BEYOND TRAJECTORIES AND JACOBIANS: DISTILLING FINITE-TIME TANGENT COCYCLES FOR CHAOTIC FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

State-space losses and local Jacobian objectives do not directly constrain how infinitesimal errors are transported over multiple steps in a chaotic system. We introduce BUNDLEDISTILL, a training-only objective that distills low-rank finite-time tangent-cocycle geometry from controlled antithetic trajectory rollouts, without requiring governing equations or derivative labels. The student matches cocycle action, probed singular growth, and transported-subspace orientation using Jacobian–vector products and small endpoint factorizations. Under matched optimizer-update budgets across five locked seeds on Lorenz–63, Lorenz–96, and Kuramoto–Sivashinsky, BUNDLEDISTILL has the smallest observed transported-subspace angle on all three systems. Against the strongest scalar-growth control, restricted-mean valid prediction time increases on Lorenz–96 by 0.018 Lyapunov times (95% percentile interval $[0.016, 0.021]$) and by 0.180 on Kuramoto–Sivashinsky ($[0.149, 0.215]$), while the Lorenz–63 interval spans zero. Five-seed ablations on Lorenz–96 and Kuramoto–Sivashinsky show that component effects depend on the system and endpoint. The auxiliary objective leaves model size unchanged and adds no inference operation, but it increases training cost; under a calibrated optimizer-update-time control, the faster Rand-Jac2 baseline is stronger on Lorenz–96 VPT and geometric endpoints. Formal out-of-regime tests yield no Holm-corrected rejection, and held-out long-run statistics vary by system. The evidence supports finite-time tangent transport as a useful, compute-dependent training target, with its clearest gains on the two higher-dimensional systems studied.

1 INTRODUCTION

Forecasting a chaotic dynamical system is not ordinary sequence extrapolation. Arbitrarily small state errors can be amplified exponentially, so point forecasts eventually decorrelate even when a learned one-step map is accurate (Lorenz, 1963; 1996). A useful learned simulator should therefore delay trajectory divergence and, separately, preserve meaningful geometric or statistical properties after divergence (Pathak et al., 2018; Vlachas et al., 2020; Gilpin, 2024). State-space losses alone do not determine local instability, which governs finite-time error growth and can influence long-run statistics without fully determining them (Mikhaeil et al., 2022; Gilpin, 2023).

Derivative-aware objectives address part of this gap by matching a Jacobian or regulating its spectrum. Finite-horizon infinitesimal perturbations, however, are transported by an ordered product. For $x_{t+1} = f(x_t)$, an infinitesimal perturbation after k steps is transported by

$$\mathbf{\Phi}^{\star}_{t,k} = Df(x_{t+k-1}) \cdots Df(x_{t+1}) Df(x_t). \tag{1}$$

The factors generally do not commute, and their singular directions rotate along the trajectory. Consequently, two models with similar one-step error, Jacobian norms, or leading local expansion can have different transient amplification and transported tangent subspaces. We call this discrepancy *cocycle misalignment*.

We introduce BUNDLEDISTILL, a training objective that supervises finite-time perturbation transport without equations or derivative labels. The access model is a resettable simulator: from a training state, it can roll out controlled nearby initial conditions. Symmetric antithetic differences estimate the

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.

actions $Df^k(x_t)v_j$ along a small orthonormal probe set. The student propagates the same probes through its observable rollout using Jacobian–vector products (JVPs). Thin QR factorizations and small SVDs yield complementary losses for tangent action, probed finite-time singular growth, and the orientation of the transported subspace. These quantities are used only during training.

Our contributions are threefold.

- We formulate cocycle misalignment as a finite-time geometric failure mode and introduce an equation-free, low-rank objective that matches perturbation action, singular growth, and transported-subspace orientation from controlled trajectory interventions.
- We establish second-order consistency of the antithetic teacher, expectation-level random-probe coverage, and finite-horizon perturbation bounds for cocycle actions and transported subspaces.
- In a locked five-seed, equal-update evaluation on three canonical chaotic systems, we find lower observed transported-subspace angles than all evaluated baselines. Prediction-horizon effect intervals against Growth-Match exclude zero on Lorenz–96 and Kuramoto–Sivashinsky. Cross-system ablations show endpoint-specific component effects; regime-shift tests yield no Holm-corrected rejection, long-run results vary by system, and the calibrated L96 timing control favors Rand-Jac2 on VPT and geometric endpoints.

The empirical gains depend on the system and endpoint. Under matched optimizer-update budgets, the largest VPT gain over Growth-Match occurs on Kuramoto–Sivashinsky; the Lorenz–63 effect interval spans zero. The objective requires more training compute than simple rollout or randomized-Jacobian baselines. Its finite-time gains coexist with mixed long-run results. The objective directly supervises tangent transport rather than invariant-measure fidelity.

2 Related work

Learning chaotic dynamics. Neural forecasters for chaotic systems include recurrent networks, neural differential equations, reservoir computers, and operator models (Funahashi & Nakamura, 1993; Chen et al., 2018; Pathak et al., 2018; Gauthier et al., 2021; Li et al., 2022). Their evaluation increasingly includes invariant measures, spectra, and attractor geometry in addition to rollout error (Gilpin, 2021; 2023; 2024). This is important because expressive unstable dynamics and stable gradient propagation can conflict in recurrent models (Pascanu et al., 2013; Mikhaeil et al., 2022). BundleDistill changes the learning signal rather than the forecaster class.

Derivative-aware learning. Trajectory loss constrains function values but leaves derivatives underdetermined. Jacobian matching, spectral regularization, derivative-informed operators, and propagated ellipsoids therefore supervise local tangent behavior (Wang, 2026; Kang et al., 2026). Our distinction is temporal rather than derivative order: we supervise ordered products of first-order tangent maps over finite horizons. Such products capture non-commutativity, non-normal transient growth, and rotating expanding directions that are not determined by an isolated Jacobian norm or scalar exponent.

Lyapunov geometry. The multiplicative ergodic theorem characterizes asymptotic growth and Oseledets subspaces (Oseledets, 1968); repeated orthogonalization estimates Lyapunov spectra (Benettin et al., 1980), and covariant Lyapunov vectors identify intrinsic tangent directions (Ginelli et al., 2007; Kuptsov & Parlitz, 2012). BundleDistill uses related finite-time geometry as supervision but does not require covariant-vector labels. Compared with the one-step JacobianMatch baseline, its teacher obtains finite-time actions from symmetric neighboring trajectories; compared with Growth-Match, its projector loss additionally supervises transported orientation.

3 Finite-time tangent geometry

Let $x_{t+1} = f(x_t)$ denote a sampled deterministic flow and let g_θ be an autoregressive forecaster. Their k-step tangent cocycles are

$$\mathbf{\Phi}^{\star}_{t,k} = Df^k(x_t), \qquad \mathbf{\Phi}^{\theta}_{t,k} = Dg^k_\theta(x_t), \tag{2}$$

where the model derivative follows the model rollout. For an orthonormal probe $V_t \in \mathbb{R}^{d\times r}$ with $r \ll d$, define the low-rank action $\Delta^\star_{t,k} = \mathbf{\Phi}^\star_{t,k} V_t$ and its thin QR factorization $\Delta^\star_{t,k} = Q^\star_{t,k} R^\star_{t,k}$. The action records finite-horizon scale, shear, and orientation without materializing a full $d \times d$ Jacobian.

Since $Q^\star_{t,k}$ has orthonormal columns, $R^\star_{t,k}$ and $\Delta^\star_{t,k}$ share singular values. We call

$$\ell^\star_{t,k} = \frac{1}{k\Delta t} \log\big(\boldsymbol{\sigma}(R^\star_{t,k}) + \eta\big) \tag{3}$$

the *probed finite-time singular-growth vector*. When $r < d$, it depends on the probe subspace and is not a standard full-tangent FTLE. Physical Lyapunov normalization in our evaluation instead uses rank-one repeated QR of the true-system tangent action to estimate the largest exponent.

The span of $Q^\star_{t,k}$ is the finite-time image of the probed subspace. We compare it with a model image $Q^\theta_{t,k}$ using the Grassmannian projector distance

$$d^2_{\mathrm{Gr}}(Q_1, Q_2) = \frac{1}{2} \left\| Q_1 Q_1^\top - Q_2 Q_2^\top \right\|_F^2 = \sum_{j=1}^{r} \sin^2 \vartheta_j, \tag{4}$$

which is invariant to sign flips and basis rotations (Edelman et al., 1998). Cocycle misalignment refers to a model having small state error but appreciable action, probed-growth, or transported-subspace error at a finite horizon. It distinguishes early geometric disagreement from eventual trajectory decorrelation.

