed}} = 0.3$, candidate fraction $\rho_{\text{cand}} = 0.2$, and minimum inter-event gap $L_{\text{sec}} = 5$, with an entry multiplier of $\mu_{\text{in}} = 0.35$ and exit multiplier of $\mu_{\text{out}} = 1.8$.

For model training, we employ a three-layer MLP backbone with a width of 100. The total number of training iterations is 12000. The first 10000 iterations are used as a warm-up stage to optimize $\mathcal{L}^{\text{DIM}}$ until convergence. The pushforward horizon for the invariant measure matching is set to $q = 2$, and the stage-1 batch size is 512. Both stage-1 and stage-2 use the Adam optimizer with a learning rate 10^{-3}. After the warm-up stage, we apply only stage-2 updates, optimizing the joint objective:

$$\mathcal{L}^{\text{DIM}} + w_2 \mathcal{L}^{\text{PRIME}},$$

with $w_2 = 3$. For $\mathcal{L}^{\text{PRIME}}$, we use recurrence blocks of length $L = 16$ and sample 16 recurrence blocks per update. The state and delay components in the recurrence coordinate are weighted by $w_{\text{state}} = 1.0$ and $w_{\text{delay}} = 0.7$. The source and target marginal weights are set to $w_0 = 0.20$, and $w_1 = 0.20$.

Flow past a cylinder

The flow past a cylinder dataset is generated using a two-dimensional D2Q9 lattice-Boltzmann solver at Reynolds number $Re = 70$. The computational grid has a size 150×50. The cylinder is centered at grid index $(30, 25)$ with a radius of 5 grid cells, and the locations of the 9 probes are shown in Figure 3. The maximum horizontal inflow velocity is set to 0.04, and the simulation is run for 200000 LBM iterations. We discard the first 3000 transient states to ensure the flow is fully developed. For training, we select the next 200 observations, where each probe location is perturbed independently by a half-grid-cell Gaussian noise while a clean dataset is used for testing.

We apply a multivariate time-delay map across all probe channels for the time-delay reconstruction. With a delay lag of $\tau = 2$ and an embedding dimension $m = 4$, the delay coordinate is given by:

$$\Psi^4_{\boldsymbol{y},\Phi}(\boldsymbol{u}) = \left(\Psi^4_{y_1,\Phi}(\boldsymbol{u})^\top, \Psi^4_{y_2,\Phi}(\boldsymbol{u})^\top, \ldots, \Psi^4_{y_9,\Phi}(\boldsymbol{u})^\top\right)^\top \in \mathbb{R}^{36},$$

Since the attractor manifold in this case is not highly chaotic and is nearly a limit cycle, we select the 9-th channel as the recurrence band and define the recurrence event using the following formulation:

$$|\tilde{y}^{(9)}_t| \le 1, \qquad \tilde{y}^{(9)}_{t+1} - \tilde{y}^{(9)}_t > 0,$$

where $\tilde{y}_t^{(9)}$ denotes the measurement of the ninth channel after standardization. We set the minimum inter-event gap to $L_{\mathrm{sec}} = 3$ with an upward crossing through a probe-based band to avoid repeated counting.

For model training, we use the same three-layer MLP backbone with a width of 100, utilizing a 20-step training objective, and the pushforward horizon is set to $q = 2$. The batch size for Stage-1 training is 512, while the loss weights for $\mathcal{L}^{\mathrm{DIM}}$ are set to $w_{\mathrm{state}} = 1.0$ and $w_{\mathrm{delay}} = 0.7$. For the PRIME training objective, we use recurrence blocks of length $L = 16$ and sample 8 recurrence blocks for each Stage-2 update, setting $w_0 = 0.2$ and $w_1 = 0.2$. Moreover, since the dimensionality of the delay-coordinate system is quite high, we project the recurrence coordinate to 16 dimensions, and the pair of recurrence events is projected to 32 dimensions. We use the Adam optimizer for both stages, setting the learning rates to 10^{-3} for Stage-1 and 10^{-5} for Stage-2. The warm-up stage for optimizing $\mathcal{L}^{\mathrm{DIM}}$ is set to 3000 iterations, after which we alternate one stage-2 update with three stage-1 updates.

E.2 Evaluation criteria

Normalized mean squared error

The normalized mean squared error (NMSE) measures pointwise rollout accuracy over a finite prediction horizon. For a prediction horizon length H, the NMSE is computed as:

$$\textbf{NMSE H} = \frac{\sum_{t=1}^{H} \|\widehat{\boldsymbol{y}}_t - \boldsymbol{y}_t\|_2^2}{\sum_{t=1}^{H} \|\boldsymbol{y}_t\|_2^2 + \varepsilon},$$

where $\widehat{\boldsymbol{y}}_t$ denotes the predicted state at time t, $\boldsymbol{y}_t$ represents the ground truth, and ε is a small numerical constant to prevent division by zero. This metric evaluates both short-term (**NMSE 20**) and long-term (**NMSE 400**) trajectory accuracy. However, in chaotic systems, a low NMSE alone may not necessarily imply correct long-term dynamics, since collapsed trajectories may sometimes remain pointwise close over finite windows.

