VIRTUAL-TRANSITION TEMPORAL-DIFFERENCE LEARNING FOR AUTOREGRESSIVE PDE DYNAMICS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive PDE models are typically trained to predict each observed next state from its input window. We study a complementary training choice: which examples should exchange prediction-error information? Building on the Markov reward process formulation for supervised learning, we use virtual-transition temporal-difference (TD) learning to couple residuals. These transitions need not follow the physical dynamics; our default pairs different trajectories at the same time index and leaves the prediction architecture and inference procedure unchanged. For linear models, we quantify the cost of imperfect inverse-covariance weighting: the bound on excess covariance relative to oracle generalized least squares (GLS) is quadratic near exact matching, and the same bound controls leading-order finite-horizon rollout mean squared error (MSE). Strict leading-order improvements over supervised least squares persist for exact rollouts at sufficiently small target noise. A local nonlinear extension characterizes when TD retains GLS efficiency and how weighting mismatch affects rollout error. Controlled experiments verify these theoretical results. Using the same pairing rule across 18 PDE domains, TD improves one-step test accuracy on 15 domains and short-horizon rollout accuracy on 15, with median error reductions of 4.0% and 5.7%; the median reduction over steps 21–60 is 10.0% across the 15 eligible domains. Original-scale experiments and pairing, backbone, and multistep studies establish the method's utility across several model and training configurations.

1 INTRODUCTION

Learning dynamics models is important in many fields, with applications ranging from scientific simulation and classical model predictive control to model-based reinforcement learning for a wide variety of decision-making and control tasks. In many of these settings, the underlying system dynamics are governed by partial differential equations (PDEs). Learned PDE solvers have therefore emerged as an attractive approach, providing fast, data-driven approximations of the evolution of complex physical systems. Recent advances span neural operators, transformer-based models, and other physics-oriented architectures (Li et al., 2021; Muser et al., 2026; Herde et al., 2024).

Existing approaches improve long-horizon dynamics prediction through several complementary mechanisms, including self-rollout or multistep training, architectural stabilization, differentiable solver unrolling, and distribution-level regularization (Brandstetter et al., 2022; Um et al., 2020; McCabe et al., 2023; Lippe et al., 2023; Schiff et al., 2024).

We investigate a different training choice: which examples should exchange prediction-error information when learning an autoregressive dynamics model? Building on the supervised-learning MRP formulation of Pan et al. (2025), we use virtual-transition TD to couple the residuals of paired training examples. The pairing rule is a training construction: examples need not be consecutive in physical time or belong to the same trajectory. Each paired residual supplies a detached correction to the current example's target, changing the training update while preserving the prediction architecture and inference procedure. This separates the design of training relationships from the real physical evolution or dynamics that the model is required to predict. Our contributions are threefold.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

First, we study virtual-transition pairing as a training-design variable for autoregressive PDE models. Such a transition specifies pairs of training examples that should be statistically coupled to compute the TD target. A simple same-time, cross-trajectory rule couples prediction residuals without following physical transitions or changing the deployed model.

Second, we quantify when this coupling preserves statistical efficiency. Our linear analysis shows that when the transition-induced weighting approximates inverse target-error covariance, the TD estimator's conditional covariance is within an explicit factor of oracle GLS. This bound transfers to first-order rollout MSE and gives a sufficient condition for improvement over supervised least squares. The local nonlinear analysis further identifies the part of a weighting mismatch that affects the model's first-order prediction directions; full inverse-covariance matching is sufficient but not necessary for local GLS efficiency.

Third, we verify the linear and local nonlinear theoretical predictions in controlled systems, then examine the empirical utility of a fixed pairing across neural PDE models. Across 18 PDE domains, a fixed pairing improves one-step test accuracy on 15 domains and short-horizon rollout accuracy on 15. Original-scale, transition-control, cross-backbone, and multistep comparisons demonstrate that the gains extend beyond the reduced Flower benchmark.

2 Background and related work

We briefly review the background on PDE dynamics learning and TD learning, together with their applications in supervised learning, which form the foundation for understanding our work.

2.1 Autoregressive Learning for PDE Dynamics

We consider a time-dependent physical field $u(t, x) \in \mathbb{R}^c$, where t denotes time, x denotes spatial location, and c is the number of field variables. Its evolution is governed by a PDE of the generic form $\partial_t u(t, x) = \mathcal{F}(u(t, x), \nabla u(t, x), \nabla^2 u(t, x), \ldots)$, together with appropriate initial and boundary conditions, where $\mathcal{F}$ denotes the underlying evolution operator. Let u_k denote the spatially discretized field observed at the k-th sampled time. Given the most recent L states, a neural dynamics model f_θ, parameterized by θ, predicts $\widehat{u}_{k+1} = f_\theta(u_{k-L+1}, \ldots, u_k)$. One-step supervised learning trains f_θ using ground-truth input windows and the next observed field u_{k+1} as the target. Below we categorize and review existing methods for learning the dynamics.

Architecture and representation design. One line of work develops architectures intended to represent PDE evolution accurately and efficiently. Fourier neural operators perform global operator learning through spectral transformations (Li et al., 2021), while message-passing models represent physical interactions through graph-structured computation (Brandstetter et al., 2022). More recent approaches include transformer-based PDE models such as Poseidon (Herde et al., 2024) and adaptive feature-warping architectures such as Flower (Muser et al., 2026). Beyond expressiveness, McCabe et al. (2023) study architectural sources of instability in autoregressive neural operators and propose design choices that improve their behavior under repeated application.

Rollout-aware training and input-distribution correction. A second family of approaches modifies the states or horizons used during training so that the model is optimized under conditions closer to autoregressive prediction. Brandstetter et al. (2022) formulate autoregressive stability as a domain-adaptation problem and introduce the pushforward trick, which trains the model using inputs perturbed by its own predictions. In hybrid neural numerical solvers, Um et al. (2020) differentiate through solver rollouts so that a learned correction model accounts for the effect of its current prediction on later states. More generally, multistep and self-rollout objectives optimize errors over several predicted steps. List et al. (2025) distinguish the effects of training on model-generated states from those of propagating gradients through time.

Long-horizon structural and statistical regularization. Other methods improve autoregressive behavior by imposing structure on the predicted fields or on the long-run dynamics. PDE-Refiner uses iterative denoising to better model low-amplitude components of the spatial frequency spectrum, which can be important for accurate long rollouts (Lippe et al., 2023). DySLIM introduces distribution-matching regularization that encourages the learned dynamics to preserve the invariant measure of a chaotic system, targeting long-run statistical fidelity rather than only pointwise short-horizon

error (Schiff et al., 2024). These methods illustrate that long-horizon quality can depend not only on immediate prediction accuracy, but also on spectral, distributional, and dynamical properties of the learned evolution map.

Our work intervenes at a different point in the training pipeline. It does not introduce a new prediction backbone, require the model to consume its own rollout states, or explicitly optimize an invariant distribution. Instead, it changes only the training update and leaves the prediction architecture and inference procedure unchanged. The closest methodological foundation for our training rule is the MRP formulation of supervised learning and its associated TD update, which we review next.

2.2 TD Learning and the MRP formulation of Supervised Learning

A Markov reward process (MRP) is specified by a state space $\mathcal{S}$, an expected reward $r(s)$, a transition distribution $P(\cdot \mid s)$, and a discount factor $\gamma \in [0, 1)$ (Szepesvari, 2010). Its value function $v(s)$ is the expected discounted sum of rewards starting from s. With linear function approximation, $v_\theta(s) = x(s)^\top \theta$, where $x(s) \in \mathbb{R}^q$ is a fixed feature representation and $\theta \in \mathbb{R}^q$. The value function satisfies the Bellman equation

$$v(s) = r(s) + \gamma \mathbb{E}_{S' \sim P(\cdot|s)}[v(S')]. \tag{1}$$

A fundamental approach to policy evaluation is temporal-difference (TD) learning (Sutton, 1988). Given a sampled transition (s_t, r_{t+1}, s_{t+1}), TD performs a stochastic fixed-point update based on (1):

$$\theta \leftarrow \theta + \alpha \left(y_{t,\mathrm{TD}} - x(s_t)^\top \theta\right) x(s_t), \tag{2}$$

where $\alpha > 0$ is the step size and $y_{t,\mathrm{TD}} \stackrel{\text{def}}{=} r_{t+1} + \gamma x(s_{t+1})^\top \theta$ is the one-step bootstrapped target. For notational convenience, we write $x_t \stackrel{\text{def}}{=} x(s_t)$. Under standard conditions, linear TD converges to the solution of the linear system $A\theta = b$ (Bradtke & Barto, 1996; Tsitsiklis & Van Roy, 1997), where

$$A = \mathbb{E}\big[x_t(x_t - \gamma x_{t+1})^\top\big] = X^\top D(I - \gamma P)X, \quad b = \mathbb{E}[r_{t+1}x_t] = X^\top Dr. \tag{3}$$

Here, $X \in \mathbb{R}^{|\mathcal{S}| \times q}$ is the feature matrix whose row corresponding to state s is $x(s)^\top$, $D \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the diagonal matrix containing the stationary state distribution, $P \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the policy-induced transition matrix with $P_{ij} = P(s_j \mid s_i)$, and $r \in \mathbb{R}^{|\mathcal{S}|}$ is the expected reward vector. Under commonly used assumptions, the TD solution is uniquely defined (Tsitsiklis & Van Roy, 1997).

TD-style bootstrapping has also been used in model-based control and predictive learning. Pong et al. (2018) learn goal- and horizon-conditioned value functions for model-based control. Knight & Achiam (2019) learn an auxiliary estimator of discounted long-horizon model error and use it to improve dynamics-model training. Ying et al. (2026) formulate diffusion-model training as policy evaluation and introduce a TD objective that encourages consistency of posterior-mean changes across noise levels.

The work most closely related to ours is Pan et al. (2025), which proposes a general framework for reformulating supervised learning as TD learning by conceptually constructing an MRP over training examples, where labels are interpreted as state values. Let y denote the vector of training labels. For a chosen transition matrix P, defining the expected reward vector as $r = (I - \gamma P)y$ makes y satisfy the corresponding Bellman equation. For a transition from example i to a sampled successor $j \sim P(i, \cdot)$, the sampled reward is $y_i - \gamma y_j$, giving the bootstrapped target $y_i - \gamma y_j + \gamma f_\theta(x_j)$. Importantly, these transitions describe statistical relationships between training examples rather than physical reachability in the real world. Their analysis relates linear TD estimation to least-squares regression and identifies correlated-noise settings under which TD can achieve greater statistical efficiency than ordinary least squares. Complementary work on graph-based regression models residual dependence through a structured precision matrix (Jia & Benson, 2020). Building on the MRP formulation, we quantify the efficiency cost of transition-induced weighting mismatch, propagate it to autoregressive rollout risk, and characterize the model-dependent mismatch that matters locally for nonlinear predictors. We then test a single pairing rule across PDE systems, model/data scales, and architectures.

3 Virtual-transition TD for dynamics models

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a dynamics dataset, where $x_i \in \mathcal{X}$ contains the conditioning information (e.g., previous states and control inputs) and $y_i \in \mathbb{R}^q$ is the next state or vectorized next-state field.

Following the supervised-learning MRP formulation of Pan et al. (2025), we regard each training example as an MRP state and its target as a vector-valued state value. For a predictor $f_\theta : \mathcal{X} \to \mathbb{R}^q$, with parameterization $\theta \in \mathbb{R}^m$, define

$$e_i(\theta) \stackrel{\text{def}}{=} f_\theta(x_i) - y_i. \tag{4}$$

A row-stochastic matrix $P \in \mathbb{R}^{n\times n}$ specifies the virtual transition. We draw the current example I from a full-support sampling distribution d and then draw its partner $J \sim P(I, \cdot)$. The transition describes a pairing rule rather than real-world physical evolution. For aligned trajectory data, our fixed default pairs (r, t) with (s, t), where s is sampled uniformly from the trajectories other than r. This rule can be implemented directly without constructing an $n \times n$ matrix; Appendix E.1 gives the sampling procedure and the alternative pairing rules used in the ablations. Section 4 analyzes generic virtual transitions for which d is stationary.

With nonlinear function approximation, we compute the TD target using a separate target network whose parameters track the online parameters through an exponential moving average (EMA). Let $\bar\theta$ denote the target-network parameters and write $\bar e_j = f_{\bar\theta}(x_j) - y_j$. Let sg denote the stop-gradient operator: it preserves its input value and blocks derivatives through that input. For the discount factor $\gamma \in [0, 1)$, define the virtual-transition TD target

$$\tilde y^{\mathrm{TD}}_{I,J} \stackrel{\text{def}}{=} y_I - \gamma y_J + \gamma\,\mathrm{sg}[f_{\bar\theta}(x_J)] = y_I + \gamma\,\mathrm{sg}[\bar e_J]. \tag{5}$$

For a sampled ordered pair (i, j), define the TD regression loss

$$\ell_\gamma(\theta, \bar\theta; i, j) \stackrel{\text{def}}{=} \frac{1}{2}\|f_\theta(x_i) - \tilde y^{\mathrm{TD}}_{i,j}\|_2^2 = \frac{1}{2}\|e_i(\theta) - \gamma\bar e_j\|_2^2. \tag{6}$$

During each online update, gradients pass only through $f_\theta(x_i)$; the paired prediction supplies a detached correction to the target. For fixed $\bar\theta$, expanding the square gives

$$\ell_\gamma = \frac{1-\gamma}{2}\|e_i(\theta)\|_2^2 + \frac{\gamma}{2}\|e_i(\theta) - \bar e_j\|_2^2 - \frac{\gamma(1-\gamma)}{2}\|\bar e_j\|_2^2. \tag{7}$$

The last term is constant with respect to θ. Thus the update combines supervised regression with an incentive for paired residuals to agree; γ controls this coupling, and $\gamma = 0$ recovers supervised training.

4 Statistical Efficiency and Rollout Risk

We analyze how virtual-transition weighting affects statistical efficiency and finite-horizon rollout error. The linear theory quantifies the cost of imperfect inverse-covariance matching; the local nonlinear extension identifies which weighting errors affect estimator efficiency through the model's first-order prediction directions. Controlled experiments test these quantitative predictions, while the PDE benchmark evaluates the training rule on neural dynamics models.

4.1 Statistical efficiency under a virtual transition

Assumption 4.1 (Fixed-design linear model)**.** Let $x_i \in \mathbb{R}^p$ be the input to the linear predictor, and let the ith rows of X and Y be $x_i^\top$ and $y_i^\top$, respectively. The data satisfy

$$Y = XW_\star^\top + E, \qquad X \in \mathbb{R}^{n\times p}, \quad Y, E \in \mathbb{R}^{n\times q}, \quad W_\star \in \mathbb{R}^{q\times p}, \quad \mathrm{rank}(X) = p. \tag{8}$$

Conditional on X, the matrix $W_\star$ is fixed, $\mathbb{E}[E \mid X] = 0$, and there are fixed matrices $C \in \mathbb{R}^{n\times n}$ and $\Omega \in \mathbb{R}^{q\times q}$, with $C \succ 0$ and $\Omega \succeq 0$, such that

$$\mathrm{Cov}(E_{:a}, E_{:b} \mid X) = \Omega_{ab} C, \qquad a, b \in \{1, \dots, q\}. \tag{9}$$

Here C describes dependence across training examples, while Ω describes dependence across output coordinates. The separable covariance in (9) allows the same sample-space weighting to be used for every output.

Throughout this section, P is row-stochastic and d is a full-support stationary distribution: $P^\top d = d$ and $\mathbf{1}^\top d = 1$. Let $D = \mathrm{diag}(d)$ and $\gamma \in [0, 1)$, and define

$$S_P \stackrel{\text{def}}{=} D(I_n - \gamma P). \tag{10}$$

The quantities P, d, γ are fixed conditional on X and do not depend on the realized target error E. We write $\mathrm{vec}(\cdot)$ for column-wise vectorization and use $\|\cdot\|_2$ for the Euclidean norm of a vector and the spectral norm of a matrix.

The linear TD fixed point. For $f_W(x) = Wx$, let $R_W = XW^\top - Y$. At an equilibrium of the linear mean semi-gradient update with an EMA target, the averaging update gives $\bar{W} = W$. The expected semi-gradient equation is therefore

$$X^\top S_P R_W = 0. \tag{11}$$

Stationarity gives $u^\top X^\top S_P X u \geq (1-\gamma)\|Xu\|_D^2 > 0$ for every $u \neq 0$, where $\|v\|_D^2 = v^\top D v$. Thus $X^\top S_P X$ is nonsingular, including for non-reversible P; Appendix A.1 gives the derivation. Define the TD and oracle generalized least-squares (GLS) estimators by

$$L_P \stackrel{\text{def}}{=} (X^\top S_P X)^{-1} X^\top S_P, \qquad \widehat{W}_P^\top \stackrel{\text{def}}{=} L_P Y, \tag{12}$$

$$L_{\mathrm{GLS}} \stackrel{\text{def}}{=} (X^\top C^{-1} X)^{-1} X^\top C^{-1}, \qquad \widehat{W}_{\mathrm{GLS}}^\top \stackrel{\text{def}}{=} L_{\mathrm{GLS}} Y. \tag{13}$$

Both are conditionally unbiased. Write

$$\Sigma_\nu \stackrel{\text{def}}{=} \mathrm{Cov}\Big(\mathrm{vec}(\widehat{W}_\nu^\top - W_\star^\top) \mid X\Big), \qquad \nu \in \{P, \mathrm{GLS}\}. \tag{14}$$

When P satisfies detailed balance, $DP = P^\top D$, the matrix S_P is symmetric positive definite. Under exact covariance matching, $S_P = cC^{-1}$ for some $c > 0$, the linear TD fixed point equals GLS (Pan et al., 2025). It is therefore a best linear unbiased estimator (BLUE) conditional on X (Aitken, 1936); Proposition A.1 states the corresponding multi-output covariance decomposition. The following theorem quantifies its covariance efficiency under approximate matching using a matrix Kantorovich inequality (Liu et al., 2010).

Theorem 4.2 (Near-GLS covariance and prediction risk)**.** *Under Assumption 4.1, suppose that P satisfies detailed balance and, for some $c > 0$ and $\eta \in [0, 1)$,*

$$(1-\eta)cC^{-1} \preceq S_P \preceq (1+\eta)cC^{-1}. \tag{15}$$

Then

$$\Sigma_{\mathrm{GLS}} \preceq \Sigma_P \preceq \frac{1}{1-\eta^2}\Sigma_{\mathrm{GLS}}. \tag{16}$$

For every fixed feature $x \in \mathbb{R}^p$,

$$\mathbb{E}\Big[\|(\widehat{W}_P - W_\star)x\|_2^2 \mid X\Big] \leq \frac{1}{1-\eta^2}\mathbb{E}\Big[\|(\widehat{W}_{\mathrm{GLS}} - W_\star)x\|_2^2 \mid X\Big]. \tag{17}$$

The risk inequality also holds after averaging over a random x that is conditionally independent of E given X and has finite conditional second moment.

Equivalently, the excess covariance satisfies $0 \preceq \Sigma_P - \Sigma_{\mathrm{GLS}} \preceq \frac{\eta^2}{1-\eta^2}\Sigma_{\mathrm{GLS}}$. Thus the excess is controlled in every coefficient direction, not only in total variance. For example, $\eta = 0.2$ permits a covariance inflation of at most $1/(1-0.2^2) - 1 \approx 4.2\%$ relative to GLS in every coefficient direction. Remark A.2 gives the scale-invariant form of the bound.