4 Equation-free cocycle distillation

Access model and probes. Figure 1 summarizes the training graph. The core setting assumes a deterministic transition source that can be reset at a training state and rolled out from controlled nearby initial conditions. It need not expose equations or derivatives. Let $z = S^{-1}(x - \mu)$ be standardized coordinates, with S and μ fit on the training split. At every anchor z_t, sample

$$V_t = [v_1, \ldots, v_r] \sim \mathrm{Haar}(\mathrm{St}(d, r)), \qquad 1 \le r < d, \tag{5}$$

implemented by the thin Q factor of a Gaussian matrix. Independent probes rotate across anchors; Proposition 2 gives their expectation-level coverage.

Antithetic trajectory teacher. For radius δ and horizon k, the symmetric finite-difference action is

$$\Delta^\star_{t,k}(\delta) = \left[\frac{\widetilde{f}^k(z_t + \delta v_j) - \widetilde{f}^k(z_t - \delta v_j)}{2\delta} \right]_{j=1}^{r}, \tag{6}$$

where $\widetilde{f}$ is the standardized transition. We use the stopped-gradient fine-radius estimate $\bar{\Delta}^\star{}_{t,k} = \mathrm{sg}[\Delta^\star_{t,k}(\varepsilon/2)]$ and downweight cross-radius disagreement:

$$e_{t,k} = \frac{\left\| \Delta^\star_{t,k}(\varepsilon) - \Delta^\star_{t,k}(\varepsilon/2) \right\|_F}{\left\| \Delta^\star_{t,k}(\varepsilon/2) \right\|_F + \eta}, \qquad w_{t,k} = \mathrm{sg}[\exp(-e_{t,k}/\tau_c)]. \tag{7}$$

The score measures numerical consistency across radii; it is not a certificate that nonlinear separation is absent.

Student action. Let $\widehat{F}_{\theta,k}(z_t; c_t^-)$ be the model's k-step observable prediction with preceding context c_t^- fixed. The student action

$$U^\theta_{t,k} = D_{z_t} \widehat{F}_{\theta,k}(z_t; c_t^-) V_t \tag{8}$$

is evaluated with r batched JVPs. For a Markov model, $U^\theta_{t,k+1} = Dg_\theta(\hat{z}_{t,k}) U^\theta_{t,k}$, so no full Jacobian is formed. Appendix B gives the corresponding lifted map for recurrent forecasters.

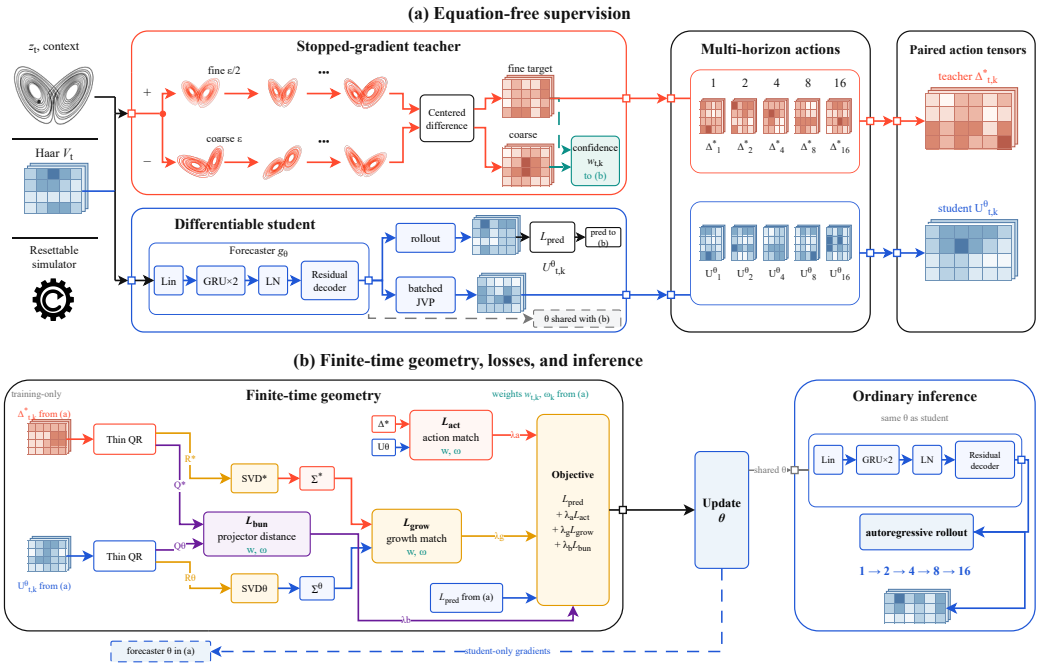

Figure 1: Schematic training graph of BUNDLEDISTILL. The same low-rank Haar probe feeds a stopped-gradient controlled-trajectory teacher and a differentiable student JVP branch. Fine-radius antithetic rollouts form the target, while fine/coarse disagreement supplies the consistency weight. Direct action matching and endpoint thin-QR/small-SVD summaries yield the three geometry losses, which join rollout prediction. The blue dashed return path marks gradients; the curriculum schedule is omitted for clarity. All BUNDLEDISTILL auxiliary operations are discarded at inference.

Finite-horizon losses. For each $k \in \mathcal{H}$, factor $\bar{\Delta}^\star{}_{t,k} = Q^\star_{t,k} R^\star_{t,k}$ and $U^\theta_{t,k} = Q^\theta_{t,k} R^\theta_{t,k}$. Let $\ell^\nu_{t,k} = \log(\boldsymbol{\sigma}(R^\nu_{t,k}) + \eta)/(k\Delta t)$ and let nonnegative ω_k sum to one. We optimize

$$\mathcal{L}_{\text{act}} = \sum_{k\in\mathcal{H}} \omega_k w_{t,k} \frac{\left\| U^\theta_{t,k} - \bar{\Delta}^\star{}_{t,k} \right\|_F^2}{\left\| \bar{\Delta}^\star{}_{t,k} \right\|_F^2 + \eta}, \tag{9}$$

$$\mathcal{L}_{\text{grow}} = \sum_{k\in\mathcal{H}} \omega_k w_{t,k} \left\| \ell^\theta_{t,k} - \ell^\star_{t,k} \right\|_2^2, \tag{10}$$

$$\mathcal{L}_{\text{bun}} = \sum_{k\in\mathcal{H}} \frac{\omega_k w_{t,k}}{2r} \left\| Q^\theta_{t,k} Q^{\theta\top}_{t,k} - Q^\star_{t,k} Q^{\star\top}_{t,k} \right\|_F^2. \tag{11}$$

Action constrains scale, shear, and orientation on the sampled subspace; growth isolates basis-invariant amplification; and the projector term directly aligns transported orientation.

The prediction loss is a weighted rollout error,

$$\mathcal{L}_{\text{pred}}(t) = \sum_{k\in\mathcal{H}_p} \alpha_k d^{-1} \left\| \widehat{F}_{\theta,k}(z_t; c_t^-) - z_{t+k} \right\|_2^2, \tag{12}$$

and the complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_a \mathcal{L}_{\text{act}} + \lambda_g \mathcal{L}_{\text{grow}} + \lambda_b \mathcal{L}_{\text{bun}}. \tag{13}$$

We introduce horizons progressively and ramp the geometric weights after the forecast-only warm-up. Teacher quantities are detached, while gradients pass through student JVPs and their small QR/SVD factors. Rank r yields $\mathcal{O}(Kr)$ directional propagations up to maximum horizon K; the two-radius

teacher uses $4r$ simulator rollouts per anchor. Inference retains the original forecaster and its ordinary recurrent state, with no BUNDLEDISTILL-specific probes, teacher queries, decompositions, auxiliary parameters, or losses.

5 ANALYSIS

We summarize the properties that justify the trajectory teacher, random low-rank probes, and finite-time objective. Proofs are in Appendix A.

Proposition 1 (Antithetic teacher consistency). *Let $F_k = \widetilde{f}^k$ be C^3 on the closed ε-neighborhood of z, let $V \in \mathbb{R}^{d\times r}$ have orthonormal columns, and suppose $\sup_z \left\|D^3 F_k(z)\right\|_{\mathrm{op}} \le M_{3,k}$. Then*

$$\left\|\Delta^\star_{t,k}(\varepsilon) - DF_k(z)V\right\|_F \le \frac{\sqrt{r}}{6} M_{3,k}\varepsilon^2. \tag{14}$$

A one-sided difference has an $\mathcal{O}(\varepsilon)$ remainder. The bound controls finite-radius truncation error; it does not account for round-off.

Proposition 2 (Random-probe coverage). *For $V \sim \mathrm{Haar}(\mathrm{St}(d,r))$ and fixed $A \in \mathbb{R}^{m\times d}$,*

$$\mathbb{E}_V \left\|AV\right\|_F^2 = \frac{r}{d}\left\|A\right\|_F^2. \tag{15}$$

Thus averaging the unnormalized squared exact-action error over probes gives a scaled squared Frobenius norm of the full-cocycle error. This is expectation-level coverage, not a guarantee that one low-rank probe captures every unstable direction.