Time-correlation metric

We use a time-correlation metric (TCM) based on the integrated autocorrelation time. For each observed channel c, we compute the normalized autocorrelation function up to $L_{\max}$:

$$\rho_c(\ell) = \frac{\frac{1}{H-\ell}\sum_{t=1}^{H-\ell}(y_{t+\ell}^{(c)} - \bar{y}^{(c)})(y_t^{(c)} - \bar{y}^{(c)})}{\frac{1}{H}\sum_{t=1}^{H}(y_t^{(c)} - \bar{y}^{(c)})^2 + \varepsilon}, \qquad \ell = 0, 1, \ldots, L_{\max},$$

where $\bar{y}^{(c)}$ denotes the temporal mean of channel c, and ε is a small numerical constant. The integrated autocorrelation time of channel c is then approximated by:

$$\tau_c = \Delta t \left(1 + 2 \sum_{\ell=1}^{L_{\max}} \rho_c(\ell)\right).$$

We compute τ_c^{GT} and τ_c^{pred} for the ground-truth and predicted trajectories, respectively, and define the TCM as the absolute difference of their spatial averages:

$$\mathrm{TCM} = \left| \frac{1}{d}\sum_{c=1}^{d} \tau_c^{\mathrm{GT}} - \frac{1}{d}\sum_{c=1}^{d} \tau_c^{\mathrm{pred}} \right|.$$

In our implementation, $L_{\max} = 200$ unless the rollout is shorter, in which case the maximum lag is truncated to half of the available trajectory length. A lower TCM indicates better agreement of the temporal correlation scale between the generated and ground-truth trajectories.

Distance of recurrence time distributions

To assess recurrence-level temporal organization, we compute the distance between recurrence time distributions (DRTD). Given a recurrence section Σ, we extract the time indices of recurrence events from the ground-truth trajectory as

$$\mathcal{T} = \{t_1, t_2, \ldots, t_N\},$$

and from the rollouts,

$$\widehat{\mathcal{T}} = \{\widehat{t}_1, \widehat{t}_2, \ldots, \widehat{t}_{\widehat{N}}\}.$$

The corresponding inter-event times are calculated by multiplying the index differences by the observation interval Δt:

$$\tau_i = (t_{i+1} - t_i)\Delta t, \qquad \widehat{\tau}_j = (\widehat{t}_{j+1} - \widehat{t}_j)\Delta t.$$

We then define the DRTD as:

$$\mathrm{DRTD}_\Sigma = \mathcal{W}_1\left(\frac{1}{N-1}\sum_{i=1}^{N-1}\delta_{\tau_i}, \frac{1}{\widehat{N}-1}\sum_{j=1}^{\widehat{N}-1}\delta_{\widehat{\tau}_j}\right),$$

where $\mathcal{W}_1$ denotes the one-dimensional empirical Wasserstein-1 distance and δ represents the Dirac measure. A lower DRTD indicates better recurrence-level temporal consistency on the selected Poincaré section.

Discrepancy of attractors

To evaluate long-term geometric fidelity, we compare the empirical attractor distributions of the ground-truth trajectories and predicted rollouts. Let

$$\mu = \frac{1}{H}\sum_{t=1}^{H}\delta_{\boldsymbol{y}_t}, \qquad \widehat{\mu} = \frac{1}{H}\sum_{t=1}^{H}\delta_{\widehat{\boldsymbol{y}}_t}.$$

We report the Sinkhorn divergence with squared Euclidean cost as a computational approximation to the Wasserstein-2 geometry:

$$S_\varepsilon(\widehat{\mu}, \mu) = \mathrm{OT}_\varepsilon(\widehat{\mu}, \mu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\widehat{\mu}, \widehat{\mu}) - \frac{1}{2}\mathrm{OT}_\varepsilon(\mu, \mu),$$

where OT_ε denotes the entropy-regularized optimal transport cost. This metric measures whether the generated trajectory occupies the same region of the observed attractor as the ground truth. It is insensitive to pointwise phase alignment, but strongly penalizes collapsed, divergent, or geometrically distorted rollouts.

Log-PSD

Finally, we quantify spectral consistency using a log-power spectral density (**Log-PSD**) distance via Welch's method. For each observed channel c, we remove the temporal mean and estimate the power spectral density as follows:

$$P_c(f) = \mathrm{Welch}\big(y_{\cdot}^{(c)} - \bar{y}^{(c)}\big), \qquad \widehat{P}_c(f) = \mathrm{Welch}\big(\widehat{y}_{\cdot}^{(c)} - \bar{\widehat{y}}^{(c)}\big).$$

The Log-PSD distance is defined as

$$D_{\text{Log-PSD}} = \frac{1}{d}\sum_{c=1}^{d}\frac{1}{|\Omega|}\sum_{f\in\Omega}\left|\log\left(\widehat{P}_c(f) + \varepsilon\right) - \log\left(P_c(f) + \varepsilon\right)\right|,$$

where Ω denotes the set of evaluated frequency bins. The logarithm reduces the dominance of high-power frequency bands and makes the metric sensitive to discrepancies across both dominant and weaker spectral components.

F Ablation and robustness results

In this section, we conduct ablation studies of the loss terms, $\mathcal{L}^{\mathrm{DIM}}$, recurrence events and three additional robustness and sensitivity analyses on the Lorenz63 benchmark. We vary the speed-quantile threshold q_{speed}, which directly controls the active candidate pool for constructing the data-driven recurrence section. Then, we perturb the empirical recurrence schedule K_i to test the robustness of the recurrence-conditioned multi-shooting mechanism. Finally, we report diagnostics for the recurrence section construction under variations of q_{speed}, delay dimension m, and entry threshold $\varepsilon_{\mathrm{in}}$. These diagnostics are computed before model training and are intended to assess whether the extracted recurrence events are sufficiently stable for constructing the data-driven recurrence section.

Table 6: Comparison of attractor and spectral reconstruction performance.

Model	W2	**Log-PSD**
PRIME	**1.84191**	**0.413584**
Fixed-lag ($K = 3$)	58.0751	2.8142
Fixed-lag ($K = 6$)	70.7059	5.82627
Fixed-lag ($K = 9$)	6.86586	0.490735
Fixed-lag ($K = 12$)	49.9528	2.43378
Empirical variable-K	32.028358	0.932246
Sequential empirical-K	52.521202	2.284807

Ablation study on recurrence events

We conduct the following experiments to assess whether the performance of PRIME compared with DIM-only arises from longer rollout horizons instead of the recurrence events.

