Terminal Failure Is Not Computational Failure: A Pre-Collapse Readout Regime in Nonlinear Dynamical Learners

Rohan Maskrey*
Independent Researcher
rohanmaskrey@dunelm.org.uk

Abstract

Iterative dynamical learners perform inference by running a state trajectory, usually decoded at its endpoint. We show that this convention can fail: terminal representations can degrade while earlier trajectory prefixes still support accurate prediction, defining a pre-collapse regime. We formalise the usable prefix as an accuracy-defined grace period and compare fixed early-window readout, norm-ratio stopping, and GRACE (Growth-Regulated Adaptive State Extraction) as potential decoders for this regime. In complex modReLU reservoirs, non-terminal readouts preserve near-stable MNIST accuracy across an unstable spectral sweep while terminal readout collapses. Decoder effectiveness depends on the dynamics, with simple early windows strongest in the trained LeakyReLU and LSTM checks. The gap also appears on CIFAR-10 with frozen features. On Tiny ImageNet-200 with frozen ConvNeXt features, extending eight-step-trained recursive SwiGLU and attention models to 32 steps reduces terminal accuracy by 5.11 and 2.42 percentage points, respectively, while early prefixes remain predictive. Matched 32-step training removes this deterioration in both tested architectures. Terminal failure is therefore not necessarily computational failure: readout timing and the horizon used during training jointly determine endpoint validity.

1 Introduction

Iterative dynamical learners are almost always evaluated at the endpoint of their computation. Echo state networks read out a terminal reservoir state [1]; deep equilibrium models solve for a fixed point [2]; neural ODEs integrate to a final time [3]. This shared design choice hides a strong assumption: the terminal state is the right object of inference.

We show that this assumption can fail in a *pre-collapse learning regime*: terminal-state inference degrades before computational utility is exhausted (some timestep probe remains task-predictive; Definition 1). We formalise the usable prefix before terminal failure as an accuracy-defined *prefix grace period*, observe monotone shortening across the tested spectral sweep, and give a conditional growth argument explaining why norm-driven failure shortens usable prefixes. The central contribution is the identification and characterisation of this regime. We compare fixed early-window readout, adaptive stopping, and GRACE as potential decoders for its usable prefixes; their relative strengths depend on the dynamics. Cross-system experiments establish that, in the tested taxonomy, the regime appears with nonlinear state distortion rather than linear radial growth alone; the same endpoint-vs.-trajectory gap appears on CIFAR-10, and training at the unstable operating point postpones but does not suppress the regime under further gain shifts. On Tiny ImageNet-200, recursive SwiGLU and attention heads also lose endpoint accuracy when run beyond their training horizon;

*Code and saved results: https://github.com/maskrys-dev/terminal-failure.

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

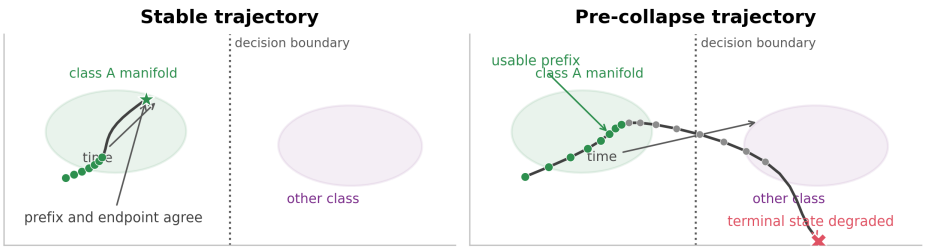


Figure 1: **Conceptual picture of the pre-collapse regime.** In a stable trajectory, prefix states and the endpoint agree on the class label. In a pre-collapse trajectory, nonlinear amplification distorts the late state while predictive prefix states remain inside the correct class region. The failure is one of readout timing, not of the system's underlying computation.

matched-horizon training restores endpoint validity in both tested architectures. Unlike early-exit work, where intermediate predictions are used to save computation, our setting concerns a dynamical failure mode: additional computation can actively destroy endpoint validity even while earlier trajectory prefixes remain predictive (Figure 1).

Classical reservoir practice often sets the spectral radius $\rho(\mathbf{W}) < 1$ for stability, but that prescription is neither always available nor sufficient. In trained iterative systems, effective gain is learned; in physical and analog in-memory substrates, variability, drift, input power, and nonlinearity change rollout dynamics [23–26]; and non-normal amplification can produce large transient growth even when spectral radius suggests stability [18, 30, 33].

Contributions.

1. We identify the *pre-collapse learning regime* and formalise an accuracy-defined prefix grace period $G(\rho, \alpha)$: terminal-state inference fails before computational utility is lost, with empirically monotone shortening across the tested spectral sweep (Section 3 and Section 4).
2. We characterise the regime across unbounded, bounded, and linear systems, and replicate the endpoint-vs.-trajectory mismatch on CIFAR-10 (with FashionMNIST in the appendix) (Section 6).
3. We show the regime persists in a trained LeakyReLU classifier even when trained at the unstable operating point: training postpones but does not suppress endpoint failure, with early-window recovering up to $+26.5$ pp; a trained row-sequential CIFAR-10 LSTM shows the same smaller bounded-gate pattern (Section 6.2 and Appendix O).
4. On Tiny ImageNet-200, recursive SwiGLU and shared-attention heads exhibit endpoint deterioration under inference-horizon extension, while matched-horizon training restores endpoint validity (Section 6.3).
5. We distinguish the accuracy-defined prefix from a practical norm-ratio proxy and compare fixed early-window readout, adaptive stopping, and GRACE as potential decoders, characterising their strengths and limitations across dynamical settings (Sections 5 and 6).

Scope and non-claims. We do not claim edge-of-chaos optimality, competitive CIFAR-10 performance, universal dominance of GRACE over simpler prefix readouts, or DEQ convergence theory. Why not simply keep $\rho(\mathbf{W}) < 1$? We use spectral radius as a controlled probe, not an operating recommendation; the broader issue is endpoint validity under finite-time instability.

Relation to prior work. Transient recurrent dynamics can carry information beyond fixed-point readout [21, 1], and reservoir recipes already use averaged or windowed states [9]. Our contribution is different in kind: a supercritical endpoint-validity failure regime, operationalised through a prefix grace period, a norm-ratio proxy, and a trained-system validation. Adaptive computation and test-time compute ask how much inference compute to allocate [10, 11, 34, 35]; we ask whether the terminal readout remains valid once that computation enters an unstable regime. Early-exit and overthinking work studies depth-indexed intermediate heads [31, 32, 36]; here the axis is time and the stop signal is the trajectory norm ratio. Work on criticality, signal propagation, and non-normal

amplification explains when recurrent and deep systems preserve or amplify information [13–16, 18, 30, 33]; we ask when endpoint-based inference fails before that transient information is exhausted.

2 Background and Setup

We make the endpoint convention precise and isolate the temporally asymmetric failure of endpoint readout. Complex-valued and unitary dynamics [5, 6] provide rich substrates; complex modReLU makes the regime especially legible by preserving phase while permitting magnitude growth.

In Sections 3 through 5 and the fixed-substrate experiments of Section 6, we study dynamical learners of the form

$$\mathbf{z}_{t+1} = \sigma(\mathbf{W}\mathbf{z}_t + \mathbf{U}\mathbf{x} + \mathbf{b}), \qquad \mathbf{z}_0 = \mathbf{0}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a fixed input, $\mathbf{W} \in \mathbb{C}^{d\times d}$ is the recurrent weight matrix with spectral radius $\rho(\mathbf{W})$, $\mathbf{U} \in \mathbb{C}^{d\times n}$ is the input projection, $\mathbf{b} \in \mathbb{C}^d$ is a bias, and σ is a pointwise nonlinearity. The weights $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{b}$ are fixed (not trained), and learning occurs solely in a downstream linear probe. This controlled setting isolates representational dynamics from training dynamics: the regime, if present, is a property of rollout geometry rather than of optimisation. Section 6.2 relaxes the fixed-substrate assumption with a weight-tied iterative classifier whose recurrent weights, input projection, bias, and head are jointly trained, testing whether the regime survives when training shapes the terminal representation.

Rolling the system for a fixed horizon T produces a trajectory $(\mathbf{z}_1, \ldots, \mathbf{z}_T)$; in the main experiments, $T = 40$. We compare four readout strategies. The *final-state* readout $\phi_{\text{final}} = f(\mathbf{z}_T)$ is the standard baseline. The *early-window* readout $\phi_{\text{early}} = \frac{1}{K}\sum_{t=1}^{K} f(\mathbf{z}_t)$ averages the first $K \ll T$ steps. The *adaptive stopping* readout $\phi_{\text{stop}} = f(\mathbf{z}_{\hat{t}})$ halts at the first timestep where the safeguarded norm ratio $\|\mathbf{z}_t\| / \max(\|\mathbf{z}_1\|, \epsilon_{\text{num}})$ exceeds a threshold τ_{stop}. Finally, GRACE (*Growth-Regulated Adaptive State Extraction*) computes a per-sample norm-ratio-weighted average over the full trajectory, down-weighting timesteps whose norm ratio exceeds a threshold (Section 5.2). Here $f(\mathbf{z}) = [\mathrm{Re}(\mathbf{z}), \mathrm{Im}(\mathbf{z})]$ maps complex states to real features; for real-valued systems $f(\mathbf{z}) = \mathbf{z}$.

Definition 1 (Prefix grace period)**.** *For a fixed dataset split, probe-training protocol, and dynamical system with spectral radius $\rho(\mathbf{W})$ and accuracy threshold α, the* grace period *is the usable prefix length*

$$G(\rho(\mathbf{W}), \alpha) = \max(\{0\} \cup \{T' \in \{1, \ldots, T\} : \mathrm{acc}(\phi_t) \geq \alpha \text{ for all } 1 \leq t \leq T'\}), \tag{2}$$

where $\phi_t = f(\mathbf{z}_t)$ is the single-timestep readout and $\mathrm{acc}(\phi_t)$ is measured by training and evaluating a separate linear probe using only features from timestep t. The definition is intentionally prefix-based: if the first timestep fails, $G = 0$; if accuracy dips below α and later recovers, the grace period still ends at the first failure. Thus G, $\mathrm{acc}(\cdot)$, and acc_0 below are empirical finite-sample quantities conditional on the stated evaluation protocol.

Definition 2 (Pre-collapse learning regime)**.** *Fix margins $\Delta > 0$ and $0 < \epsilon < \Delta$. A dynamical system with spectral radius $\rho(\mathbf{W})$ operates in the* pre-collapse learning regime *when terminal accuracy has degraded from the stable baseline acc_0, $\mathrm{acc}(\phi_{\text{final}}) < \mathrm{acc}_0 - \Delta$, while a usable prefix remains: $G(\rho(\mathbf{W}), \mathrm{acc}_0 - \epsilon) > 0$. Here acc_0 is measured by the same readout/probe protocol at a stable reference configuration (in the main spectral sweep, $\rho(\mathbf{W}) = 0.5$; for trained-system checks, the corresponding training operating point).*

We use $\Delta = 5$ pp and $\epsilon = 1$ pp throughout.

Remark 3. The pre-collapse regime is distinct from the edge-of-chaos regime [13]: it concerns systems operating *past* the critical boundary ($\rho(\mathbf{W}) > 1$), not at it, and it characterises temporally asymmetric representational quality in the supercritical phase rather than optimality at criticality.

3 Empirical Discovery of the Pre-Collapse Regime

We instantiate Equation (1) on flattened MNIST [39] with a complex modReLU reservoir ($d = 128$, $T = 40$, bias $b = -0.5$), sweeping $\rho(\mathbf{W}) = 0.5$ to 1.6 across 10 random seeds. $\mathbf{W}$ is drawn from the complex Ginibre ensemble [7] and rescaled to the target radius; $\mathbf{U}$ is complex Gaussian and column-normalised. A RidgeClassifier probe [42] is trained/evaluated on a 5k/5k split; values are mean $\pm$ std. The early-window baseline uses $K = 10$, with sensitivity in Appendix I.

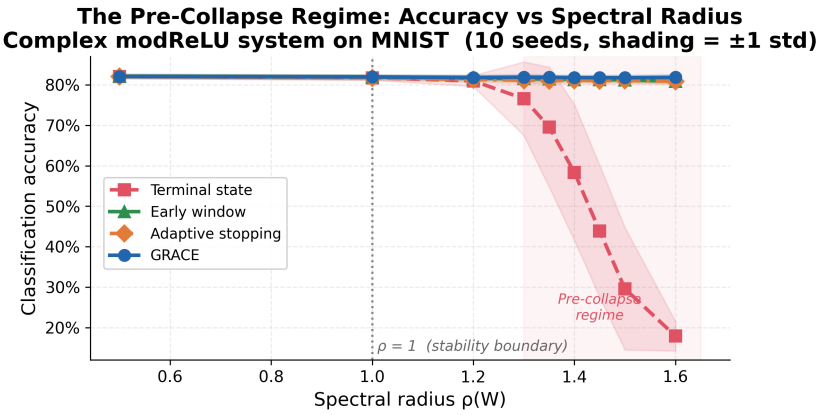


Figure 2: **The pre-collapse regime in the complex modReLU system.** As spectral radius increases, terminal-state accuracy collapses while early-window, adaptive stopping, and GRACE remain near-flat (10 seeds; shading $= \pm 1$ std). The pink region marks the pre-collapse band ($\rho \geq 1.30$) where endpoint-based evaluation fails. At $\rho = 1.6$, GRACE exceeds final-state readout by 64 pp, and all cross-seed pair comparisons favour GRACE.

Table 1: Readout accuracy (mean $\pm$ std, 10 seeds) across spectral radii on MNIST. Non-terminal readouts maintain near-stable accuracy across the unstable sweep. Cohen's d is reported for GRACE over final-state readout; the largest values are driven by the combination of mean separation and GRACE's reduced cross-seed standard deviation. At $\rho = 1.6$, Cliff's $\delta = 1.0$.