Theorem 1 (Finite-horizon perturbation). *Let A_i and B_i be true and model tangent factors along aligned length-k trajectories, with $\|A_i\|_2 \le L$, $\|B_i\|_2 \le M$, and $\|B_i - A_i\|_2 \le \delta_i$. Then*

$$\left\|B_{k-1}\cdots B_0 - A_{k-1}\cdots A_0\right\|_2 \le \sum_{i=0}^{k-1} M^{k-1-i} L^i \delta_i. \tag{16}$$

The telescoping bound shows why small local derivative error can accumulate coherently in an expanding product.

Finally, let $A = \mathbf{\Phi}^\star_{t,k} V$, $\widehat{A} = \mathbf{\Phi}^\theta_{t,k} V$, and $E = \widehat{A} - A$. If $\|E\|_2 < \sigma_r(A)$, standard singular-subspace perturbation gives

$$\left\|\widehat{Q}\widehat{Q}^\top - QQ^\top\right\|_F \le \frac{\sqrt{2r}\left\|E\right\|_2}{\sigma_r(A) - \left\|E\right\|_2}, \tag{17}$$

and $|\sigma_j(\widehat{A}) - \sigma_j(A)| \le \|E\|_2$ (Wedin, 1972). This connects action alignment to both finite-time growth and transported orientation while showing how the bound deteriorates near rank collapse. Appendix A also gives the limiting smooth-conjugacy identity

$$Dg^k(h(x))Dh(x) = Dh(f^k(x))Df^k(x), \tag{18}$$

which is an interpretation rather than an assumption of the finite-sample method.

6 EXPERIMENTS

We ask whether finite-time geometric distillation improves prediction time and transported tangent geometry (Q1), whether the gains persist under fixed regime shifts (Q2), how its three losses contribute across L96 and KS (Q3), and whether the gains can be separated from training cost (Q4).

6.1 PROTOCOL

Systems and models. We evaluate Lorenz–63 (L63), 40-dimensional Lorenz–96 (L96), and the $N = 64$, $L = 22$ Kuramoto–Sivashinsky equation (KS) (Lorenz, 1963; 1996; Kassam & Trefethen, 2005). Each method uses the same two-layer width-128 GRU, residual decoder, and 10,000 optimizer updates. BUNDLEDISTILL uses ranks $(1, 4, 4)$, radius 10^{-3}, and horizons $\{1, 2, 4, 8, 16\}$ on L63/L96/KS. Baselines are one-step teacher forcing, matched multi-step rollout, JacobianMatch, the scalar finite-time Growth-Match control, and randomized derivative supervision (Rand-Jac2) (Kang et al., 2026). Appendix C gives data splits, solvers, hyperparameters, and baseline definitions.

Table 1: In-distribution finite-time results over five locked seeds. VPT is the restricted mean in independently estimated Lyapunov times; lower is better for growth MAE and angle. Bold and underline denote the best and second-best observed means within each system. Paired effect intervals are reported in the text and Appendix D.1.

System	Method	VPT↑	Growth MAE↓	Angle (°)↓
L63	One-step	0.129	2.081	27.335
	Multi-step	2.797	0.980	5.463
	JacobianMatch	2.667	0.243	1.831
	Growth-Match	<u>2.987</u>	<u>0.018</u>	<u>0.213</u>
	Rand-Jac2	2.832	0.300	2.179
	BUNDLEDISTILL	**2.999**	**0.017**	**0.184**
L96	One-step	0.235	0.471	42.369
	Multi-step	1.165	0.361	21.059
	JacobianMatch	1.168	0.358	20.950
	Growth-Match	<u>1.169</u>	<u>0.317</u>	<u>20.650</u>
	Rand-Jac2	1.049	0.482	24.326
	BUNDLEDISTILL	**1.187**	**0.311**	**20.225**
KS	One-step	0.148	1.524	68.962
	Multi-step	0.715	1.527	66.124
	JacobianMatch	0.759	0.143	28.961
	Growth-Match	<u>0.777</u>	<u>0.076</u>	<u>24.562</u>
	Rand-Jac2	0.603	0.421	43.179
	BUNDLEDISTILL	**0.957**	**0.035**	**8.950**

Evaluation and inference. All results use locked seeds 2024–2028 and held-out trajectories. Valid prediction time (VPT) is the first 0.4-nRMSE crossing within a 512-step horizon; we report the horizon-capped restricted mean. The artifacts record right-censor fractions, whose maximum is 0.12% in distribution and zero over the formal OOD shifts. Lyapunov-time conversion uses separately cached rank-one repeated-QR estimates of the true system's largest exponent for the in-distribution and OOD regimes. Probed-growth MAE and the RMS principal angle compare matched finite-time teacher and model actions. For in-distribution VPT and angle on each system, 10,000 hierarchical paired-bootstrap replicates resample seeds, trajectories, and anchors with common indices across methods. We use these resamples for pointwise percentile effect intervals, not null-hypothesis testing; they are not simultaneous or multiplicity-adjusted. Long-run evaluations use four learned and four disjoint reference ensembles per seed, discard five Lyapunov times, and retain 100 Lyapunov times for statistics. The OOD analysis separately uses pointwise paired seed–trajectory bootstrap intervals and exact sign-flip tests over five seed blocks, assuming independent seed blocks and sign symmetry under the null, with a single Holm family over all 30 comparisons. Component intervals are exploratory paired-t intervals over five seeds, without confirmatory p-values.

6.2 Q1: PREDICTION AND TRANSPORTED GEOMETRY

Table 1 summarizes the in-distribution endpoints. BUNDLEDISTILL provides its clearest prediction gains on the higher-dimensional systems. Against Growth-Match, restricted-mean VPT increases by 0.018 Lyapunov times on L96 (95% percentile interval $[0.016, 0.021]$) and by 0.180 on KS ($[0.149, 0.215]$), corresponding to observed relative gains of 1.5% and 23.2%. On L63, the difference is 0.012 and the interval crosses zero ($[-0.048, 0.069]$).

Transported-subspace alignment is the most consistent result. BUNDLEDISTILL has the smallest mean angle on all three systems, and every paired percentile interval against a baseline excludes zero in the favorable direction. Relative to Growth-Match, the angle decreases by $0.029°$ on L63 (interval $[-0.034, -0.025]$), $0.425°$ on L96 ($[-0.438, -0.413]$), and $15.612°$ on KS ($[-16.167, -15.075]$). The magnitude varies substantially; KS shows both the strongest geometric separation and the largest VPT gain. Figure 2 visualizes the Growth-Match contrasts; full paired effects against all baselines appear in Appendix D.1.

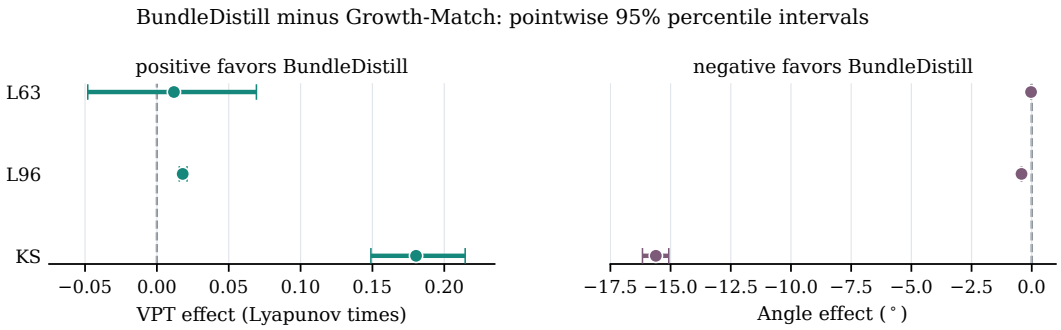

Figure 2: Audited BUNDLEDISTILL-minus-Growth-Match effects for restricted-mean VPT and transported-subspace angle. Points are observed hierarchical differences; bars are pointwise 95% percentile intervals from 10,000 paired seed–trajectory–anchor resamples. The intervals are uncertainty summaries, not simultaneous or null-calibrated tests.

Table 2: Out-of-regime restricted-mean VPT over five seeds. Models are trained only at $(\rho, F, L) = (28, 8, 22)$ and evaluated without retraining. Bold denotes the largest observed mean. Paired hierarchical intervals and exact seed-block sign-flip tests are summarized in the text; no comparison rejects after global Holm correction.

