Unrolled gradients in disguise: bridging interpolation-based and Jacobian regularization for stable neural dynamics

Maya Janvier
LOCEAN
Sorbonne University
Paris, France
maya.janvier@locean.ipsl.fr

Etienne Meunier
Inria
Paris, France
etienne.meunier@inria.fr

Abstract

Hybrid models and neural differential equations (NDEs) are becoming increasingly important for modeling physical systems; however, they often encounter stability and accuracy issues during long-term integration. Training on unrolled trajectories is known to limit these divergences but quickly becomes too computationally expensive due to the need to backpropagate through the solver, making this method impractical for large-scale physical systems (e.g. computational fluid mechanics, weather and climate forecasting). Two popular paradigms for improving the stability of learned dynamics are to train on interpolated or noisy samples or to regularize the Jacobian of the learned representation. In this work, we demonstrate that these paradigms optimize equivalent objectives as they both correspond to an approximation of the gradients obtained in an unrolled learning strategy. Building on this understanding, we derive a principled framework to regularize training and demonstrate its impact on the stability of learned dynamical systems in both supervised and unsupervised settings.

1 Introduction

Autoregressive models are ubiquitous in the deep learning literature and are central tools in many applications, from forecasting [26, 22, 55, 37, 46] to video analysis [58, 24]. Autoregressive models typically perform two distinct tasks at once: learning an underlying dynamic and learning a solver to advance this ODE in time. These two tasks are intertwined in a way that we cannot decipher the contribution of one or the other. Neural ODE models [5], on the other hand, leave the solving part to numerical integration schemes, while a neural network (NN) is used as the dynamical function or right-hand side of the ODE. This separation has many benefits as it provides a more interpretable and flexible representation while encompassing the literature of autoregressive models (recovering ResNets when using Euler forward). It allows for example to use different timesteps at inference, while most autoregressive models can only predict a fixed number of timesteps ahead. Neural ODEs gained popularity across a wide range of deep learning tasks, from the modeling of physical systems [52, 56, 25, 35], to classification tasks [54, 17] and generative tasks [8, 45]. There is a particular and renewed interest in them for climate and weather forecasting as they provide a natural framework for hybrid modeling [1, 29, 21].

There are two settings for the training of Neural ODEs: the supervised case, where we have access to the true dynamic, and the unsupervised case where this dynamic is unknown and we only have access to trajectories. The former is interesting for building fast emulators of models (to accelerate General Circulation Models predictions where it takes months to perform a hundred years of simulations for example) or in hybrid models where the learned dynamic is tasked only to correct the biases of

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

the existing one (parameterizing high resolution phenomena in low resolution models [47]). The unsupervised case on the other hand corresponds to more realistic scenarios (observed trajectories).

In the supervised case, the Neural ODE can be directly trained to match the right-hand side dynamic in what is called offline or a-priori training [9]. However, a core challenge that arises is that such models quickly become unstable at inference when integrated over long-rollouts [9, 46]. Training on unrolled sequences through the solver (online or a-posteriori training) is known to mitigate this problem and no longer requires the knowledge of F (unsupervised case); however, this method quickly becomes too computationally expensive due to the need to compute gradients over an iterative process while also suffering from an explosion of the gradients during unrolling [27]. Although methods have been explored to mitigate the latter [11, 31], the issue of the cost remains prominent. On real-world, large scale applications that involve timescales of months to years (weather forecasting), the cost of long rollouts becomes entirely prohibitive while offline training fails in producing stable models for long-horizon predictions [2, 57, 42, 38]. This limits Neural ODEs usage to simple dynamical systems in practice, such as robotic simulations [6, 52].

Among the various methods developed to improve stability (see Section 2), the approaches that can be applied in an offline setting are thus the preferred solutions to this challenge. In particular, methods that either regularize the Jacobian of the system [32, 53] or train on interpolated or noisy samples [4, 12] give promising results on complex physical systems, including chaotic ones. In this paper, we show that these paradigms are not inherently different but rather correspond to an approximation of the gradients in the state-of-the-art online training (Section 3). We then exploit this common framework to derive principled regularization loss functions (Section 4) and investigate their efficiency on various physical problems (Section 5) in both supervised and unsupervised settings.

2 Related work

2.1 One challenge, various strategies

We identify four categories of approaches to mitigate the stability issue of Neural ODEs.

Projection and energy conservation. Early attempts at stabilizing Neural ODEs focused on energy conservation by introducing dissipation [25]. Projecting onto an underlying manifold (e.g., algebraic constraints) allows stabilizing long-term dynamics learned on short rollouts [52, 35] and thus alleviating the computational cost of unrolling. However, these methods are inherently limited to conservative systems and projections are prohibitively expansive for large-scale systems. Autoregressive models trained to correct the accumulation of errors in the form of denoising diffusion processes are efficient [26, 37], but can be cumbersome to train and rely on a number of denoising steps, which is a sensitive and hard parameter to tune in order not to break the temporal coherence of the predicted trajectories.

Controlling the Lipschitz constant. Stability in autoregressive models is closely linked to robustness against adversarial perturbations, that is to say ensuring that similar inputs (for a certain norm) produce similar outputs. The Lipschitz constant encompasses this property (annex A) and can be leveraged to enforce robustness. Unfortunately, estimating a computationally efficient Lipschitz constant for training is an open research question [40, 44, 16, 14, 36, 7, 39]. Hard constraints in the form of Lipschitz layers [30, 49, 14], energy conserving architectures [41] or penalizations to ensure a contracting Lipschitz constant ($L < 1$) [54] directly enforce robust predictions of steady states and stationarity, which is no longer valid when it comes to model chaotic dynamics that are constantly evolving and never reach steady states. Furthermore, hard constraints limit the expressiveness of the model which may harm its performances and generalization capacities.

Jacobian regularization. Softer constraints focusing on the Jacobian (annex A) have been deployed, such as spatially-adaptive Jacobian regularization in the real space (JAWS) [32] or third-order time derivative penalization obtained by finite differences in the latent space (Jerk) [53], however these methods still need online training to really improve stability. Spectral constraints acting on the eigenvalues of the Jacobian can ensure stability around equilibrium points [17] or within integration schemes [51] (A-stability) and are known to stabilize training [48], accelerate convergence [33, 8, 18] and promote continual training of models by maintaining gradient diversity throughout training

[23].These observations are qualitative and made on a limited set of physical systems, the lack of theoretical understanding limits their generalizability.

Data augmentation and interpolation. Robustness can also be achieved through adversarial training on noisy samples [43, 28]. Such methods would also result in a contracting representation (proof in annex B), limiting their efficiency in modeling chaotic physical systems. The pushforward trick [4] achieves a form of noisy training by unrolling for two steps and backpropagating gradients only through the last step, but its efficiency still lacks theoretical understanding. Continuous Flow Operators [12] (CFO) fit temporal splines using finite-difference estimates of time derivatives at true trajectory points to closely approximate the true PDE dynamics and train in an offline fashion on new data points along the spline.

2.2 Our work: bridging interpolation-based and Jacobian regularization.

Overall, the energy and Lipschitz approaches often are system-dependent and lack flexibility. Jacobian regularization and interpolation approaches on the other hand are more systemic/general methods in tackling the issue of stability in Neural ODEs, but still lack theoretical understanding. In this paper, we bridge this understanding gap, focusing on Jacobian regularization (JAWS) and interpolation (CFO) methods as they have a particular advantage: both can be performed not only in an offline fashion (that is, the training is performed outside the solver) but also without having access to the true dynamic function F: the former leverages the learned Jacobian J_{F_θ}; the latter constructs probability paths whose velocities closely approximate the true F, also relying on finite differences estimation. We will: (i) prove that these approaches can be understood as regularization of gradients obtained in an unrolled learning strategy (involving the Jacobian in particular), (ii) derive a principled framework to investigate all of the terms that appear from unrolling, and (iii) analyze its efficiency on different physical systems in both supervised and unsupervised settings.

3 Unrolled gradients in disguise: on the importance of the Jacobian

Let us consider a general Partial Differential Equation (or PDE) in the form of

$$\frac{\partial \mathbf{x}}{\partial t} = F(t, \mathbf{x}, \nabla \mathbf{x}, \nabla^2 \mathbf{x}, ...) \tag{1}$$

with $\mathbf{x}$ the prognostic variables evolving temporally through the dynamic function F. To simulate trajectories of $\mathbf{x}$, we numerically integrate the PDE forward in time using a temporal integration scheme or solver. Training a neural network F_θ to approximate F from observed trajectories is a central task in scientific ML. The length of the trajectories we feed to the neural network, measured in number of integration steps performed through the solver, is called the training rollout N. Since non-autonomous systems described by $\frac{dx}{dt} = F(x(t), t)$ can always be rewritten as autonomous ones by using the augmented state $X = (x(t), t)$ and $\frac{dX}{dt} = (F(X), 1)$, we will use the autonomous formulation in the paper without losing generality.

3.1 General influence of the Lipschitz constant

Neural ODEs [5] combine data-driven estimators and numerical solvers. Consider a neural network F_θ representing the right-hand side or dynamics of an autonomous ODE $\frac{dx_\theta(t)}{dt} = F_\theta(x_\theta(t))$, associated with an integration scheme (or solver) and a step-size Δt, that aims to produce the same trajectories as the true ODE $\frac{dx(t)}{dt} = F(x(t))$, namely $X \in \mathcal{T}_F = \{x_i = x_0 + \int_{t_0}^{t_i} F(x(s))ds, 0 \leq i \leq N\}$. We consider both F and F_θ to be differentiable and Lipschitz continuous.

In offline training, the model is tasked to minimize $\|F(x) - F_\theta(x)\|_2^2$, so that learning F is consequently achieved in a direct way, without using the solver. However, this method needs access to the true F which is unknown in most scenarios. Furthermore, when applied to real world problems where the true dynamic is approximated by a high resolution simulation, this method leads to diverging trajectories for long-term integration at inference [9, 27]. Looking at bounds on the trajectory error [56, 54], it is easy to explain the failure of offline learning. Using Grönwall's lemma C.2, we get for the trajectory error starting from the same initial conditions (proof in appendix C):

$$\forall t \geq 0,\ \|x(t) - x_\theta(t)\|_2 \leq \frac{\|\epsilon_\theta\|_\infty}{L_{F_\theta}} (\exp(L_{F_\theta} t) - 1) \tag{2}$$

with $\epsilon_\theta = F - F_\theta$ the offline error and L_{F_θ} the Lipschitz constant of F_θ. Thus, a model with a low $\|\epsilon_\theta\|_\infty$ (minimized in offline training) but a large L_{F_θ} can still lead to diverging trajectories as t grows, driven by the exponential term in the upper bound. This bound shows that there is a trade-off between the smoothness of a neural network and its expressiveness, i.e., between ϵ_θ and L_{F_θ}, that is not trivial to achieve in practice [14, 50, 7].

On the other hand, performing the training via unrolled trajectories (or online learning) only requires access to trajectory points, and no longer to F (unsupervised case). It is achieved by minimizing the trajectory loss [5, 56, 25]:

$$\mathcal{L}^N_{\text{traj}}(\theta) = \sum_{X \in \mathcal{T}_F} \sum_{i=0}^{N} \|x(t_i) - \tilde{x}_\theta(t_i)\|_2^2, \quad t_i = i\Delta t \tag{3}$$

with $\tilde{x}_\theta(t_i)$ obtained by numerically integrating the ODE forward in time for $0 \leq i \leq N$ steps. As seen in the introduction, unrolling for long enough sequences results in much more stable rollouts [9, 27]. However, there is still little theoretical understanding of the reasons behind this property. This is what we investigate in the next section.

3.2 What is actually hidden in unrolled trajectories ?

Examining the terms that arise during unrolling, we better understand where the stability of online training stems from. We use the Taylor-Young formula D.1 to linearize around the data points and see the different terms at play. Starting from some $x(t) = x_t$ (and similarly for $x_\theta(t) = x^\theta_t$), we derive the trajectory error for $N = 2$ steps under Euler forward (full proof in annex D):

$$x_{t+2\Delta t} - x^\theta_{t+2\Delta t} = \underbrace{(x_t - x^\theta_t)}_{\text{initial error}} \tag{4}$$

$$\begin{aligned}
&+ 2\,\Delta t\,[\ \underbrace{\epsilon_\theta(x_t)}_{\text{offline error}} + \underbrace{J_{F_\theta}(x_t)\cdot(x_t - x^\theta_t)}_{\text{Grönwall drift}} + \ o(\|x_t - x^\theta_t\|_2)] \\
&+ \Delta t^2\,[\ \underbrace{J_{\epsilon_\theta}(x_t)\cdot F(x_t)}_{\text{Jacobian match in true direction}} + \underbrace{J_{F_\theta}(x_t)\cdot\epsilon_\theta(x_t)}_{\text{Lipschitz stretch in direction of error}} \\
&+ \underbrace{J_{F_\theta}(x_t)\cdot(F_\theta(x_t) - F_\theta(x^\theta_t))}_{\text{Grönwall drift on derivatives}} + \ o(\|x_t - x^\theta_t\|_2)] \\
&+ \ o(\Delta t^2)
\end{aligned} \tag{5}$$

In this formula, we first recover the offline error and the Lipschitz contribution from the derivation of the Grönwall bounds (annex C), terms in $O(\Delta t)$ that appear naturally with the first unrolling step (t to $t + \Delta t$). Unrolling for a second step gives terms in $o(\Delta t)$ that involve directional derivatives of the true and the learned Jacobians in specific directions. Equation 4 shows that even if we train the model with the trajectory loss by unrolling until t, the error for a longer inference (up to $t + 2\Delta t$ in eq. 4) will involve those terms that need to be appropriately learned, which is a new insight on the trajectory loss.