$\rho(\mathbf{W})$	Final	Early	Adaptive	GRACE	Cohen's d
0.50	0.821±0.005	0.821±0.005	0.821±0.005	0.821±0.005	0.0
1.00	0.818±0.006	0.820±0.005	0.818±0.006	0.819±0.006	0.3
1.20	0.810±0.014	0.817±0.006	0.813±0.006	0.818±0.006	0.8
1.30	0.766±0.091	0.815±0.006	0.811±0.007	0.819±0.007	0.8
1.35	0.696±0.148	0.814±0.007	0.811±0.008	0.819±0.006	1.1
1.40	0.584±0.168	0.813±0.006	0.811±0.008	0.818±0.006	1.9
1.45	0.439±0.163	0.813±0.006	0.810±0.009	0.818±0.006	3.1
1.50	0.296±0.151	0.812±0.006	0.811±0.008	0.818±0.006	4.6
1.60	0.179±0.037	0.809±0.007	0.809±0.008	0.819±0.007	23.0

Figure 2 visualises the regime in the complex modReLU sweep. In this sweep, the accuracy-defined grace period $G(\rho(\mathbf{W}), 0.70)$ shortens as spectral radius increases, remaining at the full $T = 40$ horizon through $\rho(\mathbf{W}) \leq 1.20$, then falling to 38.4 at $\rho(\mathbf{W}) = 1.30$, 31.1 at $\rho(\mathbf{W}) = 1.40$, 20.2 at $\rho(\mathbf{W}) = 1.50$, and 15.2 at $\rho(\mathbf{W}) = 1.60$ (Appendix H). Separately, a norm-threshold stopping-time proxy also shortens and provides an online indicator of usable prefix length.

The quantitative summary in Table 1 matches the visual transition in Figure 2. Terminal-state accuracy collapses from $82.1\% \pm 0.5$ at $\rho(\mathbf{W}) = 0.5$ to $17.9\% \pm 3.7$ at $\rho(\mathbf{W}) = 1.6$, while non-terminal readouts remain near the stable baseline; GRACE reaches $81.9\% \pm 0.7$ at $\rho(\mathbf{W}) = 1.6$. Cohen's d for GRACE over final-state grows from $d = 0.30$ at $\rho(\mathbf{W}) = 1.0$ to $d = 23.0$ at $\rho(\mathbf{W}) = 1.6$. This very large value should be read together with the raw means and standard deviations: it reflects both the 64 pp mean separation and the fact that GRACE sharply reduces cross-seed standard deviation. A rank-based check gives Cliff's $\delta = 1.0$ at $\rho(\mathbf{W}) = 1.6$ (all 100 cross-seed pairs favour GRACE over final-state).

We treat the coarse regime radii ($\rho(\mathbf{W}) \in \{1.3, 1.4, 1.5, 1.6\}$) as the primary inferential family because each satisfies Definition 2 under $\Delta = 5$ pp and $\epsilon = 1$ pp, and these are the main-text operating points used for the headline claims. The intermediate radii 1.35 and 1.45 are corroborating measurements included in the appendix-wide audit. Across the primary family, one-sided exact paired Wilcoxon signed-rank tests confirm that GRACE significantly outperforms final-state readout (all raw $p = 0.001$, the minimum achievable for $n = 10$ non-tied paired samples under this one-sided test) and early-window readout (raw $p \leq 0.007$ at all four radii), with effect sizes ranging

from $d = 0.60$ to $d = 1.33$ for the GRACE-vs-early comparison. After Holm-Bonferroni correction across the eight primary tests (4 radii $\times$ 2 comparisons), these remain significant. Broader sweeps across all radii and auxiliary readout comparisons are reported as secondary analyses in Appendix C.

4 Mechanism: Why Terminal Failure Precedes Transient Failure

The pre-collapse regime can be understood through a separation between two timescales: the timescale at which discriminative structure is imprinted by the input projection, and the timescale at which nonlinear amplification destroys it. We formalise this below.

Consider two inputs $\mathbf{x}, \mathbf{x}'$ from distinct classes, producing trajectories $\mathbf{z}_t, \mathbf{z}'_t$ under Equation (1). Define inter-class separation $\delta_t = \|\mathbf{z}_t - \mathbf{z}'_t\|$. Since both trajectories start from $\mathbf{z}_0 = \mathbf{z}'_0 = \mathbf{0}$, at $t = 1$,

$$\delta_1 = \|\sigma(\mathbf{U}\mathbf{x} + \mathbf{b}) - \sigma(\mathbf{U}\mathbf{x}' + \mathbf{b})\|,$$

and the pre-activation difference is $\mathbf{U}(\mathbf{x} - \mathbf{x}')$. Thus the initial class separation is set by the input projection rather than the recurrent gain. For later timesteps, the local Jacobian dynamics along each input-conditioned trajectory are governed by the effective Jacobians

$$J_t = D\sigma(\mathbf{W}\mathbf{z}_t + \mathbf{U}\mathbf{x} + \mathbf{b})\,\mathbf{W}.$$

For complex modReLU, $D\sigma$ denotes the real Jacobian acting on stacked $(\operatorname{Re} z, \operatorname{Im} z)$ coordinates, not a holomorphic complex derivative. In supercritical regimes, products of these Jacobians can amplify perturbations sufficiently that late-time distortion overtakes the input-imprinted class structure. This creates the observed temporal asymmetry: early timesteps are dominated by useful input-driven separation, while later timesteps become increasingly governed by recurrent amplification and nonlinear distortion. The linearisation is local; for the smooth tanh branch, Appendix U.4 makes the finite-separation claim precise by comparing accumulated nonlinear remainder along the Jacobian cocycle with the linear probe margin.

In a linear system ($\sigma = \mathrm{id}$), $\mathbf{z}_{t+1} = \mathbf{W}\mathbf{z}_t + \mathbf{U}\mathbf{x} + \mathbf{b}$, trajectory differences remain linear images of the input difference. Recurrent growth can rescale or rotate features, but it cannot introduce activation-dependent gating or saturation that progressively erodes class-discriminative structure. Empirically, our linear null experiment exhibits no meaningful grace period: early and terminal readouts remain nearly identical across the spectral sweep, indicating that in the tested systems nonlinear distortion, rather than radial growth alone, drives the regime. The pre-collapse regime therefore requires more than spectral growth alone in our taxonomy; it appears with nonlinear state distortion.

Nonlinear activations introduce the critical asymmetry. modReLU is phase-preserving but magnitude-gated, zeroing components when $|z| + b \leq 0$; with $b = -0.5$, this means $|z| \leq 0.5$. This is relatively benign early, while input projection dominates and the gated features remain separable, but under supercritical recurrence the changing gating pattern compounds over time and can erode class separability (Appendix M). The same appendix shows that finite-time operator amplification tracks grace-period shortening and terminal failure. Bounded activations like tanh introduce a milder version: saturation compresses large activations toward ± 1, reducing inter-class variance gradually.

Proposition 4 (Grace-period upper bound)**.** *Let $\bar{r}_t$ be a deterministic reference norm-ratio curve, such as the dataset mean of $\|\mathbf{z}_t^{(i)}\| / \max(\|\mathbf{z}_1^{(i)}\|, \epsilon_{\mathrm{num}})$, with $\epsilon_{\mathrm{num}} > 0$. Fix $\alpha \in (0, 1)$. Assume there exist $t_0 \in \{1, \ldots, T\}$ and $\kappa > 1$ such that*

$$\bar{r}_t \geq \bar{r}_{t_0}\, \kappa^{t - t_0} \qquad \textit{for all } t \geq t_0.$$

Assume also that there exists $R_{\max} > 1$ with $0 < \bar{r}_{t_0} \leq R_{\max}$ such that $\mathrm{acc}(\phi_t) < \alpha$ whenever $\bar{r}_t > R_{\max}$. Then

$$G(\rho(\mathbf{W}), \alpha) \leq t_0 + \left\lceil \frac{\log(R_{\max}/\bar{r}_{t_0})}{\log \kappa} \right\rceil. \tag{3}$$

The proof is in Appendix U.

Remark 5. Proposition 4 is a conditional growth argument, not a spectral theorem. It says that once accuracy fails beyond a norm threshold and the reference norm-ratio curve grows at least geometrically after t_0, the prefix grace period is bounded. The proposition is deliberately stated at the dataset-aggregate level, matching the aggregate probe accuracy in Definition 1, rather than

as a per-sample stopping theorem. We do not identify κ with $\rho(\mathbf{W})$ in the non-normal setting; empirically, larger ρ coincides with faster norm-ratio growth and shorter measured $G(\rho, \alpha)$ under post-$t_0 = 5$ fits (Appendix H). For smooth activations, Appendix U.4 replaces κ with a linearised amplification envelope and the terminal cutoff with a margin-vs.-remainder criterion.

5 Regime-Aware Decoding

The pre-collapse regime does not prescribe one decoder: fixed early-window tests whether any usable prefix remains after terminal collapse, while adaptive stopping and GRACE add norm-ratio timing when the usable-prefix boundary is sample-dependent or noisy. Computationally, adaptive stopping can reduce inference cost by halting before the full horizon. GRACE still evaluates the trajectory to length T, but it can be streamed with memory $O(\dim f(\mathbf{z}))$ per sample, constant in T, by accumulating $\sum_t \tilde{w}_t^{(i)} f(\mathbf{z}_t^{(i)})$ and $\sum_t \tilde{w}_t^{(i)}$, rather than storing all states.

5.1 Adaptive Stopping

A stopping rule based on the norm ratio provides a simple, effective baseline:

$$\mathcal{T}_{\text{stop}}^{(i)} = \left\{ t \in \{1, \dots, T\} : \frac{\|\mathbf{z}_t^{(i)}\|}{\max(\|\mathbf{z}_1^{(i)}\|, \epsilon_{\text{num}})} > \tau_{\text{stop}} \right\} \cup \{T\}, \qquad \hat{t}^{(i)} = \min \mathcal{T}_{\text{stop}}^{(i)}. \tag{4}$$

Here the superscript (i) indexes samples, $\mathcal{T}_{\text{stop}}^{(i)}$ is the set of threshold-crossing timesteps with fallback T, $\hat{t}^{(i)}$ is the selected stopping timestep, τ_{stop} is the hard stopping threshold, and $\epsilon_{\text{num}} = 10^{-10}$ keeps the ratio defined if the first-step norm vanishes. We keep the rule minimal, with no learning or auxiliary state, to isolate feature quality from the stopping policy. We fix $\tau_{\text{stop}} = 1.5$ for the main experiments; Appendix J audits nearby thresholds and a reference-radius selection at $\rho = 1.3$. The rule maintains $\sim$81% accuracy across $\rho(\mathbf{W}) = 0.5$ to 1.6 (Figure 3), recovering $+63$ pp over terminal readout at $\rho(\mathbf{W}) = 1.60$. If the threshold is never crossed, it defaults to T.

5.2 GRACE: Soft Norm-Weighted Trajectory Readout

Rather than selecting a single stopping time, GRACE computes a smooth weighting over the entire trajectory:

$$g_t^{(i)} = \frac{\|\mathbf{z}_t^{(i)}\|}{\max(\|\mathbf{z}_1^{(i)}\|, \epsilon_{\text{num}})}, \tag{5}$$

$$\tilde{w}_t^{(i)} = \exp\big(-\alpha \max(0,\, g_t^{(i)} - \tau)\big), \tag{6}$$

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{s=1}^{T} \tilde{w}_s^{(i)}}, \tag{7}$$

$$\phi_{\text{GRACE}}^{(i)} = \sum_{t=1}^{T} w_t^{(i)}\, f(\mathbf{z}_t^{(i)}), \tag{8}$$

where $g_t^{(i)}$ is the per-sample norm ratio at timestep t, $\tilde{w}_t^{(i)}$ is the unnormalised growth-penalised weight, $w_t^{(i)}$ is its normalised version, τ is a threshold parameter, and $\alpha \geq 0$ controls penalty sharpness.

The interpretation is simple. In stable regimes, where $g_t \lesssim 1$ throughout the rollout, all unnormalised weights equal one and GRACE reduces to the transient mean. In unstable regimes, timesteps whose norm ratio exceeds τ are exponentially down-weighted, so weight shifts toward the usable prefix. Crucially, this weighting is per-sample rather than global: different inputs can receive different temporal profiles.

The following proposition formalises the non-interference property:

Proposition 6 (Threshold non-interference)**.** *For any trajectory $(\mathbf{z}_1, \dots, \mathbf{z}_T)$, if the safeguarded ratios $g_t \leq \tau$ for all $t = 1, \dots, T$, then* GRACE *assigns uniform weights $w_t = 1/T$ for all t, and therefore reduces exactly to transient-mean readout.*

Proof. If $g_t \leq \tau$, then $\max(0, g_t - \tau) = 0$, so each unnormalised weight equals $\exp(0) = 1$. Normalisation yields $w_t = 1/T$. □

In the stable regime of our experiments ($\rho(\mathbf{W}) \leq 1.0$), the observed norm ratios remain close to unity, so GRACE behaves nearly identically to uniform trajectory averaging there.

Figure 3 makes this behaviour visible. The x-axis is timestep and curve colour indicates spectral radius. At $\rho(\mathbf{W}) = 1.0$, weights are close to uniform; at $\rho(\mathbf{W}) = 1.3$, they are front-loaded but still broad; at $\rho(\mathbf{W}) = 1.5$, the first timestep receives about one third of the total weight and the first ten receive roughly 55%. Thus GRACE does not estimate the grace period and then read uniformly across it; the norm ratio constructs a conservative per-sample schedule inside the usable prefix. The profile is also a regime diagnostic, shifting from full-trajectory averaging toward prefix extraction as the substrate enters the pre-collapse band, consistent with the reduction in cross-seed variability in Appendix D.

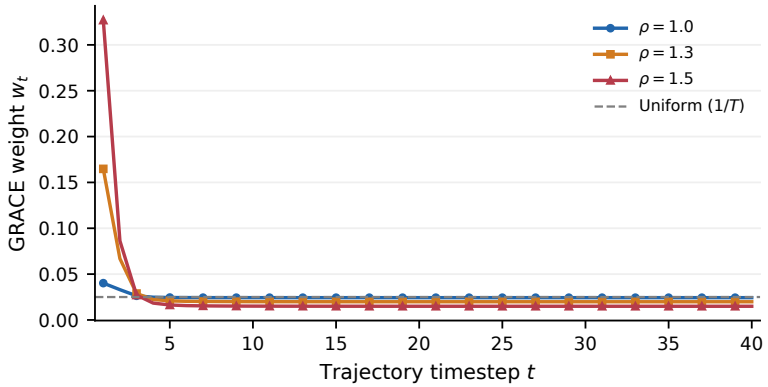


Figure 3: GRACE **as a soft regime-aware prefix selector.** Temporal weight profiles for three spectral radii; curve colour indicates ρ and the x-axis is trajectory timestep. Weights shift from near-uniform ($\rho = 1.0$) to front-loaded (1.3) to prefix-concentrated (1.5; $\tau = 1.2, \alpha = 10$).