	L63		L96		KS	
Method	ρ=32	ρ=36	F=10	F=12	L=23	L=24
One-step	0.129	0.121	0.268	0.250	0.025	0.074
Multi-step	0.255	0.126	0.589	0.324	0.100	0.228
JacobianMatch	0.271	0.135	0.588	0.324	0.103	0.227
Growth-Match	0.271	0.135	0.602	0.333	0.101	0.230
Rand-Jac2	**0.275**	**0.137**	0.470	0.250	0.093	0.210
BUNDLEDISTILL	0.271	0.135	**0.611**	**0.338**	**0.120**	**0.239**

Finite-time versus long-run behavior. On L96, BUNDLEDISTILL has the smallest observed long-run sliced Wasserstein distance (SWD) and power-spectrum error, but differences from Growth-Match are inconclusive. It is not best on L63, and on KS it is substantially worse than Multi-step for both SWD (difference $+168.843$, percentile interval $[121.859, 216.617]$) and power-spectrum error ($+4.214$, $[3.749, 4.867]$). Appendix D.3 reports all values. The finite-time improvements coexist with mixed long-run results; the objective does not directly supervise the invariant measure.

6.3 Q2: FIXED REGIME SHIFTS

Without retraining, BUNDLEDISTILL attains the largest observed mean VPT in four of six prespecified shifts: both L96 forcing shifts and both KS domain-length shifts. It does not lead either L63 parameter shift, where Rand-Jac2 is slightly higher. The paired hierarchical intervals for BUNDLEDISTILL relative to the strongest observed control are $[0.0075, 0.0095]$ at $F = 10$, $[0.0047, 0.0056]$ at $F = 12$, $[0.0127, 0.0224]$ at $L = 23$, and $[0.0008, 0.0182]$ at $L = 24$; the L63 intervals instead favor Rand-Jac2. With five seed blocks, the smallest attainable two-sided exact sign-flip p-value is $2/32 = .0625$. Consequently, none of the 30 OOD comparisons rejects after global Holm correction (all adjusted $p = 1$). Table 2 reports these observed, shift-dependent VPT differences; Appendix D.2 reports the paired contrasts.

6.4 Q3: COMPONENT-ALIGNED ABLATIONS

Table 3 shows system- and endpoint-dependent effects. On L96, removing the action or bundle term reduces VPT by 0.0089 (exploratory 95% CI $[0.0079, 0.0100]$) and 0.0085 ($[0.0069, 0.0101]$), respectively. Removing growth has a near-zero VPT effect but raises growth MAE by 0.0409 ($[0.0380, 0.0439]$). On KS, removing action or bundle lowers VPT by 0.0520 and 0.0740 and

Table 3: Five-seed component ablations on L96 and KS (mean±sample SD). Every row uses the same 10,000-update budget, data, and seed-matched evaluation anchors. Intervals in the text are exploratory paired-t intervals; no confirmatory component-level p-value is reported.

System	Variant	VPT↑	Growth MAE↓	Angle (°)↓
L96	Full BUNDLEDISTILL	1.187±0.032	0.311±0.005	20.225±0.359
	$-\mathcal{L}_{\text{act}}$	1.178±0.032	0.314±0.006	20.438±0.359
	$-\mathcal{L}_{\text{grow}}$	1.185±0.033	0.352±0.004	20.578±0.351
	$-\mathcal{L}_{\text{bun}}$	1.178±0.032	0.315±0.005	20.424±0.359
KS	Full BUNDLEDISTILL	0.957±0.044	0.035±0.001	8.950±0.179
	$-\mathcal{L}_{\text{act}}$	0.905±0.036	0.045±0.002	11.006±0.218
	$-\mathcal{L}_{\text{grow}}$	1.049±0.065	0.041±0.002	9.270±0.222
	$-\mathcal{L}_{\text{bun}}$	0.883±0.041	0.044±0.002	12.096±0.201

Table 4: Frozen L96 calibrated optimizer-update-time comparison over five paired seeds (mean±sample SD). Actual synchronized update and end-to-end process times are reported; calibration targeted, but did not attain, exact realized equality.

Method	Updates	Update time (s)	Process time (s)
BUNDLEDISTILL	10,000	1,976±73	2,090±75
Rand-Jac2	37,915	2,163±71	2,437±75

Method	VPT↑	Mean nRMSE↓	Growth MAE↓	Angle (°)↓
BUNDLEDISTILL	1.187±0.032	1.269±0.004	0.311±0.005	20.225±0.359
Rand-Jac2	1.798±0.022	1.292±0.003	0.289±0.003	13.988±0.300

worsens angle by $2.057°$ and $3.146°$. In contrast, removing growth *increases* VPT by 0.0919 ($[0.0482, 0.1355]$) while worsening growth MAE by 0.0060 and angle by $0.320°$. These ablations show endpoint-dependent contributions, including a prediction–geometry trade-off on KS. Appendix D.8 also reports an implementation sanity check against exact JVPs at two L63 anchors, with rank one and horizons one and two.

6.5 Q4: COMPUTATIONAL COST AND A CALIBRATED CONTROL

Within each system, every method has the same architecture and parameter count, and measured batch-one inference latency is similar; the geometric objective is training-only. BUNDLEDISTILL has 299,523/323,240/338,624 parameters and mean single-step latency of 365/361/369 μs on L63/L96/KS. Its peak allocated training memory is 92.5–162.2 MiB, and its end-to-end training time is 3.9–9.3 times that of Multi-step. Appendix Table 9 reports all 90-checkpoint measurements. These trainer times can reflect hardware contention, and VPT per GPU-hour is descriptive rather than a controlled causal comparison.

Table 4 reports a calibrated optimizer-update-time comparison on L96. A frozen timing calibration allocates Rand-Jac2 37,915 updates versus 10,000 for BUNDLEDISTILL, although realized optimizer-update time is still 9.4% longer for Rand-Jac2 and process time is not matched. With 3.79 times as many updates, Rand-Jac2 attains VPT 1.798 versus 1.187, lower growth MAE, and a smaller angle, whereas BUNDLEDISTILL retains lower mean nRMSE (1.269 versus 1.292). The paired BUNDLEDISTILL-minus-Rand-Jac2 differences are -0.611 for VPT (descriptive 95% CI $[-0.623, -0.598]$), -0.0230 for nRMSE ($[-0.0271, -0.0200]$), $+0.0216$ for growth MAE ($[0.0184, 0.0246]$), and $+6.237°$ for angle ($[6.152, 6.321]$). These descriptive intervals are not confirmatory tests. Under this calibration, the L96 VPT ranking reverses relative to the equal-update comparison.

7 CONCLUSION AND LIMITATIONS

We introduced BUNDLEDISTILL, an equation-free training objective for matching low-rank finite-time perturbation transport in chaotic forecasters. Under a common 10,000-update protocol, it has the smallest observed transported-subspace angle on all three canonical systems; hierarchical percentile intervals support prediction-horizon gains on L96 and KS, with the largest effect on KS, while the L63 interval against Growth-Match spans zero. Five-seed L96 and KS ablations show endpoint-dependent contributions and a trade-off: removing growth matching on KS improves VPT while worsening growth MAE and transported-subspace angle. The method leaves model size unchanged and adds no inference operation, yet costs more to train than simple rollout baselines; after the faster Rand-Jac2 objective receives more updates in a calibrated L96 timing control, it leads on VPT and geometric endpoints while BUNDLEDISTILL retains lower mean nRMSE. Formal regime-shift tests yield no Holm-corrected rejection with five seed blocks, and held-out long-run statistics are mixed. These results support finite-time cocycle supervision as a training signal whose benefits depend on the system, endpoint, and compute budget. The method requires resettable controlled trajectories, uses partial low-rank coverage, and incurs second-order differentiation cost. The experiments cover one GRU forecaster family and one calibrated-time comparison (L96 versus Rand-Jac2); noisy, partially observed, stochastic, and non-smooth systems are not evaluated.

AI USE STATEMENT

We used generative AI tools to assist with code optimization, manuscript editing, proof revision, scientific figure preparation, and interpretation of experimental results. The authors take responsibility for the final code, figures, text, and scientific claims.

ETHICS STATEMENT

This work studies deterministic synthetic dynamical systems and a public simulation benchmark. It uses no human-subject, personal, or sensitive data. We report negative and inconclusive comparisons alongside positive results and distinguish the prespecified core analysis from descriptive follow-up experiments.

REPRODUCIBILITY STATEMENT

Appendix C specifies the systems, data splits, architecture, optimization budget, baselines, evaluation thresholds, censoring rule, and hierarchical inference. The anonymous supplementary ZIP contains the frozen method and baseline implementation, effective core configurations and locked seeds, deterministic data-generation code, environment and hardware specifications, selected tests and a CPU smoke run, an opt-in 90-cell command generator, and compact path-free extracts of the reported aggregates and intervals. Raw trajectory tensors, checkpoints, scheduler state, and machine-specific logs are excluded; the supplied commands regenerate core data and runs, while full reconstruction of trajectory-level inference requires rerunning the corresponding evaluations. The 90-run core aggregates and in-distribution intervals trace to the audited core manifest; separately mapped extracts cover long-run uncertainty, formal trajectory-level OOD inference, five-seed L96/KS ablations, the calibrated optimizer-update-time control, compute audit, descriptive Dysts–12 evaluation, and the limited teacher sanity check.