In particular, we now understand why regularizing the norm of the Jacobian of the learned dynamic J_{F_θ} as done in JAWS (annex E.2) is efficient in stabilizing the long-term dynamics: controlling it limits the scale of the propagation of initial errors. However, regularizing the learned Jacobian does not capture the Jacobian matching term $J_{\epsilon_\theta}(x_t)\cdot F(x_t) = (J_F(x_t) - J_{F_\theta}(x_t))\cdot F(x_t)$ ensuring that J_F and J_{F_θ} resemble each other in realistic directions of the trajectories: it can explain why in practice JAWS still need short-online training ($N = 5$) to successfully stabilize long-term dynamics.

We now analyze equation 4 in the training regime, i.e., when the trajectories have the same initial conditions $x_t = x^\theta_t$:

$$x_{t+2\Delta t} - x^\theta_{t+2\Delta t} = 2\,\Delta t\,\epsilon_\theta(x_t) + \Delta t^2\,(J_{\epsilon_\theta}(x_t)\cdot F(x_t) + J_{F_\theta}(x_t)\cdot\epsilon_\theta(x_t)) + o(\Delta t^2) \tag{6}$$

$$= 2\,\Delta t\,\epsilon_\theta(x_t) + \Delta t^2(\underbrace{J_F(x_t)\cdot F(x_t)}_{d^2x/dt^2} - J_{F_\theta}(x_t)\cdot F_\theta(x_t)) + o(\Delta t^2) \tag{7}$$

This equation shows that the unrolled loss, i.e., the trajectory loss $N = 2$ can be approximated, accurate to $o(\Delta t^2)$, by an expression involving not only the offline error ϵ_θ but also another offline

term, the acceleration error $\eta_\theta(x_t) := J_F(x_t) \cdot F(x_t) - J_{F_\theta}(x_t) \cdot F_\theta(x_t)$ involving the Jacobians of the systems. This formula can be generalized to $N \geq 2$ steps and to other solvers such as RK2 (further detailed in annex D). Under sufficient regularity assumptions on F, F_θ and their derivatives, the gradients of the approximation terms in eq. 7 (right side) with respect to the parameter θ remain close to the online gradients (left side), with a relative error of $O(\Delta t^2)$ (proof in D.4). The benefit is that this approximation comes at much lower cost than doing direct online training since it does not backpropagate through the auto-regressive steps and doesn't require a differentiable solver.

In addition, we found that CFO is actually acting on the offline error and Jacobian matching term during training by linearizing its loss:

$$\mathcal{L}_{\text{CFO}}(\theta) = \sum_{X \in \mathcal{T}_\text{I}} \|F_\theta(x(t_i)) - \partial_t I(t_i, x(t_i))\|_2^2 \tag{8}$$

where I is the fitted spline and $\mathcal{T}_\text{I}$ the ensemble of data points sampled along I. In the linear spline fitting, CFO consists in sampling new points on the line between x_t and $x_{t+\Delta t}$: for $\gamma \in [0, \Delta t]$, $x_t^\gamma = x_t + \gamma F(x_t)$. If the fit was perfect or if we had access to the true F, the CFO loss on linearly interpolated points would yield with a first order Taylor-Young approximation:

$$\begin{aligned}
F(x_t^\gamma) - F_\theta(x_t^\gamma) &= F(x_t + \gamma F(x_t)) - F_\theta(x_t + \gamma F(x_t)) && (9)\\
&= F(x_t) + \gamma J_F(x_t) \cdot F(x_t) - F_\theta(x_t) - \gamma J_{F_\theta}(x_t) \cdot F(x_t) + o(\Delta t) && (10)\\
&= \underbrace{\epsilon_\theta(x_t)}_{\text{offline error}} + \gamma \underbrace{J_{\epsilon_\theta}(x_t) \cdot F(x_t)}_{\text{Jacobian match in true directions}} + o(\Delta t) && (11)
\end{aligned}$$

As in equation 6, the CFO loss also constrains the offline error and the Jacobian matching term (eq. 11). CFO sampling acts partly as unrolling by characterizing the Jacobians: it explains its success in long-horizon stability. It also explains the stabilization and acceleration of the training process observed in [12] as it is a known consequence of a better characterization of the Jacobian documented in [33, 8, 18, 23, 48]. Contrary to JAWS, the norm of J_{F_θ} is not constrained at all in the CFO formulation, and thus not controlling the complete term involved when unrolling: the acceleration error $J_F(x_t) \cdot F(x_t) - J_{F_\theta}(x_t) \cdot F_\theta(x_t)$ in equation 7.

In the end, we see that the efficiency of both Jacobian regularization such as JAWS and interpolation-based strategies such as CFO is rooted in the different Jacobian characterizations that appear when linearizing the unrolled trajectory loss and that are simply absent in offline training. This derivation also exhibits offline approximations of the online trajectory loss, in particular the acceleration matching formulation in eq. 7, which we leverage in the next section either as a training objective in an offline setting or regularization term in an online one.

4 Methods

It comes naturally from the previous section that minimizing the terms at play in equation 4 should improve long-term stability. We focus on two challenging cases: **(1) the offline setting**, where we directly use the offline formulations of the losses as described in equations 7 and 11, and **(2) the short-rollout training** where the model is trained online on short rollouts ($N \leq 5$) with regularization terms derived from equation 4.

4.1 Offline learning via spline fitting

From equations 7 and 11, we can derive offline versions of AccMatch and JacMatch to compare to plain offline training, with γ a hyperparameter that we tune on a held out validation set. We can evaluate them in both supervised and unsupervised settings, by using either the true F or the spline fitting technique of CFO. Compared to using finite differences approximations, this approach avoids the noise amplification issue (when $\Delta t \to 0$, more in next section) by directly fitting a cubic spline to F from trajectory points, inducing a smooth representation at the cost of a less accurate one.

We summarize the offline methods we will test in Table 1. To give a better intuition of what the losses represent, Figure 1 illustrates the discrepancies they target in an unrolled trajectory: offline acts on the direction divergence the trajectories x and x^θ take in the first step from the same initial conditions, while AccMatch looks at reducing this direction gap after a second step. JacMatch and CFO, considering $\gamma = 1$, also target a single step divergence, but starting from the later $x_{t+\Delta t}$ state.

Table 1: **Summary of the methods**. **(i) Offline :** The original offline method only uses the exact F and CFO the fitted spline version of it. CFO differs from JacMatch by the upsampling (10 more x_t^γ data points), although they come from the same theoretical formula. **(ii) Short-online:** the regularizations are tested within the training loss: $\mathcal{L}(\theta) = \mathcal{L}^N_{\text{traj}}(\theta) + \lambda \mathcal{L}_{\text{reg}}(\theta)$. The Lipschitz method refers to the Jacobian regularization approach, in the spirit of JAWS. To build an unsupervised version of JacMatch, we use a first-order finite differences (FD) approximation (eq. 16). A FD version of AccMatch is possible but not tested here. **Both:** the γ and λ hyperparameters are tuned on a held-out validation set.

METHOD	EXACT FORM	LINEARIZED FORM (OFFLINE)	REGULARIZATION FORM (SHORT-ONLINE)
OFFLINE	$\|F(x_t) - F_\theta(x_t)\|_2^2$	$\|\epsilon_\theta(x_t)\|_2^2$ EXACT F	X
LIPSCHITZ	$\|J_{F_\theta}(x_t)\|_F^2$	X	$\frac{1}{V}\sum_{j=1}^V \|J_{F_\theta}(x_t)\cdot v_j\|_2^2$
CFO	$\|F(x_t^\gamma) - F_\theta(x_t^\gamma)\|_2^2$	$\|F(x_t^\gamma) - F_\theta(x_t^\gamma)\|_2^2$ SPLINE FIT F VARIOUS γ (UPSAMPLING X10)	X
JACMATCH	$\|F(x_t^\gamma) - F_\theta(x_t^\gamma)\|_2^2$	$\|\epsilon_\theta(x_t) + \gamma J_{\epsilon_\theta}(x_t)\cdot F(x_t)\|_2^2$ EXACT OR SPLINE FIT F, J_F SINGLE γ	$\|J_{\epsilon_\theta}(x_t)\cdot F(x_t)\|_2^2$ EXACT OR FD F, J_F [16] SINGLE λ
ACCMATCH	$\|F(x_t^\gamma) - F_\theta(x_t^{\theta,\gamma})\|_2^2$	$\|\epsilon_\theta(x_t) + \gamma\eta_\theta(x_t)\|_2^2$ EXACT OR SPLINE FIT F, J_F SINGLE γ	$\|\eta_\theta(x_t)\|_2^2$ EXACT (OR FD) F, J_F SINGLE λ

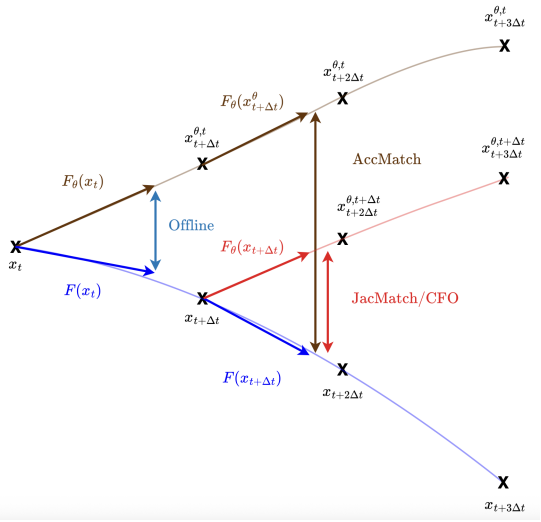

Figure 1: **Losses interpretation.** When $\gamma = 1$, $x_t^\gamma = x_{t+\Delta t}$ and we get that JacMatch is equivalent to the error of a delayed F_θ trajectory starting from $x_{t+\Delta t}$, i.e., only accounting for the error between the Jacobians while AccMatch also accounts for the error after the first timestep.

4.2 Regularizing short-online training

The goal of this experiment is to study, understand and compare the limitations and efficiencies of this family of losses when they are used as regularizations in the following training loss:

$$\mathcal{L}(\theta) = \mathcal{L}^N_{\text{traj}}(\theta) + \lambda \mathcal{L}_{\text{reg}}(\theta) \tag{12}$$

where N is short-term to avoid prohibitive computational costs, and λ a hyperparameter that we tune on a held out validation set. This extends the offline methods to situations where short-online training

is possible, under both supervised and unsupervised settings. The regularizations are summarized in Table 1.

4.2.1 Supervised approaches

From the Jacobian matching term in equation 4, we derive:

$$\mathcal{L}^{sup}_{\text{Jacmatch}}(\theta) = \sum_{X \in \mathcal{T}_F} \sum_{i=0}^{N} \|(J_F(x_{t_i}) - J_{F_\theta}(x_{t_i})) \cdot F(x_{t_i})\|_2^2 \tag{13}$$

From equation 7, we define the acceleration matching term:

$$\mathcal{L}^{sup}_{\text{Accmatch}}(\theta) = \sum_{X \in \mathcal{T}_F} \sum_{i=0}^{N} \|J_F(x_{t_i}) \cdot F(x_{t_i}) - J_{F_\theta}(x_{t_i}) \cdot F_\theta(x_{t_i})\|_2^2 \tag{14}$$

$\mathcal{L}_{\text{Jacmatch}}$ and $\mathcal{L}_{\text{Accmatch}}$ are first evaluated in a supervised setting (denoted with sup exponent) where F is accessible, which corresponds to cases where we want to build a fast emulator of the system. This setting represents some gold standard to assess the potential of the different regularizations. We also believe that they will yield similar results in a short-rollout context. Indeed, the difference between the two is only the $J_{F_\theta}(x_t) \cdot \epsilon_\theta(x_t)$ term in equation 6, and $\epsilon_\theta(x_t)$ will quickly become small for short rollouts N as $\mathcal{L}^N_{\text{traj}}$ is optimized during training.

4.2.2 Unsupervised approaches

The Jacobian regularization approach is natively unsupervised as it only requires J_{F_θ}:

$$\mathcal{L}_{\text{Lip}}(\theta) = \sum_{X \in \mathcal{T}_F} \sum_{i=0}^{N} \frac{1}{V} \sum_{j=1}^{V} \|J_{F_\theta}(x_{t_i}) \cdot v_j\|_2^2 \approx \sum_{X \in \mathcal{T}_F} \sum_{i=0}^{N} \|J_{F_\theta}(x_{t_i})\|_F^2 \tag{15}$$

This loss (similar to JAWS, see E.2) penalizes the norm of the Jacobian of the learned dynamic, computed via the Hutchinson trace estimator [15]. We use $V = 10$ directions drawn from $\mathcal{N}(0, I)$ at each epoch, to resemble noise injection strategies (annex B). This Monte-Carlo estimation is unbiased (eq. 20), with its variance decreasing in $\frac{1}{V}$ [13]. We found that $V = 10$ was a good compromise between computational cost and accuracy in practice.