4.2 Consequences for rollout prediction

Set $p = q$ and $W_\star = A_\star \in \mathbb{R}^{q \times q}$, and consider the linear dynamics $z_{t+1} = A_\star z_t$. Let $\Delta = \widehat{A} - A_\star$ be the coefficient error. For an initial state $z_0 \in \mathbb{R}^q$ and an integer horizon $H \geq 1$, define

$$\mathcal{E}_H^{\mathrm{exact}}(\Delta; z_0) \stackrel{\text{def}}{=} (A_\star + \Delta)^H z_0 - A_\star^H z_0, \mathcal{E}_H^{\mathrm{lin}}(\Delta; z_0) \stackrel{\text{def}}{=} \sum_{k=0}^{H-1} A_\star^{H-1-k} \Delta A_\star^k z_0. \tag{18}$$

For fixed $A_\star, z_0$ and integer $H \geq 1$, expanding the matrix product gives $\|\mathcal{E}_H^{\mathrm{exact}}(\Delta; z_0) - \mathcal{E}_H^{\mathrm{lin}}(\Delta; z_0)\|_2 = O(\|\Delta\|_2^2)$ as $\|\Delta\|_2 \to 0$. The leading term is linear in the coefficient error. Consequently, a coefficient covariance comparison transfers through this map without enlarging its multiplicative factor (Proposition A.3). For comparison with supervised training under the same sampling weights, define

$$L_{\mathrm{SL}} \stackrel{\mathrm{def}}{=} (X^\top D X)^{-1} X^\top D, \qquad \widehat{W}_{\mathrm{SL}}^\top \stackrel{\mathrm{def}}{=} L_{\mathrm{SL}} Y. \tag{19}$$

With uniform sampling, this is ordinary least squares (OLS). For $\nu \in \{P, \mathrm{GLS}, \mathrm{SL}\}$, write $\widehat{A}_\nu = \widehat{W}_\nu$, let $\Delta_\nu = \widehat{A}_\nu - A_\star$, and define $r_{H,\nu} \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\|\mathcal{E}_H^{\mathrm{lin}}(\Delta_\nu; z_0)\|_2^2 \mid X\right]$. The following direct consequence makes the comparison with supervised least squares explicit.

Corollary 4.3 (Rollout risk relative to supervised least squares)**.** *Under Theorem 4.2, with $p = q$, for every fixed z_0 and integer $H \geq 1$,*

$$0 \leq r_{H,P} - r_{H,\mathrm{GLS}} \leq \frac{\eta^2}{1-\eta^2} r_{H,\mathrm{GLS}}. \tag{20}$$

In particular, $r_{H,P} < r_{H,\mathrm{SL}}$ whenever

$$r_{H,\mathrm{SL}} - r_{H,\mathrm{GLS}} > \frac{\eta^2}{1-\eta^2} r_{H,\mathrm{GLS}}. \tag{21}$$

If (21) *holds and additionally $\mathbb{E}[\|\Delta_P\|_2^{2H} + \|\Delta_{\mathrm{SL}}\|_2^{2H} \mid X] < \infty$, this strict comparison also holds for exact rollout MSE for all sufficiently small $\alpha > 0$ in the family $Y^{(\alpha)} = XA_\star^\top + \alpha E$, with $(\widehat{A}_\nu^{(\alpha)})^\top = L_\nu Y^{(\alpha)}$ for $\nu \in \{P, \mathrm{SL}\}$ and L_P, L_{SL} held fixed as α varies.*

Condition (21) separates the gain available from inverse-covariance weighting from the cost of an imperfect match. The exact-risk conclusion follows from the small-noise expansion in Corollary A.4.

Time-aligned cross-trajectory pairing. Our default pairing retains the time index and samples another trajectory. It downweights the residual component shared across trajectories relative to trajectory-specific differences. Appendix A.4 gives a shared-error covariance model and a closed-form discount under which this weighting reproduces oracle GLS. The general bound above quantifies the efficiency cost when the weighting departs from inverse covariance.

4.3 Local nonlinear extension

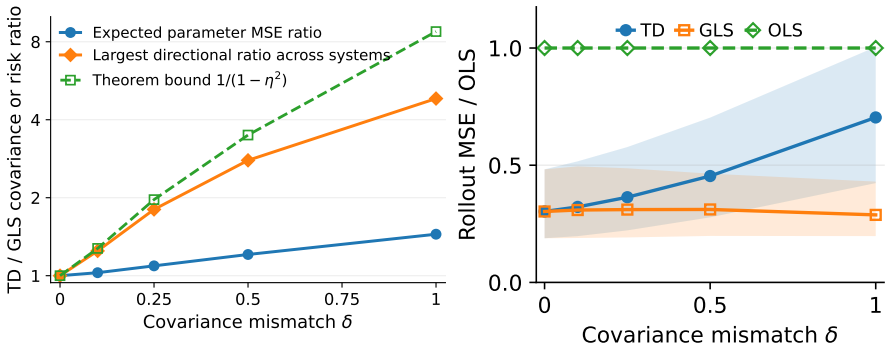
Appendix B extends the analysis to locally identifiable smooth predictors under small target perturbations around a correctly specified reference parameter. Write V_S for the leading parameter covariance under weighting S and $Q_S = V_S - V_{\mathrm{GLS}}$. Theorem B.3 gives $Q_S \succeq 0$ and an exact expression for it. Crucially, $Q_S = 0$ precisely when the covariance-rescaled weighting preserves the model's first-order prediction space; matching the full inverse covariance is not necessary. For a smooth prediction or rollout map Φ, with $K = D\Phi(\theta_\star)$, Corollary B.4 gives the leading excess risk $r_{\Phi,S} - r_{\Phi,\mathrm{GLS}} = \mathrm{tr}(KQ_SK^\top)$. Thus weighting quality depends on which error directions the model can represent and which of those directions affect the evaluated rollout.

5 Experiments

Our experiments verify the theory in controlled linear and nonlinear settings, and then evaluate broad PDE performance and residual structure, model and data scale, virtual transitions, backbone transfer, multistep compatibility and training cost. The PDE domains are drawn primarily from The Well (Ohana et al., 2024), with additional cases from PDEgym (Herde et al., 2024) and PDEBench (Takamoto et al., 2022), following the benchmark selection of Muser et al. (2026). The experimental design separates breadth from scale: an 18-domain Flower-Micro study evaluates a common pairing rule across PDE systems, while full-data Flower-Tiny experiments assess performance at the original model scale. Focused ablations use Acoustic Maze, Gray–Scott, and Shear Flow.

5.1 Controlled Theoretical Verification

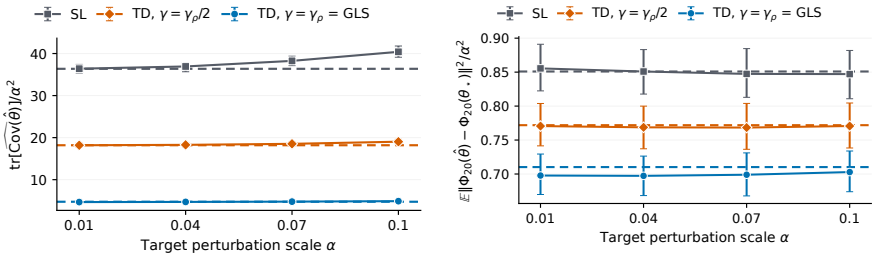
Linear controlled verification. We compare OLS, oracle GLS, and the linear TD fixed-point estimator on stable linear systems. Training inputs and test trajectories are noise-free; only training



(a) TD relative to GLS

(b) Mismatched covariance

Figure 1: **Covariance efficiency and rollout accuracy** with $q = 16$, $M = 32$, and $\sigma = 0.3$. (a) Under covariance mismatch, circles show the ratio of mean expected parameter MSE for TD versus GLS, diamonds show the largest directional covariance ratio across systems, and dashed squares show the upper bound from Theorem 4.2. These quantities are computed analytically from the conditional covariance matrices. (b) Step-40 rollout MSE relative to OLS, with 32 paired noise realizations per system. Lines are ratios of mean errors; bands are pointwise 95% paired-system bootstrap intervals.



(a) Covariance trace

(b) 20-step rollout MSE

Figure 2: **Controlled nonlinear verification.** Empirical parameter sample-covariance trace (left) and mean squared error at the 20-step rollout endpoint (right), each divided by α^2. Markers summarize 2,400 independent target-error draws; error bars are percentile 95% bootstrap intervals (500 resamples). Dashed lines show the corresponding first-order limits $\mathrm{tr}(V_S)$ and $r_{\Phi_{20},S}$ from Theorem B.3 and Corollary B.4. Matched TD and local oracle GLS coincide in this controlled setting; mismatched TD uses $\gamma = \gamma_\rho/2$.

targets are perturbed. All three estimates are obtained by solving their normal equations on the same data. We fix same-time cross-trajectory pairing and $\gamma = 31/32$, which matches the exchangeable target-error model at correlation $\rho = 0.5$. Increasing δ replaces globally shared noise with noise shared within two trajectory groups, while preserving marginal variance. The required GLS weighting therefore changes while TD remains fixed. Figure 1(a) verifies the lower and upper covariance bounds on all 100 systems at all five mismatch levels. Figure 1(b) evaluates exact step-40 rollouts of the fitted dynamics matrices. At $\delta = 0.5$, TD's rollout MSE is 45.4% of OLS, showing that exact covariance matching is not required for an advantage in this controlled family. Appendix C.2 gives the matched-correlation and discount sweeps; Appendix C.3 gives the mismatch protocol.

Nonlinear controlled verification. A 12-parameter nonlinear dynamics model tests the local predictions over 2,400 independent target-error draws. Figure 2 compares parameter covariance and 20-step MSE. Both matched and deliberately mismatched TD follow the predicted limits. Appendix C.4 gives details and an additional comparison that isolates the leading excess over local GLS.

5.2 Broad evaluation across 18 PDE domains

We evaluate the same-time cross-trajectory rule across 18 PDE domains. Within each SL/TD comparison, the architecture, selected training split, per-run epoch budget, and evaluation protocol are matched. Both methods select their learning rate from the same grid; TD additionally selects

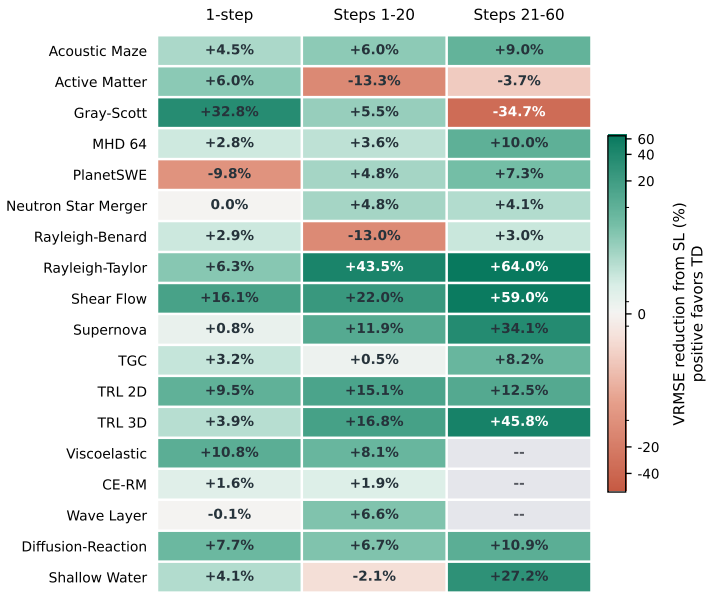


Figure 3: **Held-out test VRMSE reduction across 18 PDE domains.** Each cell reports $100(\mathrm{VRMSE}_{\mathrm{SL}} - \mathrm{VRMSE}_{\mathrm{TD}})/\mathrm{VRMSE}_{\mathrm{SL}}$, so positive values indicate lower error with TD. Columns show one-step prediction and autoregressive rollout over steps 1–20 and 21–60; the latter is available for 15 domains. "–" denotes unavailable measurements. Each column uses models selected by its corresponding validation metric. Intervals are truncated where trajectories are shorter (Table 5).

Table 1: Flower-Tiny held-out test VRMSE on the full training splits, 100 epochs; lower values within each pair are bold. Configurations are selected by validation separately for one-step prediction and each rollout interval. Here, "all" averages VRMSE over every available prediction step after the initial four-frame conditioning window. Table 8 reports the paired one-step validation measurements.

Domain	SL (1-step)	TD (1-step)	SL (1:20)	TD (1:20)	SL (21:60)	TD (21:60)	SL (all)	TD (all)
Acoustic Maze	0.00537	**0.00534**	**0.02430**	0.02570	**0.04820**	0.05330	**0.08520**	0.09210
Gray–Scott	0.00212	**0.00168**	**0.05405**	0.07161	1.13200	**0.36535**	8.51400	**0.89317**
Shear Flow	0.01640	**0.01148**	0.17300	**0.09480**	1.16000	**0.65200**	16.00000	**11.90000**

its own method-specific discount and EMA decay using validation data. This study tests prediction quality at fixed training-data and deployed-model size; implementation and selection details are given in Appendix E.

Figure 3 shows that across the 18-domain benchmark, TD lowers test VRMSE for one-step prediction on 15 of 18 domains (median gain 4.0%), over the first 20 autoregressive steps on 15 of 18 (median gain 5.7%), and on 13 of the 15 domains with long enough trajectories over steps 21–60 (median gain 10.0%). Corresponding validation measurements are reported in Appendix D.1.

5.3 Model and data scale

Original scale. We train the original Flower-Tiny configuration on the complete training splits for 100 epochs. In all three domains, SL and TD select the same original Flower learning rate. TD reduces one-step test VRMSE by approximately 21% on Gray–Scott and 30% on Shear Flow, while achieving comparable one-step VRMSE on Acoustic Maze (0.00534 vs. 0.00537). On Shear Flow, TD also lowers VRMSE in every reported rollout interval.

Scaling and training cost. On the separate 2×2 Shear Flow model/data grid, TD lowers one-step test VRMSE in all four settings by $2.5 - 9.9\%$ (Appendix D.2). TD adds one detached target-network forward evaluation per update without additional backward pass. The measured TD/SL epoch-time ratios are 1.37–1.46 on a single NVIDIA GH200 GPU. The additional computation is confined to training: evaluation uses the online model with unchanged architecture and inference cost.

5.4 Virtual-transition design

We compare same-time cross-trajectory pairing (our default, *cross*), the next window in the same trajectory (*forward-adjacent*), the previous or next window with equal probability (*symmetric-*

adjacent), a randomly sampled time in the same trajectory (*random-time*), a randomly sampled window from the full training split (*global-uniform*), and pairing an example with itself (*self*). Appendix E.1 specifies the sampling rules. Table 2 reports both validation and held-out results.

Cross gives the lowest test error on Gray–Scott and Shear Flow and is within 0.3% of the best result on Acoustic Maze. Relative to self-pairing, it lowers VRMSE by 8.8%, 34.7%, and 13.4% on the three domains, respectively. Because self-pairing retains target-network bootstrapping, this comparison supports a contribution from coupling distinct examples. On fixed SL validation residuals, the default weighting also lowers the two-mode mismatch in 14 of 16 eligible domains (Appendix B.6), characterizing how it reweights the observed shared and difference components.

Table 2: Transition comparison: Held-out one-step VRMSE (lower is better) on validation (V) and test (T). Each pairing selects its discount from the same grid, with learning rate and EMA decay fixed within each domain. All use the same online architecture and epoch budget. The lowest value in each column is bold.

Transition	Acoustic V	Acoustic T	Gray V	Gray T	Shear V	Shear T
cross (default)	**0.03667**	0.03595	0.00981	**0.00832**	**0.03969**	**0.03975**
forward-adjacent	0.03683	0.03603	0.01231	0.01194	0.04796	0.04953
global-uniform	0.03836	0.03782	0.00990	0.01451	0.04280	0.04262
random-time	0.03674	**0.03585**	**0.00892**	0.00968	0.04605	0.04767
self	0.04016	0.03942	0.01186	0.01274	0.04449	0.04589
symmetric-adjacent	0.03777	0.03712	0.01237	0.01326	0.05205	0.05470

Hyperparameter sensitivity. Validation analysis in Appendix D.6 fixes one hyperparameter at a time and retunes the others. Fixing $\gamma = 0.5$ is grid-optimal on Gray–Scott and Shear Flow and close to the best setting on Acoustic Maze; fixing the learning rate at 3×10^{-3} is within roughly 5% of the grid optimum on all three domains. The preferred EMA decay varies by domain.

5.5 Backbone transfer and compatibility with multistep training

We next ask whether the TD advantage depends on the Flower architecture or transfers to substantially different PDE backbones. We therefore replace Flower with scOT (Herde et al., 2024), FNO (Li et al., 2021), and CNextU-Net, a U-Net with ConvNeXt blocks (Ronneberger et al., 2015; Liu et al., 2022), while keeping the architecture and online trainable parameters identical between SL and TD within each backbone. For each domain, TD uses the same discount across backbones, without backbone-specific discount retuning. TD lowers held-out one-step test VRMSE in $7/9$ alternative domain–backbone pairs, with a median gain of 6.1% and a maximum of 15.5% (Table 10).

Compatibility with multistep training. We apply the same virtual-transition residual correction at each training horizon for $H \in \{2, 4\}$. With validation-selected configurations, multistep TD lowers test rollout VRMSE in 9 of 12 comparisons, including 5 of 6 over steps 1–20 and all four Shear Flow comparisons. These results demonstrate compatibility with a rollout-aware training objective. Appendix D.5 reports all comparisons.

6 Conclusion

We studied virtual-transition design as a complementary way to train autoregressive PDE models. A same-time cross-trajectory rule couples prediction residuals while leaving the deployed predictor unchanged. The analysis quantifies the efficiency cost of weighting mismatch and connects it to finite-horizon rollout risk; the local nonlinear extension identifies the model and rollout directions through which this cost acts. Controlled experiments verify these theoretical results, and PDE studies examine the same pairing across 18 domains, original-scale models, alternative backbones, and multistep training.

Limitations and future directions. The theory characterizes estimator efficiency in linear and local nonlinear regimes, while the neural PDE studies evaluate the training rule across domains and configurations. These results motivate adapting virtual transitions to residual structure and jointly designing the pairing rule, prediction architecture, and training horizon.

Reproducibility statement

Theoretical claims are stated with their assumptions in Section 4; complete proofs for the linear results appear in Appendix A, and the local nonlinear extension with its proofs in Appendix B. The controlled linear and nonlinear protocols, including the data-generating process and the estimators compared, are specified in Appendix C. For the PDE experiments, Appendix E specifies the datasets and their processing, the architectures and training setup, the virtual-transition sampling procedure, and the checkpoint-selection and evaluation protocol (Appendix E.2); the search grids are listed in Appendix E.5 and the selected configurations in Appendix E.6. Hyperparameters and checkpoints are selected using validation data only, while held-out test data are evaluated only after model selection. The code adapts the Flower implementation (Muser et al., 2026). Upon publication, all code will be released to enable reproduction of the empirical results reported in this paper.

AI Use Statement

Generative AI tools were used to assist with English-language editing, figure preparation, literature searches, and general discussions during the development of this work. The authors are fully responsible for the final manuscript, claims, and reported results.

References

Alexander C Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, pp. 42–48, 1936.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, pp. 33–57, 1996.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message Passing Neural PDE Solvers. *International Conference on Learning Representations*, 2022.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel De Bezenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. *Advances in Neural Information Processing Systems*, 2024.

Junteng Jia and Austin R. Benson. Residual Correlation in Graph Neural Network Regression. *Knowledge Discovery and Data Mining*, 2020.

Takeaki Kariya and Hiroshi Kurata. *Generalized least squares*. John Wiley & Sons, 2004.