In the fixed complex-modReLU sweep, this soft prefix selection is useful. With $\tau = 1.2$ and $\alpha = 10$, GRACE stays in the 81.8–82.1% range across the full spectral sweep, while terminal-state readout falls to 17.9%. Its gain over early-window is modest but consistent, growing from $+0.4$ pp at $\rho(\mathbf{W}) = 1.30$ to $+1.0$ pp at $\rho(\mathbf{W}) = 1.60$.

The more important benefit in this setting is stability. In the transition band ($\rho(\mathbf{W}) = 1.40$), the standard deviation of terminal-state accuracy reaches 0.168, while GRACE reduces this to 0.006, a $27\times$ reduction (Appendix D). In the same fixed complex-modReLU reservoir under injected dynamical noise, GRACE also remains stable while adaptive stopping degrades by 3–4 pp at $\rho(\mathbf{W}) \in \{1.4, 1.5\}$ (Appendix K).

6 Generality Across Systems, Datasets, and Training Regimes

To assess generality, we replicate the core experiments across complex modReLU (unbounded), real-valued tanh ESN (bounded), and linear dynamics ($\sigma = \mathrm{id}$), with additional curves in Appendix L. At $\rho(\mathbf{W}) = 1.6$, the endpoint-vs.-trajectory gap is large in complex modReLU, moderate in the real ESN, and essentially zero in the linear null. Figures 4, 5 and 8 show the same question on CIFAR-10 and trained iterative systems.

The decoder taxonomy is nonlinearity-dependent. In complex modReLU dynamics, GRACE is strongest at $\rho(\mathbf{W}) = 1.6$ ($+0.7$ pp over both early-window and adaptive stopping). In bounded tanh ESNs, adaptive stopping and the best GRACE configuration are essentially tied (84.5% vs. 84.3%). In linear systems, the gap is negligible ($+0.002$), confirming that spectral instability alone does not create a grace period.

6.1 Harder-Benchmark Generality: CIFAR-10

To test whether the regime is an artefact of simple class structure, we repeat the complex modReLU experiment on CIFAR-10 [41] (d=512, T=40, 10 seeds, 5k/5k split). On raw pixels, terminal readout falls to 10.6% at $\rho(\mathbf{W}) = 1.6$ while GRACE holds at $34.2 \pm 0.7\%$ ($+23.7$ pp). On frozen ImageNet ResNet-18 features [4], a direct ridge probe reaches $85.2\%\pm0.1$; at $\rho(\mathbf{W}) = 1.6$, terminal readout falls to $26.2\%\pm2.3$ while early-window and GRACE remain at $84.5\%\pm0.1$ and $84.0\%\pm0.1$. Thus the gap survives translation to an 85%-accuracy substrate. FashionMNIST and frozen-feature details are in Appendices E and F.

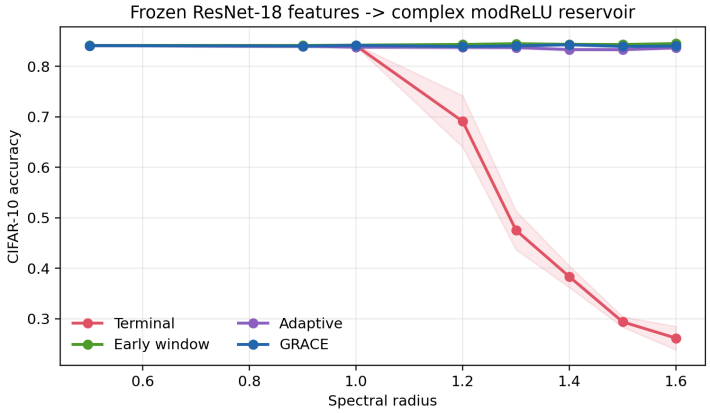


Figure 4: **Pre-collapse regime on CIFAR-10 with frozen ResNet-18 features.** A direct ridge probe on ImageNet-pretrained ResNet-18 features reaches $85.2\% \pm 0.1$ accuracy. Passing those 512-dimensional features through the same complex modReLU reservoir preserves the endpoint-vs.-trajectory gap: at $\rho(\mathbf{W}) = 1.6$, terminal readout falls to $26.2\% \pm 2.3$, while early-window and GRACE remain near 84% (5 seeds; shading $= \pm1$ std).

6.2 Trained Iterative Systems

A natural objection is that fixed reservoirs make endpoint failure too easy, so we test whether the mismatch survives when terminal representations are shaped by optimisation. Our first trained check is a weight-tied LeakyReLU iterative classifier on flattened MNIST, with recurrence

$$\mathbf{z}_{t+1} = \text{LeakyReLU}\big(\lambda \mathbf{W}\mathbf{z}_t + \mathbf{U}\mathbf{x} + \mathbf{b}\big), \qquad \mathbf{z}_0 = \mathbf{0}, \tag{9}$$

where $\mathbf{W} \in \mathbb{R}^{256\times256}$, $\mathbf{U} \in \mathbb{R}^{256\times784}$, and the recurrent weights, input projection, bias, and linear head are jointly trained from terminal-state cross-entropy. Across 5 seeds, we compare $T_{\text{train}} = 10, \lambda_{\text{train}} = 0.85$ with two at-operating-point conditions ($T_{\text{train}} = 40, \lambda_{\text{train}} \in \{1.20, 1.30\}$); all reach $\geq 96\%$ training accuracy. Details are in Appendix A.

Figure 5 summarises the main results; full statistics are in Appendix B. In A, final accuracy drops from 92.0% to 79.4% by $\lambda = 2.0$, while early-window remains near 95.2% ($+15.8$ pp). In B1/B2, terminal readout is locally optimal near the training gain but collapses to 69.4% and 77.3% at $\lambda = 2.0$, while early-window remains 95.9% ($+26.5$ and $+18.6$ pp). Training delays but does not suppress the regime. A standard row-sequential CIFAR-10 LSTM shows the same smaller bounded-gate pattern: early-window is $+3.5$ pp over terminal at $\lambda = 2.0$, and a single-timestep probe audit gives accuracy-defined prefix lengths $40, 27, 2$ at $\lambda \in \{1.0, 1.6, 2.0\}$ (Appendix O).

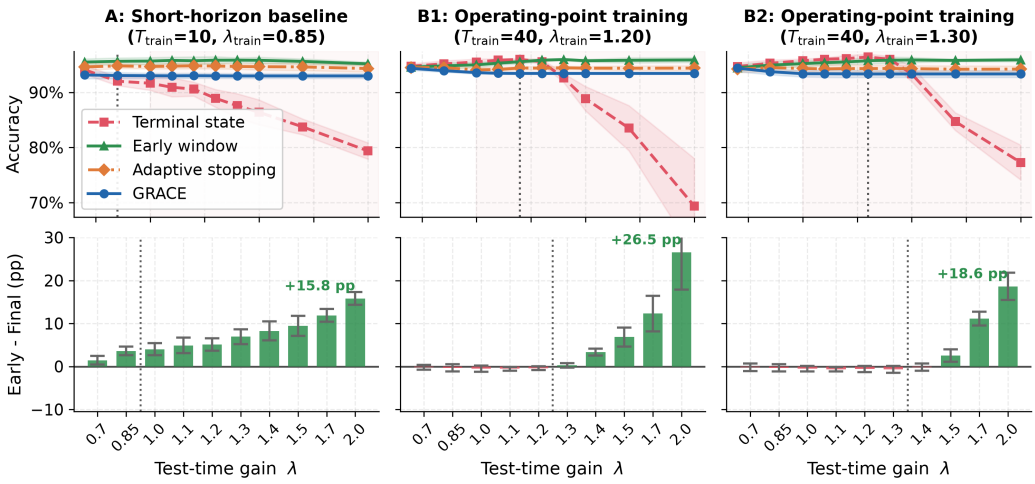


Figure 5: **Training at the operating point postpones but does not suppress the pre-collapse regime.** Columns are training conditions; top row shows accuracy vs. test-time gain λ, and bottom row shows early-window minus terminal accuracy. Peak gains are $+15.8$, $+26.5$, and $+18.6$ pp.

6.3 Modern Recursive Models on Tiny ImageNet

We test horizon extension on Tiny ImageNet-200 (100k training/10k labelled validation images), using frozen ConvNeXt-Tiny features [43] and weight-tied SwiGLU and shared-attention heads. Both use terminal cross-entropy, $T_{\text{train}} \in \{8, 32\}$, and paired seeds 101–103. Evaluation uses $T_{\text{test}} \in \{8, 16, 24, 32\}$ at gain 1, with residual scale $1/T_{\text{train}}$ fixed. Matched controls change training depth and scale together; architecture details are in Appendix P.

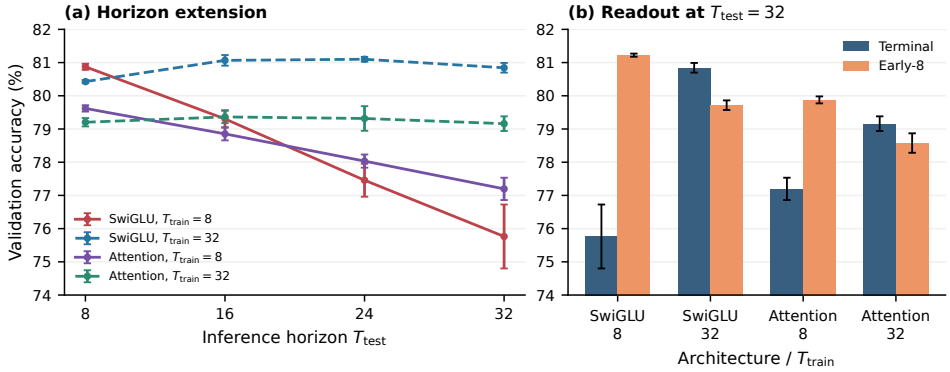


Figure 6: **Horizon extension and early-prefix readout on Tiny ImageNet.** *(a)* Terminal accuracy over inference horizons for eight- and 32-step-trained SwiGLU and attention heads. *(b)* Terminal versus early-eight readout at $T_{\text{test}} = 32$. Early readout preserves accuracy under horizon mismatch, while matched training restores a competitive endpoint. Means and sample standard deviations across three paired seeds; frozen ConvNeXt-Tiny backbone.

In Figure 6, extending the eight-step-trained SwiGLU model to 32 steps lowers terminal accuracy from $80.87 \pm 0.09\%$ to $75.76 \pm 0.96\%$, a 5.11 ± 0.94 pp loss; early-eight readout retains $81.22 \pm 0.05\%$. Shared attention reproduces the pattern: terminal accuracy falls from $79.62 \pm 0.10\%$ to $77.20 \pm 0.34\%$ (2.42 ± 0.29 pp), while early-eight readout remains at $79.88 \pm 0.11\%$. Fixed-regularisation probes also lose 2.23 ± 0.17 and 1.28 ± 0.08 pp, respectively, indicating reduced linear decodability beyond shared-head mismatch.

Matched 32-step training removes the strong endpoint loss: SwiGLU reaches $80.84 \pm 0.15\%$ at step 32, and attention loses only 0.04 ± 0.17 pp from steps 8 to 32. The paired difference in terminal loss between training conditions is 5.53 ± 0.92 pp for SwiGLU and 2.38 ± 0.43 pp for attention, positive

in every seed; all states and logits remain finite. These statistics are descriptive, without significance claims (complete readouts and robustness analyses in Appendices Q and R).

The SwiGLU effect depends on checkpoint age: retained final-epoch checkpoints lose only 0.59–0.68 pp, compared with 4.27–6.13 pp at the selected checkpoints. Attention degradation persists across its retained checkpoint audit (Appendix R).

6.4 Inference Time and Memory

We benchmark six readouts on a separate seed-0 SwiGLU model with cached ConvNeXt features (Table 2). Early-eight readout takes 1.80 ms versus terminal's 6.73 ms by executing fewer steps. Streaming GRACE reduces additional peak allocation from 48.56 to 8.51 MiB but is slower; adaptive stopping incurs active-set overhead. Lower storage does not guarantee lower latency. Appendix S details the protocol and complexity.

Table 2: **Measured recurrent-head inference cost.** RTX 5080 Laptop GPU, BF16, batch 512, $d = 512$, $T = 32$, gain 1.3; median of 25 repetitions after 10 warm-ups. Memory is additional peak allocated memory above the warmed-up baseline. Image decoding and backbone inference are excluded.

Readout	Latency (ms)	Additional peak (MiB)
Terminal	6.727	8.00
Early window ($K = 8$)	1.803	11.50
Adaptive stopping	13.840	9.01
Uniform streaming	6.889	8.50
GRACE, stored trajectory	6.948	48.56
GRACE, streaming	9.681	8.51

7 Discussion

Decoder effectiveness depends on the dynamics: early windows are strongest in the trained LeakyReLU and LSTM checks, while GRACE is useful under unbounded growth and injected noise (Appendix K). Matched training restores modern recursive endpoints. Early readout is therefore a diagnostic and possible remedy, with benefits conditional on the dynamics and training horizon.

Physical substrates face similar timing questions because device variability, temperature, and wear can shift operating points [23–25]. Norm ratios are measurable without knowing $\rho(\mathbf{W})$ [22, 26–28]. Practically, early readout saves recurrent steps, whereas streaming GRACE saves storage without guaranteeing lower latency (Section 6.4).

Related questions arise in continuous-depth graph models [44], a natural setting for studying readout validity over integration time. Graph diffusion connects depth to smoothing of node features [45], suggesting the analogous question of whether intermediate states retain task information lost at the endpoint. Diffusive homogenisation differs from the amplification-driven distortion studied here; neither our norm-growth proxy nor our nonlinear taxonomy transfers automatically. Trajectory readouts for graph ODEs remain an open empirical direction.

Limitations include controlled gain and horizon shifts, frozen vision backbones, and three seeds per modern condition. Tiny ImageNet validation also selects checkpoints. These laptop-GPU experiments establish conditional endpoint failure. Assessing its prevalence in large deployed models remains future work, alongside extensions to graph ODEs and state-space models and tighter bounds.

8 Conclusion

Usable prefixes can persist after terminal collapse. Gain-shift experiments and horizon-extension tests in modern recursive heads expose endpoint-validity failures; matched training restores endpoint validity in the latter setting. Readout timing and the training horizon are therefore part of the algorithm.

Acknowledgments and Disclosure of Funding

I thank Oneirix LLC for providing compute resources in support of this work. **Funding in direct support of this work:** Oneirix LLC provided computing resources. **Competing interests:** The author is the founder of Oneirix LLC.

References

[1] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report GMD Report 148, German National Research Center for Information Technology, 2001.

[2] S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. In *NeurIPS*, 2019.

[3] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[5] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal. Deep complex networks. In *ICLR*, 2018.