We also derive an unsupervised version of JacMatch using finite differences (FD):

$$\mathcal{L}^{unsup}_{\text{Jacmatch}}(\theta) = \sum_{X \in \mathcal{T}_F} \sum_{i=0}^{N-2} \| \frac{F_\theta(x_{t_i+\Delta t}) - F_\theta(x_{t_i})}{\Delta t} - \frac{x_{t_i+2\Delta t} - 2x_{t_i+\Delta t} + x_{t_i}}{\Delta t^2} \|_2^2 \tag{16}$$

The fraction on the right is simply a first order FD approximation of $\frac{d^2x}{dt^2} = J_F(x_t) \cdot F(x_t)$, the acceleration of the true trajectory in x_t. In addition, using the Taylor-Young expansion (D.1) we have for the left term that:

$$\frac{F_\theta(x_{t_i+\Delta t}) - F_\theta(x_{t_i})}{\Delta t} = J_{F_\theta}(x_{t_i}) \cdot \frac{x_{t_i+\Delta t} - x_{t_i}}{\Delta t} + \frac{1}{\Delta t} O((x_{t_i+\Delta t} - x_{t_i})^2) \tag{17}$$

$$= J_{F_\theta}(x_{t_i}) \cdot F(x_{t_i}) + O(\Delta t) \tag{18}$$

as $x_{t_i+\Delta t} - x_{t_i} = \Delta t F(x_{t_i}) + O(\Delta t^2)$. We effectively compute a first order approximation of $J_{F_\theta}(x_t) \cdot F(x_t)$ and we recover an unsupervised version of $\mathcal{L}_{\text{Jacmatch}}$. The bias of the FD approximation could be reduced by using centered stencils in order to get a $O(\Delta t^2)$, while the choice of the FD order is to be balanced with the noise amplification that occurs when $\Delta t \to 0$ (caused by Δt terms in the denominator).

5 Results

Both experimental settings feature training on short rollout/short-horizon trajectories and evaluating on longer ones (annex F for the implementation details). The code is available on GitHub for both the offline and short-online experiments [1].

[1] `https://github.com/mayajanvier/jax-phynity`

5.1 Offline training

In this setting, models are trained offline and evaluated on long-term rollouts. Experiments are led on the Van der Pool oscillator: a classical benchmark for Neural ODE approaches presenting a non-linear dynamic with a transient phase before reaching a periodic attractor (annex F). We used 3 different seeds and report the mean long-term MSE accross them here.

Table 2 shows that in the plain offline setting, there are no major differences between using the exact F or the fitted spline when the Δt of the model is small, however the spline becomes more efficient when it increases. We hypothesise that this potentially smoother version of F may make the training objective easier to optimize and include a native regularization on the smoothness of the learned regularization. This can be linked to previous works [8, 18] that penalize high-order derivatives to prevent learning overly complex dynamics and encourage smoother paths.

Table 2: **Impact of spline fitting**

F / Δt	0.01	0.1
EXACT	1.354E-4	3.556E-4
SPLINE	1.458E-4	1.789E-4

Table 3: **Long-term MSE (mean accross 3 seeds).** All methods use spline fitting. Best is in bold.

METHOD	OFFLINE	ACCMATCH	JACMATCH	CFO
SCORE	1.789E-4	**8.330E-5**	1.180E-4	1.443E-4

When using spline fitting, AccMatch performs best, in particular outperforming CFO, while all methods outperform plain offline. CFO and JacMatch perform similarly as expected from eq. 11. We believe that the small difference between the two methods is due to the enhanced identifiability that comes when optimizing on the explicit Jacobian of the neural network to learn (akin to Hermite interpolation), although the losses' values will be similar.

5.2 Short-online training

We test our regularizations on common one-dimensional ODE and PDE benchmarks treated in other papers that cover a variety of dynamics: a first order non-chaotic autonomous system with Rigid Body [52] as well as a chaotic fourth-order partial derivative equation with Kuramoto-Sivashinksy [25, 26]. Both present either conservation laws or holonomic constraints that should be preserved during integration. As in White [52], we are interested in the relative error between the predicted state $x_\theta(t)$ and the ground truth $x(t)$ over time, averaged over many trajectories with various initial conditions, referred as $R_e(\theta, t)$, as well as temporal evolution of their conservative properties.

In our experiments, we first diagnose the difference between Neural ODEs trained on very short rollouts (producing exploding trajectories in the long-term) and one trained on longer rollouts (stable), with the classical trajectory loss (eq. 3), representing the gold standard. We then apply our Jacobian regularizations during training in the unstable setting (short rollouts), in order to show the improvement that our losses provide on the long-term stability of the system at inference: following [52], we train and validate on chunks of size N of short-horizon trajectories of duration T_{train} and evaluate on long-term trajectories of duration T_{test} (5 to 8 times longer than T_{train}). We compare them to a projection method SNODE [52] that achieves stable dynamics by projecting onto the manifold of algebraic constraints of the system (annex E.1). Implementation details are available in annex F.

Performing the validation on short-horizon and chunked trajectories is a particularly challenging setting for our experiments. Indeed, divergence is hard to predict when looking only at early step error, as it can happen very late during the unrolling. We chose this very restrictive protocol on purpose to be in the most realistic and reproducible scenarios possible (very expensive long-term rollouts). Finally, even though our losses require Jacobian-vector products, they avoid expensive long rollout backpropagation. In practice, training remained comparable to standard short-rollout Neural ODEs training (annex H). Additional results on the TwoBody problem experiment (annex G) confirm our results.

As expected from [9, 27], training on longer rollouts improves long-term stability, although several trajectories still diverge after 4000 rollout steps for Rigid Body (Figure 2, in orange). This enhanced stability also translates into better conservation of the Casimir and Mass properties (bottom rows 2, 3). As identified in section 2, it can be hard to reach an appropriate smoothness/expressiveness balance when penalizing only the norm of J_{F_θ} in $\mathcal{L}_{\text{Lip}}$ (green), especially when validating on very

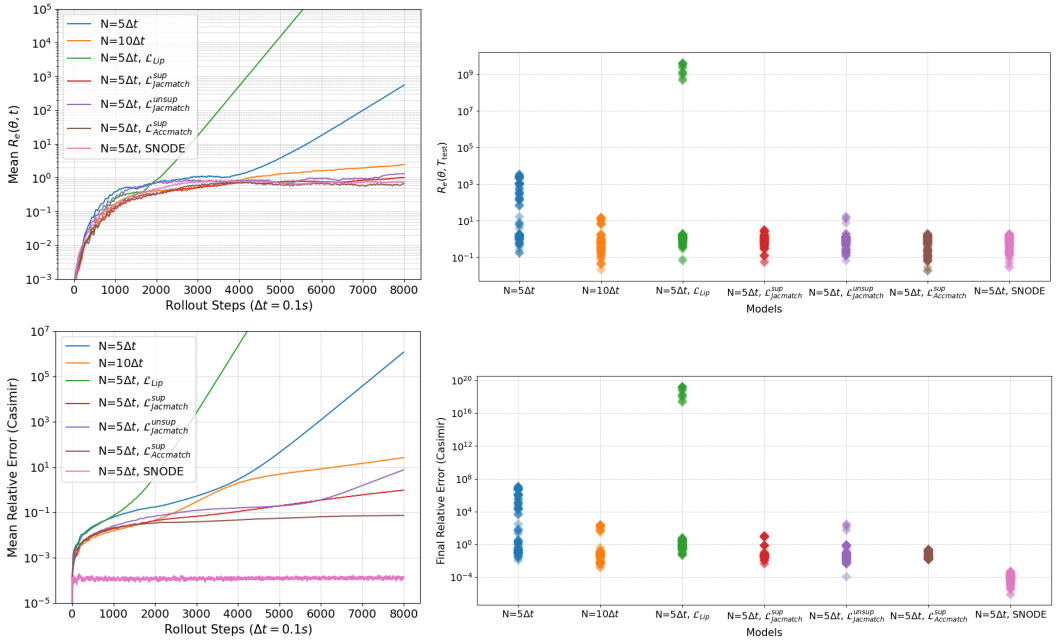

Figure 2: **Rigid Body.** For each row, we have the evolution of the relative error over time (left) and the distributions of final values (right). The top row represents the relative trajectory error $R_e(\theta, t)$ and the bottom row represents the relative conservation error (Casimir).

short rollouts as done here: this is the method that performs the worst for Rigid Body, although it is successful yet not the best for Kuramoto-Sivashinsky.

On the other hand, the $\mathcal{L}_{\text{Jacmatch}}$ regularization (extension of CFO), is successful in stabilizing the model both in supervised (red) and unsupervised (purple) settings for both physical systems. Surprisingly, the unsupervised version is even better than the supervised one for Kuramoto-Sivashinsky and even the best among all methods (Figure 3, left panels). This is similar to what we already observed on the impact of spline fitting in the offline experiments: the finite differences yields smoother values than the ones obtained with the true F (due to the solving of spatial operations such as u_{xx} or u_{xxxx}), facilitating training and enforcing a smoother learned representation. Our acceleration matching regularization $\mathcal{L}_{\text{Accmatch}}$ performs best in terms of trajectory error for RB, even outperforming the model trained on longer rollouts (Figure 2, in orange). It is also very good at conserving the Casimir property. The loss achieves relative error performances on par with the supervised $\mathcal{L}_{\text{Jacmatch}}$ and outperforms it in terms of mass conservation (Figure 3, bottom right) for KS. This is coherent with what we expected in section 4, as the two losses are quite close in the short rollout training context (small $\epsilon_\theta(x_{t_0+N\Delta t})$ when N is small).

The concurrent method, SNODE, is performing well on the small-scale system RB as in the original paper, and is the best at preserving the algebraic constraint which SNODE specifically targets. However, for KS that is a larger-scale system, the method becomes unstable: the constraint becomes too hard to maintain in the long-horizon, introducing unwanted instabilities that eventually lead to diverging trajectories. Our methods are more successful while providing a more general setup as they don't require the system to satisfy algebraic constraints.

6 Conclusion

In this work, we demonstrated that two seemingly distinct paradigms for stabilizing neural dynamical systems — Jacobian regularization [32] and interpolation-based training [12] — share a common theoretical foundation. By linearizing the unrolled trajectory loss, we showed that both approaches can be understood as approximations of the gradient terms that emerge from online training, specifically through their respective actions on the Jacobian norm and the Jacobian matching term. This perspective not only clarifies why these methods improve long-horizon stability, but also reveals them

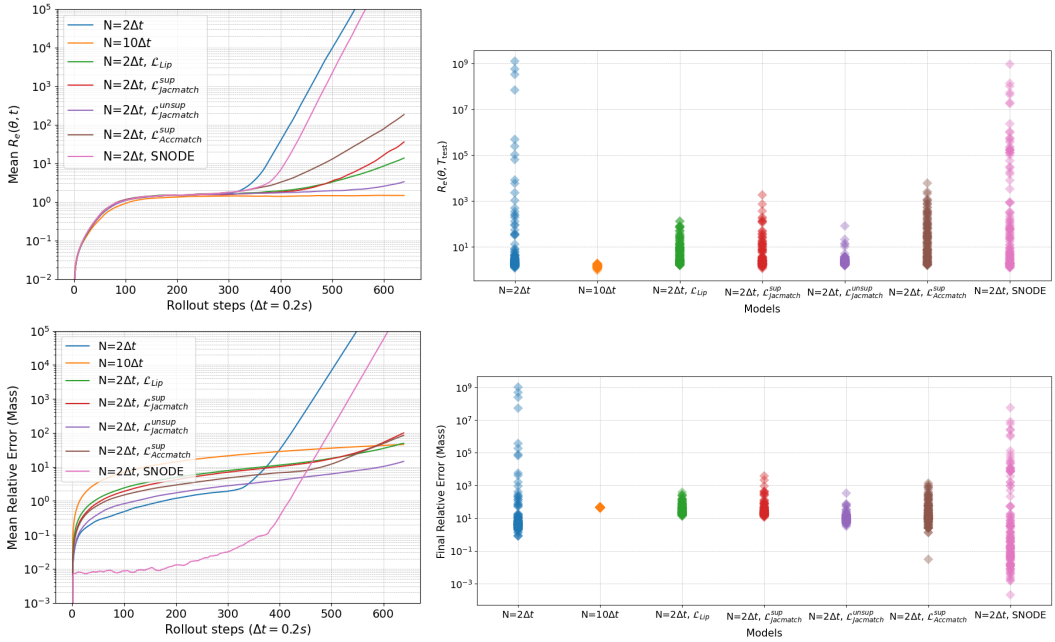

Figure 3: **Kuramoto-Sivashinsky.** For each row, we have the evolution of the relative error over time (left) and the distributions of final values (right). The top row represents the relative trajectory error $R_e(\theta, t)$ and the bottom row represents the relative conservation error (Mass).

as complementary. Furthermore, we found that the interpolation-based CFO training leaves out the correct term that focuses on acceleration matching.