Ethan Knight and Joshua Achiam. Training dynamics models for accurate long-horizon prediction. *ICLR Workshop on Task-Agnostic Reinforcement Learning*, 2019.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. *International Conference on Learning Representations*, 2021.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous Control with Deep Reinforcement Learning. *International Conference on Learning Representations*, 2016.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. *International Conference on Neural Information Processing Systems*, 2023.

Bjoern List, Li-Wei Chen, Kartik Bali, and Nils Thuerey. Differentiability in Unrolled Training of Neural Physics Simulators on Transient Dynamics. *Computer Methods in Applied Mechanics and Engineering*, 2025.

Zhibing Liu, Linzhang Lu, and Kanmin Wang. On several matrix Kantorovich-type inequalities. *Journal of Inequalities and Applications*, 2010.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards Stability of Autoregressive Neural Operators. *Transactions on Machine Learning Research*, 2023.

Till Muser, Alexandra Spitzer, Matti Lassas, Maarten V. de Hoop, and Ivan Dokmanić. Flowers: A Warp Drive for Neural PDE Solvers. *International Conference on Machine Learning*, 2026.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B Dalziel, Drummond B Fielding, et al. The Well: A large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 2024.

Yangchen Pan, Junfeng Wen, Chenjun Xiao, and Philip HS Torr. An MRP formulation for supervised learning: Generalized temporal difference learning models. *Journal of Artificial Intelligence Research*, 2025.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the Difficulty of Training Recurrent Neural Networks. *International Conference on Machine Learning*, 2013.

Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal Difference Models: Model-Free Deep RL for Model-Based Control. *International Conference on Learning Representations*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, pp. 234–241, 2015.

Yair Schiff, Zhong Yi Wan, Jeffrey B Parker, Stephan Hoyer, Volodymyr Kuleshov, Fei Sha, and Leonardo Zepeda-Núñez. DySLIM: Dynamics stable learning by invariant measure for chaotic systems. *International Conference on Machine Learning*, 2024.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 1988.

Csaba Szepesvari. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2010.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An Extensive Benchmark for Scientific Machine Learning. *Advances in Neural Information Processing Systems*, 2022.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 1997.

Kiwon Um, Robert Brand, Yun Raymond Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-loop: Learning from differentiable physics to interact with iterative pde-solvers. *Advances in neural information processing systems*, 2020.

Qizhen Ying, Yangchen Pan, Victor Adrian Prisacariu, and Junfeng Wen. Temporal difference learning for diffusion models. *International Conference on Machine Learning*, 2026.

This appendix contains the complete theoretical results and proofs, additional experiments, and all experimental details omitted from the main text.

1. Appendix A contains all missing theoretical results and proofs in the linear setting.
2. Appendix B contains the local nonlinear theory and proofs.
3. Appendix C gives the controlled linear and nonlinear experiments and additional results.
4. Appendix D contains additional PDE results.
5. Appendix E gives PDE implementation, evaluation, and hyperparameter details.

A Additional Theoretical Results and Proofs

A.1 The linear fixed point

Here the data are fixed and expectations average over the sampled pair $I \sim d$, $J \sim P(I, \cdot)$. For a fixed target parameter $\bar{W}$, the expected derivative of the sampled loss with respect to $W^\top$ is

$$X^\top D(R_W - \gamma P R_{\bar{W}}).$$

At an equilibrium of the mean semi-gradient update and the EMA update with $0 \le \beta < 1$, the latter gives $\bar{W} = W$. The former then gives (11).

To verify nonsingularity, Jensen's inequality and stationarity give, for every $v \in \mathbb{R}^n$,

$$\|Pv\|_D^2 = \sum_i d_i \left(\sum_j P_{ij} v_j \right)^2 \le \sum_{i,j} d_i P_{ij} v_j^2 = \sum_j d_j v_j^2 = \|v\|_D^2.$$

Cauchy–Schwarz therefore yields

$$v^\top S_P v = \|v\|_D^2 - \gamma v^\top D P v \ge \|v\|_D^2 - \gamma \|v\|_D \|Pv\|_D \ge (1-\gamma)\|v\|_D^2.$$

For $v = Xu$ and $u \neq 0$, full column rank of X makes the last expression positive. Thus $X^\top S_P X$ cannot have a nonzero null vector and is nonsingular. If $DP = P^\top D$, then S_P is also symmetric, so the same inequality proves $S_P \succ 0$.

We use the classical GLS covariance decomposition (Kariya & Kurata, 2004), stated below for the conditional multi-output model considered here.

A.2 Covariance comparison with GLS

Proposition A.1 (Exact covariance decomposition)**.** *Under Assumption 4.1, the estimators in* (12)–(13) *are conditionally unbiased. Define*

$$V_P \overset{def}{=} L_P C L_P^\top, \qquad V_{\mathrm{GLS}} \overset{def}{=} (X^\top C^{-1} X)^{-1}.$$

Then the covariances defined in (14) *satisfy*

$$\Sigma_P = \Omega \otimes V_P, \qquad \Sigma_{\mathrm{GLS}} = \Omega \otimes V_{\mathrm{GLS}}. \tag{22}$$

For any L fixed conditional on X and satisfying $LX = I_p$,

$$LCL^\top - V_{\mathrm{GLS}} = (L - L_{\mathrm{GLS}}) C (L - L_{\mathrm{GLS}})^\top \succeq 0. \tag{23}$$

In particular, if $R_P = L_P - L_{\mathrm{GLS}}$, then

$$\Sigma_P - \Sigma_{\mathrm{GLS}} = \Omega \otimes (R_P C R_P^\top) \succeq 0.$$

For a fixed test feature $x \in \mathbb{R}^p$,

$$\mathrm{Cov}\Big((\widehat{W}_P - W_\star)x \mid X \Big) = (x^\top V_P x)\Omega. \tag{24}$$

More generally, let $\mathcal{L} \in \mathbb{R}^{pq \times nq}$ be fixed conditional on X and satisfy $\mathcal{L}(I_q \otimes X) = I_{pq}$. For the linear unbiased estimator $\widehat{w}_{\mathcal{L}} = \mathcal{L}\,\mathrm{vec}(Y)$ of $\mathrm{vec}(W_\star^\top)$, define $\mathcal{R} = \mathcal{L} - I_q \otimes L_{\mathrm{GLS}}$. Then

$$\mathrm{Cov}(\widehat{w}_{\mathcal{L}} \mid X) - \Sigma_{\mathrm{GLS}} = \mathcal{R}(\Omega \otimes C)\mathcal{R}^\top \succeq 0. \tag{25}$$

Consequently, $\widehat{W}_{\mathrm{GLS}}$ is a BLUE conditional on X among estimators linear in $\mathrm{vec}(Y)$ and unbiased for every $W_\star$. Reversibility is not needed for these identities.

Proof. Since $L_P X = L_{\mathrm{GLS}} X = I_p$, conditional unbiasedness follows from $\mathbb{E}[E \mid X] = 0$. Column-wise vectorization gives

$$\mathrm{Cov}(\mathrm{vec}(E) \mid X) = \Omega \otimes C, \qquad \mathrm{vec}(LE) = (I_q \otimes L)\,\mathrm{vec}(E),$$

which proves (22). For any such L, write $R = L - L_{\mathrm{GLS}}$. Then $RX = 0$ and $CL_{\mathrm{GLS}}^\top = X(X^\top C^{-1} X)^{-1}$, so $RCL_{\mathrm{GLS}}^\top = 0$. Expanding $LCL^\top$ proves (23). Taking $L = L_P$ and a Kronecker product with $\Omega \succeq 0$ gives the coefficient covariance difference. Finally, $(\widehat{W}_P - W_\star)x = (I_q \otimes x^\top)\,\mathrm{vec}(\widehat{W}_P^\top - W_\star^\top)$, which proves (24).

For the general linear estimator, set $\mathcal{L}_{\mathrm{GLS}} = I_q \otimes L_{\mathrm{GLS}}$. Then

$$\mathcal{R}(I_q \otimes X) = 0, \qquad (\Omega \otimes C)\mathcal{L}_{\mathrm{GLS}}^\top = (I_q \otimes X)(\Omega \otimes V_{\mathrm{GLS}}).$$

Thus $\mathcal{R}(\Omega \otimes C)\mathcal{L}_{\mathrm{GLS}}^\top = 0$. Expanding the covariance of $(\mathcal{L}_{\mathrm{GLS}} + \mathcal{R})\,\mathrm{vec}(Y)$ proves (25). □

Theorem 4.2 (Near-GLS covariance and prediction risk)**.** *Under Assumption 4.1, suppose that P satisfies detailed balance and, for some $c > 0$ and $\eta \in [0, 1)$,*

$$(1 - \eta)cC^{-1} \preceq S_P \preceq (1 + \eta)cC^{-1}. \tag{15}$$

Then

$$\Sigma_{\mathrm{GLS}} \preceq \Sigma_P \preceq \frac{1}{1 - \eta^2}\Sigma_{\mathrm{GLS}}. \tag{16}$$

For every fixed feature $x \in \mathbb{R}^p$,

$$\mathbb{E}\Big[\|(\widehat{W}_P - W_\star)x\|_2^2 \mid X\Big] \le \frac{1}{1 - \eta^2}\mathbb{E}\Big[\|(\widehat{W}_{\mathrm{GLS}} - W_\star)x\|_2^2 \mid X\Big]. \tag{17}$$

The risk inequality also holds after averaging over a random x that is conditionally independent of E given X and has finite conditional second moment.

Proof. Let $Z = C^{-1/2}X$, $G = Z^\top Z = X^\top C^{-1} X$ and $U = ZG^{-1/2}$. Then $U^\top U = I_p$. Define

$$B = C^{1/2} S_P C^{1/2}, \qquad A = U^\top B U, \qquad a = (1 - \eta)c, \quad b = (1 + \eta)c.$$

The spectral match gives $aI_n \preceq B \preceq bI_n$ and $aI_p \preceq A \preceq bI_p$. Direct substitution yields

$$V_P = G^{-1/2} A^{-1} U^\top B^2 U A^{-1} G^{-1/2}. \tag{26}$$

The matrix Kantorovich bound can be obtained here as follows. Because B is symmetric and its spectrum is in $[a, b]$,

$$B^2 \preceq (a + b)B - abI_n, \qquad U^\top B^2 U \preceq (a + b)A - abI_p.$$

For $K = (a + b)^2/(4ab)$ and every scalar $s > 0$,

$$Ks^2 - (a + b)s + ab = K\left(s - \frac{2ab}{a + b}\right)^2 \ge 0.$$

Applying this polynomial inequality to the symmetric matrix A gives $(a + b)A - abI_p \preceq KA^2$. Congruence by A^{-1} and then $G^{-1/2}$ in (26) gives

$$V_P \preceq KG^{-1} = \frac{1}{1 - \eta^2}V_{\mathrm{GLS}}.$$

The lower bound follows from (23). Taking a Kronecker product with $\Omega \succeq 0$ proves (16). Applying the covariance order to the map $W \mapsto Wx$ and taking traces proves (17), because both estimation errors have conditional mean zero. For random x, first condition on (X, x) and then average; conditional independence and the finite conditional second moment justify this step. □

Remark A.2 (Scale-invariant form of the bound). Let λ_- and λ_+ be the smallest and largest eigenvalues of $C^{1/2} S_P C^{1/2}$, and set $\chi_P = \lambda_+/\lambda_-$. The choice $c_\star = (\lambda_+ + \lambda_-)/2$ minimizes the admissible η in (15), giving

$$\eta_\star = \frac{\chi_P - 1}{\chi_P + 1}, \qquad \frac{1}{1 - \eta_\star^2} = \frac{(\chi_P + 1)^2}{4\chi_P}.$$

To see optimality, any admissible c, η must satisfy $\lambda_+/(1 + \eta) \le c \le \lambda_-/(1 - \eta)$; this interval is nonempty only when $\eta \ge (\lambda_+ - \lambda_-)/(\lambda_+ + \lambda_-)$. The stated choice attains equality. Exact weighting correspondence $S_P = cC^{-1}$ is equivalent to $\chi_P = 1$.

A.3 From coefficient covariance to rollout risk

The term containing exactly one copy of Δ in the expansion of $(A_\star + \Delta)^H z_0$ is (18). Every other nonconstant term contains at least two copies of Δ; this gives the $O(\|\Delta\|_2^2)$ remainder for fixed $A_\star, z_0, H$. Since $\Delta v = (I_q \otimes v^\top)\,\mathrm{vec}(\Delta^\top)$, define

$$K_H(A_\star, z_0) \stackrel{\text{def}}{=} \sum_{k=0}^{H-1} A_\star^{H-1-k} \left[I_q \otimes (A_\star^k z_0)^\top\right] \in \mathbb{R}^{q \times q^2}.$$

Then

$$\mathcal{E}_H^{\mathrm{lin}}(\Delta; z_0) = K_H(A_\star, z_0)\,\mathrm{vec}(\Delta^\top). \tag{27}$$

Proposition A.3 (Transfer of coefficient covariance to rollout risk)**.** *Let $\widehat{A}_a = A_\star + \Delta_a$ and $\widehat{A}_b = A_\star + \Delta_b$ be conditionally unbiased estimators with finite conditional second moments, and define $\Sigma_\nu = \mathrm{Cov}(\mathrm{vec}(\Delta_\nu^\top) \mid X)$ for $\nu \in \{a, b\}$. If $\Sigma_a \preceq \kappa \Sigma_b$ for some $\kappa \geq 0$, then for every fixed z_0 and integer $H \geq 1$,*

$$\mathrm{Cov}\big(\mathcal{E}_H^{\mathrm{lin}}(\Delta_a; z_0) \mid X\big) \preceq \kappa \mathrm{Cov}\big(\mathcal{E}_H^{\mathrm{lin}}(\Delta_b; z_0) \mid X\big), \tag{28}$$

$$\mathbb{E}\big[\|\mathcal{E}_H^{\mathrm{lin}}(\Delta_a; z_0)\|_2^2 \mid X\big] \leq \kappa \mathbb{E}\big[\|\mathcal{E}_H^{\mathrm{lin}}(\Delta_b; z_0)\|_2^2 \mid X\big]. \tag{29}$$

The same conclusions hold after averaging over a random z_0 that is conditionally independent of the estimator errors given X and has finite conditional second moment.

Proof. Equation (27) gives covariance $K_H \Sigma_\nu K_H^\top$ for estimator ν. Congruence preserves positive-semidefinite order. Conditional unbiasedness makes the leading-order rollout errors mean zero, so taking traces gives the MSE comparison. For random z_0, condition on (X, z_0) and then average. The conditional means are zero, so the covariance given X is also the average of these conditional covariances. Conditional independence and finite second moments ensure the expressions are well defined. □

Corollary A.4 (Local comparison of exact rollout risk)**.** *In the square linear model, fix z_0 and an integer $H \geq 1$. For $\nu \in \{a, b\}$, let $L_\nu \in \mathbb{R}^{q \times n}$ be fixed conditional on X and satisfy $L_\nu X = I_q$. For $\alpha \geq 0$, define*

$$Y^{(\alpha)} = X A_\star^\top + \alpha E, \qquad (\widehat{A}_\nu^{(\alpha)})^\top = L_\nu Y^{(\alpha)}, \qquad \Delta_\nu = \widehat{A}_\nu^{(1)} - A_\star.$$

If $\mathbb{E}[\|\Delta_\nu\|_2^{2H} \mid X] < \infty$ for $\nu \in \{a, b\}$, then, as $\alpha \downarrow 0$,

$$\mathbb{E}\Big[\|(\widehat{A}_\nu^{(\alpha)})^H z_0 - A_\star^H z_0\|_2^2 \mid X\Big] = \alpha^2 \mathbb{E}\big[\|\mathcal{E}_H^{\mathrm{lin}}(\Delta_\nu; z_0)\|_2^2 \mid X\big] + O(\alpha^3). \tag{30}$$

A strict inequality between the two leading-order risks therefore implies the same strict inequality between their exact rollout risks for every sufficiently small $\alpha > 0$.

Proof. The identity $L_\nu X = I_q$ gives $\widehat{A}_\nu^{(\alpha)} - A_\star = \alpha \Delta_\nu$. The expansion of $(A_\star + \alpha\Delta_\nu)^H - A_\star^H$ is a finite polynomial in α, and its coefficient of α, applied to z_0, is $\mathcal{E}_H^{\mathrm{lin}}(\Delta_\nu; z_0)$. After squaring, the only term of degree two is the squared norm of this coefficient, and every other term has degree at least three. Each coefficient of the squared polynomial is bounded in absolute value by a finite constant times $1 + \|\Delta_\nu\|_2^{2H}$, where the constant can depend on $A_\star, z_0, H$. The conditional moment assumption thus permits taking expectations term by term. The sum of the remaining terms is $O(\alpha^3)$ for $0 \leq \alpha \leq 1$. For $H = 1$ the remainder is identically zero. Subtracting the two expansions proves the strict comparison. □

Corollary 4.3 (Rollout risk relative to supervised least squares)**.** *Under Theorem 4.2, with $p = q$, for every fixed z_0 and integer $H \geq 1$,*

$$0 \leq r_{H,P} - r_{H,\mathrm{GLS}} \leq \frac{\eta^2}{1 - \eta^2} r_{H,\mathrm{GLS}}. \tag{20}$$

In particular, $r_{H,P} < r_{H,\mathrm{SL}}$ whenever

$$r_{H,\mathrm{SL}} - r_{H,\mathrm{GLS}} > \frac{\eta^2}{1 - \eta^2} r_{H,\mathrm{GLS}}. \tag{21}$$

If (21) *holds and additionally* $\mathbb{E}[\|\Delta_P\|_2^{2H} + \|\Delta_{\mathrm{SL}}\|_2^{2H} \mid X] < \infty$*, this strict comparison also holds for exact rollout MSE for all sufficiently small* $\alpha > 0$ *in the family* $Y^{(\alpha)} = XA_\star^\top + \alpha E$*, with* $(\widehat{A}_\nu^{(\alpha)})^\top = L_\nu Y^{(\alpha)}$ *for* $\nu \in \{P, \mathrm{SL}\}$ *and* L_P, L_{SL} *held fixed as* α *varies.*

Proof. Theorem 4.2 and Proposition A.3 give $r_{H,\mathrm{GLS}} \le r_{H,P} \le r_{H,\mathrm{GLS}}/(1-\eta^2)$. Subtracting $r_{H,\mathrm{GLS}}$ proves (20); comparing this bound with (21) proves $r_{H,P} < r_{H,\mathrm{SL}}$. Both $L_P X$ and $L_{\mathrm{SL}} X$ equal I_q, so Corollary A.4, with $a = P$ and $b = \mathrm{SL}$, proves the exact-risk statement under the stated conditional moment condition. □

A.4 A TIME-ALIGNED CROSS-TRAJECTORY INSTANCE

For this instance, suppose that the dataset contains $M \ge 2$ trajectories and $T \ge 1$ aligned physical-time indices, so $n = MT$. Write (r,t) for trajectory $r \in \{1,\dots,M\}$ at time $t \in \{1,\dots,T\}$. Order the rows by time, with the M trajectories in each time block, and sample the current example uniformly from these n examples.

Definition A.5 (Time-aligned cross-trajectory pairing)**.** Retain the current time index and choose the partner trajectory uniformly from the other $M-1$ trajectories. Thus

$$P\big((r,t),(s,u)\big) = \frac{\mathbf{1}\{u = t,\ s \ne r\}}{M-1}, \qquad d_{r,t} = \frac{1}{MT}.$$

Equivalently,

$$P = I_T \otimes P_M, \qquad P_M = \frac{\mathbf{1}\mathbf{1}^\top - I_M}{M-1}, \qquad D = I_n/n.$$

The matrix P is symmetric and row-stochastic, so uniform d is stationary and detailed balance holds.