REFERENCES

Giancarlo Benettin, Luigi Galgani, Antonio Giorgilli, and Jean-Marie Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems: A method for computing all of them. part 1: Theory. *Meccanica*, 15:9–20, 1980.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jean-Pierre Eckmann and David Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57(3):617–656, 1985.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801–806, 1993.

Daniel J. Gauthier, Erik Bollt, Aaron Griffith, and Wendson A. S. Barbosa. Next generation reservoir computing. *Nature Communications*, 12(1):5564, 2021.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=enYjtbjYJrf.

William Gilpin. Model scale versus domain knowledge in statistical forecasting of chaotic systems. *Physical Review Research*, 5(4):043252, 2023.

William Gilpin. Generative learning for nonlinear dynamics. *Nature Reviews Physics*, 6(3):194–206, 2024.

Francesco Ginelli, Paolo Poggi, Andrea Turchi, Hugues Chaté, Roberto Livi, and Antonio Politi. Characterizing dynamics with covariant lyapunov vectors. *Physical Review Letters*, 99(13):130601, 2007.

Shinhoo Kang, Hai V. Nguyen, and Tan Bui-Thanh. Learning chaotic dynamics through second-order geometric supervision. *arXiv preprint arXiv:2606.01596*, 2026.

Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005.

Pavel V. Kuptsov and Ulrich Parlitz. Theory and computation of covariant lyapunov vectors. *Journal of Nonlinear Science*, 22:727–762, 2012.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16768–16781, 2022.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963.

Edward N. Lorenz. Predictability: A problem partly solved. In *Seminar on Predictability, Vol. I*, pp. 1–18. European Centre for Medium-Range Weather Forecasts, 1996.

Jonas M. Mikhaeil, Zahra Monfared, and Daniel Durstewitz. On the difficulty of learning chaotic dynamics with RNNs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11297–11312, 2022.

Valery I. Oseledets. A multiplicative ergodic theorem. lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:197–231, 1968.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1310–1318, 2013.

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

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pp. 366–381. Springer, 1981.

Pantelis R. Vlachas, Jaideep Pathak, Brian R. Hunt, Themistoklis P. Sapsis, Michelle Girvan, Edward Ott, and Petros Koumoutsakos. Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126: 191–217, 2020.

Qilin Wang. Ellipsoidal time series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=CoAHlJuMdh.

Per-Å ke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

A Proofs and additional analysis

A.1 Proof of Proposition 1

Set $F = F_k = \widetilde{f}^k$ and $x = z$, and fix a unit vector v. Taylor's theorem with integral remainder, applied along the line segments $x \pm t\varepsilon v$, gives

$$F(x \pm \varepsilon v) = F(x) \pm \varepsilon DF(x)v + \frac{\varepsilon^2}{2} D^2F(x)[v, v] + R_{\pm}, \tag{19}$$

$$R_{\pm} = \pm \frac{\varepsilon^3}{2} \int_0^1 (1-t)^2 D^3F(x \pm t\varepsilon v)[v, v, v]\, dt. \tag{20}$$

The assumed C^3 regularity and derivative bound hold on both segments. Since $\int_0^1 (1-t)^2\, dt = 1/3$, each remainder satisfies $\|R_{\pm}\|_2 \le M_{3,k}\varepsilon^3/6$. Subtracting the two expansions and dividing by 2ε yields

$$\left\| \frac{F(x+\varepsilon v) - F(x-\varepsilon v)}{2\varepsilon} - DF(x)v \right\|_2 = \left\| \frac{R_+ - R_-}{2\varepsilon} \right\|_2 \le \frac{M_{3,k}\varepsilon^2}{6}. \tag{21}$$

Applying this bound at $x = z$ to each of the r orthonormal columns of V and summing their squared errors proves Eq. (14). A one-sided difference retains the $\varepsilon D^2F(x)[v, v]/2$ term and is only first-order accurate in general. □

A.2 Proof of Proposition 2

Haar invariance implies that $\mathbb{E}[VV^\top]$ commutes with every orthogonal matrix, hence $\mathbb{E}[VV^\top] = cI_d$. Taking traces gives $cd = \mathbb{E}\,\mathrm{tr}(V^\top V) = r$, so $c = r/d$. Therefore

$$\mathbb{E}_V \|AV\|_F^2 = \mathbb{E}_V\, \mathrm{tr}(A^\top AVV^\top) = \frac{r}{d}\, \mathrm{tr}(A^\top A) = \frac{r}{d}\, \|A\|_F^2 . \tag{22}$$

Applying the identity to $A = \mathbf{\Phi}^{\theta}_{t,k} - \mathbf{\Phi}^{\star}_{t,k}$ proves the stated consequence. □

A.3 Proof of Theorem 1

Write $A_{j:i} = A_j A_{j-1} \cdots A_i$ and similarly for B, with an empty product equal to the identity. The exact telescoping identity is

$$B_{k-1:0} - A_{k-1:0} = \sum_{i=0}^{k-1} B_{k-1:i+1}(B_i - A_i)A_{i-1:0}. \tag{23}$$

Taking spectral norms and using submultiplicativity gives Eq. (16). If $M, L \le \alpha$ and $\delta_i \le \delta$, every summand is at most $\alpha^{k-1}\delta$. □

The aligned-trajectory assumption separates derivative composition from state-distribution shift. In practice, $B_i = Dg_\theta(\hat{x}_i)$ and $A_i = Df(x_i)$ are evaluated on different paths. If both Jacobians are Lipschitz with constant L_J, then

$$\|Dg_\theta(\hat{x}_i) - Df(x_i)\|_2 \le \|Dg_\theta(x_i) - Df(x_i)\|_2 + L_J \|\hat{x}_i - x_i\|_2 , \tag{24}$$

so the same theorem applies with a local derivative term plus rollout deviation. This motivates retaining the trajectory loss.

A.4 TRANSPORTED-SUBSPACE PERTURBATION

Let $A = QR$ and $\widehat{A} = \widehat{Q}\widehat{R}$ be thin QR factorizations. Since A has rank r, its column space is its leading r-dimensional left singular subspace. Wedin's sin-Θ theorem (Wedin, 1972) gives

$$\left\|\sin\Theta(\widehat{Q}, Q)\right\|_F \leq \frac{\sqrt{r}\,\|E\|_2}{\sigma_r(A) - \|E\|_2} \tag{25}$$

under the stated separation condition. For equal-dimensional subspaces,

$$\left\|\widehat{Q}\widehat{Q}^\top - QQ^\top\right\|_F = \sqrt{2}\left\|\sin\Theta(\widehat{Q}, Q)\right\|_F, \tag{26}$$

which proves Eq. (17). Weyl's inequality yields the singular-value claim. □

The singular values of $A = QR$ and R coincide because Q has orthonormal columns. The method therefore trains on probed endpoint growth, which generally changes with the probe subspace and is not a full-tangent FTLE when $r < d$. The largest physical Lyapunov exponent used for evaluation is estimated separately by rank-one repeated QR of the true-system tangent action (Benettin et al., 1980; Eckmann & Ruelle, 1985).

A.5 SMOOTH-CONJUGACY ARGUMENT

Differentiate $g \circ h = h \circ f$ at x and apply the identity recursively to obtain Eq. (18). Uniform upper and lower bounds on Dh and Dh^{-1} contribute bounded multiplicative constants to tangent-vector norms; their logarithms vanish after division by k. Corresponding asymptotic exponents are therefore equal, and Dh maps invariant Oseledets subspaces wherever the splitting exists. □

A.6 WHY SCALAR GROWTH IS INSUFFICIENT

Consider

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{bmatrix}, \qquad B = R(\varphi)AR(-\varphi), \qquad \lambda > 1. \tag{27}$$

The maps have identical singular values, eigenvalues, and Lyapunov exponents, but their unstable directions differ by φ. Alternating differently rotated factors can also create or suppress non-normal transient growth despite matched factor spectra. This motivates an explicit transported-subspace term rather than scalar expansion alone.