[6] M. Arjovsky, A. Shah, and Y. Bengio. Unitary evolution recurrent neural networks. In *ICML*, 2016.

[7] J. Ginibre. Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics*, 6(3):440–449, 1965.

[8] T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of ESDs and the circular law. *Annals of Probability*, 38(5):2023–2065, 2010.

[9] M. Lukoševičius. A practical guide to applying echo state networks. In *Neural Networks: Tricks of the Trade*, pages 659–686. Springer, 2012.

[10] A. Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[11] A. Banino, J. Balaguer, and C. Blundell. PonderNet: Learning to ponder. *arXiv preprint arXiv:2107.05407*, 2021.

[12] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *ICLR*, 2021.

[13] N. Bertschinger and T. Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004.

[14] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. Information processing capacity of dynamical systems. *Scientific Reports*, 2:514, 2012.

[15] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *NeurIPS*, 2016.

[16] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. In *ICLR*, 2017.

[17] I. B. Yildiz, H. Jaeger, and S. J. Kiebel. Re-visiting the echo state property. *Neural Networks*, 35:1–9, 2012.

[18] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

[19] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *ICML*, pages 1310–1318, 2013.

[20] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[21] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

[22] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose. Recent advances in physical reservoir computing: A review. *Neural Networks*, 115:100–123, 2019.

[23] M. Le Gallo, R. Khaddam-Aljameh, M. Stanisavljevic, et al. A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics*, 6:680–693, 2023.

[24] W. Wan, R. Kubendran, C. Schaefer, et al. A compute-in-memory chip based on resistive random-access memory. *Nature*, 608:504–512, 2022.

[25] M. J. Rasch, F. Carta, O. Fagbohungbe, et al. Fast and robust analog in-memory deep neural network training. *Nature Communications*, 15:7133, 2024.

[26] A. Momeni, B. Rahmani, B. Scellier, et al. Training of physical neural networks. *Nature*, 645:53–61, 2025.

[27] C. C. Wanjura and F. Marquardt. Fully non-linear neuromorphic computing with linear wave scattering. In *Proceedings of SPIE*, volume 13375, page 133750H, 2025.

[28] H. Wang, J. Hu, A. Morandi, et al. Large-scale photonic computing with nonlinear disordered media. *Nature Computational Science*, 4:429–439, 2024.

[29] C. Gallicchio and A. Micheli. Echo state property of deep reservoir computing networks. *Cognitive Computation*, 9(3):337–350, 2017.

[30] G. Hennequin, T. P. Vogels, and W. Gerstner. Non-normal amplification in random balanced neuronal networks. *Physical Review E*, 86(1):011909, 2012.

[31] S. Teerapittayanon, B. McDanel, and H. T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *ICPR*, 2016.

[32] Y. Kaya, S. Hong, and T. Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *ICML*, 2019.

[33] G. Kerg, K. Goyette, M. Puelma Touzel, G. Gidel, E. Vorontsov, Y. Bengio, and G. Lajoie. Non-normal recurrent neural network (nnRNN): learning long time dependencies while improving expressivity with transient dynamics. In *NeurIPS*, 2019.

[34] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *ICLR*, 2025.

[35] J. Geiping, S. M. McLeish, N. Jain, J. Kirchenbauer, S. Singh, B. R. Bartoldson, B. Kailkhura, A. Bhatele, and T. Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *NeurIPS*, spotlight, 2025.

[36] M. Elhoushi, A. Shrivastava, D. Liskovich, B. Hosmer, B. Wasti, L. Lai, A. Mahmoud, B. Acun, S. Agarwal, A. Roman, A. Aly, B. Chen, and C.-J. Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In *ACL*, 2024.

[37] M. Gabor, T. Piotrowski, and R. L. G. Cavalcante. Positive concave deep equilibrium models. In *ICML*, 2024.

[38] J. Lin, Z. Ling, J. Xu, and R. C. Qiu. Consistency deep equilibrium models. *arXiv preprint arXiv:2602.03024*, 2026.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[41] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[43] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *CVPR*, 2022.

[44] M. Poli, S. Massaroli, J. Park, A. Yamashita, H. Asama, and J. Park. Graph neural ordinary differential equations. *arXiv preprint arXiv:1911.07532*, 2019.

[45] B. Chamberlain, J. Rowbottom, M. I. Gorinova, M. Bronstein, S. Webb, and E. Rossi. GRAND: Graph neural diffusion. In *ICML*, pages 1407–1418, 2021.

A Additional Experimental Details

System parameters. All complex modReLU experiments use $d = 128$, $n = 784$, modReLU bias $b = -0.5$, rollout $T = 40$, magnitude clipping at $|\mathbf{z}| > 50$, and no injected noise ($\sigma_{\text{noise}} = 0$) unless otherwise stated. The clipping threshold is a numerical safeguard; our claim is that non-terminal readouts retain separability before terminal degradation, not that clipping is the mechanism of terminal collapse. The input projection $\mathbf{U}$ is drawn i.i.d. complex Gaussian and normalised. The recurrent weight $\mathbf{W}$ is drawn from the complex Ginibre ensemble and rescaled to the target spectral radius. Real ESN uses the same architecture with $\sigma = \tanh$, unit leak rate, and real-valued weights of matching dimension. MNIST, FashionMNIST, and the cross-system sweep use $d = 128$ reservoirs. These are robustness checks rather than exhaustive capacity studies; the real-ESN comparison is a qualitative decoder-taxonomy check rather than a matched real-valued capacity benchmark.

A dedicated clipping audit on the same 10 seeds and 5k held-out MNIST samples confirms that the usable early prefix is not created by the hard threshold: no state entry is clipped in the first 10 timesteps at any spectral radius. Across the full 40-step rollout, the mean clipped-entry fraction is 0.25% at $\rho = 1.40$, 1.33% at $\rho = 1.50$, and 4.06% at $\rho = 1.60$; the earliest clipping event across all runs occurs at $t = 12$. Thus the late terminal state can encounter the numerical safeguard at the largest radii, but the early-window evidence and the prefix selected by GRACE lie before clipping is active.

The stronger CIFAR-10 feature experiment in Appendix F uses frozen ImageNet-pretrained ResNet-18 penultimate features (512 dimensions), a stratified 10k subset of the CIFAR-10 training split, and the full 10k test split. The downstream complex modReLU reservoir uses $d = 512$, $T = 40$, 5 seeds, and exact eigenvalue spectral-radius rescaling.

Figure 7 visualises the random-matrix substrate used in the complex modReLU sweep. The unscaled complex Ginibre ensemble has the standard circular-law spectral geometry in high dimension [7, 8], and our experimental sweep rescales this cloud radially to the requested spectral radius. This diagnostic is only a substrate check: the pre-collapse regime does not follow from the circular law alone, but from radial scaling interacting with finite-time amplification, input imprinting, and nonlinear modReLU distortion.

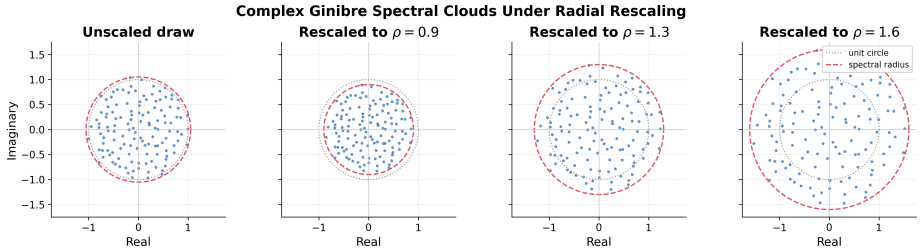


Figure 7: **Complex Ginibre spectral clouds before and after spectral-radius rescaling.** Each panel shows one $d = 128$ complex Ginibre draw rescaled to the target spectral radius. The figure documents the canonical circular-law substrate used in the spectral sweep; it is not a mechanism claim for the pre-collapse regime by itself.

Probe. All aggregate readout results use `RidgeClassifierCV` from scikit-learn with leave-one-out cross-validation for the regularisation parameter [42]. This provides a closed-form solution with no iterative optimisation, ensuring that all accuracy differences reflect the quality of the dynamical features rather than probe optimisation artefacts. The dense single-timestep LSTM audit in Appendix O uses fixed ridge regularisation $\alpha_{\text{ridge}} = 1.0$ for computational tractability.

Trained iterative systems. The LeakyReLU trained-system experiments use flattened MNIST with 10k training examples and 2k test examples. The model has hidden dimension 256, LeakyReLU negative slope 0.2, and state clipping to $[-100, 100]$. It is trained end-to-end from terminal-state cross-entropy with Adam (learning rate 10^{-3}, weight decay 10^{-4}), cosine annealing, batch size 256, and gradient-norm clipping (1.0 in condition A, 0.5 in B1/B2). Condition A trains for 40 epochs at $T_{\text{train}} = 10$, $\lambda_{\text{train}} = 0.85$; conditions B1/B2 train for 60 epochs at $T_{\text{train}} = 40$ with $\lambda_{\text{train}} \in \{1.20, 1.30\}$. Evaluation uses $T_{\text{test}} = 40$ over $\lambda_{\text{test}} \in \{0.7, 0.85, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 2.0\}$ and 5 random seeds. After training, readout accuracies are measured with the same RidgeClassifierCV protocol on extracted readout features, fitting on the first 1k held-out test examples and scoring on the remaining 1k; this keeps final, early-window, adaptive, and GRACE readouts under a common probe.

The row-sequential CIFAR-10 LSTM check in Appendix O uses 10k/2k CIFAR-10, a 512-hidden-unit LSTM trained end-to-end with AdamW for 30 epochs, and readouts from a 40-step zero-input post-image settling trajectory; the appendix gives the full training and timestep-audit protocol.

Readout hyperparameters. All reservoir experiments use early-window $K = 10$; trained-system experiments use $K = 8$. Adaptive stopping uses the fixed default $\tau_{\text{stop}} = 1.5$ in the main paper.

Datasets and usage terms. MNIST is used via torchvision's standard benchmark distribution derived from the original NIST release [39]. FashionMNIST is distributed by Zalando under the MIT License [40]. CIFAR-10 is used from the Alex Krizhevsky / University of Toronto benchmark distribution, which requests citation of the original technical report [41]. The raw-pixel, frozen-feature, and row-sequential LSTM CIFAR-10 checks all use standard train/test splits; the frozen-feature CIFAR-10 check uses torchvision's ImageNet-pretrained ResNet-18 weights as a fixed feature extractor.

GRACE **hyperparameters.** The primary configuration uses $\tau = 1.2$, $\alpha = 10$. An anchor configuration ($\tau = 1.5$, $\alpha = 5$) is reported for robustness. Appendix J gives a lightweight reference-radius audit at $\rho = 1.3$: it selects a slightly more aggressive frozen alternative, but the qualitative conclusions and method ordering remain unchanged.

Compute resources. The original reservoir, CIFAR-10, and trained LeakyReLU/LSTM experiments were run on a single local CUDA-capable GPU (NVIDIA RTX PRO 1000 Blackwell Generation Laptop GPU, 8 GB VRAM). The core 10-seed MNIST reservoir sweep took 10,737 s ($\approx$ 3.0 h), the dynamical-noise study 2,720 s ($\approx$ 45 min), the accuracy-defined grace-period appendix analysis 360 s ($\approx$ 6 min), and the readout-audit appendix sweep 13,952 s ($\approx$ 3.9 h), as recorded in the saved result files. Additional reported experiments (FashionMNIST, CIFAR-10,

trained LeakyReLU, cross-system sweeps, the DEQ relevance check, and the sequential CIFAR-10 LSTM check) were run on the same single-GPU setup; the LSTM check and per-timestep audit were hour-scale runs (the saved audit run recorded $3{,}229$ s for 3 seeds). The compute for these original experiments remains comfortably below one day of GPU time overall. Pilot and discarded exploratory runs required additional compute of comparable order, but not orders of magnitude more than the reported totals. The frozen ResNet-18 CIFAR-10 feature extraction took approximately 3 min including the one-time ImageNet-weight download, and the cached 5-seed spectral sweep took 243 s (≈ 4.1 min). The small compute footprint is useful for reproducibility, but it is also a limitation: the present experiments are stress tests of the readout regime, not evidence that the same quantitative gaps will persist unchanged in much larger, production-scale models.

The additional Tiny ImageNet experiments used an RTX 5080 Laptop GPU; their configurations, runtime, and memory measurements are reported in Appendices P and S.

B Trained-System Readout Statistics

Table 3 gives full per-condition readout accuracy for the trained LeakyReLU classifier across all three training conditions and selected test-time gains. Bold entries mark the gain at which final-state readout is locally optimal (i.e. at or just above the training gain). Training uses flattened MNIST (10k/2k), hidden dimension 256, terminal-state cross-entropy, Adam with learning rate 10^{-3} and weight decay 10^{-4}, cosine annealing, batch size 256, 40 epochs for condition A and 60 for B1/B2, and the condition-specific T_{train} and λ_{train} values listed in the table. The full gain sweep is $\lambda_{\text{test}} \in \{0.7, 0.85, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 2.0\}$; as in the script, reported readout accuracies use RidgeClassifierCV probes fit on the first half of the 2k held-out subset and scored on the second half.

To verify that the GRACE underperformance is structural rather than a tuning artefact, we swept $\tau \in \{0.5, 0.8, 1.0, 1.2, 1.5, 2.0, 3.0\}$ and $\alpha \in \{2, 5, 10, 20, 50\}$ using the per-timestep probe accuracies and norm trajectories stored from the trained-system runs. The best configuration ($\tau = 3.0$, $\alpha = 2$, which approaches uniform weighting) recovers approximately 1.0 pp over the default ($\tau = 1.2$, $\alpha = 10$) but remains 1.0–2.0 pp below early-window in conditions A and B1, and is essentially tied in condition B2 near the training gain. The gap therefore does not close under re-tuning. The mechanistic reason is that the norm ratio in the trained system is a less informative readout-timing signal: the early trajectory contains a representation-building phase in which norms increase without implying discriminability loss, while the learned head has absorbed expected state magnitudes near the training operating point. The early prefix therefore remains closer to the model's in-distribution representation manifold regardless of how τ and α are set.

Table 3: Trained LeakyReLU classifier: readout accuracy (mean, 5 seeds) at selected test-time gains λ_{test}. Bold marks the locally optimal gain for final-state readout. GRACE consistently underperforms early-window by $\sim$2.5 pp across all conditions; see Section 7 for explanation.