Assumption A.6 (Same-time exchangeable sample covariance)**.** In Assumption 4.1, take

$$C = C_{\mathrm{sample}} \overset{\text{def}}{=} I_T \otimes C_\rho, \qquad C_\rho \overset{\text{def}}{=} (1-\rho) I_M + \rho \mathbf{1}\mathbf{1}^\top, \qquad 0 \le \rho < 1. \tag{31}$$

Thus the sample covariance has exchangeable within-time blocks and zero covariance between different time blocks. The overall error scale and dependence across output coordinates are contained in Ω.

Remark A.7 (Shared and difference components). For $v \in \mathbb{R}^M$, write $v = \bar{v}\mathbf{1} + v^{\mathrm{diff}}$, where $\bar{v} = \mathbf{1}^\top v/M$ and $\mathbf{1}^\top v^{\mathrm{diff}} = 0$. Since $P_M\mathbf{1} = \mathbf{1}$ and $P_M v^{\mathrm{diff}} = -v^{\mathrm{diff}}/(M-1)$,

$$(I_M - \gamma P_M)v = (1-\gamma)\bar{v}\mathbf{1} + \left(1 + \frac{\gamma}{M-1}\right) v^{\mathrm{diff}}. \tag{32}$$

The inverse covariance C_ρ^{-1} has the same two eigenspaces. Its difference-to-shared weight ratio is $[1 + (M-1)\rho]/(1-\rho)$.

Proposition A.8 (Exact inverse-covariance match)**.** *Under Definition A.5 and Assumption A.6, choose*

$$\gamma_\rho \overset{\text{def}}{=} \frac{\rho(M-1)}{1 + \rho(M-2)}. \tag{33}$$

Then $0 \le \gamma_\rho < 1$ *and*

$$I_M - \gamma_\rho P_M = b_\rho C_\rho^{-1}, \tag{34}$$

$$S_{P,\rho} \overset{\text{def}}{=} D(I_n - \gamma_\rho P) = \frac{b_\rho}{n} C_{\mathrm{sample}}^{-1}, \tag{35}$$

where

$$b_\rho \overset{\text{def}}{=} \frac{(1-\rho)[1 + (M-1)\rho]}{1 + \rho(M-2)} > 0. \tag{36}$$

Proof. The eigenvalues of P_M on the shared and difference subspaces are 1 and $-1/(M-1)$, respectively. Those of C_ρ are $1 + (M-1)\rho$ and $1-\rho$, both positive. The value in (33) satisfies

$$1 - \gamma_\rho = \frac{1-\rho}{1 + \rho(M-2)} = \frac{b_\rho}{1 + (M-1)\rho}, \qquad 1 + \frac{\gamma_\rho}{M-1} = \frac{b_\rho}{1-\rho}.$$

These two identities prove (34). Applying it to every time block and using $D = I_n/n$ proves (35). The numerator defining γ_ρ is nonnegative and $1 - \gamma_\rho > 0$, proving its stated range. □

Corollary A.9 (Statistical and rollout consequences of the exact match)**.** *Under Assumption 4.1, Definition A.5, and Assumption A.6, let $\widehat{W}_{\gamma_\rho}$ be the TD estimator with $\gamma = \gamma_\rho$. Define $\widehat{W}_0^\top \overset{\text{def}}{=} (X^\top X)^{-1} X^\top Y$; under the uniform sampling used here, $\widehat{W}_0 = \widehat{W}_{\mathrm{SL}}$. Then*

$$\widehat{W}_{\gamma_\rho} = \widehat{W}_{\mathrm{GLS}}, \tag{37}$$

and

$$\mathrm{Cov}\Big(\mathrm{vec}(\widehat{W}_{\gamma_\rho}^\top - W_\star^\top) \mid X\Big) \preceq \mathrm{Cov}\Big(\mathrm{vec}(\widehat{W}_0^\top - W_\star^\top) \mid X\Big). \tag{38}$$

The same order holds for the MSE at every fixed test feature. If $p = q$ and $W_\star = A_\star$, it also holds for the first-order rollout MSE at every fixed $z_0 \in \mathbb{R}^q$ and integer $H \geq 1$. If, additionally, the two coefficient errors have finite conditional $2H$th moments, a strict leading-order rollout risk inequality implies a strict exact-risk improvement in the small-noise family of Corollary A.4.

Proof. Equation (35) makes the TD and GLS estimator matrices identical. Apply (23) to $L = (X^\top X)^{-1} X^\top$ to obtain the OLS comparison. The prediction statement follows by a linear map and taking traces. Proposition A.3 and, under the extra moment condition, Corollary A.4 give the two rollout conclusions. □

B Local Nonlinear Efficiency

This section extends the linear efficiency result to nonlinear dynamics models by studying what happens near a correctly fitted model when the training targets are slightly perturbed. Locally, the nonlinear model behaves like a linear model through its Jacobian, which allows us to compare the uncertainty induced by SL, TD, and oracle GLS. The key result is that TD does not need to reproduce the full inverse covariance everywhere: it only needs to weight the error directions that the model can locally respond to appropriately. This also gives a corresponding comparison for finite-horizon rollout error.

Here "oracle GLS" denotes inverse-covariance-weighted nonlinear least squares; after linearizing the predictor around the reference parameter, this estimator reduces to GLS for the resulting local linear regression in parameter perturbations.

B.1 The nonlinear mean fixed-point equation

We use the predictor f_θ and parameter dimension m from the method section. The sampling assumptions, the convention of conditioning on fixed training inputs, and the requirement that P, d, γ do not depend on realized target errors are inherited from Section 4. We also retain its S_P and column-wise vectorization convention; no linear form for f_θ is imposed here. All expectations and covariances below are conditional on these inputs.

For a differentiable predictor, let $F_\theta \in \mathbb{R}^{n \times q}$ have ith row $f_\theta(x_i)^\top$. Define

$$\begin{aligned} N &\overset{\text{def}}{=} nq, \qquad \mathbf{f}(\theta) \overset{\text{def}}{=} \mathrm{vec}(F_\theta), \qquad \mathbf{e}(\theta) \overset{\text{def}}{=} \mathrm{vec}(F_\theta - Y), \\ J_\theta &\overset{\text{def}}{=} \frac{\partial \mathbf{f}(\theta)}{\partial \theta^\top} \in \mathbb{R}^{N \times m}, \qquad \mathcal{S}_P \overset{\text{def}}{=} I_q \otimes S_P, \qquad \mathcal{S}_{\mathrm{SL}} \overset{\text{def}}{=} I_q \otimes D. \end{aligned} \tag{39}$$

The mean derivative of the sampled TD loss, holding the target parameters fixed, is

$$g(\theta, \bar{\theta}) = J_\theta^\top (I_q \otimes D)\big[\mathbf{e}(\theta) - \gamma (I_q \otimes P)\mathbf{e}(\bar{\theta})\big]. \tag{40}$$

Indeed, if $G_i(\theta) \overset{\text{def}}{=} \partial f_\theta(x_i)/\partial \theta^\top$ then the sampled derivative is $G_i(\theta)^\top (e_i(\theta) - \gamma e_j(\bar{\theta}))$. Averaging it with weights $d_i P_{ij}$ gives (40). This expectation averages the sampled pair, not the random target errors.

Proposition B.1 (Nonlinear mean TD–EMA fixed points)**.** *For an unregularized mean online step $\theta^+ = \theta - a\, g(\theta, \bar{\theta})$ with $a > 0$, together with the paper's EMA target update, the joint fixed points are exactly the pairs (θ, θ) satisfying*

$$J_\theta^\top \mathcal{S}_P \mathbf{e}(\theta) = 0. \tag{41}$$

Thus their set is independent of the EMA decay within its stated range.

Proof. At a joint fixed point the online step gives $g = 0$, and the EMA equation becomes $\bar{\theta} = \beta\bar{\theta} + (1-\beta)\theta$. Its nonzero target-update coefficient implies $\bar{\theta} = \theta$. Substitution into (40) proves (41). Conversely, a solution of (41), with $\bar{\theta} = \theta$, leaves both updates unchanged. □

The statistical results below concern symmetric positive-definite weightings. For TD we therefore impose detailed balance, as in Theorem 4.2; Section A.1 already proves $S_P \succ 0$, and hence $\mathcal{S}_P \succ 0$. Proposition B.1 itself does not require detailed balance. It characterizes equilibria, not convergence, stability, or finite-time behavior. In particular, it does not identify the outputs of a regularized optimizer with these unregularized equilibria.

B.2 Local target perturbations and estimator covariance

Retain the zero-mean target-error convention for E in Assumption 4.1, but replace its linear mean by the nonlinear mean below. Let $\theta_\star \in \mathbb{R}^m$ be a reference parameter fixed conditional on the inputs. Write

$$Y^{(\alpha)} = F_{\theta_\star} + \alpha E, \qquad \varepsilon \stackrel{\text{def}}{=} \operatorname{vec}(E), \qquad \Gamma \stackrel{\text{def}}{=} \operatorname{Cov}(\varepsilon), \qquad \alpha \geq 0. \tag{42}$$

The nonlinear argument allows any full target-error covariance Γ. When the separable covariance (9) is retained, $\Gamma = \Omega \otimes C$. The linear mean and the rank assumption on X are not assumed in this appendix.

Assumption B.2 (Additional local regularity)**.** The map $\mathbf{f}$ is twice continuously differentiable on an open neighborhood of $\theta_\star$, and $J_\star \stackrel{\text{def}}{=} J_{\theta_\star}$ has full column rank. For the probability expansions, $\Gamma \succ 0$ and $\|\varepsilon\|_2 \leq M_\varepsilon$ almost surely for a fixed finite constant M_ε.

The rank condition means that the training predictions distinguish infinitesimal changes in the analyzed parameters; in particular, $m \leq N$. It must be checked for the parameterization being analyzed. An explicitly specified lower-dimensional model can also satisfy the assumption, but a change of coordinates does not automatically remove rank deficiency of a neural network. Any reduced coordinates used for a rollout result must also determine that rollout. The mean-update statement above is in the actual update coordinates and asserts no invariance of EMA under nonlinear reparameterization.

Let $\mathcal{S} \succ 0$ be fixed conditional on the inputs and independent of α and the realized target errors. Define the local solution $\widehat{\theta}_{\mathcal{S}}^{(\alpha)}$ by

$$J_\theta^\top \mathcal{S}\big[\mathbf{f}(\theta) - \operatorname{vec}(Y^{(\alpha)})\big] = 0, \tag{43}$$

whose existence and local uniqueness are proved below. The choices $\mathcal{S} = \mathcal{S}_P$, $\mathcal{S}_{\mathrm{SL}}$, and Γ^{-1} give local TD, supervised least squares, and oracle GLS, respectively. Introduce

$$\begin{gathered} \mathcal{L}_{\mathcal{S}} \stackrel{\text{def}}{=} (J_\star^\top \mathcal{S} J_\star)^{-1} J_\star^\top \mathcal{S}, \qquad \mathcal{V}_{\mathcal{S}} \stackrel{\text{def}}{=} \mathcal{L}_{\mathcal{S}} \Gamma \mathcal{L}_{\mathcal{S}}^\top, \\ G \stackrel{\text{def}}{=} J_\star^\top \Gamma^{-1} J_\star, \qquad \mathcal{L}_{\mathrm{GLS}} \stackrel{\text{def}}{=} G^{-1} J_\star^\top \Gamma^{-1}, \qquad \mathcal{V}_{\mathrm{GLS}} \stackrel{\text{def}}{=} G^{-1}. \end{gathered} \tag{44}$$

Here $\mathcal{L}_{\mathcal{S}} \in \mathbb{R}^{m \times N}$ and $\mathcal{V}_{\mathcal{S}} \in \mathbb{R}^{m \times m}$; these are local nonlinear quantities, distinct from the sample-space matrices L_P, V_P in the linear analysis. Matrix remainder bounds below use the spectral norm.

Theorem B.3 (Local nonlinear expansion and exact excess covariance)**.** *Under Assumption B.2, for each fixed $\mathcal{S} \succ 0$ there is a fixed neighborhood of $\theta_\star$ in which* (43) *has a unique solution for all sufficiently small α, uniformly over the allowed target errors. This solution is a strict local minimizer of $\frac{1}{2}[\mathbf{f}(\theta) - \operatorname{vec}(Y^{(\alpha)})]^\top \mathcal{S}[\mathbf{f}(\theta) - \operatorname{vec}(Y^{(\alpha)})]$, and*

$$\widehat{\theta}_{\mathcal{S}}^{(\alpha)} - \theta_\star = \alpha \mathcal{L}_{\mathcal{S}} \varepsilon + O(\alpha^2), \tag{45}$$

$$\mathbb{E}[\widehat{\theta}_{\mathcal{S}}^{(\alpha)} - \theta_\star] = O(\alpha^2), \qquad \operatorname{Cov}(\widehat{\theta}_{\mathcal{S}}^{(\alpha)} - \theta_\star) = \alpha^2 \mathcal{V}_{\mathcal{S}} + O(\alpha^3). \tag{46}$$

To characterize its first-order covariance, define

$$\begin{gathered} U \stackrel{\text{def}}{=} \Gamma^{-1/2} J_\star G^{-1/2}, \qquad B \stackrel{\text{def}}{=} \Gamma^{1/2} \mathcal{S} \Gamma^{1/2}, \qquad A \stackrel{\text{def}}{=} U^\top B U, \\ \mathcal{R} \stackrel{\text{def}}{=} (I_N - U U^\top) B U, \qquad \delta_{\tan} \stackrel{\text{def}}{=} \|\mathcal{R} A^{-1}\|_2. \end{gathered} \tag{47}$$

The square roots here are the symmetric positive-definite square roots. The columns of U are orthonormal and span the possible first-order prediction changes after rescaling errors by $\Gamma^{-1/2}$. Then

$$\mathcal{V}_{\mathcal{S}} - \mathcal{V}_{\mathrm{GLS}} = G^{-1/2} A^{-1} \mathcal{R}^\top \mathcal{R} A^{-1} G^{-1/2} \succeq 0, \tag{48}$$

and

$$\begin{aligned} \mathcal{V}_{\mathrm{GLS}} \preceq \mathcal{V}_{\mathcal{S}} \preceq (1 + \delta_{\tan}^2) \mathcal{V}_{\mathrm{GLS}}, \\ \lambda_{\max}(G^{1/2} \mathcal{V}_{\mathcal{S}} G^{1/2}) = 1 + \delta_{\tan}^2. \end{aligned} \tag{49}$$

Consequently, first-order covariance equals that of local oracle GLS if and only if $\mathcal{R} = 0$, equivalently if B maps the column space of U into itself. The global match $\mathcal{S} = c\Gamma^{-1}$ for $c > 0$ is sufficient, but not necessary.

Proof. **A common neighborhood and the expansion.** Treat the entire target perturbation as a vector $z \in \mathbb{R}^N$, and set

$$h(\theta, z) \stackrel{\text{def}}{=} J_\theta^\top \mathcal{S}[\mathbf{f}(\theta) - \mathbf{f}(\theta_\star) - z].$$

Because $\mathbf{f}$ is twice continuously differentiable, h is continuously differentiable. At $(\theta_\star, 0)$,

$$h(\theta_\star, 0) = 0, \qquad D_\theta h(\theta_\star, 0) = J_\star^\top \mathcal{S} J_\star \succ 0.$$

The implicit-function theorem provides a single continuously differentiable solution $\theta = \psi(z)$ on a neighborhood of $z = 0$, with uniqueness in a fixed parameter neighborhood and $\psi(0) = \theta_\star$. After shrinking these neighborhoods, the Hessian of the weighted squared-residual objective remains positive definite by continuity, which also gives the strict local minimum. The derivative of ψ at zero is $\mathcal{L}_{\mathcal{S}}$, and continuity of this derivative on a smaller closed ball gives $\|\psi(z) - \theta_\star\|_2 = O(\|z\|_2)$.

For completeness, twice continuous differentiability already gives the quadratic remainder; a third derivative is unnecessary. Write $u = \psi(z) - \theta_\star$ and $H_0 = J_\star^\top \mathcal{S} J_\star$. Bounded second derivatives on a smaller closed parameter ball give

$$\mathbf{f}(\theta_\star + u) - \mathbf{f}(\theta_\star) = J_\star u + O(\|u\|_2^2), \qquad J_{\theta_\star + u} = J_\star + O(\|u\|_2).$$

Substitution in $h(\psi(z), z) = 0$ yields

$$0 = H_0 u - J_\star^\top \mathcal{S} z + O(\|u\|_2^2 + \|u\|_2 \|z\|_2).$$

Since $u = O(\|z\|_2)$ and H_0 is invertible,

$$\psi(z) - \theta_\star = \mathcal{L}_{\mathcal{S}} z + O(\|z\|_2^2),$$

with one remainder constant on this ball. Setting $z = \alpha\varepsilon$ and using $\|\varepsilon\|_2 \le M_\varepsilon$ proves uniform existence and (45) for a single sufficiently small range of α.

Write the remainder as u_α, where $\|u_\alpha\|_2 \le c_0 \alpha^2$ for a fixed c_0. The inherited zero-mean assumption gives bias $O(\alpha^2)$. The two cross terms in the second moment of $\alpha \mathcal{L}_{\mathcal{S}} \varepsilon + u_\alpha$ are $O(\alpha^3)$, and $\mathbb{E}[u_\alpha u_\alpha^\top] = O(\alpha^4)$. Subtracting the outer product of the bias proves (46).

The exact first-order comparison. We have $U^\top U = I_m$ and $J_\star = \Gamma^{1/2} U G^{1/2}$. Hence

$$J_\star^\top \mathcal{S} J_\star = G^{1/2} A G^{1/2}, \qquad \mathcal{V}_{\mathcal{S}} = G^{-1/2} A^{-1} U^\top B^2 U A^{-1} G^{-1/2}.$$

Using the orthogonal decomposition $I_N = UU^\top + (I_N - UU^\top)$ gives

$$U^\top B^2 U = (U^\top B U)^2 + U^\top B (I_N - UU^\top) B U = A^2 + \mathcal{R}^\top \mathcal{R}.$$

Substitution proves (48). The normalized covariance is

$$G^{1/2} \mathcal{V}_{\mathcal{S}} G^{1/2} = I_m + (\mathcal{R} A^{-1})^\top (\mathcal{R} A^{-1}),$$

which proves both assertions in (49). Since A and G are invertible, the excess vanishes exactly when $\mathcal{R} = 0$. This is precisely the stated invariance of the column space of U. □

Interpretation and nondegeneracy. The column space of U is the covariance-rescaled model tangent space: the prediction changes available to the model to first order. The matrix $\mathcal{R}$ measures mixing between this space and its orthogonal complement. The covariance penalty depends on this mixing, not on unequal weights within either space. This is a refinement of the linear GLS covariance decomposition in Proposition A.1, not a different optimality principle from GLS (Kariya & Kurata, 2004). The comparisons concern the coefficient of α^2, not an exact finite-noise covariance ordering or a finite-noise BLUE claim.

The expansion is around a model that represents the noiseless targets. At a reference fit with nonzero residual, the derivative of the estimating equation also contains second derivatives of the predictor multiplied by that residual, so the displayed covariance formula need not apply. Bounded target errors supply a common local neighborhood for all noise realizations; finite covariance alone is not being used as a substitute for this localization condition.