B STATEFUL MODELS AND PARTIAL OBSERVATION

B.1 OBSERVABLE MAP INDUCED BY A RECURRENT FORECASTER

Let $y_t \in \mathbb{R}^{d_y}$ be the standardized observation and $p_t = y_{t-L+1:t-1}$ the preceding context, held fixed when y_t is perturbed. A recurrent forecaster initializes $h_t = H_\theta(p_t, y_t)$ and evolves

$$h_{q+1} = F_\theta(h_q, \hat{y}_q), \qquad \hat{y}_{q+1} = D_\theta(h_{q+1}). \tag{28}$$

Define $\zeta_q = [\hat{y}_q; h_q]$ and

$$G_\theta(\zeta_q) = \begin{bmatrix} D_\theta(F_\theta(h_q, \hat{y}_q)) \\ F_\theta(h_q, \hat{y}_q) \end{bmatrix}. \tag{29}$$

The induced observable map and its action are

$$\widehat{F}_{\theta,k}(y_t; p_t) = P_y G_\theta^k([y_t; H_\theta(p_t, y_t)]), \tag{30}$$

$$U_{t,k}^\theta = P_y DG_\theta^k(\zeta_t) \begin{bmatrix} V_t \\ D_{y_t} H_\theta(p_t, y_t) V_t \end{bmatrix}. \tag{31}$$

Thus teacher and student actions have the same observable dimension. Gradients are retained through hidden-state initialization and the lifted rollout. In the full-state Markov experiments, the preceding context is empty.

B.2 Passive observations are outside the empirical scope

For a partially observed deterministic system, one may form a delay state $s_t = [y_t, y_{t-\tau}, \ldots, y_{t-(m-1)\tau}]$ (Takens, 1981). When controlled resets are impossible, a possible extension is the local recurrence-neighbor regression

$$\widehat{\Phi}_{t,k} = \arg\min_{A} \sum_{j \in \mathcal{N}(t)} \kappa_{tj} \left\| (s_{j+k} - s_{t+k}) - A(s_j - s_t) \right\|_2^2 + \lambda \left\| A \right\|_F^2 . \tag{32}$$

Temporal exclusion would be required to avoid overlapping windows. This extension introduces sampling, curvature, and hidden-confounder error. It is not evaluated on passive or partially observed data in this paper.

C Experimental protocol and reproducibility

C.1 Dynamical systems and splits

Lorenz–63. We integrate

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z \tag{33}$$

with $(\sigma, \rho, \beta) = (10, 28, 8/3)$. Broad-box initial conditions receive a 50-time-unit burn-in. Fixed-step RK4 uses internal step 0.01 and saves every second step. The train/validation/test splits contain 256/64/128 independent trajectories with 1,001 saved states (20 time units). OOD evaluation uses $\rho \in \{32, 36\}$ and preserves the training normalization.

Lorenz–96. For cyclic indices $i = 1, \ldots, 40$,

$$\dot{x}_i = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F. \tag{34}$$

Training uses $F = 8$. RK4 uses internal step 0.01 and sample interval 0.05. We generate 128/32/64 independent trajectories, each with 2,001 saved states (100 time units), after a 100-time-unit burn-in. OOD forcing values are $F \in \{10, 12\}$.

Kuramoto–Sivashinsky. We solve $u_t + uu_x + u_{xx} + u_{xxxx} = 0$ with periodic boundary conditions on $[0, L)$, $L = 22$, and $N = 64$ equispaced values. A spectral ETDRK4 step of 0.25 is also the observation interval (Kassam & Trefethen, 2005). Each trajectory discards 500 steps and retains 10,000 states. The splits contain 64/16/32 trajectories. OOD domain lengths are $L \in \{23, 24\}$ at the same N.

All simulation is performed in float64 and stored in float32. Split-specific initial-condition seeds prevent trajectory overlap, and standardization statistics use only the training split. The data files and generation source are SHA-256 locked before final training.

C.2 Model and optimization

The common forecaster has a linear input projection, two width-128 GRU layers, layer normalization, and a residual decoder with two width-256 SiLU layers. Each final run uses 10,000 AdamW updates, batch size 64 on L63/L96 and 32 on KS, initial learning rate 3×10^{-4} with cosine decay to 10^{-6}, weight decay 10^{-5}, and gradient clipping at 1.0. Checkpoints are validated every 2,500 updates. Seeds 2024–2028 are locked final seeds.

The prediction curriculum starts with horizon $\{1\}$, adds $\{2, 4\}$ at 20% of training, horizon 8 at 35%, and horizon 16 at 50%. All methods except One-step share this multi-step prediction curriculum. Geometric supervision is disabled during the first 20%; at 20% it activates horizons $\{1, 2, 4\}$, adds horizon 8 at 35% and horizon 16 at 50%, while its loss weights ramp linearly from 20% to 60% of training. The final BundleDistill weights are $(\lambda_a, \lambda_g, \lambda_b) = (0.03, 0.01, 0.03)$. The teacher radius is 10^{-3} and confidence temperature is 0.2. Probe rank is one on L63 and four on L96/KS. All primary methods receive the same optimizer-update budget. Because geometric supervision has higher per-update cost, the primary experiment is not equal-wall-clock. We report a full compute audit and a separate calibrated optimizer-update-time control; the latter does not match end-to-end process time.

C.3 BASELINES

One-step and Multi-step. One-step uses teacher forcing without autonomous rollout loss. Multi-step uses the same prediction horizons and curriculum as BUNDLEDISTILL but no tangent losses, controlling for additional unrolling.

JacobianMatch. This baseline matches one-step reference and model Jacobian actions on equation-known systems using full identity probes with action weight 0.03. It receives the shared multi-step prediction curriculum, architecture, and update budget but no multi-step cocycle loss.

Growth-Match. This control isolates singular-growth supervision: it uses the same intervention-generated finite-time teacher, confidence, horizons, and prediction curriculum as BUNDLEDISTILL but matches only the probed singular-growth vector with weight 0.01, not action or orientation.

Rand-Jac2. The randomized derivative baseline combines one-step model/reference JVP matching on Haar probes (weight 0.03) with centered second differences of the one-step model and reference transition maps along those same probes for a directional second-derivative term (weight 0.10). It shares the multi-step prediction curriculum and is an inspired control rather than a verbatim reproduction of recent second-order geometric supervision (Kang et al., 2026). It is available for the three differentiable reference systems but not for the Dysts adapter, whose Torch RHS/RK4 path explicitly disables autograd.

C.4 METRICS AND STATISTICAL INFERENCE

At each held-out anchor, nRMSE is normalized with training statistics. VPT is the first time its curve crosses 0.4. If no crossing occurs within 512 saved steps, the observation is right-censored at the horizon. We report the restricted mean including horizon-capped observations; the artifacts also record censor fractions (at most 0.12% in distribution and zero for the formal OOD shifts), while the Dysts table reports them explicitly because they can be substantial. Lyapunov-time VPT multiplies physical time by a separately cached regime-specific largest Lyapunov exponent from rank-one repeated QR of the true-system tangent action. For the three in-distribution regimes, the estimate uses eight anchors, a 1,000-step burn-in, and 10,000 subsequent steps; the six OOD regimes use four anchors, a 200-step burn-in, and 2,000 subsequent steps. No student or training probe supplies this normalization.

Geometry evaluation uses fixed, matched probes at horizons $\{1, 2, 4, 8, 16\}$. Probed-growth MAE averages absolute errors in Eq. (3); transported-subspace angle is the RMS principal angle. Test evaluation uses four anchors per trajectory and up to 128 geometry anchors per seed.

For each system, the prespecified core analysis uses 10,000 hierarchical paired-bootstrap replicates for VPT and angle with seed 2027. Each replicate resamples the five locked model seeds, trajectories within selected seeds, and anchors within trajectories. Common resample indices preserve method pairing, and pointwise percentile 95% intervals summarize BUNDLEDISTILL-minus-baseline effects without simultaneous or multiplicity adjustment. The archived core analyzer also computed two-sided tail masses from these ordinary bootstrap draws and then applied Holm correction. Because those draws estimate the sampling distribution around the observed effect rather than a distribution generated under the null, we do not interpret or report the resulting p-values. Core claims therefore rest on effect magnitudes and pointwise percentile intervals, not a multiplicity-adjusted rejection decision.

C.5 LONG-RUN AND OOD EVALUATION

Long-run SWD and power-spectrum error are computed after trajectory pairing is no longer meaningful. For every model seed, four learned ensembles and four disjoint reference ensembles use separate initial conditions. Both discard at least five regime-specific Lyapunov times and retain at least 100 Lyapunov times. SWD uses 1,024 fixed projections; uncertainty resamples seed-level ensemble aggregates, never paired time indices.

For OOD inference, we deterministically reevaluate all 90 immutable best checkpoints at two shifts per system, retaining 32 matched trajectory-level VPT values for each seed, method, and shift. This

produces 5,760 rows with common initial conditions and trajectory hashes across methods. Each BUNDLEDISTILL-minus-baseline interval uses 10,000 paired hierarchical bootstrap replicates that resample seed blocks and trajectories within seeds. Two-sided p-values enumerate all $2^5 = 32$ sign flips of the five seed-level mean paired differences, under independent seed blocks and sign symmetry under the null. Holm correction is applied once across the 30 comparisons. Five seed blocks imply a minimum attainable two-sided $p = 2/32 = .0625$, so this design can quantify paired uncertainty but cannot reject at level .05.