First, we uniformly sampled timestamps s and matched the recurrence graph measure of (U_s, U_{s+K}) and $(\hat{U}_s, \hat{U}_{s+K})$ using fixed lags $K \in \{3, 6, 9, 12\}$, where the median Algorithm 2 inter-event interval is 6 and U_t is the delay coordinate. Second, we retained the empirical interval distribution $K_i = h_{i+1} - h_i, \quad \tilde{K} \sim \frac{1}{M} \sum_{i=0}^{M-1} \delta_{K_i}$, but independently assigned each $\tilde{K}$ to a uniformly sampled timestamp s. This removes the original recurrence timestamps and the pairing between each state and its subsequent interval. Third, we retained the ordered sequence $K_j, \ldots, K_{j+L-2}$ and unrolled the model from a randomly sampled timestamp, without intermediate state resets. Thus, this experiment preserves both interval ordering and sequential rollout, while removing the original recurrence timestamps extracted by Algorithm 2. The experimental results are reported in Table 6.

Despite sufficient optimization of stage-2, all controls collapsed during long-term autonomous rollout and failed to preserve the attractor, which demonstrates the importance of recurrence events as well as the recurrence intervals. We think the reason might be the complex dynamics of chaotic attractors, which mix different dynamical phases. Landmark recurrence events selected by Algorithm 2 provide long term stable graph-measure information for training while maximally covering the dynamical diversity.

Ablation study on $\mathcal{L}^{\mathrm{DIM}}$

To assess the contribution of retaining the DIM loss after warm-up, we conduct an ablation study on the Lorenz63 system, comparing the full PRIME training scheme ($\mathcal{L}^{\mathrm{DIM}} + \mathcal{L}^{\mathrm{PRIME}}$) with a variant that optimizes only $\mathcal{L}^{\mathrm{PRIME}}$ following DIM warm-up.

Table 7: Performance comparison across different random seeds

Method	Seed 33	Seed 123	Seed 1000	Seed 13532	Seed 123456	Mean $\pm$ std	Median	Q25	Q75
Full PRIME (DIM + PRIME)	1.25	18.42	1.51	2.16	2.88	5.24 ± 6.61	2.16	1.51	2.88
PRIME-only (after DIM warm-up)	23.81	13.15	19.06	4.92	53.66	22.92 ± 16.61	19.06	13.15	23.81
$\mathcal{L}^{\mathrm{DIM}} + \mathcal{L}^{\mathrm{MSE}}$	37.46	25.24	66.88	18.81	3.47	30.37 ± 21.29	25.24	18.81	37.46

Table 7 shows that retaining the DIM loss during the second stage improves attractor reconstruction compared with optimizing $\mathcal{L}^{\mathrm{PRIME}}$ alone after warm-up. The two losses capture complementary aspects of the dynamics: $\mathcal{L}^{\mathrm{DIM}}$ constrains delay-coordinate distributions and short-horizon statistics, whereas $\mathcal{L}^{\mathrm{PRIME}}$ constrains transitions between recurrence events. Thus, $\mathcal{L}^{\mathrm{PRIME}}$ serves as a second-stage regularizer that complements, rather than replaces, $\mathcal{L}^{\mathrm{DIM}}$.

Seed-Level Results and Summary Statistics.

We provide detailed seed-level results for Tables 2 and 3. For Tables 8, 9, 10, 11, 12, and 13, each seed entry reports the mean over 32 rollout initial conditions for TCM and prediction metrics, or one DRTD value per section. Overall mean and population standard deviation are computed over 160 rollouts for TCM and prediction metrics, and over valid seeds for DRTD. Median, Q25, and Q75 are computed across the displayed seed-level results using linear interpolation. Missing DRTD values are denoted by "–" and excluded; n indicates the number of valid seeds when fewer than five. Any scale indicated beside a metric applies to all numerical entries in that row.

Table 8: Seed-level temporal consistency results on Lorenz63. Lower values are better.

Model	Metric	Seed 33	Seed 123	Seed 1000	Seed 13532	Seed 123456	Overall Mean ± Std	Median	Q25	Q75
PRIME	TCM	0.1156	0.6111	0.2392	0.1667	0.2712	0.2807 ± 0.5087	0.2392	0.1667	0.2712
PRIME	DRTD Σ_1	0.3905	0.7246	0.5196	0.4619	0.3060	0.4805 ± 0.1415	0.4619	0.3905	0.5196
PRIME	DRTD Σ_2	0.8395	4.672	0.8142	0.8358	0.8139	1.595 ± 1.538	0.8358	0.8142	0.8395
PRIME	DRTD Σ_3	0.1209	0.06617	0.04652	0.05096	0.1157	0.08005 ± 0.03194	0.06617	0.05096	0.1157
MSE	TCM	1.091	0.6063	2.083	1.504	1.699	1.397 ± 1.789	1.504	1.091	1.699
MSE	DRTD Σ_1	–	–	–	2.620	0.8116	1.716 ± 0.9044 $(n = 2)$	1.716	1.264	2.168
MSE	DRTD Σ_2	–	–	–	5.002	–	5.002 $(n = 1)$	5.002	5.002	5.002
MSE	DRTD Σ_3	0.3757	0.03493	0.5793	0.1569	0.3771	0.3048 ± 0.1899	0.3757	0.1569	0.3771
DIM	TCM	0.2329	0.8609	0.2849	0.1399	1.512	0.6062 ± 1.268	0.2849	0.2329	0.8609
DIM	DRTD Σ_1	0.3985	0.7458	0.3907	0.6918	0.9639	0.6382 ± 0.2187	0.6918	0.3985	0.7458
DIM	DRTD Σ_2	0.5742	0.8716	0.8867	1.431	0.5385	0.8604 ± 0.3200	0.8716	0.5742	0.8867
DIM	DRTD Σ_3	0.1013	0.1331	0.07717	0.2533	0.1046	0.1339 ± 0.06227	0.1046	0.1013	0.1331
DIM+MSE	TCM	0.5564	0.1244	0.4493	0.2148	0.1443	0.2978 ± 0.3995	0.2148	0.1443	0.4493
DIM+MSE	DRTD Σ_1	0.6813	0.8340	0.1401	2.040	0.5036	0.8398 ± 0.6432	0.6813	0.5036	0.8340
DIM+MSE	DRTD Σ_2	1.522	–	–	4.887	3.241	3.216 ± 1.374 $(n = 3)$	3.241	2.381	4.064
DIM+MSE	DRTD Σ_3	0.4219	0.3584	0.4897	0.4036	0.3117	0.3971 ± 0.06003	0.4036	0.3584	0.4219

Table 9: Seed-level temporal consistency results on KS.