When $m < N$, target perturbations can contain components outside the locally representable prediction directions, and their weighting can change estimation variance. When $m = N$, U is square orthogonal and $R = 0$ for every positive-definite weighting. In this locally saturated case, the inverse-function theorem gives the same exact interpolating local fit for all such weightings.

A genuinely nonlinear example with a strict gain. Let $n = 2$, $q = m = 1$ and let $v_{\mathrm{sh}} = (1,1)^\top/\sqrt{2}$ and $v_{\mathrm{diff}} = (1,-1)^\top/\sqrt{2}$. Take

$$\mathbf{f}(\theta) = \theta(v_{\mathrm{sh}} + v_{\mathrm{diff}}) + \theta^2 v_{\mathrm{diff}}, \qquad \theta_\star = 0, \qquad \varepsilon = 2\xi_1 v_{\mathrm{sh}} + \xi_2 v_{\mathrm{diff}},$$

where ξ_1, ξ_2 are independent and each equals 1 or -1 with equal probability. The errors are bounded and centered, and $\Gamma = 4v_{\mathrm{sh}}v_{\mathrm{sh}}^\top + v_{\mathrm{diff}}v_{\mathrm{diff}}^\top \succ 0$. The derivative $J_\star = v_{\mathrm{sh}} + v_{\mathrm{diff}}$ has rank one; the quadratic term has a component outside its span. For uniform sampling, positive scalar factors in the weighting cancel. Supervised least squares therefore has $\mathcal{V}_{\mathrm{SL}} = 5/4$. The two-example cross transition at $\gamma = 3/5$ has weights $2/5$ and $8/5$ on the shared and difference directions, so its weighting is proportional to Γ^{-1} and $\mathcal{V}_P = \mathcal{V}_{\mathrm{GLS}} = 4/5$. Thus local parameter MSE is strictly smaller for TD at all sufficiently small positive noise levels.

B.3 Consequences for nonlinear prediction and rollout

Corollary B.4 (Task-specific excess risk and exact local comparison)**.** *Under Theorem B.3, let Φ map a neighborhood of $\theta_\star$ into $\mathbb{R}^s$ and be twice continuously differentiable. Fix this task map independently of the target errors, define $K \overset{def}{=} D\Phi(\theta_\star) \in \mathbb{R}^{s\times m}$, and write $r_{\Phi,\mathcal{S}} \overset{def}{=} \mathrm{tr}(K\mathcal{V}_{\mathcal{S}}K^\top)$. Then*

$$\mathbb{E}\Big[\|\Phi(\widehat{\theta}_{\mathcal{S}}^{(\alpha)}) - \Phi(\theta_\star)\|_2^2\Big] = \alpha^2 r_{\Phi,\mathcal{S}} + O(\alpha^3), \tag{50}$$

and the leading excess over local oracle GLS is

$$r_{\Phi,\mathcal{S}} - r_{\Phi,\mathrm{GLS}} = \|\mathcal{R}A^{-1}G^{-1/2}K^\top\|_F^2, \tag{51}$$

where $\|\cdot\|_F$ is the Frobenius norm. For TD, evaluate $A, \mathcal{R}, \delta_{\tan}$ at $\mathcal{S}_P$. A sufficient condition for strict improvement over the local supervised estimator is

$$r_{\Phi,\mathrm{SL}} - r_{\Phi,\mathrm{GLS}} > \delta_{\tan}^2 r_{\Phi,\mathrm{GLS}}. \tag{52}$$

Whenever the actual leading risks satisfy $r_{\Phi,P} < r_{\Phi,\mathrm{SL}}$, the same strict inequality holds between the exact task MSEs for all sufficiently small $\alpha > 0$.

Proof. Taylor expansion on a fixed small neighborhood gives

$$\Phi(\widehat{\theta}_{\mathcal{S}}^{(\alpha)}) - \Phi(\theta_\star) = \alpha K\mathcal{L}_{\mathcal{S}}\varepsilon + O(\alpha^2)$$

uniformly. Squaring and averaging proves (50). Taking the trace after applying K to (48) gives (51). Equation (49) gives $r_{\Phi,P} \le (1+\delta_{\tan}^2) r_{\Phi,\mathrm{GLS}}$, which proves the sufficient condition. For the final assertion, use a common neighborhood and noise range for the two fixed weightings and subtract their expansions. If $d_0 = r_{\Phi,\mathrm{SL}} - r_{\Phi,P} > 0$, the difference in exact risks is $\alpha^2 d_0 + O(\alpha^3) > 0$ for sufficiently small positive α. □

A finite-horizon autoregressive task. Let $\Psi_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ be the state-update map used for autoregressive prediction, where d_z is the state dimension. The state may include the full input history, so this also describes window-based predictors. Fix an initial state z_0 and a finite integer $H \geq 1$, and set

$$z^\star_{t+1} = \Psi_{\theta_\star}(z^\star_t), \qquad z^\star_0 = z_0, \qquad \Phi_H(\theta) = \Psi^{\circ H}_\theta(z_0),$$

where $\Psi^{\circ H}_\theta$ denotes H successive applications. Suppose $(\theta, z) \mapsto \Psi_\theta(z)$ is twice continuously differentiable near the reference parameter and all the reference states through this horizon. Continuity and finite composition give a parameter neighborhood on which Φ_H is twice continuously differentiable. Its derivative $K_H = D\Phi_H(\theta_\star)$ can be computed by

$$K_0 = 0, \qquad K_{t+1} = D_z\Psi_{\theta_\star}(z^\star_t)K_t + D_\theta\Psi_{\theta_\star}(z^\star_t), \quad 0 \leq t < H. \tag{53}$$

This follows by the chain rule applied to each rollout step. Stacking several predicted steps gives another valid task map. Corollary B.4 therefore extends the fixed-horizon, small-noise comparison in Corollary A.4 to this nonlinear setting. To interpret the discrepancy as error relative to the physical dynamics, the reference map $\Psi_{\theta_\star}$ must describe those dynamics on the evaluated trajectory; fitting the noiseless training targets alone does not establish that property.

B.4 A SHARP BOUND INDEPENDENT OF THE MODEL DERIVATIVES

For a symmetric positive-definite matrix B, write $\mathrm{cond}_2(B) = \lambda_{\max}(B)/\lambda_{\min}(B)$. The factor below is the same scale-invariant Kantorovich factor already given in Remark A.2; the additional assertion here is its exact worst-case interpretation for the local nonlinear covariance.

Theorem B.5 (Sharp worst-case first-order covariance factor)**.** *Use the local matrices from Theorem B.3. Let $b_- = \lambda_{\min}(B)$ and $b_+ = \lambda_{\max}(B)$, and define*

$$\chi \overset{def}{=} \frac{b_+}{b_-}, \qquad \kappa(\chi) \overset{def}{=} \frac{(\chi+1)^2}{4\chi} = 1 + \frac{(\chi-1)^2}{4\chi}. \tag{54}$$

Then

$$\mathcal{V}_{\mathrm{GLS}} \preceq \mathcal{V}_{\mathcal{S}} \preceq \kappa(\chi)\mathcal{V}_{\mathrm{GLS}}. \tag{55}$$

For fixed $\Gamma, \mathcal{S} \succ 0$, $N \geq 2$, and every fixed $1 \leq m < N$, this constant is sharp over all full-column-rank Jacobians:

$$\sup_{J_\star \in \mathbb{R}^{N\times m}:\, \mathrm{rank}(J_\star)=m} \lambda_{\max}(G^{1/2}\mathcal{V}_{\mathcal{S}}G^{1/2}) = \kappa(\chi). \tag{56}$$

For $m = N$ the left side is 1, including when $N = 1$. The function κ is strictly increasing on $[1, \infty)$.

Proof. The covariance normalization in Theorem B.3 is the same as in the proof of Theorem 4.2. The matrix Kantorovich argument established there, with spectral endpoints b_-, b_+, gives

$$U^\top B^2 U \preceq \frac{(b_- + b_+)^2}{4b_-b_+}(U^\top B U)^2 = \kappa(\chi)A^2.$$

Multiplying this inequality on the left by $G^{-1/2}A^{-1}$ and on the right by $A^{-1}G^{-1/2}$ gives the upper bound; (48) gives the lower bound. This use of the earlier matrix argument requires only positive definiteness and $U^\top U = I_m$, not a new transition assumption (Liu et al., 2010).

If $b_- = b_+$, then $B = b_- I_N$, so every normalized covariance is I_m and sharpness is immediate. Otherwise choose orthonormal eigenvectors v_-, v_+ for the distinct extreme eigenvalues and set

$$u = \sqrt{\frac{b_+}{b_- + b_+}}v_- + \sqrt{\frac{b_-}{b_- + b_+}}v_+.$$

For $m = 1$, take $J_\star = \Gamma^{1/2}u$, so $G = 1$ and

$$u^\top B u = \frac{2b_-b_+}{b_- + b_+}, \qquad u^\top B^2 u = b_-b_+, \qquad \mathcal{V}_{\mathcal{S}} = \frac{u^\top B^2 u}{(u^\top B u)^2} = \kappa(\chi).$$

For $1 < m < N$, adjoin $m-1$ orthonormal eigenvectors of B perpendicular to v_-, v_+. Such vectors exist because $m - 1 \leq N - 2$. Use these and u as the columns of U, and set $J_\star = \Gamma^{1/2}U$. Then $G = I_m$; the first normalized covariance entry is $\kappa(\chi)$ and the other entries are 1, with zero cross entries. This proves (56). For $m = N$, the square-orthogonal case of Theorem B.3 applies. Finally $\kappa'(\chi) = (1 - \chi^{-2})/4$, which is positive for $\chi > 1$. □

A smaller χ thus means a smaller sharp worst-case inflation factor relative to local GLS. It does not compare the actual TD and SL risks for a particular Jacobian, nor is it a minimax theorem over arbitrary estimators or training algorithms. The exact comparison for a fixed model and task is (51). The sharpness statement allows all full-column-rank Jacobian matrices; restricting to a specific architecture can give a smaller worst-case factor.

B.5 The two-mode surrogate, the discount, and robustness

Use the trajectory count M and the cross-transition block P_M from Definition A.5. Define the orthogonal projectors

$$Q_{\mathrm{sh}} \stackrel{\mathrm{def}}{=} \frac{\mathbf{1}\mathbf{1}^\top}{M}, \qquad Q_{\mathrm{diff}} \stackrel{\mathrm{def}}{=} I_M - Q_{\mathrm{sh}}.$$

For one output coordinate and one aligned time block, consider the explicit sample-covariance surrogate

$$C_{2\mathrm{mode}} = \lambda_{\mathrm{sh}} Q_{\mathrm{sh}} + \lambda_{\mathrm{diff}} Q_{\mathrm{diff}}, \qquad \lambda_{\mathrm{sh}}, \lambda_{\mathrm{diff}} > 0. \tag{57}$$

This specifies the variance of the shared direction and an equal variance for every orthogonal difference direction, with no covariance between these directions. By Remark A.7, the corresponding TD weighting, omitting its irrelevant positive sampling scalar, is

$$S_\gamma^{\mathrm{block}} \stackrel{\mathrm{def}}{=} I_M - \gamma P_M = (1-\gamma) Q_{\mathrm{sh}} + \left(1 + \frac{\gamma}{M-1}\right) Q_{\mathrm{diff}}. \tag{58}$$

Proposition B.6 (Worst-case factor and discount sensitivity under the surrogate)**.** *Under* (57)–(58)*, define*

$$\begin{aligned} b_{\mathrm{sh}}(\gamma) &= \lambda_{\mathrm{sh}}(1-\gamma), \qquad b_{\mathrm{diff}}(\gamma) = \lambda_{\mathrm{diff}} \left(1 + \frac{\gamma}{M-1}\right), \\ \chi_{2\mathrm{mode}}(\gamma) &= \frac{\max\{b_{\mathrm{sh}}(\gamma), b_{\mathrm{diff}}(\gamma)\}}{\min\{b_{\mathrm{sh}}(\gamma), b_{\mathrm{diff}}(\gamma)\}}, \qquad \kappa_{2\mathrm{mode}}(\gamma) = \frac{(\chi_{2\mathrm{mode}}(\gamma)+1)^2}{4\chi_{2\mathrm{mode}}(\gamma)}. \end{aligned} \tag{59}$$

Then $\chi_{2\mathrm{mode}}$ is exactly the condition number of $C_{2\mathrm{mode}}^{1/2} S_\gamma^{\mathrm{block}} C_{2\mathrm{mode}}^{1/2}$, and $\kappa_{2\mathrm{mode}}$ is its sharp worst-case local covariance factor from Theorem B.5. Writing $R = \lambda_{\mathrm{sh}}/\lambda_{\mathrm{diff}}$, its unique minimizer over $\gamma \in [0,1)$ is

$$\gamma_\star = \begin{cases} 0, & R \le 1, \\ \dfrac{(M-1)(R-1)}{(M-1)R+1}, & R > 1. \end{cases} \tag{60}$$

For $R \ge 1$, the optimum gives $\chi_{2\mathrm{mode}}(\gamma_\star) = 1$ and exact first-order GLS matching for every local model in this surrogate. For $R > 1$, the following exact identity describes discount sensitivity:

$$\kappa_{2\mathrm{mode}}(\gamma) - 1 = \frac{[(M-1)R+1]^2(\gamma-\gamma_\star)^2}{4R(M-1)(1-\gamma)(M-1+\gamma)}. \tag{61}$$

In particular, the excess is quadratic near $\gamma_\star$. For a tolerance $\epsilon \ge 0$, define

$$t_\gamma \stackrel{\mathrm{def}}{=} R\frac{1-\gamma}{1+\gamma/(M-1)}, \qquad t_\pm(\epsilon) \stackrel{\mathrm{def}}{=} \left(\sqrt{1+\epsilon} \pm \sqrt{\epsilon}\right)^2.$$

Then, for every $R > 0$,

$$\kappa_{2\mathrm{mode}}(\gamma) \le 1+\epsilon \quad \Longleftrightarrow \quad t_-(\epsilon) \le t_\gamma \le t_+(\epsilon). \tag{62}$$

Proof. The two projectors are symmetric, orthogonal, and sum to I_M. They simultaneously diagonalize the covariance and the weighting, so the rescaled weighting has eigenvalues b_{sh} and b_{diff}. This proves the condition-number and sharp-factor claims by Theorem B.5. The ratio $t_\gamma = b_{\mathrm{sh}}/b_{\mathrm{diff}}$ strictly decreases from R to zero as γ increases from zero to one. Since $\chi_{2\mathrm{mode}} = \max(t_\gamma, t_\gamma^{-1})$, its unique minimum occurs at zero discount when $R \le 1$, and at the unique solution of $t_\gamma = 1$ when $R > 1$. Solving that equation gives (60); $R = 1$ is matched at $\gamma = 0$.

For any positive t_γ,

$$\kappa_{2\mathrm{mode}}(\gamma) = \frac{(t_\gamma+1)^2}{4t_\gamma} = 1 + \frac{(t_\gamma-1)^2}{4t_\gamma}.$$

For $R > 1$,

$$t_\gamma - 1 = -\frac{[(M-1)R+1](\gamma-\gamma_\star)}{M-1+\gamma},$$

which proves (61) by substitution. Its denominator is positive and smooth at the interior optimum. The Taylor coefficient of $(\gamma-\gamma_\star)^2$ is consequently $c_\star^2/4$, where $c_\star = M/[(1-\gamma_\star)(M-1+\gamma_\star)]$; the remaining terms are $O(|\gamma-\gamma_\star|^3)$. Finally the tolerance condition is $(t_\gamma-1)^2 \le 4\epsilon t_\gamma$. The two roots of the corresponding quadratic are $t_-(\epsilon)$ and $t_+(\epsilon)$, proving (62). □

Equation (62) is an explicit algebraic sensitivity interval. More directly, define $g_R(t) = (M-1)(R-t)/[(M-1)R+t]$ for $t > 0$; the allowed discounts are $[g_R(t_+(\epsilon)), g_R(t_-(\epsilon))] \cap [0,1)$, possibly empty. The quadratic expansion is local: it does not imply uniform insensitivity as $\gamma \uparrow 1$. Under the exchangeable covariance of Assumption A.6, $R = [1+(M-1)\rho]/(1-\rho)$, so (60) reduces to the already established γ_ρ in (33). It is an oracle surrogate minimizer, not an assumption that the empirical discount was chosen by this formula.

Multiple outputs and the scope of sample-space matching. For the entire training array, the same two-mode factor applies to the explicit repeated-block surrogate

$$\widetilde{\Gamma} = I_q \otimes I_T \otimes C_{2\mathrm{mode}}, \qquad \mathcal{S}_P = \frac{1}{MT} I_q \otimes I_T \otimes S_\gamma^{\mathrm{block}}. \tag{63}$$

The positive scalar cancels, and the two rescaled eigenvalues are simply repeated. This surrogate includes uncorrelated, equal-scale output coordinates and identical, uncorrelated time blocks. For the paper's more general separable covariance with $\Omega \succ 0$, however,

$$\begin{aligned} \Gamma^{1/2}\mathcal{S}_P\Gamma^{1/2} &= \Omega \otimes (C^{1/2} S_P C^{1/2}), \\ \mathrm{cond}_2(\Gamma^{1/2}\mathcal{S}_P\Gamma^{1/2}) &= \mathrm{cond}_2(\Omega)\,\mathrm{cond}_2(C^{1/2} S_P C^{1/2}). \end{aligned} \tag{64}$$

Thus sample-space matching $S_P = cC^{-1}$ alone is not sufficient for GLS efficiency for every nonlinear model with parameters shared across outputs. The exact condition remains $\mathcal{R} = 0$. This does not contradict the linear result: with $J_\star = I_q \otimes X$, the rescaled model column space is $\mathbb{R}^q \otimes \mathrm{col}(C^{-1/2}X)$ and is invariant under $\Omega \otimes I_n$; here col denotes column space. The additional $\Gamma \succ 0$ assumption is used only for the inverses in this nonlinear analysis; it does not strengthen the assumptions of the existing linear results, which allow singular Ω.

B.6 Interpretation of the residual diagnostic

The residual diagnostic uses fixed SL validation residuals to estimate shared- and difference-mode energies $\widehat{\lambda}_{\mathrm{sh}}$ and $\widehat{\lambda}_{\mathrm{diff}}$ and forms $\chi_{2\mathrm{mode}}(\gamma)$. Under the explicit two-mode covariance surrogate of Proposition B.6, $\chi_{2\mathrm{mode}}$ is exactly the condition number of the covariance-rescaled weighting, and its corresponding sharp worst-case local covariance factor is

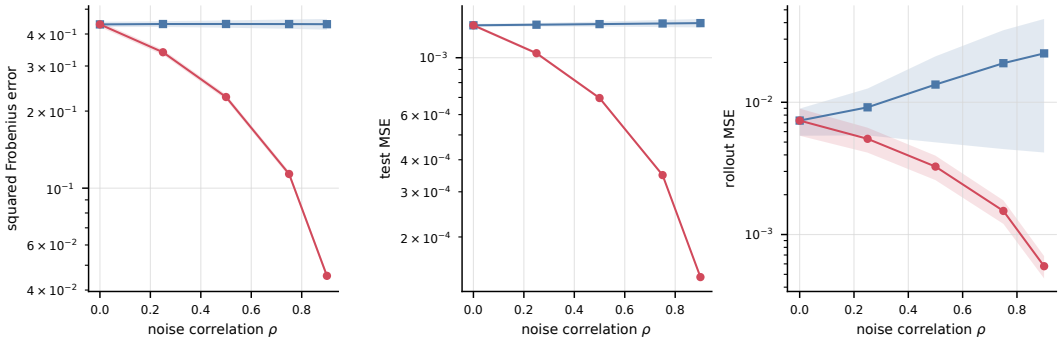
$$\kappa(\chi_{2\mathrm{mode}}) = \frac{(\chi_{2\mathrm{mode}}+1)^2}{4\chi_{2\mathrm{mode}}}.$$

Because κ is increasing for $\chi_{2\mathrm{mode}} \ge 1$, the reduction in $\chi_{2\mathrm{mode}}$ under the selected TD weighting on 14 of 16 eligible domains also reduces this surrogate efficiency factor. The residuals are held fixed in this comparison, so it isolates how the TD discount reweights the same observed residual structure.