Building on this analysis, we introduced a family of losses targeting these newly identified terms, notably generalizing CFO from an interpolation strategy into a regularization. We tested these losses in both offline/short-online as well as supervised/unsupervised settings, using either exact, spline fitted or finite differences based products. All proposed methods improve on the baseline, with different degrees of success. In particular, our JacMatch approach, which generalizes CFO, is effective in stabilizing long-term dynamics in exact, spline and FD formulations. Our newly derived term AccMatch performs best offline and similarly to JacMatch in the supervised short-online setting (relevant for hybrid parametrization of physical models [20] and fast emulation of expensive simulators), as for short rollouts the two losses remain close. The Jacobian regularization is successful for KS although the choice of the hyperparameter λ can remain difficult as observed for RB.

Crucially, the unsupervised formulations of the different losses, either obtained by fitting a spline as in CFO or derived entirely from finite differences of observed trajectories, perform competitively on the different physical systems. Because it requires neither access to the true dynamics nor expensive long rollouts, it is cheap to compute and broadly applicable, suggesting potential applicability to larger-scale systems where unrolled training remains prohibitively costly. In hybrid modelling specifically, online training is often not possible due to the non-differentiability of the solvers (ocean models [34]).

Our theoretical analysis opens the door to defining efficient losses or protocols to stabilize the long-term dynamics learned offline on larger scale systems. The current offline framework we proposed in this paper needs to be assessed more completely by performing evaluations of the losses on other systems (for example KS, RB). Another direct perspective of our study is to apply both our offline and short-online methods on larger scale systems such as Navier-Stokes 2D [55], quasi-geostrophic turbulence [9] and from other benchmarks [46] to improve the application of our method in practice as well as our understanding of it. Finally, the methods and code to reproduce them are publicly available on GitHub, and can be directly applied to any kind of autoregressive model.

7 Acknowledgments

We would like to thank Julien Salomon (Inria, Laboratoire Jacques Louis-Lions at Sorbonne University) for his help in the various proofs of this paper. We also thank the reviewers for their insightful comments that we believe greatly improved the quality of the paper. This work is part of the PhD of Maya Janvier, funded by Sorbonne University (ED 129) and EUR IPSL.

References

[1] Jordi Bolibar, Facundo Sapienza, Alban Gossard, Mathieu Le Séac'H, Lucille Gimenes, Vivek Gajadhar, Fabien Maussion, Ching-Yao Lai, Bert Wouters, and Fernando Pérez. ODINN.jl: Scientific machine learning glacier modelling. working paper or preprint, January 2026.

[2] Thomas Bolton and Laure Zanna. Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11(1):376–399, 2019.

[3] Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie Point Symmetry Data Augmentation for Neural PDE Solvers, 2022.

[4] Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers, 2023.

[5] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations, 2019.

[6] Thai Duong, Abdullah Altawaitan, Jason Stanley, and Nikolay Atanasov. Port-hamiltonian neural ODE networks on Lie groups for robot dynamics learning and control. *IEEE Transactions on Robotics*, 40:3695–3715, 2024.

[7] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks, 2023.

[8] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization, 2020.

[9] Hugo Frezat, Julien Le Sommer, Ronan Fablet, Guillaume Balarac, and Redouane Lguensat. A posteriori learning of quasi-geostrophic turbulence parametrization: an experiment on integration steps, 2021.

[10] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration. *Springer-Verlag, Berlin, 2nd edition,*, 2006.

[11] Florian Hess, Zahra Monfared, Manuel Brenner, and Daniel Durstewitz. Generalized teacher forcing for learning chaotic dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[12] Xianglong Hou, Xinquan Huang, and Paris Perdikaris. CFO: Learning continuous-time PDE dynamics via flow-matched neural operators, 2025.

[13] Zheyuan Hu, Zekun Shi, George Em Karniadakis, and Kenji Kawaguchi. Hutchinson Trace Estimation for high-dimensional and high-order Physics-Informed Neural Networks. *Computer Methods in Applied Mechanics and Engineering*, 424:116883, May 2024.

[14] Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, and Anima Anandkumar. Training Certifiably Robust Neural Networks with Efficient Local Lipschitz Bounds, 2021.

[15] M.F. Hutchinson. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, 1989.

[16] Matt Jordan and Alexandros G. Dimakis. Exactly Computing the Local Lipschitz Constant of ReLU Networks, 2021.

[17] Qiyu Kang, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable Neural ODE with Lyapunov-Stable Equilibrium Points for Defending Against Adversarial Attacks, 2021.

[18] Jacob Kelly, Jesse Bettencourt, Matthew James Johnson, and David Duvenaud. Learning differential equations that are easy to solve, 2020.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017.

[20] Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21), May 2021.

[21] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

[22] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2023.

[23] Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, András György, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C. Machado. Learning continually by spectral regularization, 2024.

[24] Jinghan Li, Yang Jin, Hao Jiang, Yadong Mu, Yang Song, and Kun Xu. Learning from next-frame prediction: Autoregressive video modeling encodes effective representations, 2025.

[25] Alec J. Linot, Joshua W. Burby, Qi Tang, Prasanna Balaprakash, Michael D. Graham, and Romit Maulik. Stabilized neural ordinary differential equations for long-time forecasting of dynamical systems. *Journal of Computational Physics*, 474:111838, February 2023.

[26] Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving Accurate Long Rollouts with Neural PDE Solvers, 2023.

[27] Bjoern List, Li-Wei Chen, Kartik Bali, and Nils Thuerey. Differentiability in Unrolled Training of Neural Physics Simulators on Transient Dynamics, 2024.

[28] Andreas Mayr, Sebastian Lehner, Arno Mayrhofer, Christoph Kloss, Sepp Hochreiter, and Johannes Brandstetter. Boundary graph neural networks for 3D simulations, 2023.

[29] Etienne Meunier, Said Ouala, Hugo Frezat, Julien Le Sommer, and Ronan Fablet. Towards fully differentiable neural ocean model with Veros, 2025.

[30] Laurent Meunier, Blaise Delattre, Alexandre Araujo, and Alexandre Allauzen. A Dynamical System Perspective for Lipschitz Neural Networks, 2022.

[31] Jonas M. Mikhaeil, Zahra Monfared, and Daniel Durstewitz. On the difficulty of learning chaotic dynamics with RNNs, 2022.

[32] Fengxiang Nie and Yasuhiro Suzuki. JAWS: Enhancing long-term rollout of neural PDE solvers via spatially-adaptive Jacobian regularization, 2026.

[33] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study, 2018.

[34] Said Ouala, Bertrand Chapron, Fabrice Collard, Lucile Gaultier, and Ronan Fablet. Online calibration of deep learning sub-models for hybrid numerical modeling systems, 2023.

[35] Avik Pal, Alan Edelman, and Christopher Rackauckas. Semi-Explicit Neural DAEs: Learning Long-Horizon Dynamical Systems with Algebraic constraints, 2025.

[36] Patricia Pauli, Anne Koch, Julian Berberich, Paul Kohler, and Frank Allgower. Training Robust Neural Networks Using Lipschitz Bounds. *IEEE Control Systems Letters*, 6:121–126, 2022.

[37] Chris Pedersen, Laure Zanna, and Joan Bruna. Thermalizer: Stable autoregressive neural emulation of spatiotemporal chaos, 2025.

[38] Christian Pedersen, Laure Zanna, Joan Bruna, and Pavel Perezhogin. Reliable coarse-grained turbulent simulations through combined offline learning and neural emulation, 2023.

[39] Moreno Pintore and Bruno Després. Computable Lipschitz Bounds for Deep Neural Networks, 2024.

[40] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples, 2018.

[41] Max Revay, Ruigang Wang, and Ian R. Manchester. Recurrent Equilibrium Networks: Flexible Dynamic Models with Guaranteed Stability and Robustness, 2023.

[42] Andrew Ross, Ziwei Li, Pavel Perezhogin, Carlos Fernandez-Granda, and Laure Zanna. Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, 15(1):e2022MS003258, 2023. e2022MS003258 2022MS003258.

[43] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks, 2020.

[44] Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation, 2019.

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations, 2021.

[46] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBENCH: An extensive benchmark for scientific machine learning, 2024.

[47] Louis Thiry, Long Li, and Etienne Mémin. *Modified (Hyper-)Viscosity for Coarse-Resolution Ocean Models*, page 273–285. Springer International Publishing, 2022.

[48] Akiyoshi Tomihari and Ryo Karakida. Recurrent self-attention dynamics: An energy-agnostic perspective from Jacobians, 2025.

[49] Ruigang Wang and Ian R. Manchester. Direct Parameterization of Lipschitz-Bounded Deep Networks, 2023.

[50] Zi Wang, Gautam Prakriya, and Somesh Jha. A Quantitative Geometric Approach to Neural-Network Smoothness, 2022.

[51] Theodor Westny, Arman Mohammadi, Daniel Jung, and Erik Frisk. Stability-Informed Initialization of Neural Ordinary Differential Equations, 2024.

[52] Alistair White, Niki Kilbertus, Maximilian Gelbrecht, and Niklas Boers. Stabilized Neural Differential Equations for Learning Dynamics with Explicit Constraints, 2024.

[53] Xiaoyu Xie, Saviz Mowlavi, and Mouhacine Benosman. Smooth and sparse latent dynamics in operator learning with Jerk regularization, 2024.

[54] Hanshu Yan, Jiawei Du, Vincent Y. F. Tan, and Jiashi Feng. On Robustness of Neural Ordinary Differential Equations, 2022.

[55] Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous PDE dynamics forecasting with implicit neural representations, 2023.

[56] Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, December 2021.

[57] Laure Zanna and Thomas Bolton. Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17):e2020GL088376, 2020. e2020GL088376 10.1029/2020GL088376.

[58] Wentao Zhang, Junliang Guo, Tianyu He, Li Zhao, Linli Xu, and Jiang Bian. Video in-context learning: Autoregressive transformers are zero-shot video imitators, 2025.

A Notations and definitions

Notations $\|\cdot\|_2$ and $\|\cdot\|_F$ denote respectively the L2 and Fröbenius norms.

Definition A.1. The Jacobian of a function $F : \mathbb{R}^n \to \mathbb{R}^m$ is defined as the matrix $J_F \in \mathbb{R}^{m\times n}$ of all its first-order partial derivatives:

$$J_F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \tag{19}$$

Definition A.2. A function F is Lipschitz continuous if there exists $L \in \mathbb{R}^+$ such that $\forall x, y \in \mathbb{R}^n \times \mathbb{R}^n$:

$$\|F(x) - F(y)\|_2 \leq L\|x - y\|_2.$$

L is called a Lipschitz constant of F and when it is differentiable, the optimal (smallest) Lipschitz constant is defined as $L_F = \sup_{x\in\mathbb{R}^n} \|J_F(x)\|_2$.

B Noise injection and Lipschitz control

Let $x \in \mathbb{R}^n$ be a training point and $F : \mathbb{R}^n \to \mathbb{R}^n$ a differentiable function.

Adversarial losses are build to resist to small perturbations or added noise (usually gaussian) $\epsilon \sim \mathcal{N}(0, I)$, that is to say to minimize $\mathcal{L}_{\text{noise}} = \mathbb{E}_\epsilon(\|F(x + \epsilon) - F(x)\|_2)$. If we use a first order Taylor-Young approximation D.1 in x, it becomes:

$$\begin{aligned} \mathcal{L}_{\text{noise}} &= \mathbb{E}_\epsilon(\|F(x) + J_F(x)\cdot\epsilon + o(\|\epsilon\|_2) - F(x)\|_2) \\ &\leq \mathbb{E}_\epsilon(\|J_F(x)\cdot\epsilon\|_2) + \mathbb{E}_\epsilon(o(\|\epsilon\|_2)) \quad \text{(triangular inequality and linearity of expectation)} \\ &= \|J_F(x)\|_F + \mathbb{E}_\epsilon(o(\|\epsilon\|_2)) \quad \text{(Hutchinson trace estimator)} \end{aligned}$$

Indeed, the Hutchinson trace estimator [15] allows us to compute the Fröbenius norm of matrices using only vector products, as done in [32, 13]. Let $v \sim P(v)$ be a random vector sampled from the $P(v)$ distribution such that $\mathbb{E}_{v\sim P(v)}[vv^T] = I$ (which is the case of the normal distribution):

$$\begin{aligned} \mathbb{E}_{v\sim P(v)}[\|J_F(x)\cdot v\|_2^2] &= \mathbb{E}_{v\sim P(v)}[Tr(\|J_F(x)\cdot v\|_2^2)] \quad (Tr(x) = x \text{ for } x \text{ scalar}) \\ &= \mathbb{E}_{v\sim P(v)}[Tr((J_F(x)\cdot v)^T(J_F(x)\cdot v))] \\ &= \mathbb{E}_{v\sim P(v)}[Tr(v^T\cdot(J_F(x)^T J_F(x))\cdot v)] \\ &= \mathbb{E}_{v\sim P(v)}[Tr((J_F(x)^T J_F(x))vv^T)] \quad (Tr(AB) = Tr(BA)) \\ &= Tr[J_F(x)^T J_F(x)\mathbb{E}_{v\sim P(v)}(vv^T)] \quad \text{(linearity expectation and trace)} \\ &= Tr(J_F(x)^T J_F(x)) \quad \text{as } \mathbb{E}_{v\sim P(v)}(vv^T) = I \\ &= \|J_F(x)\|_F^2 \quad \text{(definition)} \end{aligned} \tag{20}$$

This estimator is unbiased, and its variance is decreasing at the rate of $\frac{1}{V}$ the number of directions v used when estimated through a Monte-Carlo algorithm [13].