Model	Metric	Seed 4242	Seed 10000	Seed 13531	Seed 123123	Seed 123123123	Overall Mean ± Std	Median	Q25	Q75
PRIME	TCM	7.201	5.528	18.12	16.37	3.653	10.18 ± 18.78	7.201	5.528	16.37
PRIME	DRTD Σ_1	7.974	7.158	7.458	4.544	9.385	7.304 ± 1.577	7.458	7.158	7.974
PRIME	DRTD Σ_2	8.001	11.10	7.343	3.320	2.713	6.494 ± 3.116	7.343	3.320	8.001
PRIME	DRTD Σ_3	17.52	24.93	8.191	5.219	5.804	12.33 ± 7.695	8.191	5.804	17.52
MSE	TCM	53.42	83.75	91.98	34.84	176.2	88.03 ± 114.2	83.75	53.42	91.98
MSE	DRTD Σ_1	12.22	12.98	15.30	14.34	2.167	11.40 ± 4.739	12.98	12.22	14.34
MSE	DRTD Σ_2	13.20	10.94	14.30	11.92	5.420	11.16 ± 3.087	11.92	10.94	13.20
MSE	DRTD Σ_3	13.05	10.07	10.26	12.55	3.061	9.799 ± 3.573	10.26	10.07	12.55
DIM	TCM	308.8	46.82	4.098	48.58	50.04	91.67 ± 120.6	48.58	46.82	50.04
DIM	DRTD Σ_1	14.12	13.59	4.122	11.33	8.795	10.39 ± 3.657	11.33	8.795	13.59
DIM	DRTD Σ_2	18.78	12.45	6.011	10.07	8.572	11.18 ± 4.339	10.07	8.572	12.45
DIM	DRTD Σ_3	–	20.41	4.650	3.829	16.44	11.33 ± 7.235 $(n = 4)$	10.54	4.445	17.43
DIM+MSE	TCM	36.65	20.83	240.9	116.8	8.000	84.65 ± 103.3	36.65	20.83	116.8
DIM+MSE	DRTD Σ_1	5.964	8.443	24.76	11.75	2.602	10.70 ± 7.640	8.443	5.964	11.75
DIM+MSE	DRTD Σ_2	7.754	7.271	8.260	10.61	4.007	7.580 ± 2.123	7.754	7.271	8.260
DIM+MSE	DRTD Σ_3	–	19.25	19.51	19.92	7.605	16.57 ± 5.183 $(n = 4)$	19.38	16.34	19.61

Ablation study on the loss terms

We conduct a loss-term ablation on two components of the PRIME training objective: the combined target and source marginal terms, denoted by $\boldsymbol{\mu}$, and the empirical graph-measure term of the induced return map, denoted by $\boldsymbol{\pi}$. As shown in Table 14, the marginal-only variant is insufficient to preserve the attractor geometry or recurrence-level temporal structure, yielding performance close to the DIM baseline. In contrast, adding the graph measure term of the induced return map substantially refines these metrics, which shows that the graph measure term is the main mechanism to alleviate topological collapse. The full PRIME method achieves the best trade-off, further improving the TCM and achieving the best attractor W_2. As a result, the marginal terms are useful as stabilizers when combined with the graph measure term.

The following ablation experiments further demonstrate the importance of graph measure/recurrence transition measure matching. The relatively large standard deviation of PRIME is driven by a single outlier run (18.42), whereas the median Attractor W2 is 2.16. The results are shown in Table 15.

Robustness and sensitivity analysis

We conduct three robustness and sensitivity analyses on the partially observed Lorenz63 benchmark. First, we vary the speed-quantile threshold q_{speed}, which controls the active candidate pool used to construct the data-driven recurrence section. As this is the primary hyperparameter for the recurrence section construction, we perform a full model-level sensitivity analysis: for each value of q_{speed}, we train five models with different random seeds and evaluate NMSE over 30 rollout initial conditions for each trained model. We also report the number of extracted recurrence events N_{hits} and the mean

Table 10: Seed-level temporal consistency results for flow past a cylinder (CFD).