λ_{test}	Final	Early (K=8)	GRACE	Adaptive
Cond. A: trained T=10, λ_{train}=0.85				
0.85	0.920	0.956	0.931	0.948
1.20	0.906	0.957	0.930	0.948
1.50	0.864	0.958	0.930	0.947
2.00	0.794	0.952	0.930	0.943
Cond. B1: trained T=40, λ_{train}=1.20				
1.20	**0.960**	0.956	0.934	0.944
1.30	0.956	0.958	0.934	0.944
1.50	0.889	0.957	0.934	0.944
2.00	0.694	0.959	0.934	0.944
Cond. B2: trained T=40, λ_{train}=1.30				
1.30	**0.964**	0.957	0.933	0.943
1.50	0.933	0.959	0.933	0.943
2.00	0.773	0.959	0.933	0.942

C Full Statistical Tables

Full Wilcoxon signed-rank audit. The main text treats the four regime radii satisfying Definition 2 as the primary inferential family. Here we expand to a 27-test secondary audit over all radii and readout comparisons. All tests are one-sided (GRACE $>$ competitor). The reported p values are raw exact Wilcoxon signed-rank values; significance markers indicate which comparisons survive Holm-Bonferroni correction across this 27-test secondary family (9 radii $\times$ 3 comparisons). W is the test statistic, and the 'sig.' columns report Holm-Bonferroni significance markers.

Table 4: Wilcoxon signed-rank test: GRACE vs. each readout strategy across spectral radii (10 seeds per condition, one-sided H_1: GRACE $\succ$ competitor).

	vs. Final-state			vs. Early-window			vs. Adaptive		
$\rho(\mathbf{W})$	W	p	HB sig.	W	p	HB sig.	W	p	HB sig.
0.50	15	0.664		4	0.988		15	0.664	
1.00	53	0.003	**	18	0.819		53	0.003	**
1.20	55	0.001	***	30	0.055	(†)	55	0.001	***
1.30	55	0.001	***	44	0.004	**	54	0.002	**
1.35	55	0.001	***	52	0.005	**	55	0.001	***
1.40	55	0.001	***	43	0.006	**	55	0.001	***
1.45	55	0.001	***	44	0.004	**	55	0.001	***
1.50	55	0.001	***	51	0.007	**	55	0.001	***
1.60	55	0.001	***	55	0.001	***	55	0.001	***

'HB sig.' denotes Holm-Bonferroni significance markers: *** $p < 0.01$; ** $p < 0.02$; (†) marginal ($p < 0.06$). All marked results survive Holm-Bonferroni correction at familywise $\alpha = 0.05$.

D Cross-Seed Variability Reduction

Table 5 reports the standard deviation of accuracy across 10 seeds for final-state and GRACE readout at each spectral radius. GRACE reduces cross-seed standard deviation by up to $27\times$ in the transition band.

Table 5: Cross-seed variability: standard deviation of accuracy across 10 seeds. GRACE reduces cross-seed standard deviation by up to $27\times$ in the transition band. Ratios use unrounded standard deviations.

$\rho(\mathbf{W})$	std(Final)	std(GRACE)	Ratio
1.00	0.006	0.006	1.1×
1.30	0.091	0.007	13.7×
1.35	0.148	0.006	24.2×
1.40	0.168	0.006	**27.3**×
1.45	0.163	0.006	27.2×
1.50	0.151	0.006	24.1×
1.60	0.037	0.007	5.1×

E FashionMNIST Results

We replicate the final, early-window, GRACE, and adaptive-stopping readout sweep on FashionMNIST [40] (5,000 train /5,000 test, 3 seeds) using the same complex modReLU reservoir dimension as the main sweep ($d = 128$, $\rho(\mathbf{W})$ swept 0.5–1.6), changing the dataset while preserving substrate scale. GRACE uses the fixed main-paper configuration ($\tau = 1.2, \alpha = 10$). The pre-collapse regime emerges at the same spectral radius as in MNIST, confirming that the phenomenon is not dataset-specific.

Table 6: FashionMNIST results (mean ± std, 3 seeds). The pre-collapse regime is visible from $\rho(\mathbf{W}) \geq 1.30$, as in MNIST.

$\rho(\mathbf{W})$	Final	Early-window	GRACE	Adaptive
0.50	0.742±0.005	0.742±0.005	0.742±0.005	0.742±0.005
1.00	0.743±0.002	0.744±0.003	0.743±0.003	0.743±0.002
1.20	0.740±0.009	0.741±0.005	0.742±0.003	0.740±0.009
1.30	0.702±0.022	0.738±0.007	0.743±0.002	0.735±0.010
1.35	0.627±0.098	0.739±0.005	0.743±0.003	0.737±0.008
1.40	0.506±0.124	0.738±0.005	0.743±0.005	0.738±0.008
1.45	0.373±0.107	0.736±0.004	0.743±0.005	0.739±0.009
1.50	0.266±0.073	0.736±0.005	0.744±0.005	0.740±0.009
1.60	0.314±0.007	0.734±0.007	0.746±0.005	0.743±0.005

F Additional CIFAR-10 Results

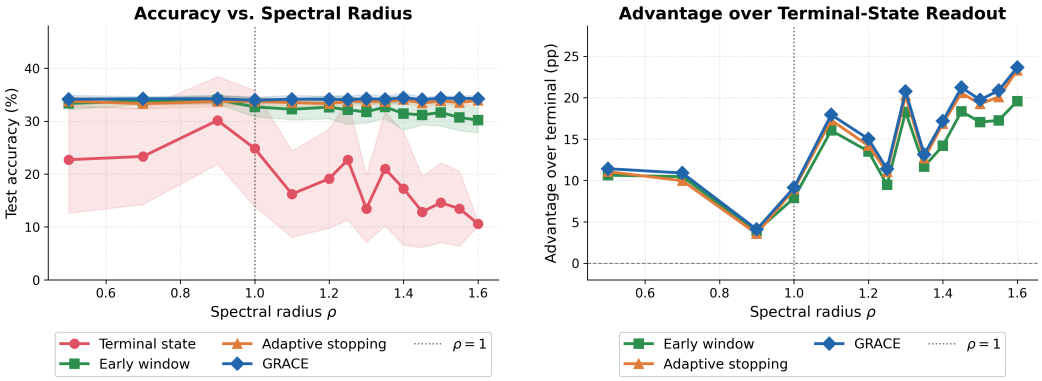


Figure 8: **Pre-collapse regime on CIFAR-10.** Complex modReLU reservoir ($d = 512$, $T = 40$, 10 seeds, 5k/5k split). *Left:* accuracy vs. spectral radius $\rho(\mathbf{W})$; shading shows ±1 std across seeds. *Right:* advantage of each non-terminal readout over the terminal state (pp). Terminal-state accuracy collapses to near chance at large $\rho(\mathbf{W})$ while GRACE and early-window remain 20–24 pp higher.

To separate the pre-collapse effect from the low absolute accuracy of the raw-pixel CIFAR-10 reservoir, we repeat the CIFAR experiment using a stronger frozen feature substrate. CIFAR-10 images are embedded with an ImageNet-pretrained ResNet-18 penultimate layer [4]; a direct ridge probe on these features achieves $85.2\% \pm 0.1$ accuracy over 5 stratified 10k-train subsets and the full 10k test set. We then feed the 512-dimensional frozen features into the same complex modReLU reservoir ($d = 512$, $T = 40$), using exact eigenvalue spectral-radius rescaling.

The stable regime preserves the strong representation: at $\rho(\mathbf{W}) \leq 1.0$, all readouts are near 84%. As $\rho(\mathbf{W})$ increases, terminal readout collapses while trajectory readouts remain near the frozen-feature baseline. At $\rho(\mathbf{W}) = 1.6$, terminal readout is $26.2\% \pm 2.3$, while early-window is $84.5\% \pm 0.1$ and GRACE is $84.0\% \pm 0.1$. The corresponding accuracy curves are shown in the main text in Figure 4; this appendix records the numerical setup and endpoint-vs.-trajectory values.

G ESN GRACE Hyperparameter Sensitivity

We evaluate four GRACE hyperparameter configurations on the real-valued tanh ESN (real ESN, $d = 128$) across spectral radii, using the cross-system results data (5 seeds per condition). The four configurations are defined by (τ, α): `t1.2_a5` ($\tau = 1.2$, $\alpha = 5$), `t1.2_a10` ($\tau = 1.2$, $\alpha = 10$), `t1.5_a5` ($\tau = 1.5$, $\alpha = 5$), and `t1.5_a10` ($\tau = 1.5$, $\alpha = 10$).

Table 7: GRACE hyperparameter sensitivity on real-valued tanh ESN (mean accuracy, 5 seeds). All four configurations substantially outperform final-state readout across the unstable regime. The gains are robust to hyperparameter choice within $\tau \in [1.2, 1.5]$, $\alpha \in [5, 10]$.

$\rho(\mathbf{W})$	Final	τ=1.2, α=5	τ=1.2, α=10	τ=1.5, α=5	τ=1.5, α=10
0.50	0.844	0.843	0.843	0.843	0.843
1.00	0.841	0.843	0.844	0.842	0.842
1.20	0.835	0.843	0.841	0.842	0.842
1.30	0.817	0.843	0.842	0.841	0.841
1.35	0.803	0.843	0.842	0.841	0.841
1.40	0.786	0.843	0.843	0.840	0.840
1.45	0.768	0.843	0.843	0.841	0.840
1.50	0.751	0.842	0.843	0.840	0.840
1.60	0.711	0.843	0.843	0.840	0.840

Key observation: all four configurations converge to similar performance, confirming that the qualitative conclusions are insensitive to the precise hyperparameter choice. The $\tau = 1.2$ configurations are marginally stronger at lower radii; the difference disappears in the deeply unstable regime ($\rho(\mathbf{W}) \geq 1.50$) where both τ values correctly identify the full usable prefix.

H Grace Period Sweep

Figure 9 distinguishes the formal quantity from the practical proxy. The left panel plots the accuracy-defined grace period $G(\rho, \alpha)$ for $\alpha \in \{0.65, 0.70, 0.75\}$; the right panel shows the norm-ratio stopping time used as an online proxy. In these sweeps, both shorten monotonically with spectral radius, but only the left panel is the formal definition from Definition 1. For $\alpha = 0.70$, the empirical grace period remains at the full horizon through $\rho \leq 1.20$, then falls to 31.1 at $\rho = 1.40$, 20.2 at $\rho = 1.50$, and 15.2 at $\rho = 1.60$. Using the mean norm-ratio proxy with $t_0 = 5$, a fitted post-t_0 growth factor rises approximately linearly from 1.013 ± 0.024 at $\rho = 1.30$ to 1.126 ± 0.012 at $\rho = 1.60$ (intermediate means: 1.056 at $\rho = 1.40$ and 1.100 at $\rho = 1.50$).

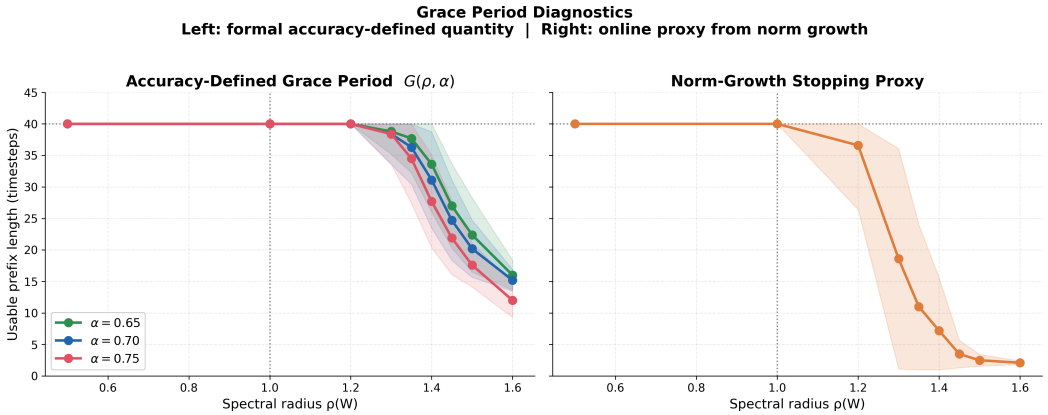


Figure 9: **Formal grace period and norm-ratio proxy both shorten with instability.** *Left:* the true accuracy-defined grace period $G(\rho, \alpha)$ for $\alpha \in \{0.65, 0.70, 0.75\}$, computed from per-timestep probe curves. *Right:* the norm-ratio stopping time used as a practical online proxy. The proxy tracks the same qualitative shortening, but it is not interchangeable with the formal quantity. Shading shows ± 1 std across 10 seeds.

I Early-Window Sensitivity

We fix $K = 10$ in the main reservoir experiments. Figure 10 shows that the main conclusion is not tied to this specific choice: smaller windows are modestly stronger deep in the unstable regime,

but $K = 10$ remains far above terminal-state readout and therefore serves as a conservative, non-optimistic baseline rather than a cherry-picked optimum. At $\rho = 1.60$, for example, $K = 1$ gives 82.2%, $K = 10$ gives 80.9%, and $K = 20$ falls to 75.9%.

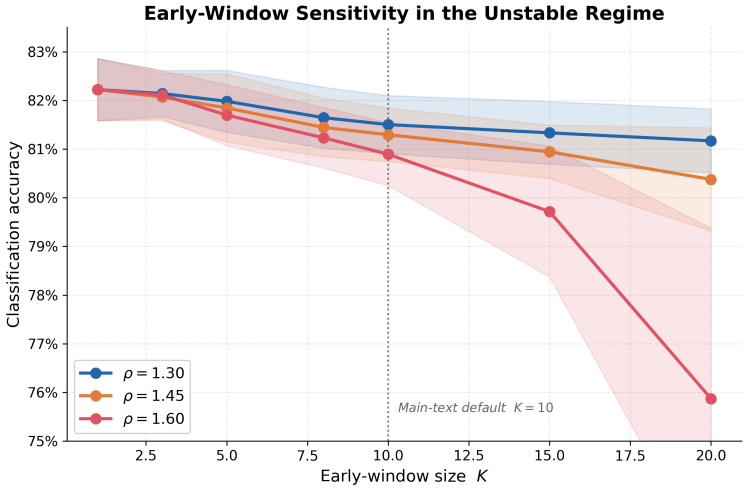


Figure 10: **Early-window sensitivity over prefix length K.** Accuracy is shown for representative unstable radii $\rho \in \{1.30, 1.45, 1.60\}$ over $K \in \{1, 3, 5, 8, 10, 15, 20\}$. The fixed default $K = 10$ is conservative: it is not optimal in the most unstable settings, but it still cleanly exhibits the pre-collapse regime.