The diagnostic summarizes metric-scaled validation residuals through two aggregate modes. It tests whether the selected weighting balances this observed residual structure in the direction identified by the surrogate; it is not an estimate of the full training-error covariance or a risk bound for the trained network.

C Controlled Experiments

This section provides additional results and details for Section 5.1.



(a) Squared parameter error (b) Clean one-step test MSE (c) Clean MSE at rollout step 40

Figure 4: Estimator diagnostics for the representative $q = 16$, $\sigma = 0.3$ slice, over $N = 100$ independent paired systems. In **(a)**–**(c)**, ■ OLS and • cross-TD at γ_ρ; lines are means and bands ± one standard error. These panels keep the absolute scale that the ratio columns of Table 3 cannot show: as the correlation grows the OLS rollout error rises while the cross-TD error falls. All vertical axes are logarithmic.

C.1 Linear Experimental Setup

Data. For each of 100 seeds and $q \in \{8, 16\}$, we draw $A_\star$ with independent standard-normal entries and rescale its spectral radius to 0.95. We generate $M = 32$ training and 1,024 test trajectories from independent $\mathcal{N}(0, I_q)$ initial states, each with 40 transitions and no process noise. Training targets receive the noise defined in Appendix A.4, with $\sigma \in \{0.03, 0.3\}$ and $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$. Inputs and test trajectories remain clean. Every estimator uses the same data within each condition. In time-major sample order, the target-error covariance for each output coordinate is $C = I_T \otimes \sigma^2[(1-\rho)I_M + \rho \mathbf{1}\mathbf{1}^\top]$.

Estimators. We solve the normal equations for OLS, oracle GLS, and TD without ridge regularization. The cross transition averages the other trajectories at the same time. We evaluate the theoretically matched discount $\gamma_\rho = \rho(M-1)/[1+\rho(M-2)]$ and the grid $\{0, 0.1, 0.3, 0.5, 0.9, 0.95, 0.99\}$.

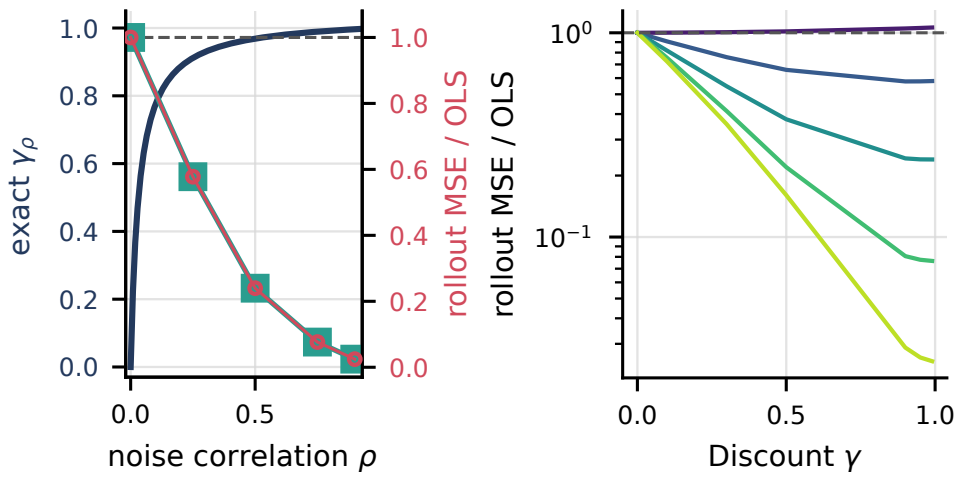
Evaluation. We measure squared Frobenius parameter error, one-step test MSE averaged over all 40 input times, and autoregressive MSE at $H \in \{1, 5, 10, 20, 40\}$. Rollout MSE averages over test trajectories and coordinates at step H, not over steps $1{:}H$. Ratios divide the mean errors over seeds; they are not means of per-seed ratios. Table 3 gives the representative error ratios, and Figure 4 shows absolute errors and standard errors.

C.2 Additional Linear Results

Table 3: Representative controlled linear results for $q = 16$, $\sigma = 0.3$, and $H = 40$. Entries are matched cross-TD error divided by OLS error; values below one favor TD. All ratios divide mean errors over the 100 paired systems.

ρ	γ_ρ	One-step/OLS	Rollout/OLS	Parameter/OLS
0.00	0.0000	1.000	1.000	1.000
0.25	0.9118	0.774	0.578	0.776
0.50	0.9688	0.515	0.240	0.519
0.75	0.9894	0.256	0.0764	0.260
0.90	0.9964	0.102	0.0246	0.104

Discount sensitivity. At each time step, target noise combines a term shared by all $M = 32$ trajectories with an independent term for each trajectory. We vary the resulting correlation ρ while keeping the marginal noise variance fixed. Choosing $\gamma_\rho = \rho(M-1)/[1+\rho(M-2)]$ makes TD's



(a) Matched discount and GLS

(b) Discount sensitivity

Figure 5: **(a)** Left y-axis: — the theoretically matched discount γ_ρ as a function of the same-time cross-trajectory noise correlation ρ. Right y-axis: rollout error relative to OLS (dashed), for ■ oracle GLS and ○ cross-TD at γ_ρ. **(b)** Discount sensitivity in the controlled linear experiment. Step-40 rollout MSE relative to OLS for $q = 16$, $M = 32$, and $\sigma = 0.3$. Each curve uses one target-noise correlation, $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$, ordered from darker to lighter. Ratios divide mean errors over 100 paired systems; the dashed line marks OLS. Larger discounts become beneficial as correlation increases. The vertical axis in (b) is logarithmic.

weighting proportional to the inverse target-error covariance, so TD and GLS produce the same fitted dynamics matrix. Their rollout errors coincide across the tested correlations in Figure 5(a). At $\rho = 0$, both reduce to OLS; at $\rho = 0.9$, their step-40 MSE is 2.46% of OLS.

Figure 5(b) varies the TD discount under the shared-noise model of Appendix A.4. Larger discounts become beneficial as the target-noise correlation increases, consistent with the increasing value of γ_ρ. This complements the estimator comparison in Figure 5(a).

C.3 Controlled covariance mismatch

Purpose and fixed protocol. This experiment tests departures from the exact covariance match while holding the TD estimator fixed. We use the data-generating protocol of Appendix C.1 with $q = 16$, $M = 32$, $T = 40$, $\sigma = 0.3$, 1,024 clean test trajectories, and no process noise. We generate an independent ensemble of 100 systems; for each system we average over 32 independent target-noise realizations. Each $A_\star$ is drawn with independent standard-normal entries and rescaled to spectral radius 0.95. Training and test initial states are independent $\mathcal{N}(0, I_q)$ draws.

Covariance perturbation. Split the trajectories into two fixed groups of 16, independently of their initial states. Let $J_m = \mathbf{1}_m \mathbf{1}_m^\top$ and set

$$C_\delta^{\mathrm{blk}} = \tfrac{1}{2} I_{32} + \tfrac{1}{2}\big[(1-\delta) J_{32} + \delta \ \mathrm{blockdiag}(J_{16}, J_{16})\big], \qquad \delta \in \{0, 0.1, 0.25, 0.5, 1\}. \tag{65}$$

The marginal variance is one, the within-group correlation remains 0.5, and the between-group correlation is $0.5(1-\delta)$. Thus $\delta = 0$ recovers the exchangeable model with $\rho = 0.5$, while $\delta = 1$ gives two independent error groups. For noise realization b at time t, draw $Z_{b,t} \in \mathbb{R}^{32\times 16}$ with independent standard-normal entries and set

$$E_{b,t} = \sigma \ \mathrm{chol}(C_\delta^{\mathrm{blk}}) Z_{b,t}, \qquad C_\delta^{\mathrm{sample}} = I_T \otimes \sigma^2 C_\delta^{\mathrm{blk}},$$

where $\mathrm{chol}(C)$ is the lower-triangular factor satisfying $C = \mathrm{chol}(C)\,\mathrm{chol}(C)^\top$. Noise is independent across times and output coordinates. We reuse the same $Z_{b,t}$ across mismatch levels, and every method receives identical data within each condition. The covariance is positive definite throughout

the sweep. For $\delta > 0$, the variance of the contrast between the two groups differs from the other trajectory-difference variances, whereas cross-TD weights all these difference directions equally. Consequently, no scalar discount can restore exact inverse-covariance matching in this family once $\delta > 0$.

Estimation and uncertainty. OLS, TD, and oracle GLS solve the unregularized normal equations. TD retains the same-time cross-trajectory transition and $\gamma = \gamma_{0.5} = 31/32 = 0.96875$ at every δ; GLS uses $(C_\delta^{\text{sample}})^{-1}$ at that δ. No discount tuning, ridge regularization, stabilization, or run exclusion is applied. Exact rollout MSE is evaluated at step 40, averaged over test trajectories and coordinates, as in Appendix C.1. For each system, errors are first averaged over the 32 noise realizations; reported ratios then divide mean errors across systems, rather than averaging per-system ratios. Pointwise 95% percentile intervals use 10,000 paired bootstrap resamples of the 100 systems. Each resampled system retains all its methods, mismatch levels, and noise realizations; the 3,200 noise realizations are not treated as independent systems.

Direct covariance-bound check. For each fixed design, compute the conditional covariance of one output row of the dynamics estimate,

$$V_\nu = L_\nu C_\delta^{\text{sample}} L_\nu^\top, \qquad \nu \in \{\text{TD}, \text{GLS}\}.$$

Here the full coefficient covariance is $I_q \otimes V_\nu$ and the expected squared Frobenius parameter error is $q\ \text{tr}(V_\nu)$. First-order rollout risk is computed analytically by applying the linear rollout-sensitivity map of Section 4 to this covariance and averaging over the clean test initial states and output coordinates. Let $S^{\text{blk}} = I_{32} - \gamma P_{32}$, and let λ_- and λ_+ be the extreme eigenvalues of $(C_\delta^{\text{blk}})^{1/2} S^{\text{blk}} (C_\delta^{\text{blk}})^{1/2}$. The omitted factors σ^2 and $1/(MT)$ do not affect the following scale-invariant quantities:

$$\eta_\delta = \frac{\lambda_+ - \lambda_-}{\lambda_+ + \lambda_-}, \qquad K_\delta = \frac{1}{1 - \eta_\delta^2}, \quad \lambda_{\max}\Big(V_{\text{GLS}}^{-1/2} V_{\text{TD}} V_{\text{GLS}}^{-1/2}\Big) \le K_\delta. \tag{66}$$

For each of the 100 fixed designs and five mismatch levels, we check both $V_{\text{TD}} \succeq V_{\text{GLS}}$ and the upper bound in Equation (66). Figure 1(a) shows the corresponding parameter-MSE and largest directional variance ratios, together with K_δ. Table 4 also reports first-order rollout risks computed from these covariance matrices.

Results and numerical checks. At $\delta = 0$, TD and GLS agree to numerical precision; the largest relative matrix difference across the new matched fits is 7.75×10^{-14} (rounded upward). All 100 designs have full column rank, all 48,000 fitted matrices yield finite errors, and the largest relative normal-equation residual is below 2.0×10^{-15}. The covariance inequalities hold in every tested condition to numerical tolerance 10^{-8}. As mismatch grows, the expected parameter-MSE ratio increases from 1 to 1.446, and the largest directional ratio reaches 4.828, below the endpoint bound 8.758. Table 4 also shows that exact matching is not necessary for a rollout gain in this family. At the strongest perturbation, the mean TD/OLS rollout ratio remains below one, but its 95% interval extends to 1.005; improvement is therefore not resolved at that endpoint.

Table 4: Covariance-mismatch results. Parameter and first-order rollout columns divide mean analytically computed risks. The exact-rollout column divides mean sampled step-40 errors and reports a pointwise 95% paired-system bootstrap interval.

δ	K_δ	Parameter TD/GLS	First-order rollout TD/GLS	Exact rollout TD/OLS [95% CI]
0	1.000	1.000	1.000	0.303 [0.188, 0.482]
0.1	1.275	1.027	1.035	0.322 [0.197, 0.517]
0.25	1.966	1.091	1.115	0.363 [0.222, 0.578]
0.5	3.491	1.207	1.256	0.454 [0.278, 0.703]
1	8.758	1.446	1.545	0.704 [0.424, 1.005]

C.4 Controlled nonlinear verification

We test the local nonlinear predictions using $\Psi_\theta(z) = W_2 \tanh(W_1 z + b_1) + b_2$, where $z \in \mathbb{R}^2$ and the hidden layer has width two. The fixed reference parameters are

$$W_{1\star} = \begin{pmatrix} 1.10 & -0.70 \\ -0.70 & -1.00 \end{pmatrix}, \quad b_{1\star} = \begin{pmatrix} 0.20 \\ -0.30 \end{pmatrix}, \quad W_{2\star} = \begin{pmatrix} 0.45 & -0.40 \\ 0.43 & 0.35 \end{pmatrix}, \quad b_{2\star} = \begin{pmatrix} 0.02 \\ -0.06 \end{pmatrix}.$$

We draw once $M = 24$ initial states independently and uniformly from $[-2, 2]^2$ and hold the resulting training design fixed across all target-noise draws. Applying $\Psi_{\theta_\star}$ gives the four clean training inputs $z_{r,t}$, $t = 0, \dots, 3$. For evaluation, we fix $z_0^{\mathrm{test}} = (1.2, -1.1)^\top$ and compare the 20-step endpoints $\Phi_{20}(\theta) = \Psi_\theta^{\circ 20}(z_0^{\mathrm{test}})$. Training inputs and the reference rollout are noise-free.

Only the one-step targets are perturbed:

$$y_{r,t,k}^{(\alpha)} = \Psi_{\theta_\star}(z_{r,t})_k + \alpha \epsilon_{r,t,k}, \qquad \epsilon_{r,t,k} = \sqrt{0.7}\, u_{t,k} + \sqrt{0.3}\, v_{r,t,k},$$

where all $u_{t,k}$ and $v_{r,t,k}$ are independent symmetric $\{-1, 1\}$ signs. The errors are centered and bounded, as required by Theorem B.3. With coordinates ordered by output, time, and trajectory, their covariance is $\Gamma = I_2 \otimes I_4 \otimes (0.3 I_{24} + 0.7 \mathbf{1}\mathbf{1}^\top)$.

Up to a positive scalar, $S_{\mathrm{SL}} = I_{192}$, $S_{\mathrm{TD}} = I_2 \otimes (I_{96} - \gamma P)$, and $S_{\mathrm{GLS}} = \Gamma^{-1}$, where $P = I_4 \otimes (\mathbf{1}\mathbf{1}^\top - I_{24})/23$. Matched TD uses $\gamma_\rho \approx 0.981707$; intentionally mismatched TD uses $\gamma = \gamma_\rho/2$. The training prediction Jacobian has full column rank (12 columns and 192 rows), with condition number 24.92.

For each $\alpha \in \{0.01, 0.04, 0.07, 0.10\}$, we use 2400 independent target-noise draws, reusing each draw across methods and perturbation scales. Each fit starts at $\theta_\star$ and uses deterministic weighted nonlinear least squares with an analytic Jacobian. All fits satisfied the deterministic solver's convergence criterion. Under the symmetric weighting used here, the gradient of $\frac{1}{2} e(\theta)^\top \mathcal{S}_P e(\theta)$ is $J_\theta^\top \mathcal{S}_P e(\theta)$, the mean TD fixed-point equation in Proposition B.1. Solving weighted nonlinear least squares therefore isolates the estimator analyzed by the local theory from the choice of iterative solver. Because the matched-TD and GLS objectives are exactly proportional in this controlled setting, they define the same local estimator; we therefore compute this estimator once per target-noise draw. Separate numerical refits confirm that the optimizer returns the same solution.

Figure 2 compares the empirical parameter sample-covariance trace and the mean squared error at the 20-step rollout endpoint, each divided by α^2, with their first-order limits. At $\alpha = 0.01$, mismatched TD gives 18.188 versus 18.189 predicted for parameter covariance and 0.7706 versus 0.7719 predicted for rollout risk. At $\alpha = 0.10$, SL parameter covariance rises to 40.408 against its small-noise limit of 36.367, while the rollout measurements remain close to their limits throughout the tested range.

Figure 6 compares fits obtained from the same target-noise draw. GLS orthogonality gives the leading limits

$$\frac{\mathbb{E}\|\hat{\theta}_S - \hat{\theta}_{\mathrm{GLS}}\|_2^2}{\alpha^2} \longrightarrow \mathrm{tr}(V_S - V_{\mathrm{GLS}}), \qquad \frac{\mathbb{E}\|\Phi_{20}(\hat{\theta}_S) - \Phi_{20}(\hat{\theta}_{\mathrm{GLS}})\|_2^2}{\alpha^2} \longrightarrow r_{\Phi_{20},S} - r_{\Phi_{20},\mathrm{GLS}}.$$

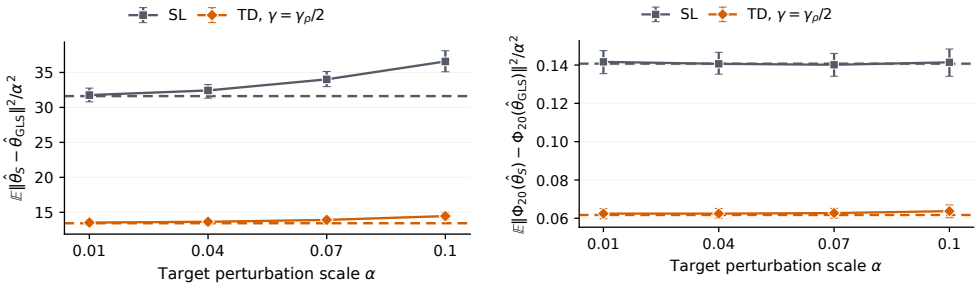
At $\alpha = 0.01$, the paired mismatched-TD distances are 13.536 in parameter space and 0.06244 at the rollout endpoint, versus predicted limits of 13.440 and 0.06169. These paired distances test the leading excess terms; at finite α, they need not equal differences between the corresponding risks.

D Additional PDE Results

This appendix provides the validation measurements and additional comparisons supporting Sections 5.2–5.5. Implementation and selection details are in Appendix E.

D.1 Broad benchmark

Table 6 reports validation error at one-step prediction for the one-step-selected models in Table 5. TD is better at one-step prediction on 17 of the 18 domains, with a median gain of 3.9%; five domains gain more than 10%, and the only regression is Wave Layer at -0.1%.



(a) Parameter distance to GLS

(b) Rollout distance to GLS

Figure 6: **Paired verification of the excess-over-GLS terms.** Markers show mean squared distances between estimators obtained from the same target-noise draw, in parameter space (left) and at the 20-step rollout endpoint (right), each divided by α^2. Error bars are percentile 95% bootstrap intervals (500 resamples). Dashed lines show the corresponding leading-order excess terms, $\mathrm{tr}(V_S - V_{\mathrm{GLS}})$ and $r_{\Phi_{20},S} - r_{\Phi_{20},\mathrm{GLS}}$, from Theorem B.3 and Corollary B.4. As $\alpha \to 0$, the paired distances converge to these excess terms; at finite α, they need not equal differences between the corresponding finite-noise risks. SL is shown as a reference.