Furthermore, as $\frac{1}{n}\|J_F(x)\|_F \leq \|J_F(x)\|_2 \leq \|J_F(x)\|_F$ (norms are equivalent in finite dimension), we get that using $\mathcal{L}_{\text{noise}}$ is equivalent to penalizing the Fröbenius norm of J_F and ultimately the Lipschitz constant of F.

C Lipschitz constant and trajectory error

In this section, we denote $\|\cdot\|_2$ as $\|\cdot\|$ and L_{F_θ} as L_θ for lighter notations.

Let us suppose we have two dynamics, the true equation with exact resolution x and dynamics F, the other x_θ from the learned dynamic F_θ:

$$\begin{cases} \dot{x}(t) = F(x(t)) \\ x(t=0) = x_0 \end{cases} \quad \text{and} \quad \begin{cases} \dot{x_\theta}(t) = F_\theta(x_\theta(t)) \\ x_\theta(t=0) = x_0 \end{cases} \tag{21}$$

We make the assumption that F and F_θ are, respectively, L and L_θ-Lipschitz continuous. We define $\epsilon_\theta = F - F_\theta$ the offline error function. We prove the following inequalities:

Proposition C.1.

$$\forall t \geq 0, \|x(t) - x_\theta(t)\| \leq \exp(L_\theta t) \int_0^t \|\epsilon_\theta(x(s))\| \exp(-L_\theta s) ds, \tag{22}$$

and if ϵ_θ is bounded by $\|\epsilon_\theta\|_\infty$:

$$\begin{aligned}
\|x(t) - x_\theta(t)\| &\leq \|\epsilon_\theta\|_\infty \exp(L_\theta t) \int_0^t \exp(-L_\theta s) ds \\
&\leq \|\epsilon_\theta\|_\infty \exp(L_\theta t) \frac{1}{L_\theta} (1 - \exp(-L_\theta t)) \\
&\leq \frac{\|\epsilon_\theta\|_\infty}{L_\theta} (\exp(L_\theta t) - 1).
\end{aligned}$$

Proof. The second inequality comes from the first one by integrating the exponential. The first one uses Grönwall's lemma:

Lemma C.2. *Grönwall's inequality:*
Let $\psi \geq 0, \phi, K$ be continuous functions.
If

$$\forall t \geq t_0, \phi(t) \leq K(t) + \int_{t_0}^t \psi(s)\phi(s) ds,$$

then

$$\forall t \geq t_0, \phi(t) \leq K(t) + \int_{t_0}^t K(s)\psi(s) \exp\left(\int_s^t \psi(u) du\right) ds.$$

As x and x_θ are continuous, with the Fundamental theorem of calculus we can write:

$$\begin{aligned}
\|x(t) - x_\theta(t)\| &= \|x_0 + \int_0^t \dot{x}(s) ds - x_0 - \int_0^t \dot{x_\theta}(s) ds\| \\
&= \| \int_0^t F(x(s)) - F_\theta(x_\theta(s)) ds\| \quad \text{using eq. 21} \\
&= \| \int_0^t F(x(s)) - F_\theta(x(s)) + F_\theta(x(s)) - F_\theta(x_\theta(s)) ds\| \\
&\leq \int_0^t \|\epsilon_\theta(x(s))\| ds + \int_0^t \|F_\theta(x(s)) - F_\theta(x_\theta(s))\| ds \quad \text{(triangular inequality)} \\
\|x(t) - x_\theta(t)\| &\leq \int_0^t \|\epsilon_\theta(x(s))\| ds + L_\theta \int_0^t \|x(s) - x_\theta(s)\| ds \quad \text{as } F_\theta \text{ is } L_\theta\text{-Lipschitz.}
\end{aligned}$$

We can apply Grönwall's inequality with $\phi(t) = \|x(t) - x_\theta(t)\|$, $\psi(t) = L_\theta \geq 0$ and $K(t) = \int_0^t \|\epsilon_\theta(x(s))\| ds$:

$$\begin{aligned}
\|x(t) - x_\theta(t)\| &\leq \int_0^t \|\epsilon_\theta(x(s))\| ds + \int_0^t L_\theta \left(\int_0^s \|\epsilon_\theta(x(u))\| du\right) \exp\left(\int_s^t L_\theta du\right) ds \\
&\leq K(t) + L_\theta \exp(L_\theta t) \int_0^t \left(\int_0^s \|\epsilon_\theta(x(u)\| du\right) \exp(-L_\theta s) ds.
\end{aligned}$$

For the integral, we can use integration by parts (IPP), with

$$\begin{cases} h(s) = \int_0^s \|\epsilon_\theta(x(u))\| du = K(s) \\ \dot{h}(s) = \|\epsilon_\theta(x(s))\| \end{cases} \quad \text{and} \quad \begin{cases} g(s) = -\frac{1}{L_\theta} \exp(-L_\theta s) \\ \dot{g}(s) = \exp(-L_\theta s) \end{cases}$$

We have:

$$\begin{aligned}\int_0^t h(s)\dot{g}(s)ds &= [h(s)g(s)]_0^t - \int_0^t \dot{h(s)}g(s)ds \\ &= \left[K(s)\frac{-1}{L_\theta}\exp(-L_\theta s)\right]_0^t - \int_0^t \frac{-1}{L_\theta}\exp(-L_\theta s)\|\epsilon_\theta(x(s))\|ds \\ \int_0^t h(s)\dot{g}(s)ds &= \frac{-1}{L_\theta}\exp(-L_\theta t)K(t) + \frac{1}{L_\theta}\int_0^t \|\epsilon_\theta(x(s))\|\exp(-L_\theta s)ds\end{aligned}$$

Plugging it back to the main inequality,

$$\begin{aligned}\|x(t) - x_\theta(t)\| &\leq K(t) + L_\theta \exp\left(L_\theta t\right)\left(\int_0^t \left(\int_0^s \|\epsilon_\theta(x(u))\|du\right)\exp\left(-L_\theta s\right)ds\right) \\ &\leq K(t) + L_\theta\exp(L_\theta t)\left(\frac{-1}{L_\theta}\exp(-L_\theta t)K(t) + \frac{1}{L_\theta}\int_0^t \|\epsilon_\theta(x(s))\|\exp(-L_\theta s)ds\right) \\ &\leq K(t) - K(t) + \exp(L_\theta t)\int_0^t \|\epsilon_\theta(x(s))\|\exp(-L_\theta s)ds \\ &\leq \exp(L_\theta t)\int_0^t \|\epsilon_\theta(x(s))\|\exp(-L_\theta s)ds.\end{aligned}$$

□

D Effect of unrolling

We consider the same notations and definitions that in annex C. We suppose that F and F_θ are differentiable, in order to use the Taylor-Young formula:

Theorem D.1. *Taylor-Young formula*

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable k times in $a \in \mathbb{R}^n$, we have

$$f(x) = \sum_{j=0}^{k} \frac{f^{(j)}(a)}{j!}(x-a)^j + o\left((x-a)^k\right), \quad x \to a \tag{23}$$

D.1 Euler forward $x_t \neq x_t^\theta$

Let's derive equation 4. Starting from x_t, if the solver is Euler forward we have:

$$\begin{aligned}x_{t+2\Delta t} &= x_{t+\Delta t} + F(x_{t+\Delta t})\,\Delta t \\ x_{t+2\Delta t} &= x_t + F(x_t)\,\Delta t + F(x_t + F(x_t)\,\Delta t)\,\Delta t\end{aligned}$$

Similarly, starting from some x_t^θ with the same solver:

$$x_{t+2\Delta t}^\theta = x_t^\theta + F_\theta(x_t^\theta)\,\Delta t + F_\theta\big(x_t^\theta + F_\theta(x_t^\theta)\,\Delta t\big)\,\Delta t$$

Thus, the error between the two is:

$$\begin{aligned}x_{t+2\Delta t} - x_{t+2\Delta t}^\theta = x_t - x_t^\theta &+ \Delta t\big(F(x_t) - F_\theta(x_t^\theta)\big) \\ &+ \Delta t\big(F(x_t + F(x_t)\,\Delta t) - F_\theta\big(x_t^\theta + F_\theta(x_t^\theta)\,\Delta t\big)\big)\end{aligned}$$

Using Taylor-Young formula we get:

$$\begin{aligned}F(x_t) - F_\theta(x_t^\theta) &= F(x_t) - F_\theta(x_t) + J_{F_\theta}(x_t)\cdot(x_t - x_t^\theta) + o(\|x_t - x_t^\theta\|_2) \\ F(x_t + F(x_t)\,\Delta t) &= F(x_t) + J_F(x_t)\cdot F(x_t)\,\Delta t + o(\Delta t) \\ F_\theta\big(x_t^\theta + F_\theta(x_t^\theta)\Delta t\big) &= F_\theta(x_t) + J_{F_\theta}(x_t)\cdot\big(x_t^\theta + F_\theta(x_t^\theta)\,\Delta t - x_t\big) + o(\Delta t) + o(\|x_t - x_t^\theta\|_2) \\ &= F_\theta(x_t) + J_{F_\theta}(x_t)\cdot(x_t^\theta - x_t) + J_{F_\theta}(x_t)\cdot F_\theta(x_t^\theta)\,\Delta t + o(\Delta t) + o(\|x_t - x_t^\theta\|_2)\end{aligned}$$

Substituting back into equation 24 :

$$\begin{aligned}
x_{t+2\Delta t} - x^\theta_{t+2\Delta t} &= x_t - x^\theta_t + \Delta t\big(\epsilon_\theta(x_t) + J_{F_\theta}(x_t)\cdot(x_t - x^\theta_t) + o(\|x_t - x^\theta_t\|_2)\big) \\
&\quad + \Delta t\Big(F(x_t) + J_F(x_t)\cdot F(x_t)\,\Delta t - F_\theta(x_t) - J_{F_\theta}(x_t)\cdot(x^\theta_t - x_t) - J_{F_\theta}(x_t)\cdot F_\theta(x^\theta_t)\,\Delta t \\
&\quad + o(\Delta t) + o(\|x_t - x^\theta_t\|_2)\Big) \\
&= x_t - x^\theta_t + 2\,\Delta t\,\big(\epsilon_\theta(x_t) + J_{F_\theta}(x_t)\cdot(x_t - x^\theta_t) + o(\|x_t - x^\theta_t\|_2)\big) \\
&\quad + \Delta t^2\big(J_F(x_t)\cdot F(x_t) - J_{F_\theta}(x_t)\cdot F_\theta(x^\theta_t) + o(\|x_t - x^\theta_t\|_2)\big) \\
&\quad + o(\Delta t^2)
\end{aligned}$$

Now we expand the Δt^2 term:

$$\begin{aligned}
&J_F(x_t)\cdot F(x_t) - J_{F_\theta}(x_t)\cdot F(x_t) + J_{F_\theta}(x_t)\cdot F(x_t) - J_{F_\theta}(x_t)\cdot F_\theta(x^\theta_t) \\
&= J_{F-F_\theta}(x_t)\cdot F(x_t) + J_{F_\theta}(x_t)\cdot\big(F(x_t) - F_\theta(x^\theta_t)\big) \\
&= J_{\epsilon_\theta}(x_t)\cdot F(x_t) + J_{F_\theta}(x_t)\cdot\big(F(x_t) - F_\theta(x_t) + F_\theta(x_t) - F_\theta(x^\theta_t)\big) \\
&= J_{\epsilon_\theta}(x_t)\cdot F(x_t) + J_{F_\theta}(x_t)\cdot\epsilon_\theta(x_t) + J_{F_\theta}(x_t)\cdot\big(F_\theta(x_t) - F_\theta(x^\theta_t)\big) \qquad (24)
\end{aligned}$$

Substituting everything back, we finally get:

$$\begin{aligned}
x_{t+2\Delta t} - x^\theta_{t+2\Delta t} &= (x_t - x^\theta_t) \\
&\quad + 2\,\Delta t\,\big(\epsilon_\theta(x_t) + J_{F_\theta}(x_t)\cdot(x_t - x^\theta_t) + o(\|x_t - x^\theta_t\|_2)\big) \\
&\quad + \Delta t^2\,\big[J_{\epsilon_\theta}(x_t)\cdot F(x_t) + J_{F_\theta}(x_t)\cdot\epsilon_\theta(x_t) \\
&\quad + J_{F_\theta}(x_t)\cdot\big(F_\theta(x_t) - F_\theta(x^\theta_t)\big) + o(\|x_t - x^\theta_t\|_2)\big] \\
&\quad + o(\Delta t^2) \qquad (25)
\end{aligned}$$