Model	Metric	Seed 100	Seed 200	Seed 13532	Seed 123456	Seed 123123123	Overall Mean ± Std	Median	Q25	Q75
PRIME	TCM ($\times 10^{-2}$)	0.8587	1.737	0.4832	1.923	0.6995	1.140 ± 0.5786	0.8587	0.6995	1.737
PRIME	DRTD Σ_1 ($\times 10^{-3}$)	1.796	2.180	3.672	1.158	4.500	2.661 ± 1.236	2.180	1.796	3.672
PRIME	DRTD Σ_2 ($\times 10^{-3}$)	4.837	0.5174	1.445	4.237	0.7009	2.347 ± 1.824	1.445	0.7009	4.237
PRIME	DRTD Σ_3 ($\times 10^{-3}$)	5.795	7.631	4.099	4.457	5.127	5.422 ± 1.248	5.127	4.457	5.795
MSE	TCM ($\times 10^{-2}$)	4.245	1.180	9.035	8.198	21.83	8.898 ± 7.061	8.198	4.245	9.035
MSE	DRTD Σ_1 ($\times 10^{-3}$)	1.947	3.419	2.565	2.128	4.241	2.860 ± 0.8576	2.565	2.128	3.419
MSE	DRTD Σ_2 ($\times 10^{-3}$)	8.253	2.528	3.010	3.059	7.507	4.871 ± 2.475	3.059	3.010	7.507
MSE	DRTD Σ_3 ($\times 10^{-3}$)	2.722	5.292	10.89	5.690	7.687	6.456 ± 2.722	5.690	5.292	7.687
DIM	TCM ($\times 10^{-2}$)	25.66	4.595	21.15	10.21	5.961	13.52 ± 8.406	10.21	5.961	21.15
DIM	DRTD Σ_1 ($\times 10^{-3}$)	11.59	1.129	3.606	5.724	1.179	4.645 ± 3.868	3.606	1.179	5.724
DIM	DRTD Σ_2 ($\times 10^{-3}$)	11.28	1.130	4.537	4.114	2.070	4.626 ± 3.558	4.114	2.070	4.537
DIM	DRTD Σ_3 ($\times 10^{-3}$)	11.77	8.498	3.956	6.695	6.737	7.532 ± 2.571	6.737	6.695	8.498
DIM+MSE	TCM ($\times 10^{-2}$)	10.24	16.18	9.996	4.401	10.24	10.21 ± 3.725	10.24	9.996	10.24
DIM+MSE	DRTD Σ_1 ($\times 10^{-3}$)	5.863	3.156	5.223	2.664	5.822	4.546 ± 1.364	5.223	3.156	5.822
DIM+MSE	DRTD Σ_2 ($\times 10^{-3}$)	4.198	3.418	3.865	1.706	4.170	3.471 ± 0.9263	3.865	3.418	4.170
DIM+MSE	DRTD Σ_3 ($\times 10^{-3}$)	8.674	7.576	7.661	7.156	8.241	7.861 ± 0.5337	7.661	7.576	8.241

Table 11: Seed-level prediction and structure metrics on Lorenz63.

Model	Metric	Seed 33	Seed 123	Seed 1000	Seed 13532	Seed 123456	Overall Mean ± Std	Median	Q25	Q75
PRIME	NMSE_{20}	0.2202	0.2174	0.1859	0.1989	0.1829	0.2011 ± 0.1284	0.1989	0.1859	0.2174
PRIME	NMSE_{100}	0.2970	0.2747	0.3082	0.2884	0.2945	0.2926 ± 0.06155	0.2945	0.2884	0.2970
PRIME	NMSE_{400}	0.3306	0.2715	0.3419	0.3156	0.3421	0.3203 ± 0.03860	0.3306	0.3156	0.3419
PRIME	SD	1.270	26.50	4.864	3.970	3.946	8.110 ± 10.43	3.970	3.946	4.864
PRIME	Log-PSD	0.6313	1.193	0.6416	0.6201	0.6264	0.7424 ± 0.2928	0.6313	0.6264	0.6416
MSE	NMSE_{20}	0.2101	0.1904	0.2119	0.2008	0.1802	0.1987 ± 0.1313	0.2008	0.1904	0.2101
MSE	NMSE_{100}	0.2959	0.2640	0.2950	0.2960	0.2790	0.2860 ± 0.05985	0.2950	0.2790	0.2959
MSE	NMSE_{400}	0.2823	0.2578	0.2808	0.2833	0.2772	0.2763 ± 0.03959	0.2808	0.2772	0.2823
MSE	SD	70.38	69.88	43.11	40.08	41.85	53.06 ± 31.61	43.11	41.85	69.88
MSE	Log-PSD	4.352	4.305	3.198	3.832	3.568	3.851 ± 3.384	3.832	3.568	4.305
DIM	NMSE_{20}	0.2502	0.2377	0.2706	0.2655	0.2758	0.2600 ± 0.1599	0.2655	0.2502	0.2706
DIM	NMSE_{100}	0.3645	0.2953	0.3844	0.3764	0.3377	0.3517 ± 0.08318	0.3645	0.3377	0.3764
DIM	NMSE_{400}	0.3969	0.2840	0.4130	0.4121	0.3315	0.3675 ± 0.06599	0.3969	0.3315	0.4121
DIM	SD	11.76	28.65	12.52	11.47	42.44	21.37 ± 24.10	12.52	11.76	28.65
DIM	Log-PSD	1.057	3.307	1.055	1.338	2.290	1.810 ± 2.092	1.338	1.057	2.290
DIM+MSE	NMSE_{20}	0.1944	0.2081	0.2042	0.2450	0.1941	0.2092 ± 0.1239	0.2042	0.1944	0.2081
DIM+MSE	NMSE_{100}	0.2594	0.2729	0.2398	0.3315	0.3143	0.2836 ± 0.06756	0.2729	0.2594	0.3143
DIM+MSE	NMSE_{400}	0.2414	0.2908	0.2326	0.3549	0.3384	0.2916 ± 0.05396	0.2908	0.2414	0.3384
DIM+MSE	SD	57.68	19.13	61.73	14.53	6.457	31.91 ± 26.83	19.13	14.53	57.68
DIM+MSE	Log-PSD	2.209	0.7027	4.963	0.7165	0.9556	1.909 ± 2.120	0.9556	0.7165	2.209

inter-event interval $\bar{K}$. As shown in Table 16, performance is stable across nearby values of q_{speed}, and the default value $q_{\text{speed}} = 0.3$ exhibits the least fluctuation for both N_{hits} and $\bar{K}$.

Next, we test the robustness of the empirical recurrence block and check whether the effectiveness of PRIME depends on the exact empirical recurrence schedule. We perturb the recurrence-conditioned multi-shooting horizon K_i within a reasonable range, for which the mean relative percentage of perturbation is calculated as:

$$\frac{1}{N_{\text{hit}} - 1} \sum_{i=1}^{N_{\text{hit}}-1} \frac{|\tilde{K}_i - K_i|}{K_i},$$

where $\tilde{K}_i$ refers to the perturbed K_i. From Table 17, we observe that the jittered variants remain close to the exact-K_i model in attractor geometry and spectral properties. This robustness test illustrates that PRIME is not overly sensitive to small perturbations of the empirical recurrence schedule.