Table 5: Flower-Micro held-out test VRMSE; lower values within each pair are bold. Rollout intervals may use different validation-selected models. Intervals stop at the available horizon: 17 steps for CE-RM and Wave Layer, 15 for Viscoelastic, 55 for Supernova, and 46 for TGC. – denotes unavailable measurements.

Domain	SL (1-step)	TD (1-step)	SL (1:20)	TD (1:20)	SL (21:60)	TD (21:60)
Acoustic Maze	0.03744	**0.03577**	0.23840	**0.22420**	0.42470	**0.38650**
Active Matter	0.21284	**0.20000**	**2.26230**	2.56380	**1.37990**	1.43050
Gray–Scott	0.01238	**0.00832**	0.62370	**0.58960**	**1.33690**	1.80060
MHD 64	0.33203	**0.32288**	1.06270	**1.02410**	1.58090	**1.42360**
PlanetSWE	**0.00614**	0.00674	0.08360	**0.07960**	0.41450	**0.38440**
Neutron Star Merger	**0.48208**	0.48216	0.85000	**0.80880**	1.02800	**0.98580**
Rayleigh–Bénard	0.24603	**0.23878**	**6.45769**	7.29512	3.92771	**3.80897**
Rayleigh–Taylor	0.21013	**0.19689**	5.66950	**3.20110**	7.03040	**2.52880**
Shear Flow	0.04735	**0.03975**	0.40020	**0.31210**	3.68150	**1.50970**
Supernova	0.40587	**0.40260**	0.98550	**0.86840**	2.06560	**1.36200**
TGC	0.24951	**0.24152**	1.37740	**1.37030**	2.56390	**2.35450**
TRL 2D	0.43694	**0.39537**	0.92600	**0.78590**	1.44910	**1.26850**
TRL 3D	0.49167	**0.47247**	0.90310	**0.75110**	1.71780	**0.93190**
Viscoelastic	0.68037	**0.60713**	0.93760	**0.86170**	–	–
CE-RM	0.22379	**0.22014**	0.34050	**0.33410**	–	–
Wave Layer	**0.03048**	0.03050	0.29560	**0.27600**	–	–
Diffusion–Reaction	0.00517	**0.00477**	0.08990	**0.08390**	0.07130	**0.06350**
Shallow Water	0.00387	**0.00371**	**0.03820**	0.03900	0.10580	**0.07700**

Table 7 reports the validation errors of the checkpoints used in Table 5. TD is better on 16 of 18 domains at 1:20 (median gain 5.4%) and on 13 of 15 at 21:60 (median gain 8.6%).

D.2 Model and data scale

Original scale. We select configurations by validation VRMSE separately for one-step prediction and each rollout interval, comparing the original learning rates with rates four times as large. All selected configurations use the original rate. Table 8 provides the one-step validation and test results, and Table 14 lists the selected hyperparameters.

Table 6: One-step validation VRMSE of the models reported in the one-step columns of Table 5. Gains are relative to SL.

Domain	SL	TD	Gain (%)
Acoustic Maze	0.03836	**0.03662**	+4.5
Active Matter	0.26217	**0.23485**	+10.4
Gray–Scott	0.01475	**0.00981**	+33.5
MHD 64	0.33576	**0.32612**	+2.9
PlanetSWE	0.01762	**0.01330**	+24.5
Neutron Star Merger	0.56213	**0.55677**	+0.9
Rayleigh–Bénard	0.24853	**0.24357**	+2.0
Rayleigh–Taylor	0.20686	**0.19391**	+6.3
Shear Flow	0.04506	**0.03969**	+11.9
Supernova	0.39030	**0.38739**	+0.8
TGC	0.25080	**0.24201**	+3.5
TRL 2D	0.46355	**0.45586**	+1.7
TRL 3D	0.49139	**0.47297**	+3.8
Viscoelastic	0.68520	**0.61203**	+10.7
CE-RM	0.22409	**0.22049**	+1.6
Wave Layer	**0.02925**	0.02927	-0.1
Diffusion–Reaction	0.00510	**0.00476**	+6.6
Shallow Water	0.00365	**0.00350**	+4.1

Table 7: Validation rollout VRMSE corresponding to the test rollout columns of Table 5. Intervals stop at the available horizon: 17 steps for CE-RM and Wave Layer, 15 for Viscoelastic, 55 for Supernova, and 46 for TGC. – denotes unavailable measurements.

Domain	SL (1:20)	TD (1:20)	SL (21:60)	TD (21:60)
Acoustic Maze	0.24190	**0.22730**	0.42200	**0.38650**
Active Matter	1.12900	**1.10360**	1.25910	**1.25270**
Gray–Scott	0.65810	**0.59380**	1.40880	**1.36840**
MHD 64	1.07730	**1.05110**	1.57570	**1.44000**
PlanetSWE	0.08550	**0.08290**	0.42180	**0.41790**
Neutron Star Merger	1.01450	**0.96300**	1.06650	**1.01460**
Rayleigh–Bénard	**6.69588**	7.88683	**4.05213**	4.07050
Rayleigh–Taylor	5.31610	**3.01710**	6.11810	**2.06620**
Shear Flow	0.39790	**0.30140**	3.69100	**1.60200**
Supernova	0.97960	**0.87870**	2.02670	**1.44770**
TGC	1.38160	**1.35670**	2.68080	**2.36500**
TRL 2D	0.91040	**0.76490**	1.39610	**1.22560**
TRL 3D	0.90400	**0.77300**	1.71280	**0.93660**
Viscoelastic	0.94810	**0.86580**	–	–
CE-RM	0.34130	**0.33520**	–	–
Wave Layer	0.30090	**0.28480**	–	–
Diffusion–Reaction	0.08920	**0.08440**	**0.06600**	0.06640
Shallow Water	**0.03570**	0.03690	0.10290	**0.07260**

Table 8: Flower-Tiny one-step validation and held-out test VRMSE at the validation-selected learning rate. Test measurements also appear in Table 1.

Domain	Selected LR	Val SL	Val TD	Test SL	Test TD
Acoustic Maze	1×10^{-3}	0.00549	**0.00544**	0.00537	**0.00534**
Gray–Scott	5×10^{-4}	0.00270	**0.00207**	0.00212	**0.00168**
Shear Flow	5×10^{-4}	0.01561	**0.01097**	0.01640	**0.01148**

Data and model scaling. Table 9 gives validation and test measurements for the eight validation-selected models in the Shear Flow capacity–data grid. TD improves all four conditions on both validation (4.01%–7.79%) and test (2.47%–9.86%).

Table 9: Shear Flow scaling: validation and held-out test one-step VRMSE of validation-selected models. Fractions denote model and training-data size.

Setting	Val SL	Val TD	Test SL	Test TD
M 5% / D 5%	0.14742	**0.13632**	0.14954	**0.13638**
M 5% / D 20%	0.05839	**0.05605**	0.06067	**0.05917**
M 20% / D 5%	0.12601	**0.11845**	0.12564	**0.11786**
M 20% / D 20%	0.04461	**0.04113**	0.04642	**0.04184**

D.3 Virtual-transition ablations

Transition family. Table 2 compares the six transition families using the grids in Appendix E.5. The cross transition gives the lowest validation error on Acoustic Maze and Shear Flow, improving on the best alternative by 0.2% (random time) and 7.3% (global uniform). On Gray–Scott the random-time transition gives the lowest validation error, 9.0% below cross. With these validation-selected discounts fixed, cross has the lowest test error on Gray–Scott and Shear Flow, while random-time is slightly better on Acoustic Maze.

D.4 Backbone transfer

Table 10 reports validation and test VRMSE for scOT, FNO, and CNextU-Net on Acoustic Maze, Gray–Scott, and Shear Flow.

Table 10: Backbone comparison: validation and held-out test one-step VRMSE. Flower reference rows use the same fixed domain-specific discounts as the alternative backbones (Appendix E.5).

Domain	Backbone	Val SL	Val TD	Test SL	Test TD
Acoustic Maze	Flower	0.03836	**0.03667**	0.03744	**0.03595**
	FNO	0.34455	**0.34343**	0.34267	**0.34219**
	scOT	**0.21701**	0.22172	**0.21299**	0.21723
	CNextU-Net	0.14924	**0.13980**	0.14593	**0.13708**
Gray–Scott	Flower	0.01475	**0.01320**	0.01238	**0.01075**
	FNO	0.15360	**0.13559**	0.12625	**0.11233**
	scOT	1.13083	**1.11546**	0.87453	**0.86306**
	CNextU-Net	0.23386	**0.20569**	0.18534	**0.16575**
Shear Flow	Flower	0.04506	**0.03969**	0.04735	**0.03975**
	FNO	0.16631	**0.16123**	0.16536	**0.15252**
	scOT	0.23990	**0.19878**	0.26455	**0.22363**
	CNextU-Net	**0.37333**	0.38189	**0.36926**	0.37509

D.5 Alternative objectives

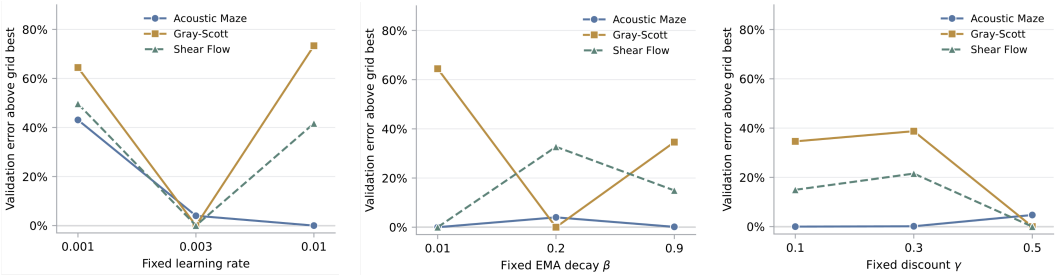
Table 11 compares one-step and multistep SL/TD on one-step prediction and rollout steps 1:20 and 21:60. For multistep training ($H = 2, 4$), TD lowers test VRMSE in four of six one-step comparisons and nine of twelve rollout comparisons. The one-step-trained reference pairs also favor TD on all three domains for one-step prediction. These comparisons use the selected configurations at each H, rather than holding all other hyperparameters fixed.

D.6 Hyperparameter sensitivity

We examine sensitivity to the three principal optimization and TD hyperparameters on Acoustic Maze, Gray–Scott, and Shear Flow. For a hyperparameter h and fixed candidate value v, let $E_{h=v}$

Table 11: Flower-Micro validation and held-out test VRMSE for one-step and multistep training. H is the training unroll length. Configurations are selected by validation separately for each evaluation metric. Selected hyperparameters are in Table 18. Bold marks the lower value within each SL/TD pair.

			Validation		Test	
Domain	H	Evaluation	SL	TD	SL	TD
Acoustic Maze	1	One-step	0.03836	**0.03662**	0.03744	**0.03577**
	1	1:20	0.24190	**0.22730**	0.23840	**0.22420**
	1	21:60	0.42200	**0.38650**	0.42470	**0.38650**
	2	One-step	0.04331	**0.03801**	0.04225	**0.03736**
	2	1:20	0.18538	**0.16303**	0.18126	**0.16127**
	2	21:60	0.30357	**0.25755**	0.29622	**0.25657**
	4	One-step	**0.05334**	0.06770	**0.05226**	0.06714
	4	1:20	**0.19051**	0.22397	**0.18650**	0.22339
	4	21:60	**0.27777**	0.39914	**0.27441**	0.40881
Gray–Scott	1	One-step	0.01475	**0.00981**	0.01238	**0.00832**
	1	1:20	0.65810	**0.59380**	0.62370	**0.58960**
	1	21:60	1.40880	**1.36840**	**1.33690**	1.80060
	2	One-step	0.02841	**0.01720**	0.02937	**0.01588**
	2	1:20	0.68018	**0.63673**	0.66983	**0.63212**
	2	21:60	**1.23386**	1.58232	**1.17044**	2.29837
	4	One-step	0.02152	**0.01784**	0.02354	**0.01993**
	4	1:20	0.71786	**0.63767**	0.68668	**0.63746**
	4	21:60	2.81251	**1.24798**	3.01577	**1.81255**
Shear Flow	1	One-step	0.04506	**0.03969**	0.04735	**0.03975**
	1	1:20	0.39790	**0.30140**	0.40020	**0.31210**
	1	21:60	3.69100	**1.60200**	3.68150	**1.50970**
	2	One-step	0.05825	**0.05034**	0.06188	**0.05158**
	2	1:20	0.36974	**0.21096**	0.35237	**0.21251**
	2	21:60	2.47992	**0.89765**	2.45313	**0.88322**
	4	One-step	**0.05774**	0.06302	**0.05825**	0.06370
	4	1:20	0.21997	**0.15917**	0.21689	**0.15630**
	4	21:60	0.76367	**0.57396**	0.73009	**0.56811**



(a) Learning rate (b) EMA decay (c) TD discount

Figure 7: **Validation sensitivity to fixed hyperparameters.** For each fixed value, we report the percentage increase in validation error relative to the best configuration over the full corresponding search grid, after minimizing over the remaining searched hyperparameters.

denote the lowest validation error among configurations with $h = v$, minimizing over the remaining searched hyperparameters, and let E_{best} denote the lowest validation error over the complete grid. We report

$$100\frac{E_{h=v} - E_{\text{best}}}{E_{\text{best}}}, \tag{67}$$

so zero denotes a grid-optimal value and smaller values indicate less sensitivity to fixing that hyperparameter.

Figure 7 shows that the middle learning rate, 3×10^{-3}, is within roughly 5% of the grid best on all three domains. Likewise, $\gamma = 0.5$ is grid-optimal for Gray–Scott and Shear Flow and remains close to the best setting on Acoustic Maze. EMA decay is more domain dependent: different domains prefer different target-network time scales.

E PDE Implementation Details

This appendix specifies the PDE experiments in Sections 5.2–5.5. Appendix D contains the additional result tables.

E.1 Data, models, and training

Our implementation adapts the Flower codebase (Muser et al., 2026) (MIT licence), retaining its model implementation and data-processing conventions, including its Well-derived data-loading components; Flower in turn adapts the data pipeline of The Well (Ohana et al., 2024), whose VRMSE and learning-rate scheduler implementations we also use. Our additions are the virtual-transition sampling, the TD training objectives, the reduced model and data configurations, the hyperparameter searches, and the checkpoint-selection protocol specified below.

Data and architecture. Following Flower's domain selection, we use 18 PDE domains primarily from The Well (Ohana et al., 2024), with additional cases from PDEgym (Herde et al., 2024) and PDEBench (Takamoto et al., 2022). SL and TD share the data split, training examples, and architecture within each comparison. We construct the reduced Flower-Micro configuration with lifting dimension 16, four levels, four heads, and four groups, trained on 50% of the training trajectories for the broad benchmark. The original-scale study retains Flower's Flower-Tiny configuration with lifting dimension 160 and uses the complete training splits. Validation and test memberships are unchanged by training-data reduction. Each input contains four consecutive ground-truth frames. For one-step training, the target is the next frame. Field normalization uses the dataset statistics specified in the run configuration, shared within each SL/TD pair. Predictions are converted back to physical units for evaluation.

Optimization. Following Flower's optimizer and scheduler choices, the PDE runs use AdamW and linear warmup followed by cosine learning-rate annealing. Learning rates, warmup lengths,

and epoch budgets follow our experiment-specific settings. The Flower and reduced-backbone configurations use weight decay 10^{-2}. The reduced ablation configurations use four warmup epochs; the original-scale configuration uses five. The schedule spans the planned training run, not the epoch of its subsequently selected checkpoint. One-step SL and TD optimize MSE in normalized training coordinates. Except for the explicit multistep controls in Appendix E.3, training predicts one step, not an autoregressive rollout. Epoch budgets and search grids are specified in Appendix E.5; selected learning rates, discounts, and EMA decays are listed in Appendix E.6.

Random-seed protocol. All PDE runs use a single common random seed 0. Reported results are those of the validation-selected trained models, and cross-domain summaries aggregate these results across PDE systems. They do not estimate variation across training seeds.

Virtual transitions: definitions and sampling. Write a training window as $i = (r, t)$, where r identifies its trajectory and t its start-time index. A transition selects a partner window j, whose target-network residual supplies the detached correction in (6). The six rules are:

Rule	Partner of (r, t)
Cross	(s, t), with s uniform among other trajectories having a window at time t.
Forward-adjacent	$(r, t+1)$; at the final window, use the source-only supervised loss.
Symmetric-adjacent	$(r, t-1)$ or $(r, t+1)$ with probability $1/2$ each; a step outside the trajectory stays at (r, t).
Random-time	(r, u), with u uniform among all other window indices of trajectory r.
Global-uniform	A uniform window from the selected training split, including the possibility $j = i$.
Self	$j = i$; the partner prediction still uses the detached target network.

Cross is fixed across the broad benchmark; the other rules are used in the transition ablation. All partner pools contain training windows only. A pairing plan is generated before each epoch and held fixed within that epoch. Partners are drawn from the selected training split, not the current minibatch, and no dense transition matrix is constructed.

For equal-length trajectories, cross sampling shuffles the trajectory IDs and applies a uniformly sampled nonzero cyclic shift to that shuffled list. This gives a one-to-one pairing without self-pairs, with a uniform partner marginal for each trajectory. The same trajectory pairing is used at every time within the epoch. For unequal lengths, the procedure is applied separately to the available trajectories at each time. The symmetric-adjacent, random-time, and global-uniform rules draw a partner for each source window. If random-time has no alternative window, it uses the source-only supervised loss; if cross has no other trajectory at a time, that window is omitted from the TD loss. In all cases, source and partner inputs are ground-truth windows: temporal pairing does not introduce a training rollout.

Target-network update. After each online parameter update, we apply a soft target-network update (Lillicrap et al., 2016):

$$\bar{\theta}_{k+1} = \beta \bar{\theta}_k + (1 - \beta)\theta_{k+1}, \qquad 0 \le \beta < 1. \tag{68}$$

Thus the reported EMA decay β is the weight retained on the previous target parameters. During the online update, the target prediction is detached from the computational graph.

E.2 Checkpoint selection and evaluation

Selection. Validation and test predictions use the online network from the selected checkpoint, not the EMA target network; the latter is used only to construct TD targets during training. Checkpoints are selected using a 20-batch one-step validation diagnostic. We compare retained candidates on one-step validation and report validation and held-out test measurements for the selected models. For rollout comparisons, candidate configurations are selected by the corresponding validation interval where applicable. Test data do not select either checkpoints or hyperparameters.

One-step and rollout samples. One-step evaluation includes every eligible ground-truth-conditioned window, including the final partial batch. Each rollout starts from the initial four

ground-truth frames of a trajectory, then feeds predictions back into the input history. We average VRMSE over prediction steps 1:20 and 21:60, using shorter intervals where required by the available trajectory length, as indicated in the tables.