D.2 Euler forward $x_t = x^\theta_t$

Now if $x_t = x^\theta_t$ (in the training context), we have:

$$\begin{aligned}
x_{t+2\Delta t} - x^\theta_{t+2\Delta t} &= 2\,\Delta t\,\epsilon_\theta(x_t) + \Delta t^2\,\big(J_F(x_t)\cdot F(x_t) - J_{F_\theta}(x_t)\cdot F_\theta(x_t)\big) + o(\Delta t^2) \qquad (26) \\
&= 2\,\Delta t\,\epsilon_\theta(x_t) + \Delta t^2\,\eta_\theta(x_t) + o(\Delta t^2) \qquad (27)
\end{aligned}$$

We can generalize the formula to N steps by induction. We define:

$$H_N : \begin{cases} x_{t+N\Delta t} &= x_t + N\Delta t F(x_t) + \frac{N(N-1)}{2} J_F(x_t)\cdot F(x_t) + o(\Delta t^2) \\ x^\theta_{t+N\Delta t} &= x_t + N\Delta t F_\theta(x_t) + \frac{N(N-1)}{2} J_{F_\theta}(x_t)\cdot F_\theta(x_t) + o(\Delta t^2) \end{cases} \qquad (28)$$

$$\text{such that } x_{t+N\Delta t} - x^\theta_{t+N\Delta t} = N\Delta t\epsilon_\theta(x_t) + \frac{N(N-1)}{2}\eta_\theta(x_t) + o(\Delta t^2), \quad n \geq 2 \qquad (29)$$

The formula is already true for $N = 2$, we suppose H_N is true for a certain $N \geq 2$. We have under Euler forward:

$$x_{t+(N+1)\Delta t} - x^\theta_{t+(N+1)\Delta t} = x_{t+N\Delta t} - x^\theta_{t+N\Delta t} + \Delta t(F(x_{t+N\Delta t}) - F_\theta(x^\theta_{t+N\Delta t}))$$

and under H_N:

$$\begin{aligned}
x_{t+N\Delta t} &= x_t + F(x_t)\Delta t + F(x_t + F(x_t)\Delta t)\Delta t + ... \\
&= x_t + \delta_N \\
\delta_N &= N\Delta t F(x_t) + \frac{N(N-1)}{2}\Delta t^2 J_F(x_t)\cdot F(x_t) + o(\Delta t^2) \\
&= N\Delta t F(x_t) + o(\Delta t)
\end{aligned}$$

By using Taylor's formula around x_t:

$$\begin{aligned} F(x_{t+N\Delta t}) &= F(x_t + \delta_N) \\ &= F(x_t) + J_F(x_t) \cdot \delta_N + o(\|\delta_N\|_2^2) \\ &= F(x_t) + J_F(x_t) \cdot (N\Delta F(x_t) + o(\Delta t)) + o(\Delta t^2) \\ &= F(x_t) + N\Delta t J_F(x_t) \cdot F(x_t) + o(\Delta t) \end{aligned}$$

and

$$\begin{aligned} x_{t+(N+1)\Delta t} &= x_t + (N+1)\Delta t F(x_t) + (\frac{N(N-1)}{2} + N) J_F(x_t) \cdot F(x_t) + o(\Delta t^2) \\ &= x_t + (N+1)\Delta t F(x_t) + \frac{N(N+1)}{2} J_F(x_t) \cdot F(x_t) + o(\Delta t^2) \end{aligned}$$

The coefficients are the same for $x^\theta_{t+(N+1)\Delta t} = x_t + F_\theta(x_t)\Delta t + F_\theta(x_t + F_\theta(x_t)\Delta t)\Delta t + ...$, only with F_θ since $x_t = x^\theta_t$, and we get H_{N+1}:

$$x_{t+(N+1)\Delta t} - x^\theta_{t+(N+1)\Delta t} = (N+1)\epsilon_\theta(x_t) + \frac{N(N+1)}{2}\eta_\theta(x_t) + o(\Delta t^2) \tag{30}$$

D.3 RK2 and higher-order explicit schemes

The precedent reasoning can be applied to other integration schemes. For RK2, we have

$$\begin{aligned} x_{t+\Delta t} &= x_t + \Delta t F(x_t + \frac{\Delta t}{2} F(x_t)) \\ &= x_t + \Delta t (F(x_t) + J_F(x_t) \cdot \frac{\Delta t}{2} F(x_t) + o(\Delta t)) \\ &= x_t + \Delta t F(x_t) + \frac{\Delta t^2}{2} J_F(x_t) \cdot F(x_t) + o(\Delta t^2) \end{aligned}$$

and

$$x_{t+\Delta t} - x^\theta_{t+\Delta t} = \Delta t \epsilon_\theta(x_t) + \frac{\Delta t^2}{2}\eta_\theta(x_t) + o(\Delta t^2) \tag{31}$$

A similar induction than the previous section will give the general formula:

$$x_{t+N\Delta t} - x^\theta_{t+N\Delta t} = N\Delta t \epsilon_\theta(x_t) + \frac{N^2\Delta t^2}{2}\eta_\theta(x_t) + o(\Delta t^2) \tag{32}$$

For higher order Runge-Kutta methods, if we limit ourselves to $o(\Delta t^2)$ order terms, RK4 will give the same terms as RK2 (but differ when we look at higher order terms). We observe that when using RK2, eq. 31 recovers the formula we would get from applying Taylor-Young to $x(t + \Delta t)$ in t at the order 2.

D.4 Unrolled gradients for RK2

We can now analyze the regularity needed for F, F_θ and their derivatives in order for our approximation to be close to the online training loss.

We denote $r_\theta = \Delta t \epsilon_\theta(x_t) + \frac{\Delta t^2}{2}\eta_\theta(x_t)$ so that we derive from eq. 31 the approximated loss $\mathcal{L}_{\text{approx}}(\theta) = \|r_\theta\|_2^2$, that aims at approaching $\mathcal{L}^{N=1}_{\text{traj}}(\theta) = \|x_{t+\Delta t} - x^\theta_{t+\Delta t}\|_2^2$ in the case where $x^\theta_t = x_t$. From eq. 31, we define $\delta_\theta = (x_{t+\Delta t} - x^\theta_{t+\Delta t}) - r_\theta$, so we have $\mathcal{L}^{N=1}_{\text{traj}}(\theta) = \|r_\theta + \delta_\theta\|_2^2$. With our previous assumptions (F, F_θ differentiable), we know that $\delta_\theta = o(\Delta t^2)$.

Assumptions For θ in a compact set Θ, and K a compact containing x_t and the trajectories starting from it when F_θ spans Θ, we suppose that:

- (H1) F is C^2 in x.
- (H2) F_θ is C^3 in x, and $D^k_x F_\theta$ ($k \leq 2$) is C^1 in θ with jointly continuous derivatives. Hence $D^3 F_\theta$, $\partial_\theta F_\theta$, $\partial_\theta D F_\theta$, $\partial_\theta D^2 F_\theta$ are continuous, thus bounded on $K \times \Theta$.

where the D, D^2 notations correspond to first, second order derivatives.

Characterization of δ_θ Since F is C^2, x is C^3 and we have with Taylor-Young:

$$x_{t+\Delta t} = x_t + \Delta t F(x_t) + \frac{\Delta t^2}{2} J_F(x_t) \cdot F(x_t) + r_F, \quad r_F = \frac{\Delta t^3}{6} x'''(t) + o(\Delta t^3) = O(\Delta t^3). \tag{33}$$

For $x^\theta_{t+\Delta t}$, since we will need to differentiate δ_θ through θ, we need to write the RK2 update using Taylor's theorem with first order integral remainder:

$$\begin{aligned} x^\theta_{t+\Delta t} &= x_t + \Delta t F_\theta(x_t + \frac{\Delta t}{2} F_\theta(x_t)) \\ &= x_t + \Delta t (F_\theta(x_t) + \frac{\Delta t}{2} J_{F_\theta}(x_t) \cdot F_\theta(x_t) + r_1^\theta) \\ &= x_t + \Delta t F_\theta(x_t) + \frac{\Delta t^2}{2} J_{F_\theta}(x_t) \cdot F_\theta(x_t) + \Delta t r_1^\theta \end{aligned} \tag{34}$$

with r_1^θ the first order integral remainder. Writing $h_\theta = \frac{\Delta t}{2} F_\theta(x_t)$, it is defined as:

$$\begin{aligned} r_1^\theta &= \int_0^1 (1-s) D^2 F_\theta(x_t + s h_\theta)[h_\theta, h_\theta] ds \\ &= \frac{\Delta t^2}{4} \int_0^1 (1-s) D^2 F_\theta(x_t + s h_\theta)[F_\theta(x_t), F_\theta(x_t)] ds. \end{aligned} \tag{35}$$

We then get a formula for $\delta_\theta = r_F - \Delta t r_1^\theta$.

Order of the gradient approximation We derive the gradient of $\mathcal{L}^{N=1}_{\text{traj}}$:

$$\begin{aligned} \frac{\partial \mathcal{L}^{N=1}_{\text{traj}}(\theta)}{\partial \theta} &= \frac{\partial \|r_\theta + \delta_\theta\|_2^2}{\partial \theta} \\ &= 2(r_\theta + \delta_\theta)^T \frac{\partial}{\partial \theta}(r_\theta + \delta_\theta) \\ &= 2 r_\theta^T \frac{\partial r_\theta}{\partial \theta} + 2\delta_\theta^T \frac{\partial r_\theta}{\partial \theta} + 2 r_\theta^T \frac{\partial \delta_\theta}{\partial \theta} + 2\delta_\theta^T \frac{\partial \delta_\theta}{\partial \theta} \\ &= \frac{\partial \mathcal{L}_{\text{approx}}}{\partial \theta} + 2\delta_\theta^T \frac{\partial r_\theta}{\partial \theta} + 2 r_\theta^T \frac{\partial \delta_\theta}{\partial \theta} + 2\delta_\theta^T \frac{\partial \delta_\theta}{\partial \theta}. \end{aligned} \tag{36}$$

Let us bound the different terms:

- $r_\theta = \Delta t(\epsilon_\theta(x_t) + \frac{\Delta t}{2}\eta_\theta(x_t))$ is clearly $O(\Delta t)$;
- $\dfrac{\partial r_\theta}{\partial \theta} = -\Delta t\Big(\partial_\theta F_\theta + \dfrac{\Delta t}{2}\,\partial_\theta\big(J_{F_\theta} F_\theta\big)\Big)$, with $\partial_\theta(J_{F_\theta}F_\theta) = (\partial_\theta J_{F_\theta})F_\theta + J_{F_\theta}\partial_\theta F_\theta$, all evaluated at x_t. By assumption (H2) this is $O(\Delta t)$;
- $\|r_1^\theta\| \le \frac{\Delta t^2}{8} \sup_K \|D^2 F_\theta\| \|F_\theta(x_t)\|^2$, so $r_1^\theta = O(\Delta t^2)$, and since $r_F = O(\Delta t^3)$, we have that $\delta_\theta = r_F - \Delta t r_1^\theta = O(\Delta t^3)$;
- r_F does not depend on θ, so $\dfrac{\partial \delta_\theta}{\partial \theta} = -\Delta t\, \partial_\theta r_1^\theta$. Under (H2) we can use Leibniz rule to differentiate under the integral in eq. 35, and using the symmetry of $D^2 F_\theta$, we get

$$\begin{aligned} \partial_\theta r_1^\theta = \frac{\Delta t^2}{4} \int_0^1 (1-s) \Big\{ &(\partial_\theta D^2 F_\theta)(x_t + s h_\theta)[F_\theta, F_\theta] + s\,\frac{\Delta t}{2}\, D^3 F_\theta(x_t + s h_\theta)[\partial_\theta F_\theta, F_\theta, F_\theta] \\ &+ 2\, D^2 F_\theta(x_t + s h_\theta)[\partial_\theta F_\theta, F_\theta] \Big\} ds, \end{aligned} \tag{37}$$

  where $F_\theta, \partial_\theta F_\theta$ are evaluated at x_t. The three terms under the integral are bounded under (H2), and are respectively $O(\Delta t^2)$, $O(\Delta t^3)$, $O(\Delta t^2)$. Hence $\partial_\theta r_1^\theta = O(\Delta t^2)$ and $\dfrac{\partial \delta_\theta}{\partial \theta} = O(\Delta t^3)$.

Conclusion Combining the bounds,

$$\left\|2\delta_\theta^T \frac{\partial r_\theta}{\partial \theta}\right\| = O(\Delta t^4), \left\|2r_\theta^T \frac{\partial \delta_\theta}{\partial \theta}\right\| = O(\Delta t^4), \left\|2\delta_\theta^T \frac{\partial \delta_\theta}{\partial \theta}\right\| = O(\Delta t^6).$$

Therefore,

$$\left\|\frac{\partial \mathcal{L}_{\text{traj}}^{N=1}}{\partial \theta} - \frac{\partial \mathcal{L}_{\text{approx}}}{\partial \theta}\right\| \leq C\,\Delta t^4. \tag{38}$$

Since $\frac{\partial \mathcal{L}_{\text{approx}}}{\partial \theta} = 2r_\theta^T \frac{\partial r_\theta}{\partial \theta} = O(\Delta t^2)$, the gradient of $\mathcal{L}_{\text{approx}}$ approximates that of $\mathcal{L}_{\text{traj}}^{N=1}$ with relative error $O(\Delta t^2)$.