Finally, we report diagnostics for the data-driven recurrence section construction in Table 18. As demonstrated in the ablation and robustness experiments, both the graph-measure term of the induced return map and the statistics of recurrence events are critical for the performance of PRIME. Instead of training models across all multi-level parameters, we vary these parameters and report diagnostics reflecting their influence. Specifically, we evaluate the number of recurrence events N_{hits} and the

Table 12: Seed-level prediction and structure metrics on KS.

Model	Metric	Seed 4242	Seed 10000	Seed 13531	Seed 123123	Seed 123123123	Overall Mean ± Std	Median	Q25	Q75
PRIME	NMSE_{20}	1.320	1.332	1.334	1.298	1.239	1.305 ± 0.5171	1.320	1.298	1.332
PRIME	NMSE_{100}	1.450	1.437	1.602	1.417	1.393	1.460 ± 0.2350	1.437	1.417	1.450
PRIME	NMSE_{400}	1.484	1.406	1.594	1.459	1.403	1.469 ± 0.1215	1.459	1.406	1.484
PRIME	SD	1.939	2.047	1.937	2.185	1.866	1.995 ± 0.7783	1.939	1.937	2.047
PRIME	Log-PSD	1.363	2.935	1.176	1.829	0.9738	1.655 ± 1.238	1.363	1.176	1.829
MSE	NMSE_{20}	2.078	1.835	1.730	2.139	1.799	1.916 ± 0.8668	1.835	1.799	2.078
MSE	NMSE_{100}	3.912	3.596	7.355	2.558	18.61	7.206 ± 30.14	3.912	3.596	7.355
MSE	NMSE_{400}	68.75	92.50	187.1	10.06	926.7	257.0 ± 988.3	92.50	68.75	187.1
MSE	SD	4319	3465	4659	1905	4.507×10^4	$(1.188 \pm 3.915) \times 10^4$	4319	3465	4659
MSE	Log-PSD	1.637	1.520	1.634	1.572	1.584	1.589 ± 0.2453	1.584	1.572	1.634
DIM	NMSE_{20}	1.677	1.834	1.523	1.757	1.832	1.725 ± 0.6088	1.757	1.677	1.832
DIM	NMSE_{100}	2.629	2.205	1.994	2.119	2.142	2.218 ± 0.9341	2.142	2.119	2.205
DIM	NMSE_{400}	61.50	9.162	1.987	5.241	2.247	16.03 ± 46.36	5.241	2.247	9.162
DIM	SD	4885	1026	1.390	71.82	3.048	1198 ± 2725	71.82	3.048	1026
DIM	Log-PSD	1.668	0.8108	0.7580	0.7630	0.8088	0.9617 ± 0.6258	0.8088	0.7630	0.8108
DIM+MSE	NMSE_{20}	1.617	1.784	1.538	1.680	1.766	1.677 ± 0.7843	1.680	1.617	1.766
DIM+MSE	NMSE_{100}	2.148	2.390	2.222	3.534	2.306	2.520 ± 1.765	2.306	2.222	2.390
DIM+MSE	NMSE_{400}	2.680	3.214	12.20	38.20	2.202	11.70 ± 31.21	3.214	2.680	12.20
DIM+MSE	SD	5.960	6.919	228.0	1438	1.800	336.1 ± 945.9	6.919	5.960	228.0
DIM+MSE	Log-PSD	1.486	1.000	1.951	2.160	0.8024	1.480 ± 0.8498	1.486	1.000	1.951

Table 13: Seed-level prediction and structure metrics for flow past a cylinder (CFD).

Model	Metric	Seed 100	Seed 200	Seed 13532	Seed 123456	Seed 123123123	Overall Mean ± Std	Median	Q25	Q75
PRIME	NMSE_{20} ($\times 10^{-4}$)	0.7659	0.5572	0.7540	0.6111	0.5048	0.6386 ± 0.2910	0.6111	0.5572	0.7540
PRIME	NMSE_{100} ($\times 10^{-4}$)	0.7793	0.6154	0.7680	0.7533	0.5411	0.6914 ± 0.3044	0.7533	0.6154	0.7680
PRIME	NMSE_{400} ($\times 10^{-3}$)	0.1159	0.2193	0.09298	0.3331	0.08595	0.1694 ± 0.1172	0.1159	0.09298	0.2193
PRIME	SD ($\times 10^{-3}$)	3.878	3.450	4.629	3.745	3.839	3.908 ± 0.3922	3.839	3.745	3.878
PRIME	Log-PSD	0.7430	0.7176	0.6657	0.9521	0.6136	0.7384 ± 0.1159	0.7176	0.6657	0.7430
MSE	NMSE_{20} ($\times 10^{-4}$)	0.7523	0.9601	1.191	1.263	1.451	1.123 ± 0.6193	1.191	0.9601	1.263
MSE	NMSE_{100} ($\times 10^{-4}$)	1.475	1.091	3.899	2.869	9.671	3.801 ± 3.851	2.869	1.475	3.899
MSE	NMSE_{400} ($\times 10^{-3}$)	1.084	0.2011	3.973	2.859	13.73	4.370 ± 5.017	2.859	1.084	3.973
MSE	SD ($\times 10^{-3}$)	5.141	4.929	6.079	6.116	6.788	5.811 ± 0.7157	6.079	5.141	6.116
MSE	Log-PSD	0.9493	0.9356	0.9830	1.469	2.327	1.333 ± 0.5988	0.9830	0.9493	1.469
DIM	NMSE_{20} ($\times 10^{-4}$)	3.157	0.6987	1.003	1.083	0.6710	1.323 ± 1.159	1.003	0.6987	1.083
DIM	NMSE_{100} ($\times 10^{-4}$)	73.90	1.801	11.53	13.96	1.590	20.56 ± 27.33	11.53	1.801	13.96
DIM	NMSE_{400} ($\times 10^{-3}$)	53.12	1.814	14.64	17.67	1.818	17.81 ± 18.82	14.64	1.818	17.67
DIM	SD ($\times 10^{-3}$)	4.618	4.097	4.006	5.057	3.900	4.335 ± 0.4499	4.097	4.006	4.618
DIM	Log-PSD	2.074	1.012	1.083	1.694	1.031	1.379 ± 0.4304	1.083	1.031	1.694
DIM+MSE	NMSE_{20} ($\times 10^{-4}$)	1.063	0.7686	0.9043	0.5282	1.037	0.8601 ± 0.4138	0.9043	0.7686	1.037
DIM+MSE	NMSE_{100} ($\times 10^{-4}$)	14.94	7.454	11.07	1.378	14.74	9.917 ± 5.363	11.07	7.454	14.74
DIM+MSE	NMSE_{400} ($\times 10^{-3}$)	18.41	9.816	14.10	1.550	18.18	12.41 ± 6.295	14.10	9.816	18.18
DIM+MSE	SD ($\times 10^{-3}$)	4.090	4.063	3.752	3.948	4.247	4.020 ± 0.1973	4.063	3.948	4.090
DIM+MSE	Log-PSD	1.686	0.8364	1.490	0.9997	1.606	1.323 ± 0.3409	1.490	0.9997	1.606