J Reference-Radius Selection Audit

The main text uses fixed, simple defaults chosen before the final audit ($\tau_{\text{stop}} = 1.5$ and $(\tau, \alpha) = (1.2, 10)$), rather than settings selected by sweeping the full spectral range. To check that the conclusion is not an artefact of those defaults, we ran a lightweight post-hoc selection audit at a single reference radius, $\rho = 1.30$. The audit selected $\tau_{\text{stop}} = 1.0$ for adaptive stopping and $(\tau, \alpha) = (1.0, 15)$ for GRACE. These settings are slightly more aggressive, but freezing them across the unstable band leaves the qualitative ranking unchanged: both adaptive stopping and GRACE remain near-flat and well above terminal-state readout.

Table 8: Frozen performance under the reference-radius selection audit. Parameters are selected at $\rho = 1.30$ and then held fixed elsewhere.

ρ	Early (K=10)	Adaptive (τ_{stop}=1.0)	GRACE (τ=1.0, α=15)
1.30	0.815	0.822	0.821
1.40	0.813	0.822	0.820
1.50	0.812	0.822	0.820
1.60	0.809	0.822	0.820

K Noise Robustness Details

Figure 11 shows per-seed accuracy in the fixed complex-modReLU reservoir as a function of additive complex Gaussian dynamical noise at two representative spectral radii ($\rho = 1.4$ and $\rho = 1.5$). At each timestep, before the modReLU activation, we add

$$\xi_t \in \mathbb{C}^d, \qquad \operatorname{Re}\xi_{t,j}, \operatorname{Im}\xi_{t,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}}),$$

using the same noise level during train- and test-feature extraction for the linear probes.

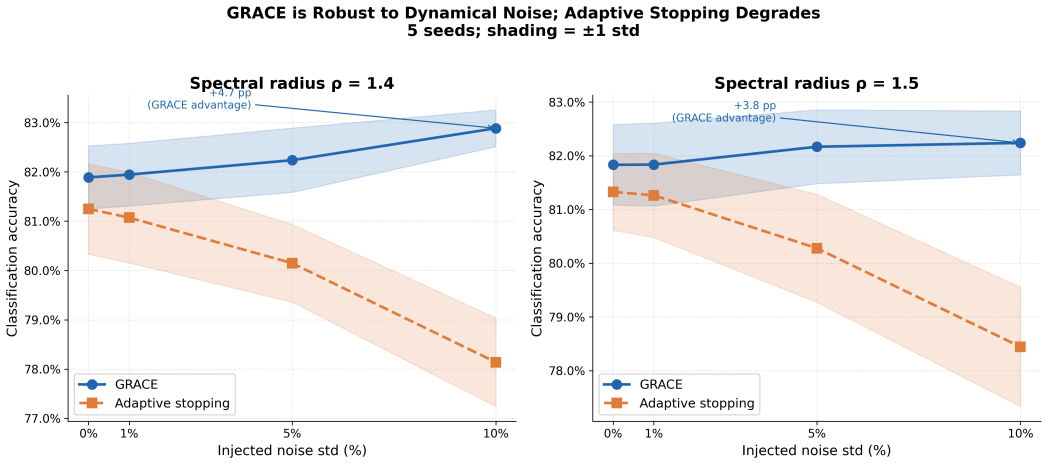


Figure 11: GRACE **is robust to dynamical noise; adaptive stopping degrades.** At both $\rho = 1.4$ (left) and $\rho = 1.5$ (right), GRACE (blue) maintains accuracy as noise increases, while adaptive stopping (orange) falls 3–4 pp at $\sigma_{\text{noise}} = 0.10$ (5 seeds; shading $= \pm 1$ std).

Adaptive stopping degrades substantially: at $\rho(\mathbf{W}) = 1.40$ and $\sigma_{\text{noise}} = 0.10$, accuracy drops from 81.2% to 78.1% (-3.1 pp). GRACE is essentially unaffected at 82.9%, a $+4.8$ pp advantage. At $\rho(\mathbf{W}) = 1.50$ the gap is $+3.8$ pp. All five GRACE runs at $\rho(\mathbf{W}) = 1.40$, $\sigma_{\text{noise}} = 0.10$ lie in $[82.2, 83.2]\%$ while all five adaptive stopping runs lie in $[76.7, 79.4]\%$, with no overlap.

L Cross-System Taxonomy: Full Results

Figure 12 shows the full per-radius accuracy curves for the two nonlinear systems (complex modReLU and real ESN/tanh). The linear null is summarised separately in Table 9: spectral growth without nonlinearity produces essentially no early-vs.-final gap.

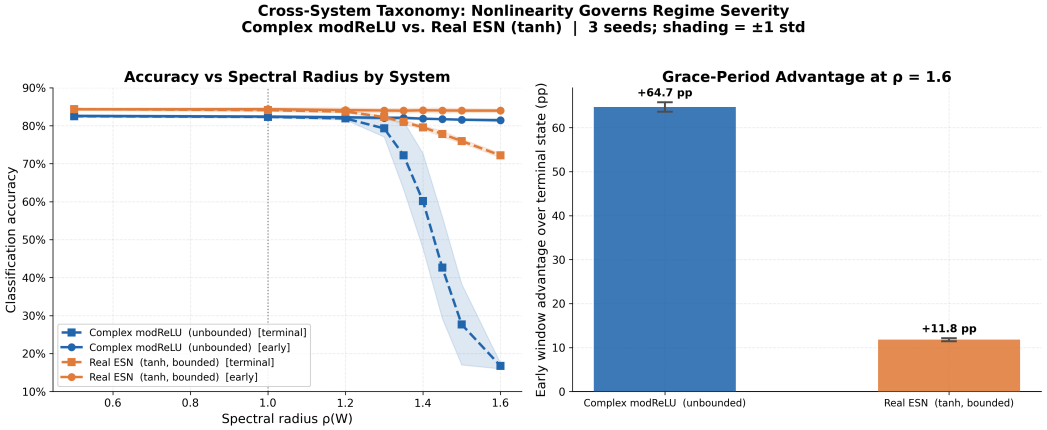


Figure 12: **Cross-system taxonomy: nonlinearity governs regime severity.** *Left:* accuracy vs spectral radius for complex modReLU (blue/dashed = terminal, blue/solid = early window) and real ESN/tanh (orange). *Right:* grace-period advantage (Early $-$ Terminal, pp) at $\rho = 1.6$. The regime is dramatic in the unbounded system ($+64.7$ pp for modReLU) and moderate in the bounded system ($+11.8$ pp for tanh ESN), confirming that nonlinearity governs regime severity.

Table 9: Cross-system summary at $\rho = 1.6$. The linear system serves as a null: spectral growth alone does not create a meaningful grace-period gap.

System	Nonlinearity	Final	Early	Gap
Complex modReLU	Unbounded	0.167	0.814	**+0.647**
Real ESN	Bounded (tanh)	0.722	0.840	+0.118
Linear	None	0.839	0.841	+0.002

M Mechanism Diagnostics: Separability and Operator Amplification

Figure 13 provides a direct diagnostic for the mechanism discussion in Section 4. In the unstable regime, per-timestep linear separability remains high over the early prefix and then collapses sharply late in the rollout; the contractive and near-critical settings are much flatter. This is consistent with the claim that useful input-imprinted structure is available early, while late nonlinear distortion erodes it.

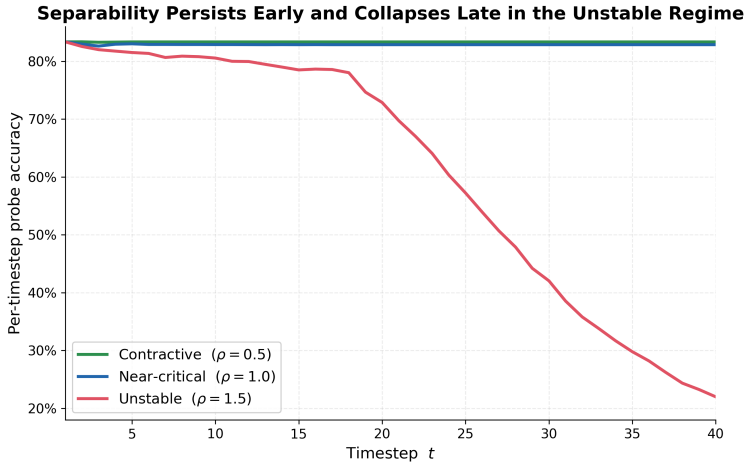


Figure 13: **Separability-over-time diagnostic.** Representative contractive ($\rho = 0.5$), near-critical ($\rho = 1.0$), and unstable ($\rho = 1.5$) settings are shown. The unstable system retains a highly separable early prefix before collapsing sharply at later timesteps, consistent with a transiently usable computation window.

Figure 14 closes the theory-experiment loop for the fixed-reservoir sweep by computing operator-level diagnostics on the actual sampled matrices. Across the 10-seed complex-modReLU sweep, the finite-horizon amplification proxy $\log_{10} \max_{t \le T} \|\mathbf{W}^t\|_2$ correlates strongly with shorter grace periods ($r = -0.75$) and larger GRACE-over-final gains ($r = 0.76$). Within fixed ρ, the scaled-Kreiss estimate $K_\rho(\mathbf{W})$ still explains residual seed-to-seed variation in the gap ($r = 0.39$), showing that radial growth dominates across the sweep while non-normality modulates residual severity. The bottom panel shows why amplification alone is not enough: the real tanh ESN and the linear null have nearly identical operator envelopes, yet only the nonlinear system develops a meaningful readout gap.

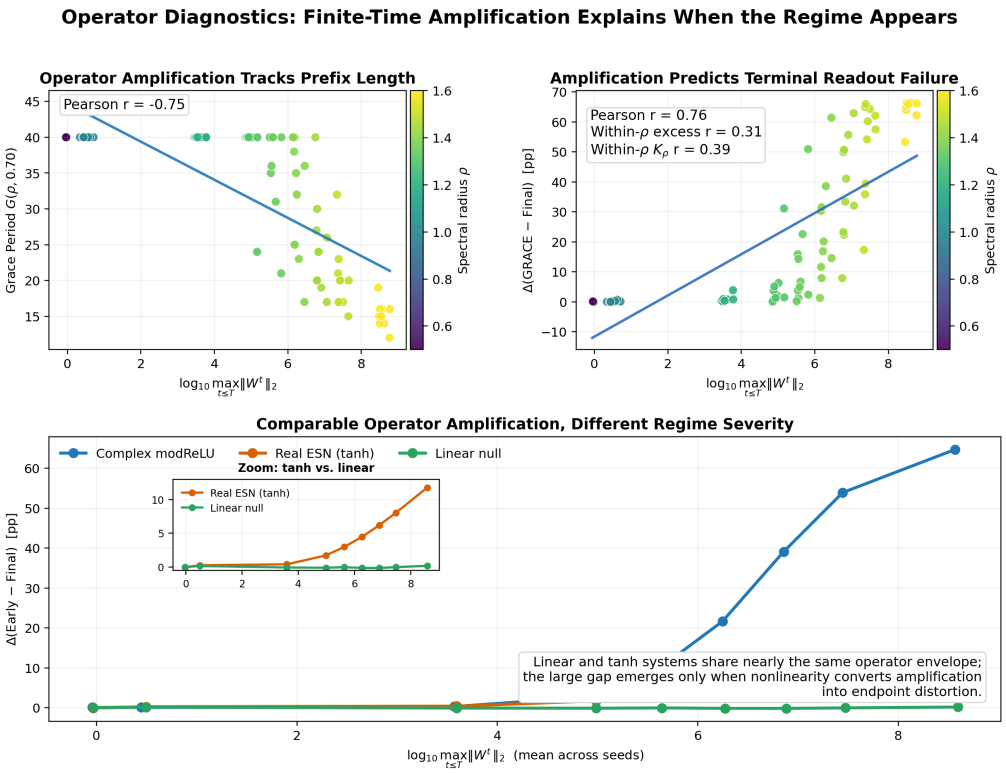


Figure 14: **Operator diagnostics on the actual reservoir sweep.** *Top left:* finite-horizon amplification tracks grace-period shortening across the 10-seed complex-modReLU sweep ($r = -0.75$). *Top right:* the same amplification proxy predicts GRACE-over-final improvement ($r = 0.76$); after factoring out spectral-radius trends, residual within-ρ variation still correlates with scaled-Kreiss non-normality ($r = 0.39$). *Bottom:* across systems at $\rho = 1.6$, the tanh ESN and linear null have nearly identical amplification envelopes, but only the nonlinear system exhibits a substantial gap. Across these tested systems, finite-time amplification tracks regime severity but is not sufficient for endpoint distortion; comparable amplification yields a substantial gap only in the nonlinear system.

N DEQ-Style Relevance Check

We include a small DEQ-style relevance check as a separate appendix experiment, not as a replot of the trained LeakyReLU sweep in Section 6.2. The model is a $d = 128$ tanh fixed-point classifier on flattened MNIST, $z_{k+1} = \tanh(\lambda \mathbf{W} z_k + \mathbf{U}x + b)$, trained on a 5k/2k split for 25 epochs with Adam at learning rate 10^{-3}, $K_{\text{train}} = 10$, and $\lambda_{\text{train}} = 0.80$. At evaluation we run $K_{\text{test}} = 50$ Picard iterations over $\lambda_{\text{test}} \in \{0.80, \ldots, 1.40\}$ and score the same final, early-window, adaptive, and GRACE readouts using RidgeClassifierCV probes on a half-split of the held-out set. This lower-capacity check is not meant to match the absolute accuracies in Table 3; it only tests whether an endpoint-vs.-trajectory mismatch can also appear near a fixed-point solver's stability boundary. Figure 15 shows a modest but consistent early-window advantage of roughly 1–1.5 pp over final-iterate readout across the tested gain range, with GRACE remaining close but slightly weaker in this setting.

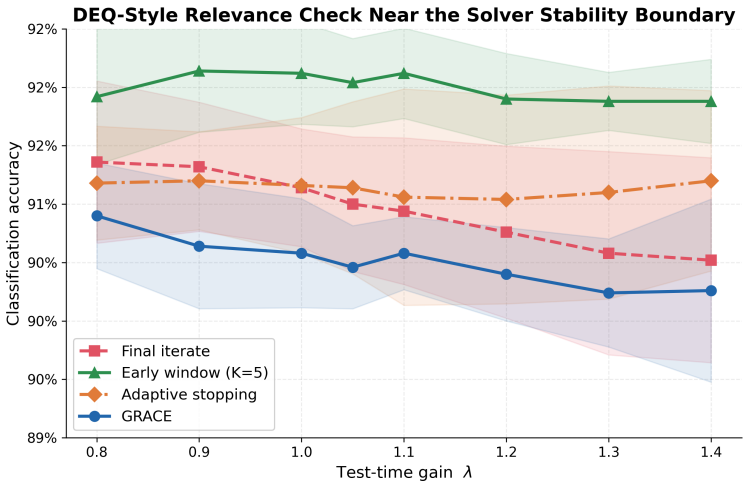


Figure 15: **Limited DEQ-style relevance check.** In a separate $d = 128$ tanh fixed-point MNIST classifier, early-window readout retains a modest advantage over the final iterate near the solver stability boundary, while adaptive stopping stays close and GRACE remains slightly weaker. We report this only as an appendix-level relevance check for iterative/fixed-point models.