C.6 AUDIT TRAIL

The formal core consists of three systems, six methods, and five seeds (90 training runs), followed by 90 checkpoint-level long-run evaluations and 90 checkpoint-level OOD tasks, each containing two shifts. Learned and reference long-run ensembles use disjoint initial conditions, while the six methods for a given system and seed share the locked anchor and reference design. The OOD supplement verifies exactly 90 task manifests, 180 shift outputs, 5,760 trajectory rows, and 30 paired comparisons. Separate audited packages cover 40 five-seed L96/KS ablation cells, a 10-run calibrated update-time control, and compute measurements for all 90 core checkpoints. Fail-closed completeness reports verify exact inventories, locked seeds, pairing, source and data hashes, bootstrap and correction settings, and output hashes. Every training run records its frozen configuration, checkpoint hash, code/source manifest, runtime fingerprint, per-anchor outputs, and scheduler return code.

D ADDITIONAL AUDITED RESULTS

D.1 FULL PAIRED EFFECTS

Table 5 reports every prespecified effect comparison used in the main analysis. L63 VPT intervals against Multi-step, Growth-Match, and Rand-Jac2 span zero, whereas the L96 and KS VPT intervals and all transported-angle intervals favor BUNDLEDISTILL. These are paired percentile uncertainty summaries, not multiplicity-adjusted hypothesis tests.

Table 5: Audited hierarchical paired-bootstrap effect estimates. Differences are BUNDLEDISTILL minus the indicated baseline. Each pointwise percentile interval uses 10,000 seed–trajectory–anchor resamples with common paired indices; intervals are not simultaneous or multiplicity-adjusted. Positive VPT and negative angle effects favor BUNDLEDISTILL. Bootstrap-tail p-values are omitted because the resamples are not generated under the null.

System	Baseline	ΔVPT [95% interval]	ΔAngle [95% interval]
L63	One-step	$2.870\ [2.476, 3.418]$	$-27.151\ [-28.259, -25.950]$
	Multi-step	$0.202\ [-0.053, 0.505]$	$-5.279\ [-6.359, -4.354]$
	JacobianMatch	$0.331\ [0.055, 0.679]$	$-1.647\ [-1.984, -1.351]$
	Growth-Match	$0.012\ [-0.048, 0.069]$	$-0.029\ [-0.034, -0.025]$
	Rand-Jac2	$0.166\ [-0.365, 0.968]$	$-1.995\ [-2.339, -1.720]$
L96	One-step	$0.952\ [0.920, 0.982]$	$-22.144\ [-22.526, -21.753]$
	Multi-step	$0.021\ [0.015, 0.028]$	$-0.834\ [-0.893, -0.775]$
	JacobianMatch	$0.018\ [0.012, 0.025]$	$-0.725\ [-0.784, -0.666]$
	Growth-Match	$0.018\ [0.016, 0.021]$	$-0.425\ [-0.438, -0.413]$
	Rand-Jac2	$0.138\ [0.125, 0.150]$	$-4.101\ [-4.308, -3.910]$
KS	One-step	$0.809\ [0.765, 0.853]$	$-60.012\ [-60.360, -59.666]$
	Multi-step	$0.242\ [0.203, 0.280]$	$-57.174\ [-57.537, -56.806]$
	JacobianMatch	$0.198\ [0.167, 0.229]$	$-20.011\ [-20.420, -19.591]$
	Growth-Match	$0.180\ [0.149, 0.215]$	$-15.612\ [-16.167, -15.075]$
	Rand-Jac2	$0.354\ [0.316, 0.391]$	$-34.229\ [-35.158, -33.037]$

D.2 FORMAL OOD PAIRED EFFECTS

Table 6 gives a compact contrast against the strongest observed baseline in each shift; this post hoc selection is only for presentation and does not define the multiplicity family. The displayed percentile intervals are pointwise and are not adjusted for either this selection or multiple comparisons. All six raw exact tests attain the smallest possible two-sided value with five seed blocks, .0625, but none can reject at .05 even before correction. The global Holm adjustment over the full 30 comparisons therefore yields $p_{\mathrm{H}} = 1$ for these rows. The hierarchical intervals summarize paired seed–trajectory sampling uncertainty, whereas the exact test uses the 32 seed-block sign flips. Their different inferential constructions allow an interval to exclude zero without a rejection by the exact test.

Table 6: Formal OOD paired effects against the strongest observed non-BUNDLEDISTILL mean in each shift. Differences are BUNDLEDISTILL minus the named comparator; positive VPT favors BUNDLEDISTILL. The displayed pointwise intervals use 10,000 paired seed–trajectory bootstrap replicates and are not adjusted for comparator selection or multiplicity. Assuming independent seed blocks and sign symmetry under the null, raw p-values enumerate the exact 2^5 seed-block sign-flip distribution; p_{H} applies one Holm correction across all 30 OOD comparisons. The complete 30-row table is included in the result package.

System	Shift	Comparator	ΔVPT [95% interval]	Exact p	p_{H}
L63	$\rho = 32$	Rand-Jac2	-0.00496 $[-0.01290, -0.00168]$	.0625	1.000
L63	$\rho = 36$	Rand-Jac2	-0.00200 $[-0.00416, -0.00068]$	.0625	1.000
L96	$F = 10$	Growth-Match	$+0.00843$ $[0.00746, 0.00950]$	.0625	1.000
L96	$F = 12$	Growth-Match	$+0.00514$ $[0.00467, 0.00562]$	.0625	1.000
KS	$L = 23$	JacobianMatch	$+0.01727$ $[0.01267, 0.02241]$	.0625	1.000
KS	$L = 24$	Growth-Match	$+0.00957$ $[0.00075, 0.01818]$	.0625	1.000

D.3 HELD-OUT LONG-RUN STATISTICS

For L96, BUNDLEDISTILL has the smallest observed SWD and power-spectrum error, but common-resample seed-level percentile intervals against Growth-Match cross zero: SWD difference -0.00110 with CI $[-0.00214, 0.00018]$, and power-spectrum difference -0.0240 with CI $[-0.1027, 0.0508]$. The resampling uses common seed indices across methods but does not pair learned or reference ensembles across methods. On L63, neither long-run endpoint is best. On KS, Multi-step has substantially lower values on both endpoints.

D.4 DYSTS–12 DESCRIPTIVE BREADTH CHECK

We additionally execute 12 systems from `dysts==0.96`: Lorenz, Rössler, Chen, Halvorsen, Thomas, Aizawa, Rabinovich–Fabrikant, Burke–Shaw, Dadras, Dequan–Li, Hadley, and `SprottA` (Gilpin, 2021). Each system has 32/8/16 train/validation/test trajectories, five model seeds, and four trajectory-capable methods, for 240 completed runs. The additional evaluation protocol uses RK4 with 50 substeps per observation, 100 observations per characteristic period, a 10-period burn-in, and 20 retained periods; these choices are frozen in the Dysts data manifest but were not part of the prespecified core. VPT is expressed in Dysts characteristic periods because a signed full-tangent Lyapunov estimate is unavailable. Exact-derivative baselines are omitted because the Dysts adapter disables autograd.

The system-balanced observed means are close for BUNDLEDISTILL (2.935 periods) and Growth-Match (2.920). Several systems have substantial censoring, especially Aizawa, so this table documents descriptive breadth and similar observed macro means rather than superiority or formal equivalence. We keep it out of the prespecified core because no hierarchical system-level analysis was specified before inspection.

D.5 FIVE-SEED CROSS-SYSTEM ABLATION DETAILS

The L96 and KS ablations reuse the locked data, 10,000-update budget, configurations, five seeds, and evaluation anchors from the full model. The audited matrix contains 40 cells: four variants, two

Table 7: Held-out long-run statistics over five seeds. Each seed aggregates four learned ensembles and four reference ensembles whose initial conditions are disjoint from the learned ensembles; methods within a system and seed share the locked reference design. Values are system-specific and must not be macro-averaged across rows; lower is better. Bold denotes the smallest observed mean within each system.

System	Method	SWD↓	Power-spectrum error↓
L63	One-step	200.214	8.160
	Multi-step	0.059	1.242
	JacobianMatch	0.095	0.814
	Growth-Match	0.057	**0.777**
	Rand-Jac2	**0.053**	1.128
	BUNDLEDISTILL	0.070	1.261
L96	One-step	160.036	8.693
	Multi-step	0.157	0.817
	JacobianMatch	0.157	0.826
	Growth-Match	0.146	0.800
	Rand-Jac2	0.171	0.798
	BUNDLEDISTILL	**0.145**	**0.776**
KS	One-step	259.631	10.750
	Multi-step	**22.302**	**6.235**
	JacobianMatch	283.605	11.188
	Growth-Match	223.029	10.695
	Rand-Jac2	384.084	11.564
	BUNDLEDISTILL	191.145	10.449

Table 8: Descriptive Dysts–12 results over five seeds. Each cell reports restricted-mean VPT in characteristic periods and the right-censor fraction in parentheses. The high censor fractions on several systems make the capped means particularly important; system-level comparisons are descriptive.