Table 14: Loss-term ablation on Lorenz63. All variants start from the same DIM checkpoint and use the same data-driven recurrence section for training. π denotes the empirical graph-measure term of the induced return map, and μ denotes the combined source and target marginal terms. All metrics are reported as mean ± std over rollouts initialized from 30 different states.

Variant	$\boldsymbol{\pi}$	$\boldsymbol{\mu}$	**TCM** ↓	**SD** ↓	**Log-PSD** ↓	**DRTD** ↓
DIM-only	–	–	1.60 ± 1.65	68.47 ± 29.20	4.22 ± 3.15	1.05 ± 1.15
Marginals only	–	✓	1.75 ± 1.15	62.23 ± 20.70	2.98 ± 2.45	0.89 ± 0.69
Graph only	✓	–	0.18 ± 0.21	4.75 ± 2.65	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.53 \pm 0.34}$
PRIME + DIM	✓	✓	$\mathbf{0.10 \pm 0.09}$	$\mathbf{3.13 \pm 0.97}$	0.65 ± 0.02	0.65 ± 0.61

mean inter-event interval $\bar{K}$. Furthermore, we compute the coefficient of variation (CV) of inter-event intervals, denoted as $CV(K)$, which quantifies the relative fluctuation of the recurrence schedule and is essential for the multi-shooting horizon. Lastly, we measure the split-half stability of the empirical graph measure (SG), which calculates the distance between empirical graph measures over recurrence pairs $(\boldsymbol{r}_i, \boldsymbol{r}_{i+1})$ extracted from the first and second halves of the sequence. A smaller SG indicates

Table 15: Ablation study on the loss terms. Lower values indicate better performance.

Method	Attractor W_2	PSD log-L1
DIM + PRIME	**5.24 ± 6.61**	**0.47 ± 0.10**
DIM only	67.92 ± 9.48	4.64 ± 2.00
DIM + recurrence marginals	61.15 ± 13.95	3.97 ± 2.22
DIM + pairwise MSE + recurrence marginals	52.48 ± 25.83	3.99 ± 2.12

Table 16: Robustness and sensitivity analysis of the speed-quantile threshold on the Lorenz63 benchmark.

q_{speed}	N_{hits}	$\bar{K}$	NMSE 20 ↓	NMSE 100 ↓	NMSE 400 ↓
0.10	75.00 ± 1.41	6.63 ± 0.13	0.21 ± 0.03	0.30 ± 0.02	0.30 ± 0.03
0.20	74.20 ± 1.60	6.70 ± 0.15	0.22 ± 0.02	0.31 ± 0.06	0.33 ± 0.08
0.30	70.60 ± 0.49	7.05 ± 0.06	0.22 ± 0.03	0.29 ± 0.02	0.29 ± 0.04
0.40	70.80 ± 1.72	7.05 ± 0.18	0.20 ± 0.03	0.29 ± 0.05	0.29 ± 0.05

Table 17: Robustness of the recurrence schedule on the Lorenz63 benchmark. All variants use the same DIM warm start and the same data-driven recurrence section with $N_{\mathrm{hits}} = 71$. The label Jittered K_i, 7.26% means that the perturbed multi-shooting horizons $\widetilde{K}_i$ have a mean relative deviation of 7.26% from the original K_i.

Variant	NMSE 100 ↓	NMSE 400 ↓	SD ↓	Log-PSD ↓
DIM-only	0.3050 ± 0.0553	0.3086 ± 0.0304	64.1711	3.5547
Exact K_i	0.3358 ± 0.0546	0.3635 ± 0.0283	3.3736	0.6503
Jittered K_i, 7.26%	0.3260 ± 0.0574	0.3460 ± 0.0289	2.4607	0.6345
Jittered K_i, 15.07%	0.2802 ± 0.0541	0.2931 ± 0.0238	3.1201	0.6941

better consistency among all recurrence events, whereas a larger SG suggests an unstable recurrence section.

Table 18: Diagnostics for the data-driven recurrence section on the Lorenz63 benchmark. All diagnostics are computed before model training and averaged over five extractor seeds from the identical initial state.