O Sequential CIFAR-10 LSTM Check

We include a second trained-system check on a standard gated recurrent architecture. CIFAR-10 images are presented as row sequences (32 timesteps, 96 input dimensions per row) to a 512-hidden-unit LSTM trained end-to-end on terminal-state cross-entropy. To avoid confounding early readout with missing pixels, all readouts are taken only after the full image has been processed: we run a 40-step zero-input post-image settling trajectory at train time ($\lambda_{\text{train}} = 1.0$) and at test time sweep the recurrent gain λ_{test}. Training uses 10k CIFAR-10 examples, 2k test examples, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 64, 30 epochs, and 3 seeds. Readout accuracies use the same RidgeClassifierCV probe protocol on final, early-window, adaptive, and GRACE features; the dense single-timestep audit uses a fixed ridge regularisation $\alpha_{\text{ridge}} = 1.0$ for computational tractability.

Figures 16 and 17 and Table 10 show the same qualitative pattern as the trained LeakyReLU sweep, at smaller magnitude: terminal readout degrades with test-time gain, while early-window readout degrades more slowly. At $\lambda_{\text{test}} = 2.0$, final accuracy is 0.361 ± 0.007, early-window is 0.396 ± 0.009 ($+3.5$ pp), and GRACE is 0.384 ± 0.009 ($+2.3$ pp). Adaptive stopping mostly matches the terminal state, consistent with a bounded gated recurrent system in which the norm-threshold signal is weak. A separate single-timestep probe audit confirms an accuracy-defined prefix shortening: for accuracy threshold 0.40, the measured prefix lengths at $\lambda_{\text{test}} \in \{1.0, 1.6, 2.0\}$ are 40, 27, and 2 post-input settling steps, respectively.

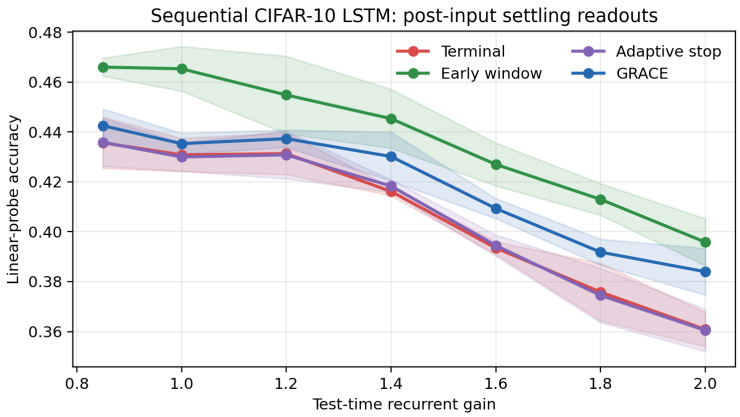


Figure 16: **Sequential CIFAR-10 LSTM post-input settling check.** After the full image has been processed, a trained 512-unit LSTM is run for a 40-step zero-input settling trajectory while recurrent gain is swept at test time. Terminal readout degrades with gain; early-window and GRACE remain higher at large gains (3 seeds).

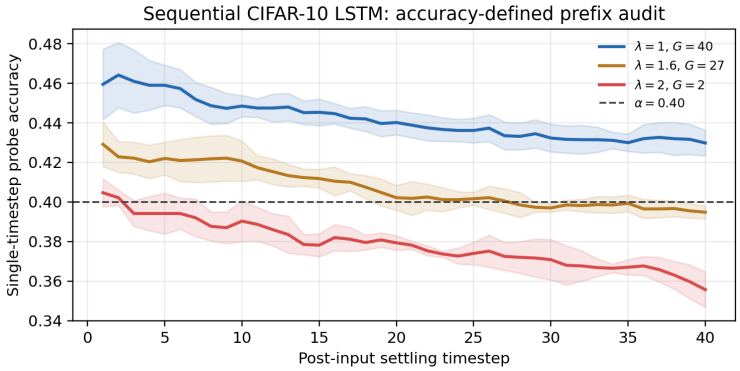


Figure 17: **Accuracy-defined prefix audit for the sequential CIFAR-10 LSTM.** We fit a separate linear probe at each post-input settling timestep for three test-time gains. At accuracy threshold 0.40, the usable prefix shortens from the full 40-step trajectory at $\lambda = 1.0$ to 27 steps at $\lambda = 1.6$ and 2 steps at $\lambda = 2.0$ (3 seeds).

Table 10: Sequential CIFAR-10 LSTM readout accuracy (mean $\pm$ std, 3 seeds). The effect is modest, as expected for a bounded gated recurrent architecture, but the endpoint-vs.-trajectory gap grows at high test-time gain.

λ_{test}	Final	Early	Adaptive	GRACE	Early−Final
0.85	0.436±0.010	0.466±0.004	0.436±0.010	0.443±0.007	+3.0 pp
1.00	0.431±0.007	0.465±0.009	0.430±0.006	0.435±0.004	+3.5 pp
1.20	0.431±0.008	0.455±0.016	0.431±0.010	0.437±0.003	+2.4 pp
1.40	0.416±0.002	0.445±0.012	0.418±0.002	0.430±0.010	+2.9 pp
1.60	0.394±0.003	0.427±0.009	0.395±0.004	0.409±0.004	+3.4 pp
1.80	0.376±0.011	0.413±0.006	0.375±0.011	0.392±0.005	+3.7 pp
2.00	0.361±0.007	0.396±0.009	0.361±0.008	0.384±0.009	+3.5 pp

P Tiny ImageNet: Architecture and Evaluation Protocol

Data and frozen representations. We use the standard Tiny ImageNet-200 training and labelled validation splits: 100,000 training images (500 per class) and 10,000 validation images (50 per class), available from the standard source distribution. The frozen backbone is torchvision's ImageNet-pretrained ConvNeXt-Tiny [43], with the preprocessing supplied with its default pretrained weights; no additional training augmentation is applied. The residual head receives 768-dimensional pooled features. The attention head receives the 7×7 spatial feature map after the frozen backbone's final channel LayerNorm, stored as 49 tokens of dimension 768. Backbone weights are never updated. The torchvision implementation is distributed under the BSD 3-Clause licence; the Tiny ImageNet and ImageNet source terms apply to the images and pretrained weights. We do not redistribute these assets in the manuscript source package.

Residual SwiGLU head. A learned projection maps pooled features to $h_0 \in \mathbb{R}^{512}$. The tied recurrence is

$$h_{t+1} = h_t + \frac{\lambda}{T_{\text{train}}} D(W_o\,[\mathrm{SiLU}(W_g\,\mathrm{LN}(h_t) + b_g) \odot (W_v\,\mathrm{LN}(h_t) + b_v)] + b_o)\,,$$

where the gate and value projections have width 2048 and D is dropout with probability 0.1 during training. A shared LayerNorm and linear 200-class head classify the resulting state. There is no input reinjection after h_0 and no intermediate supervision.

Shared-attention head. A learned projection maps each spatial token to width 256. A learned class token and fixed two-dimensional sinusoidal spatial positions are added once before recursion. Each tied block applies pre-normalised eight-head self-attention, followed by a pre-normalised SwiGLU feed-forward sublayer of width 1024. Each residual branch is scaled by λ/T_{train}; attention probabilities, attention outputs, and feed-forward outputs use dropout 0.1. Attention is unmasked over the 50 tokens. A shared LayerNorm and linear classifier read the class token. Parameters are shared across all recurrent steps, without intermediate supervision or input reinjection.

Training and checkpoint selection. Each architecture is trained separately at $T_{\text{train}} = 8$ and 32, with paired seeds 101, 102, and 103: 12 completed training runs. All runs use terminal cross-entropy, AdamW (learning rate 3×10^{-4}, weight decay 0.01), 30 epochs, two warm-up epochs followed by cosine decay, label smoothing 0.1, gradient clipping at norm 1, and nominal batch size 512. Attention uses physical batches of 128 and four accumulation steps, retaining the final partial accumulation group. Runs use BF16 on an RTX 5080 Laptop GPU. Checkpoints are selected by terminal validation accuracy at the training horizon and gain 1, with the earliest maximum retained. The same validation set supplies reported evaluation accuracies; there is no additional untouched test split. These are controlled representation experiments, not benchmark leaderboard estimates.

Horizon comparison. All final-panel evaluations fix gain $\lambda = 1$ and use horizons 8, 16, 24, and 32. Within a trained model, the step scale remains fixed as the evaluation horizon changes. Between the eight- and 32-step training conditions, both training depth and the prescribed residual scale $1/T_{\text{train}}$ change. The matched condition therefore tests the complete horizon-dependent training configuration; it does not isolate training depth from step size. Early-eight features average the first eight post-update states (class-token states for attention) and are decoded by the shared trained head. Terminal loss is $D = A(8) - A(32)$ in percentage points. The paired interaction is $D_{T_{\text{train}}=8} - D_{T_{\text{train}}=32}$ for each seed. All $\pm$ values in these additional experiments denote sample standard deviation across the three seeds, with denominator $n - 1$. No significance test is used at $n = 3$.

Table 11: **Final four-condition panel.** Accuracy is in percent; loss is $A(8) - A(32)$ in percentage points. Means and sample standard deviations across three seeds.

Model	T_{train}	Terminal 8	Terminal 32	Early 8	Loss (pp)
SwiGLU	8	80.87 ± 0.09	75.76 ± 0.96	81.22 ± 0.05	5.11 ± 0.94
SwiGLU	32	80.43 ± 0.05	80.84 ± 0.15	79.72 ± 0.14	-0.42 ± 0.18
attention	8	79.62 ± 0.10	77.20 ± 0.34	79.88 ± 0.11	2.42 ± 0.29
attention	32	79.20 ± 0.13	79.16 ± 0.22	78.58 ± 0.29	0.04 ± 0.17

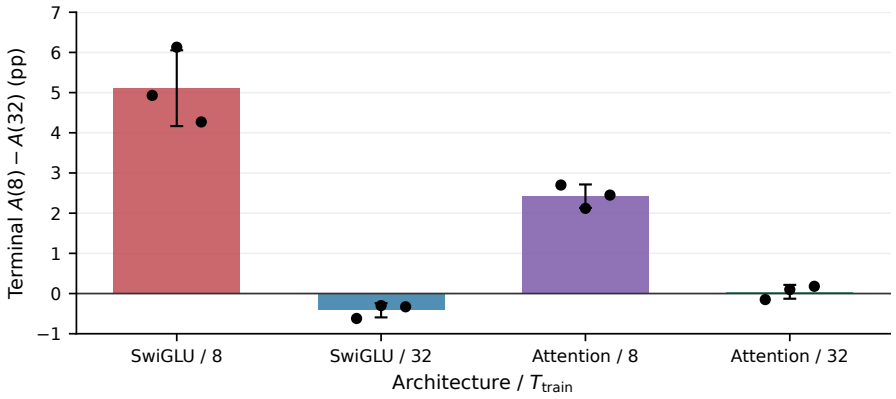


Figure 18: **Matched training removes the strong horizon-extension loss.** Terminal $A(8) - A(32)$ for the four recursive conditions. Negative values mean the endpoint improves with additional steps. Points show individual seeds; bars and error bars show means and sample standard deviations ($n = 3$). This is the difference summary of the curves in Figure 6.

Q Tiny ImageNet: Readouts and Representation Probes

Frozen readout settings. Early-window uses $K = 8$, adaptive stopping uses norm-ratio threshold 1.5, and GRACE uses $(\tau, \alpha) = (1.2, 10)$. These settings are not tuned for attention. Uniform averaging and GRACE use the available trajectory through each evaluation horizon. The shared classification head is applied to the aggregated state, without retraining. The table below retains all attention readouts, including underperformance of the norm-weighted rule in the matched 32-step condition. The modern experiments support an endpoint-validity diagnosis rather than universal dominance of GRACE.

Table 12: **All frozen attention readouts.** Validation accuracy (%, mean ± sample standard deviation, three seeds).

T_{train}	T_{test}	Terminal	Early 8	Uniform	Adaptive	GRACE
8	8	79.62 ± 0.08	79.88 ± 0.11	79.88 ± 0.11	79.40 ± 0.20	77.16 ± 0.36
8	16	78.85 ± 0.20	79.88 ± 0.11	79.41 ± 0.07	79.40 ± 0.20	77.16 ± 0.36
8	24	78.03 ± 0.20	79.88 ± 0.11	78.92 ± 0.21	79.40 ± 0.20	77.16 ± 0.36
8	32	77.20 ± 0.33	79.88 ± 0.11	78.49 ± 0.19	79.40 ± 0.20	77.16 ± 0.36
32	8	79.19 ± 0.11	78.58 ± 0.29	78.58 ± 0.29	76.81 ± 0.33	57.47 ± 3.94
32	16	79.36 ± 0.19	78.58 ± 0.29	79.40 ± 0.16	76.81 ± 0.33	57.47 ± 3.94
32	24	79.32 ± 0.37	78.58 ± 0.29	79.41 ± 0.26	76.81 ± 0.33	57.47 ± 3.94
32	32	79.15 ± 0.22	78.58 ± 0.29	79.41 ± 0.24	76.81 ± 0.33	57.47 ± 3.94

The horizon headline uses the saved full-trajectory classification results; the readout table uses the separately classified feature aggregates. Their attention terminal scores differ by a few examples because BF16 classifier matrix multiplications use different tensor shapes. Each series is retained consistently; these small differences do not alter the reported pattern.

Fixed-regularisation representation probes. We fit separate RidgeClassifiers with $\alpha_{\text{ridge}} = 1$ on step-8, early-eight, and step-32 features, using the same fixed 20,000-example training subset (subset seed 2026), no feature standardisation, and all 10,000 validation examples. The regularisation is fixed, while each representation receives its own fitted coefficients. This distinguishes a loss of linear decodability from failure of the shared trained head alone.