E More on the competing methods

E.1 SNODE [52]

The idea of SNODE is to directly modify the Neural ODE $\frac{du}{dt} = f_\theta(u)$ with a term that "nudges" back the trajectory to the manifold of algebraic constraint $\mathcal{M} = \{u/g(u) = 0\}$:

$$\frac{du}{dt} = f_\theta(u) - \gamma F(u)g(u) \tag{39}$$

Equation 39 is called the Stabilized NODE. They prove that the manifold $\mathcal{M}$ is asymptotically Lyapunov-stable in the SNODE under certain constraints for the stabilization matrix F, which they choose to be the Moore-Penrose pseudoinverse of the Jacobian of g at u. Pal et al. (2025) [35] showed that this setting was equivalent to only doing the first iteration of the projection (solved as a fixed point problem) onto the manifold.

This approach is limited to systems that respect algebraic constraints that are verified for each timestep (ex: a system under a temporal forcing). Secondly, when the system is chaotic like KS, it becomes hard to find a value for γ that allows the stability criteria to be valid throughout long rollouts (see eq.10 in the original paper and our experiment in this paper). Another limitation comes from the computational cost of the pseudoinverse in $O(m^2 n)$ (where m is the number of constraints and n is the dimension of the problem) which becomes prohibitively expensive to compute in higher dimensions, whereas our methods only require evaluations of F_θ and F.

E.2 JAWS [32]

JAWS is cited in this paper as a representative of the Jacobian regularization methods, that typically involve the core $\mathcal{L}_{\text{Lip}} = \|J_{F_\theta}(x_t)\|_2^2$ term obtained via the Hutchinson estimator. The JAWS loss is :

$$\mathcal{L} = \sum_{x\in\Omega} \frac{1}{2} e^{-s1} \|x_{t+1} - x_{t+1}^\theta\|^2 + \frac{1}{2} e^{-s2} \|J_{F_\theta}(x)\|_F^2 + \frac{1}{2}(s_1 + s_2) + \lambda \sum_{k=2}^{K} \|x_{t+k} - x_{t+k}^\theta\|^2 \tag{40}$$

It combines the trajectory loss (term weighted by λ) and an adaptative reconstruction/Jacobian regularization loss with balanced s_1, s_2 weights. The method is equivalent to $\mathcal{L}_{\text{Lip}}$ except for the adaptative coefficient strategy.

E.3 CFO [12]

CFO is a fully offline training method that belongs to interpolation based/data augmentation strategies, and relies on 3 concepts:

- building a spline approximation of F (done once) based on FD of trajectory points
- super-sample (data augmentation) using this spline during training: introduces the JacMatch regularization as demonstrated in eq. 11
- noise scheduler: the points are sampled along a noisy path (path = spline + $\gamma(t)\epsilon$ with $\epsilon \approx N(0, I)$). This penalises $||J_{\epsilon_\theta}||_F$ (Hutchinson) thus bringing the Lipschitz constants of F, F_θ closer (annex B)

Its loss is:

$$\mathcal{L}_{\text{CFO}}(\theta) = \sum_{X \in \mathcal{T}_\text{I}} \|F_\theta(x(t_i)) - \partial_t I(t_i, x(t_i))\|_2^2 \tag{41}$$

where I is the fitted spline and $\mathcal{T}_\text{I}$ the ensemble of data points sampled along I. The novelty in CFO lies mostly in this fitting+super-sampling that we linked to JacMatch in 11. The cost of the method can grow with the number of points used for the upsampling.

F Implementation details

This section presents the different parameters and protocols to reproduce our experiments. The code is implemented in jax and available on GitHub.

F.1 Data generation

Van der Pol The Van der Pol (VdP) oscillator is a non-conservative, oscillating system with non-linear damping. It evolves in time according to the second order differential equation:

$$\frac{dx^2}{dt^2} - \mu(1 - x^2)\frac{dx}{dt} + x = 0 \tag{42}$$

with $\mu = 1$. The initial conditions $x_0 = (u_0, v_0)$ are randomly sampled from the uniform distribution between $[-2.5, 2.5]^2$, and the ground truth trajectories simulated with $\Delta t = 0.01s$. The majority of the results are reported for a model timestep of $0.1s$, 10 different trajectories for each train/val/test sets and $T_{\text{train}} = T_{\text{test}} = 10s$.

Rigid Body The Rigid Body (RB) equations describe the evolution of the angular momentum vector X of a rigid body under Euler's motion laws [10]:

$$U = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \dot{U} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1/I_1 \\ y_2/I_2 \\ y_3/I_3 \end{pmatrix} \tag{43}$$

with I_1, I_2, I_3 the principal moments of inertia, where the coordinate axes are the principal axes of the body and the origin of the coordinates fixed at the center of mass of the body [52]. The Casimir function $C(U) = \frac{1}{2}(y_1^2 + y_2^2 + y_3^2)$ is conserved along the trajectory and constitutes a holonomic constraint. Similarly to White [52], we train on 40 trajectories with initial conditions $X_0 = (\cos\phi, 0, \sin\phi)$ with ϕ sampled from $\mathcal{U}(0.5, 1.5)$. We use data from trajectories of $T_{\text{train}} = 15s$ integrated from F_{RB} with a time-step $\Delta t = 0.01s$. We then divide these trajectories into non-overlapping chunks of $N = 5\Delta t$ length. The ground truth and the models both use RK4 but with different time-steps: $\Delta t_{\text{model}} = 10\Delta t = 0.1s$ for faster training. We then evaluate the long-term stability of our models on 100 test trajectories of duration $T_{\text{test}} = 800s$. We use the same parameter values $(I_1, I_2, I_3) = (1.6, 1.0, 2/3)$ as in White [52].

Kuramoto-Sivashinksy The 1D Kuramoto-Sivashinsky (KS) equation is a fourth order chaotic PDE describing the instabilities of a laminar flame front:

$$U_t + U_{xx} + U_{xxxx} + UU_x = 0 \tag{44}$$

Apart from being non linear, the biharmonic term is particularly difficult to model. Along a trajectory, we have $\int_0^L U(x,t) = 0$, L being the length of the spatial domain with periodic boundary conditions. To generate ground truth data, we follow the setups of [3] and [26], using a uniform discretization of the domain into 256 points. We do not sample across different L and Δt but keep them fixed at $L = 64, \Delta t = 0.2$, and generate 512 training trajectories of duration $T_{\text{train}} = 140\Delta t$. We alter the generation script of [3] by fixing $\Delta t = 0.2s$ and $L = 64$, instead of varying them. The code can then be used to produce our data with the following arguments:

- Training set: `-train_samples=512 -nt=500`
- Validation set: `-valid_samples=128 -nt=500`
- Test set: `-test_samples=128 -nt=1000 -nt_effective=640 -end_time=200`

As in [26], the first 360 steps are truncated and considered as a warmup for the solver (i.e., $t_0 = 360\Delta t$). The initial conditions are sampled from a distribution over truncated Fourier series with random coefficients $\{A_m, l_m, \phi_m\}$ as $U_0(x) = \sum_{m=1}^{10} A_m \sin(2\pi l_m x/L + \phi_m)$. The integration of the ground truth data is done with an implicit Runge-Kutta integration method of the Radau IIA family (order 5) [3], and our models use the RK4 solver with the same Δt. Finally, as for our previous experiments, we divide the training trajectories into non-overlapping chunks of $N = 2\Delta t$ steps and evaluate the long-term stability on 128 test trajectories of duration $T_{\text{test}} = 640\Delta t$.

F.2 Experimental setting

Models are trained using Titan RTX GPU (cuda 13.2), each training taking approximately one hour to be achieved. The models for the short-online experiments are trained using one seed, while the models for the online experiments are trained using 3 different seeds.

We use the MLP architecture from [56] for the Neural ODEs (i.e., for F_θ). It is composed of three linear layers with activation ReLU and hidden size 200 for RB and KS, hidden size 64 for Van der Pol. All Neural ODEs use RK4 as integration scheme as well, and are trained using Adam optimizer [19] with learning rate $l_r = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the $\mathcal{L}_{\text{Lip}}$ loss, we sample $V = 10$ directions $\{v_i\}_{i=1}^V$ from $\mathcal{N}(0, I)$ at each epoch. Table 4 summarizes the parameters of the different models and experiments.

Table 4: Summary of model and experiment parameters

PARAMETERS	RIGID BODY	KURAMOTO-SIVASHINSKY	VAN DER POL
U DIMENSION	3	256	2
$\Delta t/\Delta t_{\text{MODEL}}$ (S)	0.01 / 0.1	0.2 / 0.2	0.01/ 0.1 OR 0.01
$T_{\text{TRAIN}}/T_{\text{VAL}}/T_{\text{TEST}}$ (S)	15 / 15 / 800	28 / 28 / 128	10/10/10
$N_{\text{TRAIN}}/N_{\text{VAL}}/N_{\text{TEST}}$ (S)	5 / 5 / 8000	2 / 2 / 640	1/1/100 OR 1000
NB TRAJ TRAIN/VAL / TEST	40 / 40 / 100	512 / 128 / 128	10/10/10
NB EPOCHS	4500	1000	10000
BATCH SIZE	40	128	100
SEEDS	1	1	3

The best model is determined based on the mean trajectory loss on a validation set of chunked trajectories. The optimal regularization coefficient λ or γ is determined by grid-search for each experiment, also based on the same metric for the last epoch, then take the best model for this metric of this λ/γ run. Figure 4 shows its evolution across the tested hyperparameters λ, between 1E-2 to 1E-13 for RB and 1E-3 to 1E-13 for KS. The ranges of λ for SNODE are the ones from White et al. [52].

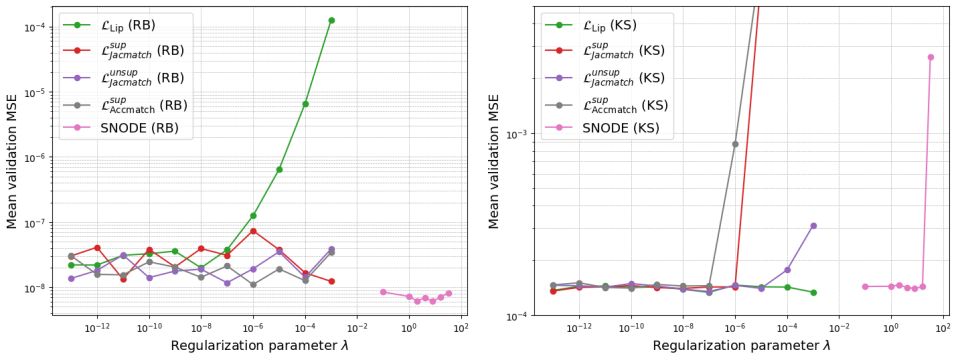

Figure 4: Mean MSE on validation set for the optimization of the λ parameter, for the Rigid Body Problem (left) and KS (right) experiments.

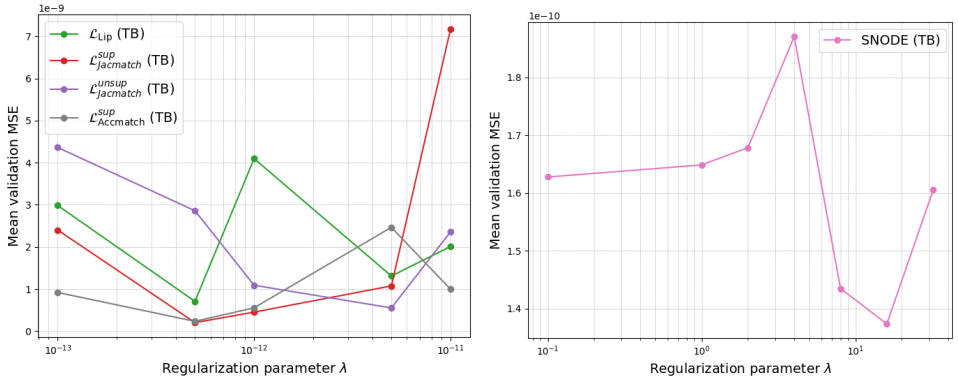

Figure 5: Mean MSE on validation set for the optimization of the λ parameter, for the Two-Body Problem (Right: SNODE, Left: the other methods)

G Additional experiment: Two-Body Problem

Two-Body Problem The Two-Body (TB) Problem describes the motion of two bodies attracting each other with a force inversely proportional to their squared distance, such as the gravitational force [10]. It is described by the following system of equations:

$$U = \begin{pmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{pmatrix}, \dot{U} = \begin{pmatrix} \dot{x} \\ \dot{y} \\ \frac{-x}{(x^2+y^2)^{3/2}} \\ \frac{-y}{(x^2+y^2)^{3/2}} \end{pmatrix} \tag{45}$$

with a body fixed at the origin and (x, y) the Cartesian coordinates of the other body in the orbit plane [52]. The angular momentum L_0 is conserved here: $x\dot{y} - \dot{x}y = L_0$ for $(x, y) \in \mathcal{T}_F^2$. As in White [52] experiments, we train on 40 trajectories with initial conditions $X_0 = (1 - e, 0, 0, \sqrt{\frac{1+e}{1-e}})$ with the eccentricity e sampled from $\mathcal{U}(0.5, 0.7)$. To reach a challenging context for Neural ODEs, we use data from trajectories of $T_{\text{train}} = 8s$ which is less than a full period, integrated from F_{TB} with a time-step $\Delta t = 0.01s$. We then divide the training trajectories into non-overlapping chunks of $N = 2\Delta t$ length. The ground truth and the models both use fourth order Runge-Kutta (RK4) with the same time-step Δt as integration scheme to produce their trajectories. We then evaluate the long-term stability of our models on 100 test trajectories of duration $T_{\text{test}} = 100s$.