Metric aggregation. We use The Well's per-sample VRMSE implementation and aggregate its outputs over the evaluation samples and horizons specified here. For evaluation sample i and output channel c, let $y_{i,c}(s)$ and $\widehat{y}_{i,c}(s)$ denote the ground-truth and predicted fields in physical units at spatial grid point s. With S grid points and $\bar{y}_{i,c} = S^{-1}\sum_s y_{i,c}(s)$, the per-sample, per-channel VRMSE is

$$v_{i,c} = \sqrt{\frac{S^{-1}\sum_s (\widehat{y}_{i,c}(s) - y_{i,c}(s))^2}{(S-1)^{-1}\sum_s (y_{i,c}(s) - \bar{y}_{i,c})^2 + 10^{-7}}}.$$

One-step VRMSE is the arithmetic mean of $v_{i,c}$ over windows and output channels. Rollout VRMSE is computed per trajectory, prediction step, and channel, then averaged over trajectories, channels, and the stated step interval. Batch contributions are weighted by their sample counts, rather than taking an unweighted mean of batch means. Thus the metric averages normalized RMS errors; it does not pool squared errors before taking one final square root. We report the percentage gain

$$100\,\frac{\mathrm{VRMSE}_{\mathrm{SL}} - \mathrm{VRMSE}_{\mathrm{TD}}}{\mathrm{VRMSE}_{\mathrm{SL}}}, \tag{69}$$

where positive values favor TD.

E.3 Alternative training objectives

Write $e_i^{(h)}(\theta)$ for the residual of window i after h autoregressive steps, so that $e_i^{(1)}$ is the one-step residual of (4), and let j be the partner of i under the cross transition. The variants compared here modify (6) as follows.

Multistep objectives. Both unroll the model for H steps and average over horizons,

$$\ell_H^{\mathrm{SL}}(\theta; i) \stackrel{\text{def}}{=} \frac{1}{2H}\sum_{h=1}^{H} \|e_i^{(h)}(\theta)\|_2^2, \tag{70}$$

$$\ell_{H,\gamma}^{\mathrm{TD}}(\theta, \bar{\theta}; i, j) \stackrel{\text{def}}{=} \frac{1}{2H}\sum_{h=1}^{H} \|e_i^{(h)}(\theta) - \gamma\ \mathrm{sg}\left[\bar{e}_j^{(h)}\right]\|_2^2, \tag{71}$$

so multistep TD applies the same residual correction at every horizon of the rollout, with $\bar{e}_j^{(h)}$ taken from the EMA target network. At $H = 1$ they reduce to the one-step supervised loss and to (6).

E.4 Residual diagnostic

We use the one-step-selected SL checkpoints and complete validation trajectories. Each residual is the predicted next-state field minus its observed target. For trajectory r, time t, and output channel c, let $\widehat{\mathrm{Var}}_s(y_{r,t,c})$ be the target field's spatial sample variance over its S grid points, with denominator $S-1$. We divide the residual field by the square root of this variance plus 10^{-7}, using the same denominator as VRMSE:

$$\widetilde{e}_{r,t,c}(s) = \frac{\widehat{y}_{r,t,c}(s) - y_{r,t,c}(s)}{\sqrt{\widehat{\mathrm{Var}}_s(y_{r,t,c}) + 10^{-7}}}.$$

This expresses each error relative to the spatial variation in its own target field. We subtract one global mean per channel and flatten the fields:

$$b_c = \mathrm{mean}_{r,t,s}\,\widetilde{e}_{r,t,c}(s), \qquad e_{r,t} = \mathrm{vec}_{c,s}(\widetilde{e}_{r,t,c}(s) - b_c) \in \mathbb{R}^q,$$

where q is the total number of channel–spatial coordinates. Within this subsection, $e_{r,t}$ denotes the metric-scaled diagnostic residual, not the residual in the training objective. The global mean uses all eligible validation trajectories, times, and spatial locations; the cross-trajectory mean at each time is retained so that the shared component remains measurable.

Throughout this diagnostic, $M = M_{\text{val}}$ is the number of validation trajectories. At each time t, let $E_t \in \mathbb{R}^{M\times q}$ contain $e_{r,t}^\top$ as row r, and let $\bar{e}_t = M^{-1}\sum_{r=1}^{M} e_{r,t}$ be the mean residual field across trajectories at time t. Each residual can therefore be written as

$$e_{r,t} = \bar{e}_t + (e_{r,t} - \bar{e}_t).$$

The first term is the shared component; the second is that trajectory's deviation from the average. Writing avg_t for the arithmetic mean over eligible times, the shared and mean difference-mode energies are

$$\widehat{\lambda}_{\text{sh}} = \frac{M}{q}\,\mathrm{avg}_t\,\|\bar{e}_t\|_2^2, \qquad \widehat{\lambda}_{\text{diff}} = \frac{\mathrm{avg}_t(\|E_t\|_F^2 - M\|\bar{e}_t\|_2^2)}{q(M-1)}.$$

They measure energy along $\mathbf{1}/\sqrt{M}$ and average energy per direction in its orthogonal complement. We report $R = \widehat{\lambda}_{\text{sh}}/\widehat{\lambda}_{\text{diff}}$. This ratio summarizes residual-energy structure; the difference component averages over $M-1$ directions rather than estimating their full covariance.

By (32), the cross-TD weighting assigns weight $1-\gamma$ to the shared component and $1+\gamma/(M-1)$ to the difference component. We use the discount selected by one-step validation for each domain, as listed in Table 12, and define

$$A_\gamma = \widehat{\lambda}_{\text{sh}}(1-\gamma), \qquad B_\gamma = \widehat{\lambda}_{\text{diff}}\left(1+\frac{\gamma}{M-1}\right), \qquad \chi_{2\text{mode}}(\gamma) = \frac{\max(A_\gamma, B_\gamma)}{\min(A_\gamma, B_\gamma)}. \tag{72}$$

The ratio equals one when the two weighted energies agree. Figure 8 compares $\gamma = 0$ with the validation-selected discount on the same fixed SL residuals; it does not compare residuals from separately trained SL and TD models. The diagnostic is not used to select the discount or any other hyperparameter; it is computed only after model selection from the fixed SL residuals.

We include complete, equal-length validation trajectory exports with at least two trajectories. Viscoelastic is excluded because its trajectories have unequal lengths, and Neutron Star Merger because it has one validation trajectory, leaving 16 domains.

E.5 Experiment-specific settings and search grids

This section describes hyperparameter search ranges. Appendix E.6 records their selected configurations.

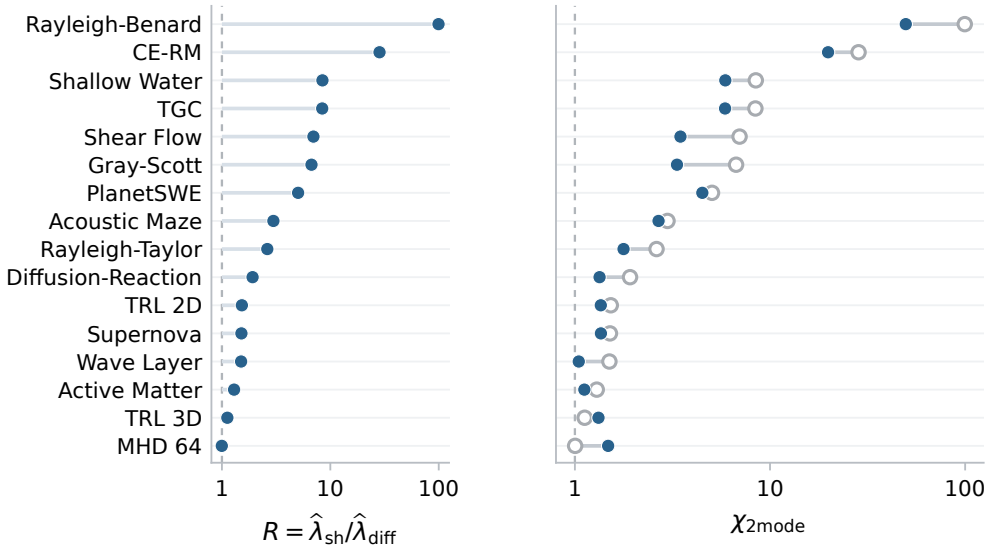
Focused ablations use Acoustic Maze, Gray–Scott, and Shear Flow, spanning wave propagation and scattering in heterogeneous media, nonlinear reaction–diffusion pattern formation, and incompressible shear-flow dynamics with advective transport, respectively.

Broad benchmark (Section 5.2). Flower-Micro runs last 50 epochs. Both TD and SL scan learning rates $\{10^{-3}, 3\times10^{-3}, 10^{-2}\}$; TD searches its method-specific $\gamma \in \{0.1, 0.3\}$ and $\beta \in \{0.01, 0.2, 0.9\}$. The discount also includes 0.5 on six domains: Acoustic Maze, Gray–Scott, Rayleigh–Bénard, Shear Flow, Wave Layer, and Neutron Star Merger.

Original scale (Section 5.3). Flower-Tiny is trained for 100 epochs on the full training splits of Acoustic Maze, Gray–Scott, and Shear Flow. Following Flower (Muser et al., 2026), we use learning rates of 10^{-3}, 5×10^{-4}, and 5×10^{-4}, respectively; we also evaluate four times each rate. TD searches method-specific $\gamma \in \{0.3, 0.5\}$ and $\beta \in \{0.01, 0.2, 0.9\}$. Training batch size is 408.

Data and model scale (Section 5.3). On Shear Flow, the grid crosses $\{5\%, 20\%\}$ training-data fractions with nominal $\{5\%, 20\%\}$ Flower-Tiny model sizes. These models use lifting dimensions 36 and 72, respectively, with four levels, 9 and 18 heads, and 9 and 18 groups. Each condition trains for 50 epochs. Both SL and TD search learning rates $\{10^{-4}, 5\times10^{-4}, 10^{-3}\}$. TD searches the method-specific $\gamma \in \{0.1, 0.3, 0.5\}$ and $\beta \in \{0.01, 0.2, 0.9\}$.

Transition study (Section 5.4). Cross, forward-adjacent, symmetric-adjacent, random-time, global-uniform, and self are each tested at $\gamma \in \{0.1, 0.3, 0.5\}$ for 50 epochs. Learning rate and EMA decay are fixed within each domain: $(10^{-2}, 0.9)$ for Acoustic Maze, $(3\times10^{-3}, 0.2)$ for Gray–Scott, and $(3\times10^{-3}, 0.01)$ for Shear Flow.



(a) Residual-mode energy ratio

(b) Mismatch under SL and TD weighting

Figure 8: **Cross-TD weighting of fixed SL residuals.** Results use fixed SL validation residuals from 16 eligible Flower-Micro domains, normalized and globally centered as specified in Appendix E.4. **(a)** Shared-component energy relative to average trajectory-difference energy, $R = \widehat{\lambda}_{\mathrm{sh}}/\widehat{\lambda}_{\mathrm{diff}}$. **(b)** Ratio of the larger to the smaller weighted component energy, $\chi_{2\mathrm{mode}}$, under SL weighting ($\gamma = 0$, open gray points) and cross-TD weighting (one-step validation-selected γ, blue points). Under the two-mode covariance surrogate of Appendix B.5, a smaller $\chi_{2\mathrm{mode}}$ also implies a smaller sharp worst-case local covariance factor $\kappa(\chi_{2\mathrm{mode}})$; cross-TD gives a smaller ratio in 14 of 16 domains. Both panels use logarithmic horizontal axes, the same domain order, and the validation trajectory count M_{val}. Dashed lines mark one. The residuals remain fixed; only the weighting changes.

Backbone transfer (Section 5.5). On the same three domains, scOT, FNO, and CNextU-Net each use a 50-epoch budget. Our CNextU-Net is Flower's implementation (Muser et al., 2026) of The Well's ConvNeXt-based U-Net (Ohana et al., 2024), which combines the U-Net architecture (Ronneberger et al., 2015) with ConvNeXt blocks (Liu et al., 2022). Each backbone retains its full-scale stage and block structure but is narrowed to a Micro configuration with approximately 0.5–1% of the full-scale parameter count: CNextU-Net 113K against 18.6M, FNO 85K against 18.9M, and scOT 177K against 17.7M, beside Flower-Micro's 176K against 17.3M. This study evaluates transfer across compact architectures, matching architecture and per-run epoch budget within each SL/TD pair; the Flower-Tiny study separately evaluates the original model and data scale. Both SL and TD search learning rates in $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$. TD searches method-specific $\beta \in \{0.01, 0.2, 0.9\}$. The TD discount is fixed across backbones within each domain: $\gamma = 0.3$ on Acoustic Maze, 0.1 on Gray–Scott, and 0.5 on Shear Flow.

Multistep training (Appendix D.5). Flower-Micro is trained on Acoustic Maze, Gray–Scott, and Shear Flow for 50 epochs with $H \in \{2, 4\}$. For these domains, respectively, SL learning rates are fixed at $(0.003, 0.001, 0.003)$, while TD uses learning rates $(0.01, 0.003, 0.003)$ and EMA decays $(0.01, 0.2, 0.01)$. TD searches $\gamma \in \{0.1, 0.3, 0.5\}$ at each training horizon. Both methods use gradient-norm clipping at 1.0 (Pascanu et al., 2013). These are targeted objective comparisons, not equal-total-compute searches.

E.6 Hyperparameters of reported models

The tables below give the hyperparameters of the models used for the reported results. Learning rates are in units of 10^{-3}; EMA decay β is defined in (68).

Table 12: Flower-Micro one-step settings for Tables 5 and 6. Learning-rate entries are in units of 10^{-3}.

Domain	SL lr	TD lr	TD γ	TD β
Acoustic Maze	10	10	0.1	0.01
Active Matter	10	10	0.3	0.9
Gray–Scott	1	3	0.5	0.2
MHD 64	10	10	0.3	0.2
PlanetSWE	10	10	0.1	0.9
Neutron Star Merger	10	10	0.5	0.2
Rayleigh–Bénard	1	3	0.5	0.2
Rayleigh–Taylor	3	10	0.3	0.01
Shear Flow	3	3	0.5	0.01
Supernova	3	3	0.1	0.2
TGC	3	10	0.3	0.9
TRL 2D	10	10	0.1	0.01
TRL 3D	3	3	0.3	0.01
Viscoelastic	3	3	0.3	0.9
CE-RM	10	10	0.3	0.2
Wave Layer	10	10	0.3	0.01
Diffusion–Reaction	10	10	0.3	0.01
Shallow Water	10	10	0.3	0.2

Table 13: Flower-Micro rollout settings for Tables 5 and 7. Learning-rate entries are in units of 10^{-3}. – denotes an unavailable interval. Intervals are truncated to the available horizon as in the result tables.

	1:20				21:60			
Domain	SL lr	TD lr	γ	β	SL lr	TD lr	γ	β
Acoustic Maze	3	3	0.1	0.2	3	3	0.1	0.2
Active Matter	10	10	0.3	0.2	10	10	0.1	0.01
Gray–Scott	1	3	0.5	0.2	1	1	0.5	0.2
MHD 64	10	10	0.1	0.9	10	10	0.1	0.9
PlanetSWE	10	10	0.3	0.9	10	10	0.3	0.9
Neutron Star Merger	10	10	0.3	0.01	1	1	0.3	0.2
Rayleigh–Bénard	1	3	0.5	0.2	1	3	0.5	0.2
Rayleigh–Taylor	3	10	0.3	0.9	10	10	0.3	0.9
Shear Flow	3	1	0.5	0.01	3	3	0.3	0.9
Supernova	1	3	0.1	0.9	1	3	0.1	0.9
TGC	10	10	0.3	0.9	10	10	0.3	0.9
TRL 2D	3	1	0.3	0.9	10	10	0.1	0.9
TRL 3D	1	3	0.1	0.9	1	10	0.3	0.2
Viscoelastic	3	10	0.1	0.2	–	–	–	–
CE-RM	10	10	0.1	0.9	–	–	–	–
Wave Layer	10	10	0.3	0.9	–	–	–	–
Diffusion–Reaction	3	10	0.3	0.01	10	10	0.3	0.01
Shallow Water	10	3	0.1	0.9	10	3	0.1	0.9

Table 14: Flower-Tiny validation-selected settings for Tables 1 and 8. SL and TD share the listed learning rate (in units of 10^{-3}).

Domain	Result	SL/TD lr	TD γ	TD β
Acoustic Maze	One-step	1	0.3	0.2
	Rollout $1:20$	1	0.5	0.01
	Rollout $21:60$	1	0.5	0.9
	Rollout all	1	0.5	0.2
Gray–Scott	One-step	0.5	0.5	0.2
	Rollout $1:20$	0.5	0.5	0.01
	Rollout $21:60$	0.5	0.3	0.01
	Rollout all	0.5	0.3	0.01
Shear Flow	One-step	0.5	0.5	0.01
	Rollout $1:20$	0.5	0.3	0.2
	Rollout $21:60$	0.5	0.5	0.01
	Rollout all	0.5	0.5	0.2

Table 15: Settings for the backbone comparisons in Table 10, including the Flower references. Learning-rate entries are in units of 10^{-3}.

Domain	Backbone	SL lr	TD lr	TD γ	TD β
Acoustic Maze	FNO	10	10	0.3	0.9
	scOT	10	10	0.3	0.01
	CNextU-Net	10	10	0.3	0.01
	Flower	10	10	0.3	0.9
Gray–Scott	FNO	3	3	0.1	0.9
	scOT	1	1	0.1	0.01
	CNextU-Net	10	3	0.1	0.9
	Flower	1	3	0.1	0.9
Shear Flow	FNO	10	10	0.5	0.2
	scOT	10	3	0.5	0.01
	CNextU-Net	3	3	0.5	0.01
	Flower	3	3	0.5	0.01

Table 16: Shear Flow scaling settings for Table 9. Fractions refer to model size and training data. Learning-rate entries are in units of 10^{-3}.

Model	Data	SL lr	TD lr	TD γ	TD β
5%	5%	1	1	0.1	0.2
5%	20%	1	1	0.3	0.01
20%	5%	1	1	0.5	0.9
20%	20%	1	1	0.3	0.01

Table 17: Transition-ablation settings for Table 2. The learning rate (in units of 10^{-3}) and EMA decay are fixed within each domain; the remaining columns give the validation-selected γ for each transition.

			Selected γ					
Domain	TD lr	β	Cross	Fwd.	Sym.	Self	Global	Rand.
Acoustic Maze	10	0.9	0.3	0.1	0.3	0.1	0.5	0.1
Gray–Scott	3	0.2	0.5	0.5	0.1	0.5	0.1	0.1
Shear Flow	3	0.01	0.5	0.1	0.3	0.5	0.1	0.1

Table 18: Selected settings for Table 11, by training unroll length H and evaluation metric. Learning rates are in units of 10^{-3}; β is the EMA decay and γ is the TD discount.

Domain	H	Evaluation	SL lr	TD lr	TD β	TD γ
Acoustic Maze	1	One-step	10	10	0.01	0.1
	1	1:20	3	3	0.2	0.1
	1	21:60	3	3	0.2	0.1
	2	One-step	3	10	0.01	0.1
	2	1:20	3	10	0.01	0.1
	2	21:60	3	10	0.01	0.1
	4	One-step	3	10	0.01	0.1
	4	1:20	3	10	0.01	0.1
	4	21:60	3	10	0.01	0.5
Gray–Scott	1	One-step	1	3	0.2	0.5
	1	1:20	1	3	0.2	0.5
	1	21:60	1	1	0.2	0.5
	2	One-step	1	3	0.2	0.5
	2	1:20	1	3	0.2	0.1
	2	21:60	1	3	0.2	0.5
	4	One-step	1	3	0.2	0.5
	4	1:20	1	3	0.2	0.1
	4	21:60	1	3	0.2	0.1
Shear Flow	1	One-step	3	3	0.01	0.5
	1	1:20	3	1	0.01	0.5
	1	21:60	3	3	0.9	0.3
	2	One-step	3	3	0.01	0.5
	2	1:20	3	3	0.01	0.5
	2	21:60	3	3	0.01	0.5
	4	One-step	3	3	0.01	0.5
	4	1:20	3	3	0.01	0.5
	4	21:60	3	3	0.01	0.3