Family	Configuration	N_{hit}	$\bar{K}$	CV(K)	SG
Baseline	Default	70.40 ± 1.50	7.10	0.63	0.28
Speed threshold	$q_{\mathrm{speed}} = 0.10$	75.00 ± 2.60	6.66	0.34	0.30
Speed threshold	$q_{\mathrm{speed}} = 0.20$	74.00 ± 1.80	6.75	0.51	0.27
Speed threshold	$q_{\mathrm{speed}} = 0.30$	70.40 ± 1.50	7.10	0.63	0.28
Speed threshold	$q_{\mathrm{speed}} = 0.40$	68.40 ± 1.00	7.31	0.58	0.29
Delay dimension	$m = 3$	72.40 ± 1.40	6.89	0.52	0.30
Delay dimension	$m = 5$	70.40 ± 1.50	7.10	0.63	0.28
Delay dimension	$m = 7$	72.00 ± 1.40	6.90	0.57	0.26
Delay dimension	$m = 9$	70.80 ± 1.70	6.99	0.54	0.28
Landmarks	$M = 50$	46.60 ± 2.40	10.73	0.76	0.40
Landmarks	$M = 100$	52.40 ± 2.70	9.51	0.72	0.35
Landmarks	$M = 200$	70.40 ± 1.50	7.10	0.63	0.28
Landmarks	$M = 400$	77.80 ± 1.60	6.41	0.47	0.23
Entry threshold	$\mu_{\mathrm{in}} = 0.25$	70.00 ± 1.70	7.14	0.63	0.28
Entry threshold	$\mu_{\mathrm{in}} = 0.35$	70.40 ± 1.50	7.10	0.63	0.28
Entry threshold	$\mu_{\mathrm{in}} = 0.50$	75.00 ± 2.00	6.66	0.49	0.28

G Computational cost

We report the computational costs of all objective-controlled methods in Table 19. All methods use the same three-layer MLP predictor within each benchmark, so the parameter counts are identical across objectives. The additional computational cost therefore stems from the training objective rather than the model architecture. Compared to pointwise MSE, DIM and DIM + MSE incur a higher per-iteration cost due to the measure-matching loss. PRIME is computationally more expensive because it additionally optimizes recurrence-level return blocks on the induced section. All runtime measurements are based on a single training run.

Table 19: Computational cost of objective-controlled training. Experiments were run on a single NVIDIA GeForce RTX 4060 GPU with 8 GB VRAM using PyTorch.

Benchmark	Method	Iterations	Train time (s)	Avg. iter. (ms)	Peak mem. (MB)
Lorenz63	MSE	10000	12.85	1.28	18.77
Lorenz63	DIM	10000	107.99	10.80	30.29
Lorenz63	DIM+MSE	10000	105.46	10.55	30.29
Lorenz63	PRIME	4000	701.20	175.30	32.72
KS	MSE	10000	12.11	1.21	18.80
KS	DIM	10000	172.34	17.23	46.26
KS	DIM+MSE	10000	167.40	16.74	46.26
KS	PRIME	12000	2179.05	181.59	46.36

The considered settings use limited training trajectories consisting of 200 to 500 observations, depending on the benchmark. We load the models trained with five different random seeds, as reported in our paper, and evaluate their inference time. Since all loss terms share the same network architecture and have the same number of parameters, their inference times are nearly identical. Results are shown in the following Table 20.

Table 20: Model complexity and inference time comparison.

Method	Parameters	Time per step (μs)
DIM	20702	48.00 ± 5.23
DIM + MSE	20702	50.05 ± 5.78
NODE	20702	49.39 ± 6.19
PRIME	20702	53.20 ± 3.78

We report the average training time for all methods in Table 21. However, history-window-based methods, i.e., LSTM and Koopman-AE, are not applicable to our main setting.

Table 21: Average training update time per iteration.

Method	Avg. update time (ms)
MSE	1.28
DIM + MSE	10.55
DIM	10.80
PRIME	175.30
LSTM	1.93
Koopman-AE	1.73
DySLIM-style	103.27

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contribution of PRIME that the recurrence-level alignment improves the temporal consistency and other metrics of the learned autonomous dynamics from partial observations. The scope of the claims is restricted to the considered chaotic and dissipative benchmark systems, and the assumptions and limitations are discussed in the paper and appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion, including the dependence on the empirical roof time, finite-sample surrogate of recurrence blocks, and the additional computational cost of PRIME.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The appendix lists the assumptions and provides a proof for the local contraction criterion. The informal theorem in the main text is linked to the corresponding formal statement in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports the model architecture, objectives, training procedure, datasets, evaluation metrics, and hyperparameters needed to reproduce the main experimental results. Additional implementation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will provide data, code, and instructions for the final paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental sections and appendix specify the benchmark systems, observation settings, train/test splits, model architecture, optimizer, training steps, loss weights, recurrence-section construction, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main quantitative results are reported with mean and standard deviation over multiple random seeds and rollout initial conditions, as described in the experimental setup. Single run diagnostic results, such as computational cost, are explicitly identified as representative runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports computational cost, including training time, average iteration time, and peak CUDA memory. The hardware and software environment used for the experiments are also specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses synthetic or standard dynamical system benchmarks and does not involve human subjects, private data, surveillance, or safety-critical deployment. We have reviewed the NeurIPS Code of Ethics and do not identify any violation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper develops a learning objective for dynamical systems and evaluates it on controlled benchmark systems. We do not identify a direct path to societal impact, since the work does not involve human subjects, personal data, deployed systems, or safety-critical decision making.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, generative models, or systems with clear misuse potential requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the benchmark systems, baseline methods, software packages, and external code resources used in the experiments. In particular, we credit the MIT-licensed Machine Learning and Simulation repository used for parts of the flow past a cylinder dataset generation, and we cite PyTorch and other software packages used in the implementation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, model checkpoint, or benchmark asset as a primary contribution.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, experiments, or analysis. Any use for writing, editing, or formatting does not affect the core methodology, scientific claims, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.