Table 5 summarizes the parameters of this experiment and Figure 5 shows the results of the gridsearch performed for λ in $\{1e-11, 5e-12, 1e-12, 5e-13, 1e-13\}$.

Table 5: Summary of model and experiment parameters

PARAMETERS	TWO-BODY
U DIMENSION	4
$\Delta t / \Delta t_{\text{MODEL}}$ (S)	0.01 / 0.01
$T_{\text{TRAIN}}/T_{\text{VAL}}/T_{\text{TEST}}$ (S)	8 / 8 / 100
$N_{\text{TRAIN}}/N_{\text{VAL}}/N_{\text{TEST}}$	2 / 2 / 10000
NB TRAJ TRAIN/VAL / TEST	40 / 40 / 100
NB EPOCHS	4500
BATCH SIZE	40

Results For this periodic system, all of the methods succeed in stabilizing the $N = 2\Delta t$ baseline, with SNODE remaining the state-of-the-art model in this case, followed by the supervised JacMatch. The unsupervised version of $\mathcal{L}_{\text{Jacmatch}}$ only degrades a little bit the supervised version.

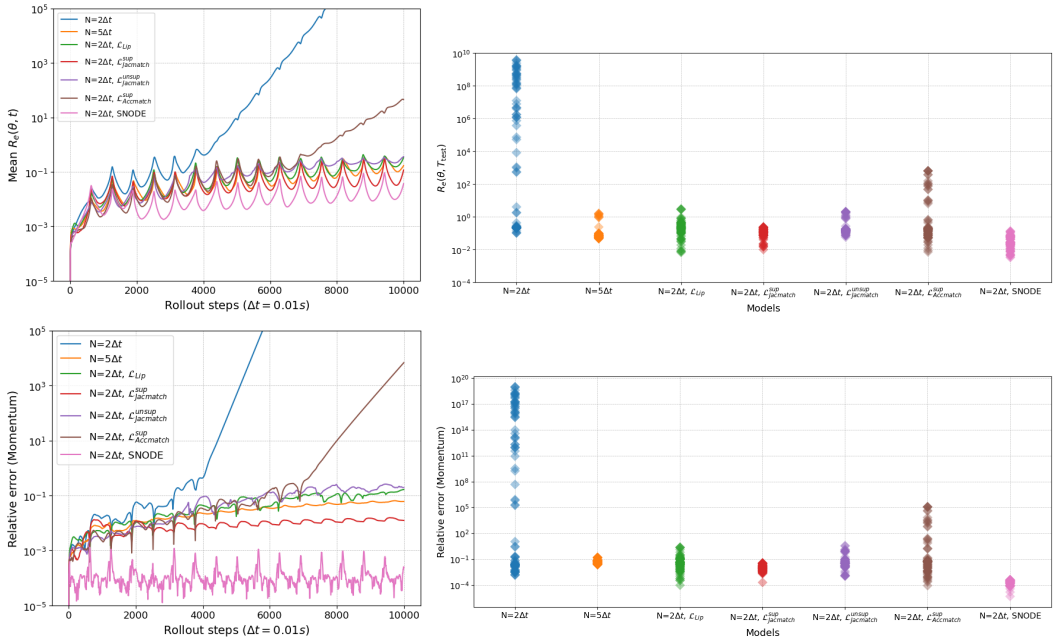

Figure 6: **Two Body.** For each row, we have the evolution of the relative error over time (left) and the distributions of final values (right). The top row represents the relative trajectory error $R_e(\theta, t)$ and the bottom row represents the relative conservation error (Momentum).

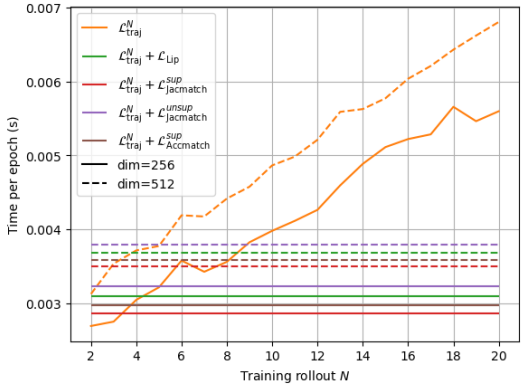

Figure 7: Computational cost on Kuramoto-Sivashinsky averaged over 1000 runs. For regularized methods, the cost for $N = 2$ is shown to compare with the $N = 10$ baseline.

H Computational cost of methods

To compute the cost of our methods, we averaged the cost of single epochs over 1000 runs. We compare the regularized training losses for $N = 2$ as in the main paper. $\mathcal{L}_{\text{Lip}}$ is computed with $V = 10$. We also analyze the cost for larger-scale systems by testing of KS with dimensions 256 and 512.

In Figure 7, we first observe the computational bottleneck of unrolling [27]: the longer the training rollout, the larger the computational cost, and even more when going to higher dimensional systems. Being able to produce stable simulations when training on short rollouts appears very advantageous.

Furthermore, all of the methods only introduce a small computational overhead, while remaining cheaper than training on $N = 10$ (stable baseline), for both dimensions.

I Trajectories

We add some plots of the predicted trajectories to have a look at their qualitative aspect, in Figure 9 for Rigid Body and in Figure 8 for KS.

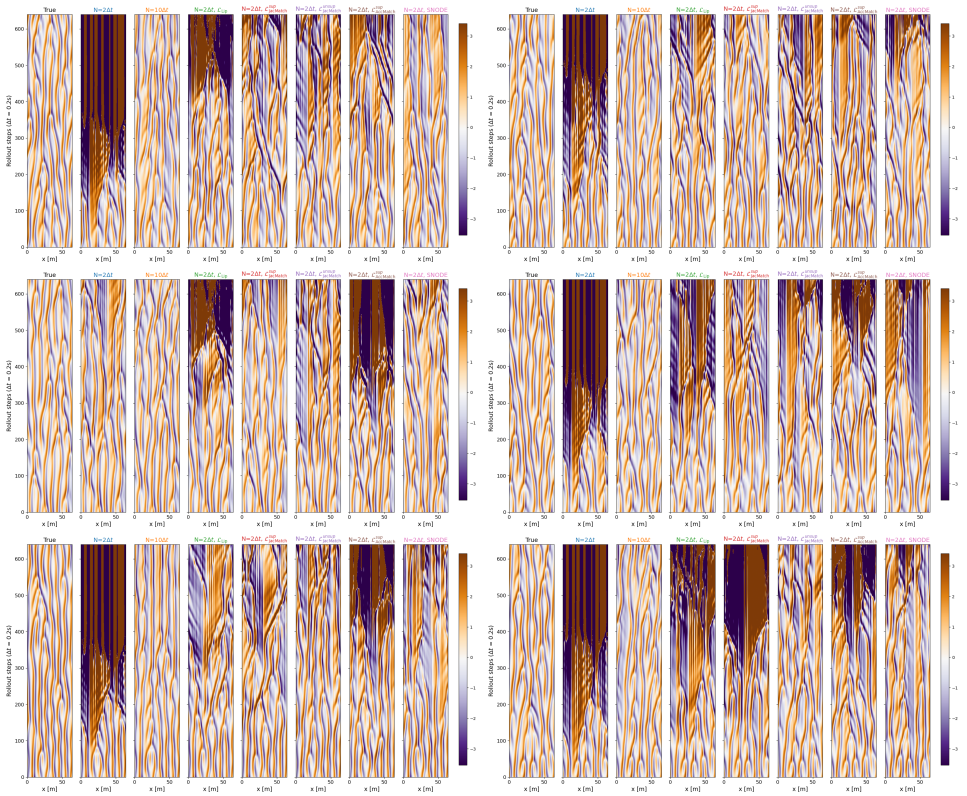

Figure 8: Kuramoto-Sivashinksy trajectories

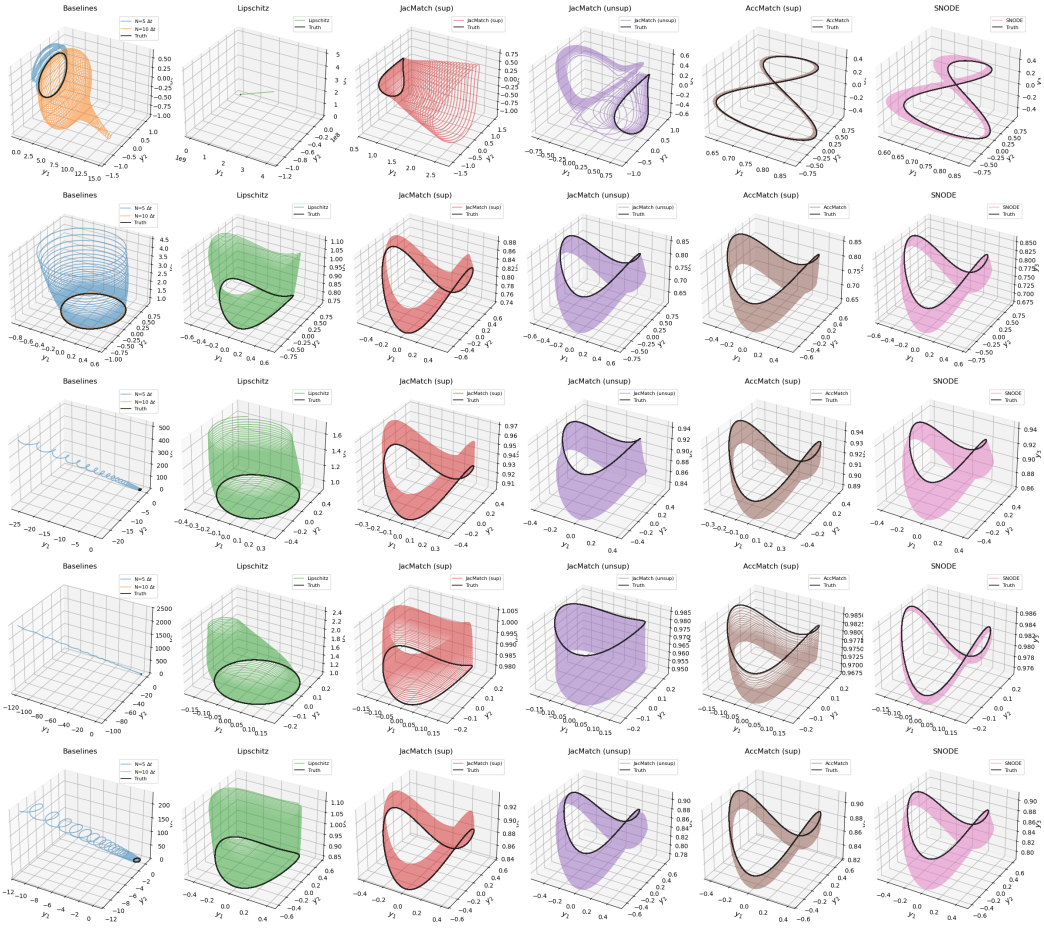

Figure 9: Rigid Body trajectories

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 claims made in the abstract and introduction reflect the paper contribution and scope, that are summarized in the conclusion 6.

   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 authors included the limitations of their work in the conclusion 6, opening the door to future work.

   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: All the theorems, formulas, and proofs in the paper are numbered and cross-referenced. Every proof is detailed in the annexes and states the necessary assumptions to its application. Furthermore, the mathematical tools at the center of the proofs are mentioned in the main paper to provide intuition (e.g., "Grönwall's lemma", "Taylor-Young formula").

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: We provide the details of the experiments in both the main paper 5 and in annex F, from the data generation protocols to the validation and testing setups or the architectures used.

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 share our code in the submission to ensure reproducibility. The data can either be generated from our code or from other open-source codes (instructions to do so being detailed in annex F).

   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: Training and test details are provided in annex F along with the validation protocol we performed for our experiments.

   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: Figures 2 and 3 report the distribution of final metric values over 100/128 test trajectories with diverse initial conditions (right panels), which captures the variance induced by sensitivity to initial conditions, the primary source of variability in chaotic systems.

   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 type of computer resources used and the mean runtime of the experiments are mentioned in annex F.

   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 paper follows the NeurIPS Code of Ethics.

   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 scope of this work is mostly tied to climate or other physical modeling.

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 pose any risk for misuse and does not release nor use sensitive data.

    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: Usage of existing assets is limited to open-source code from GitHub for generating data. Authors are cited.

    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 release new assets apart from the code available on GitHub.

    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 nor 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 crowdsourcing nor research with human subjects.

    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 important, original or non-standard components of this specific paper.

    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.