System	One-step	Multi-step	Growth-Match	BUNDLEDISTILL
Aizawa	0.145 (0.000)	5.109 (0.972)	5.119 (0.994)	5.116 (0.988)
Burke–Shaw	0.098 (0.000)	2.351 (0.006)	2.508 (0.006)	2.526 (0.006)
Chen	0.108 (0.000)	2.211 (0.003)	2.298 (0.009)	2.311 (0.006)
Dadras	0.087 (0.000)	0.851 (0.000)	0.837 (0.000)	0.843 (0.000)
Dequan–Li	0.413 (0.000)	4.276 (0.619)	3.981 (0.444)	3.991 (0.409)
Hadley	0.114 (0.000)	4.343 (0.506)	4.534 (0.581)	4.503 (0.553)
Halvorsen	0.061 (0.000)	1.834 (0.003)	2.245 (0.006)	2.254 (0.009)
Lorenz	0.089 (0.000)	1.956 (0.009)	2.095 (0.013)	2.073 (0.019)
Rabinovich–Fabrikant	0.103 (0.000)	3.908 (0.303)	3.715 (0.203)	3.724 (0.200)
Rössler	0.082 (0.000)	2.243 (0.000)	2.623 (0.075)	2.770 (0.053)
SprottA	0.110 (0.000)	4.295 (0.419)	4.626 (0.609)	4.662 (0.641)
Thomas	0.062 (0.000)	0.406 (0.000)	0.456 (0.000)	0.450 (0.000)
System-balanced mean	0.123 (0.000)	2.815 (0.237)	2.920 (0.245)	2.935 (0.240)

Table 9: Compute audit over five seeds. All methods use the same architecture within a system. Latency is batch-one, single-step CUDA-event latency (mean±SD across checkpoints); memory is the trainer-recorded peak allocated CUDA memory (not reserved); time is end-to-end trainer time from the immutable formal runs. "Reach" counts seeds whose logged validation VPT first reaches 0.5 Lyapunov times within 10,000 updates. VPT/GPU-h is descriptive final test VPT divided by trainer hours, not an equal-wall-clock causal comparison.

Method	Params	Latency (μs)	Memory (MiB)	Time (min)	Reach	VPT/GPU-h
Lorenz–63						
One-step	299,523	363.0±13.0	24.3	4.1	0/5	2.06
Multi-step	299,523	375.1±5.3	37.3	8.9	5/5	18.96
JacobianMatch	299,523	361.2±8.8	43.8	12.9	5/5	12.53
Growth-Match	299,523	362.4±9.5	105.3	46.9	5/5	3.94
Rand-Jac2	299,523	378.5±38.1	43.8	13.0	5/5	13.09
BUNDLEDISTILL	299,523	365.1±9.6	105.3	34.7	5/5	5.19
Lorenz–96						
One-step	323,240	361.0±5.5	25.1	3.6	0/5	4.19
Multi-step	323,240	364.5±7.6	38.1	4.7	5/5	15.03
JacobianMatch	323,240	361.9±8.3	95.2	37.3	5/5	1.88
Growth-Match	323,240	368.2±11.7	162.2	43.0	5/5	1.63
Rand-Jac2	323,240	361.4±6.7	55.6	14.6	5/5	4.30
BUNDLEDISTILL	323,240	361.2±6.0	162.2	43.4	5/5	1.64
Kuramoto–Sivashinsky						
One-step	338,624	368.5±6.4	25.5	2.1	0/5	4.26
Multi-step	338,624	371.1±7.8	30.2	4.2	5/5	10.32
JacobianMatch	338,624	358.3±7.5	75.7	34.3	5/5	1.33
Growth-Match	338,624	368.3±12.7	92.5	32.2	5/5	1.45
Rand-Jac2	338,624	365.8±12.6	38.8	11.7	5/5	3.09
BUNDLEDISTILL	338,624	369.3±8.4	92.5	31.6	5/5	1.82

systems, and five seeds. On L96, the full-minus-ablated VPT effects are $+0.00892$ without action (exploratory paired-t CI $[0.00788, 0.00996]$), $+0.00136$ without growth ($[-0.00261, 0.00532]$), and $+0.00851$ without the bundle term ($[0.00693, 0.01009]$). The corresponding reductions in angle are $0.213°$, $0.353°$, and $0.199°$. Growth removal primarily affects its aligned endpoint: growth MAE increases by 0.04095.

On KS, action and bundle removal reduce VPT by 0.05205 and 0.07396, while increasing angle by $2.057°$ and $3.146°$. Growth removal instead increases VPT by 0.09185 even as growth MAE increases by 0.00600 and angle increases by $0.320°$. These five-seed intervals are exploratory and no component-level multiplicity-adjusted test was specified. The observed effects show a prediction–geometry trade-off under the tested configurations.

D.6 COMPUTE AUDIT

Table 9 evaluates the immutable best checkpoint from every core run. Parameter count and inference latency are properties of the shared forecaster and therefore remain essentially unchanged by the training loss. Peak memory is the trainer-recorded maximum allocated CUDA memory (not reserved memory); trainer elapsed time comes from the original formal run and can include contemporaneous hardware load. "Reach" is interval-censored to the 2,500-update validation cadence. The VPT/GPU-hour column divides final test VPT by observed trainer time and is descriptive: methods may follow different learning curves, and this ratio is not a randomized equal-time estimand.

D.7 CALIBRATED OPTIMIZER-UPDATE-TIME CONTROL

The timing protocol freezes a common 2,992-s calibration cap from eight calibration repetitions (four per method), which maps to 10,000 BUNDLEDISTILL updates and 37,915 Rand-Jac2 updates

on L96. Validation, testing, checkpointing, logging, and I/O are excluded from the calibration but included in separately reported process time. In the final five paired seeds, neither method exhausts the conservative cap: mean update times are 1,976 and 2,163 s, and process times are 2,090 and 2,437 s. Thus the result is a calibrated control, not exact realized equality. The paired BUNDLEDISTILL-minus-Rand-Jac2 differences are -0.61121 VPT (descriptive bootstrap CI $[-0.62251, -0.59826]$), -0.022996 mean nRMSE ($[-0.027120, -0.020012]$), $+0.02163$ growth MAE ($[0.01840, 0.02456]$), and $+6.2365°$ angle ($[6.1517, 6.3213]$). Under this calibration, Rand-Jac2 has higher VPT and lower geometric errors, whereas BUNDLEDISTILL has lower mean nRMSE.

D.8 TEACHER IMPLEMENTATION SANITY CHECK

The fine-radius relative error is below 4×10^{-4} at both tested horizons. Because the audit contains only two L63 anchors, rank one, and horizons one and two, it checks implementation consistency but does not establish cross-system teacher fidelity or downstream gain retention.

Table 10: L63 antithetic-teacher sanity check against exact JVP actions at two anchors, with rank one and horizons one and two.

Quantity	Value
Fine-radius mean relative error	3.593×10^{-4}
Coarse-radius mean relative error	2.852×10^{-4}
Fine error, horizon 1	3.523×10^{-4}
Fine error, horizon 2	3.663×10^{-4}
Mean confidence	0.9987

E FAILURE MODES AND SCOPE

Finite radius and numerical cancellation. At large radii or long horizons, nonlinear separation can dominate the tangent response; at very small radii, round-off and solver tolerance dominate symmetric differences. Cross-radius confidence detects disagreement but is not a proof of local linearity.

Low-rank coverage. Proposition 2 is an expectation statement. A rank-r sample can miss an important direction at an individual anchor, and Eq. (17) becomes ill-conditioned near rank collapse.

Access model. The core method needs resettable controlled trajectories. Passive observations, stochastic forcing, discontinuous events, and hidden confounding require different estimators and are not covered by the experiments.

Compute. The two-radius teacher and differentiation through student JVPs increase training cost. All primary comparisons match optimizer updates rather than time; the calibrated L96 control favors Rand-Jac2 on VPT and geometric endpoints when its cheaper updates are reinvested, while BUNDLEDISTILL retains lower mean nRMSE. There is no auxiliary inference cost because the learned forecaster architecture is unchanged.

Long-run interpretation. In chaotic systems, amplification of initial-state or model errors can limit trajectory prediction. Paired VPT and tangent metrics quantify finite-time behavior; held-out SWD and spectral summaries address long-run distributions. Table 7 shows that finite-time gains coexist with mixed long-run distributional accuracy.