For eight-step training, both architectures lose accuracy under both the shared head and refitted probes. Matching the training horizon removes the large probe loss: the SwiGLU probe improves by 0.24 ± 0.08 pp and the attention probe loses only 0.35 ± 0.10 pp. This is evidence about the tested linear decoders, not a proof that all task information is irrecoverably lost.

Table 13: **Fixed-alpha probe accuracy.** Separately fitted probes share ridge $\alpha = 1$ and a fixed training subset. Accuracy in percent; loss in pp.

Model	T_{train}	Step 8	Early 8	Step 32	Loss (pp)
SwiGLU	8	81.31 ± 0.14	81.48 ± 0.12	79.08 ± 0.29	2.23 ± 0.17
SwiGLU	32	81.12 ± 0.01	80.75 ± 0.06	81.36 ± 0.09	-0.24 ± 0.08
Attention	8	79.59 ± 0.06	79.79 ± 0.07	78.31 ± 0.08	1.28 ± 0.08
Attention	32	79.33 ± 0.35	79.39 ± 0.30	78.98 ± 0.30	0.35 ± 0.10

R Tiny ImageNet: Robustness and Preliminary Controls

Classwise robustness. The classwise loss is the mean step-8 accuracy minus mean step-32 accuracy across seeds, evaluated separately in each of the 200 classes. All classes are included. SwiGLU has a median loss of 2.67 pp (IQR 0.00–8.67), with 144/200 classes losing accuracy and 124/200 losing at least 2 pp. Attention has a median loss of 2.00 pp (IQR 0.67–4.00), with 153/200 classes losing accuracy and 105/200 losing at least 2 pp. The macro-mean losses are 5.11 and 2.42 pp, respectively. The degradation is therefore distributed across many classes.

Checkpoint sensitivity. We evaluate existing saved states without changing the headline selection rule. Every audited checkpoint has a positive step-8-minus-step-32 loss, but its magnitude depends on checkpoint age. SwiGLU's selected checkpoints (epochs 3–4) lose 4.27–6.13 pp, while final-epoch checkpoints lose only 0.59–0.68 pp. Continued eight-step training therefore substantially reduces the loss even without matched 32-step training. For attention, the top three saved improving checkpoints per seed lose 2.12–3.49 pp; the selected checkpoint has the smallest loss among those saved alternatives in each seed. Intermediate SwiGLU checkpoints and final-epoch attention states were not retained, limiting the comparisons possible from saved artifacts.

Table 14: **Checkpoint sensitivity of eight-step-trained models.** Accuracy in percent; loss in pp. Only retained checkpoints are compared.

Model	Seed	Checkpoint	Epoch	Step 8	Step 32	Loss
SwiGLU	101	Selected	3	80.98	76.05	4.93
SwiGLU	101	Final epoch	30	80.71	80.12	0.59
SwiGLU	102	Selected	4	80.82	74.69	6.13
SwiGLU	102	Final epoch	30	80.62	79.96	0.66
SwiGLU	103	Selected	3	80.82	76.55	4.27
SwiGLU	103	Final epoch	30	80.67	79.99	0.68
Attention	101	Selected	5	79.53	76.83	2.70
Attention	101	Saved rank 2	4	78.91	76.08	2.83
Attention	101	Saved rank 3	3	78.52	75.03	3.49
Attention	102	Selected	7	79.61	77.49	2.12
Attention	102	Saved rank 2	5	79.31	76.76	2.55
Attention	102	Saved rank 3	4	79.27	76.28	2.99
Attention	103	Selected	5	79.72	77.27	2.45
Attention	103	Saved rank 2	4	79.30	76.63	2.67
Attention	103	Saved rank 3	3	78.38	75.06	3.32

Seed-0 gain-sweep pilots. Before the paired panel, two SwiGLU pilots used gain values $\{0.75, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6\}$. The first trained for eight steps and evaluated for 32. Its early-minus-terminal gap was already 4.93 pp at nominal gain, so it failed the predeclared requirement for a competitive nominal endpoint before gain stress. The aligned 32-step pilot restored the nominal endpoint but had a maximum early-window advantage of only 0.27 pp, failing the criterion of at least 1 pp at three consecutive gains. Each pilot was stopped after seed 0; neither is a three-seed confirmation. These outcomes motivated the subsequently frozen paired horizon comparison with fresh seeds. The direct fixed-alpha probe on pooled ConvNeXt features scored 80.52%.

The pilots also compare per-sample GRACE with a fixed temporal profile, formed by averaging normalised training-sample GRACE weights at gain 1.3 and then freezing that profile across gains.

Table 15: **Seed-0 SwiGLU pilot readouts at $T_{\text{test}} = 32$.** Accuracy in percent. Fixed denotes the training-derived global GRACE profile. Neither pilot continued to further seeds.

T_{train}	Gain	Terminal	Early 8	Uniform	Adaptive	GRACE	Fixed
8	0.75	77.72	80.76	79.71	80.34	80.70	80.36
8	1.00	75.92	80.85	78.84	80.38	80.70	80.39
8	1.10	75.09	80.89	78.52	80.35	80.69	80.56
8	1.20	74.47	80.78	78.31	80.36	80.72	80.59
8	1.30	73.77	80.82	77.92	80.32	80.68	80.68
8	1.40	73.08	80.87	77.65	80.28	80.71	80.78
8	1.50	72.44	80.83	77.41	80.37	80.69	80.80
8	1.60	71.86	80.82	77.13	80.29	80.76	80.81
32	0.75	80.77	79.25	80.63	80.77	80.50	80.06
32	1.00	80.75	79.72	80.85	80.71	80.64	80.33
32	1.10	80.62	79.77	80.78	80.64	80.52	80.33
32	1.20	80.54	79.84	80.77	80.59	80.54	80.39
32	1.30	80.33	79.95	80.71	80.57	80.59	80.40
32	1.40	80.15	80.07	80.80	80.43	80.63	80.51
32	1.50	80.15	80.15	80.77	80.57	80.67	80.70
32	1.60	79.93	80.20	80.79	80.49	80.66	80.68

They show no consistent advantage for per-sample weighting. At gain 1.6 in the aligned pilot, uniform averaging, GRACE, and the fixed profile score 80.79%, 80.66%, and 80.68%, respectively. A saved-trajectory horizon audit of the eight-step pilot at gain 1 gives terminal accuracy 80.82% at step 8 and 75.92% at step 32, consistent with the later paired panel. No new canonical-DEQ experiment was completed; the DEQ-style evidence in Appendix N remains the original limited check.

S Tiny ImageNet: Model Scale, Runtime, and Memory

Table 16: **Scale of the modern recursive systems.** The executed frozen backbone has 27,820,128 parameters. One multiply-add counts as two FLOPs.

Quantity	SwiGLU	Attention
Input	pooled 768-d	49×768 tokens
Recursive width	512	256
Intermediate width	2,048	1,024
Attention heads	—	8
Trainable head parameters	3,648,712	1,301,960
Executed total parameters	31,468,840	29,122,088
Dominant recurrent FLOPs/step	6,291,456	107,517,600

For state width d, intermediate width m, n tokens, and h attention heads, recurrent FLOPs are counted as $6dm$ for SwiGLU and $8nd^2 + 4n^2d + 5hn^2 + 6ndm$ for attention ($n = 50$ including the class token). These counts exclude the frozen backbone, one-time input projection, final classifier, LayerNorm, and lower-order elementwise operations.

Readout benchmark. The measurements in Table 2 use the first seed-0 SwiGLU pilot, independently of its scientific continuation outcome. Settings are BF16, batch 512, $d = 512$, $T = 32$, gain 1.3, and early window 8, on an RTX 5080 Laptop GPU. CUDA events measure elapsed device time after 10 warm-ups, with the median of 25 timed repetitions reported. Full recurrent-head inference starts with cached backbone features and includes recursion and classification; it excludes image loading and backbone execution. Additional peak allocated memory is measured relative to the warmed-up baseline for each repetition and summarised by its median. It excludes baseline model, input, and cached-tensor allocations and is not CUDA reserved or total process memory. Readout-only measurements start with an already allocated trajectory.

Table 17: **Full recurrent-head and readout-only benchmark.** Latencies in ms and additional peak allocations in MiB; the precomputed trajectory is part of the baseline in readout-only measurements.

Readout	Full ms	Readout ms	Full MiB	Readout MiB
Terminal	6.727	0.113	8.00	2.00
Early window	1.803	0.135	11.50	2.50
Adaptive	13.840	0.208	9.01	2.52
Uniform streaming	6.889	0.117	8.50	2.50
GRACE, stored	6.948	0.317	48.56	32.06
GRACE, streaming	9.681	3.653	8.51	3.01

Table 18: **Readout complexity per example.** Arithmetic excludes the recurrent update and classifier. $t \leq T$ is the adaptive stopping step.

Readout	Arithmetic	State storage	Halts recurrence
Terminal	$O(d)$	$O(d)$	No
Early window	$O(Kd)$	$O(d)$	At K
Adaptive	$O(td)$	$O(d)$	Yes
Uniform streaming	$O(Td)$	$O(d)$	No
GRACE, stored	$O(Td)$	$O(Td)$	No
GRACE, streaming	$O(Td)$	$O(d)$	No

Streaming GRACE maintains a weighted numerator and denominator while discarding each state after use. This explains its memory benefit without implying a speedup: the measured eager loop costs 9.68 ms versus 6.95 ms for stored-trajectory GRACE. Adaptive active-set compaction costs 13.84 ms for this dense batch. These are implementation- and hardware-specific measurements, not architecture-independent latency guarantees.

Training and evaluation allocations. The attention training runs separately record absolute CUDA allocator peaks. For eight-step training, allocated/reserved peaks are 871.56/1,100 MiB; for 32-step training they are 3,054.91/3,594 MiB. Evaluation peaks are 219.53/298 MiB in both conditions. The peaks are identical across the three seeds within each condition. Completed training wall times are 499.42/498.15/505.00 s for eight steps and 1,582.40/1,635.76/2,312.70 s for 32 steps. These absolute allocator peaks measure a different quantity from the additional-memory benchmark above.

Execution provenance. Protocols, configurations, seeds, and checkpoint selection were frozen for each experiment before its results. The final panel contains 12 completed scientific runs, plus the two separate seed-0 pilots. Two interrupted 32-step attention attempts were retained and restarted from the same frozen seed/configuration because optimizer and RNG states were unavailable for an exact resume; they are not extra scientific replicates. Scientific artifacts retain SHA-256 configuration and checkpoint records. The robustness analyses use existing checkpoints, with no post-result changes to architectures, gains, readout thresholds, probe regularisation, or the headline selection rule.

T Extended Related Work

Reservoir computing and the echo state property. Echo state networks [1] and liquid state machines [21] established the reservoir computing paradigm. The echo state property (ESP), the requirement that reservoir states be uniquely determined by input history, is traditionally ensured by constraining $\rho(\mathbf{W}) < 1$, though this condition is neither necessary nor sufficient [17]. Our work characterises what happens when $\rho(\mathbf{W})$ is intentionally pushed past the ESP boundary: the substrate remains useful, but only within a temporal window formalised as the prefix grace period.

Edge of chaos and criticality. Bertschinger and Natschläger [13] showed that reservoir computation is optimised at the critical boundary $\rho(\mathbf{W}) \approx 1$. Subsequent work formalised memory-nonlinearity trade-offs at criticality [14]. In nearby deep-network mean-field analyses, transient chaos and signal propagation likewise sharpen at the ordered/chaotic boundary [15, 16]. Our regime

lies beyond this boundary, showing that useful computation persists in the supercritical phase provided one reads out from the right portion of the trajectory.

Non-normal amplification. Trefethen and Embree [18] established that non-normal matrices exhibit transient growth even when spectral radius predicts asymptotic stability. Hennequin et al. [30] separate non-normal amplification from near-critical slowing in balanced networks. Recent nnRNN work likewise uses Schur-structured non-normal connectivity and shows that training preserves or suppresses transient structure according to task demand [33]. Our norm-ratio criterion (Equation (5)) directly measures this amplification and uses it as a readout signal.

Physical neuromorphic substrates. Recent physical neural-network and photonic-scattering systems show that dynamical media, wave propagation, and hardware readout can serve as trainable computational resources [26–28]. These settings make readout timing especially salient because the measured state and sampling time are part of the physical computation, not merely an implementation detail.

Vanishing and exploding gradients. The classical gradient pathology in RNNs [19, 20] motivates architectural solutions (LSTMs, GRUs, unitary RNNs [6]). These address training stability; our work addresses inference-time readout. Gradient clipping stabilises learning; trajectory-aware readouts stabilise decoding in the regimes studied.

Edge of stability in optimisation. Cohen et al. [12] identify an "edge of stability" in gradient descent where the Hessian's largest eigenvalue hovers near $2/\eta$. Our phenomenon is distinct: it concerns readout timing of a fixed dynamical substrate, not optimiser dynamics.

Deep equilibrium and neural ODE models. DEQ models [2] and neural ODEs [3] are the purest instantiations of the endpoint assumption. Near bifurcation boundaries, trajectory-level readout could improve robustness. Consistency DEQs cast DEQ inference as movement along a fixed ODE trajectory and train intermediate states to map directly to the equilibrium [38]. Our setting asks a different question: whether the endpoint remains the best object of inference at all. GRACE is therefore a training-free trajectory-weighted diagnostic rather than a distillation method for predicting the fixed point.

Adaptive computation, early exit, and overthinking. Adaptive Computation Time and PonderNet learn how many recurrent steps to use for each input [10, 11]. Early-exit systems and overthinking analyses show that intermediate predictions can outperform deeper computations [31, 32, 36]. Our setting differs in both axis and failure mode: the intermediate object is a time-indexed dynamical state rather than a shallower network layer or a learned halting distribution, and the issue is not only saving computation but detecting when extra computation destroys endpoint validity while a usable prefix remains.

Deep reservoir computing. Gallicchio and Micheli [29] show that global spectral radius increases monotonically with reservoir depth, linking architectural depth to shrinking stability margins. Our result is complementary: once a system is driven past those margins, decoder choice determines whether the remaining transient computation is still recoverable.

U Proofs

U.1 Proof of Proposition 4

Proof. Let $\bar{r}_t$ be the reference norm-ratio curve from the proposition. By assumption,

$$\bar{r}_t \geq \bar{r}_{t_0}\, \kappa^{t-t_0} \qquad \text{for all } t \geq t_0.$$

If

$$t \geq t_0 + \left\lceil \frac{\log(R_{\max}/\bar{r}_{t_0})}{\log \kappa} \right\rceil + 1,$$

then $t - t_0 > \log(R_{\max}/\bar{r}_{t_0})/\log \kappa$, and therefore

$$\bar{r}_t \geq \bar{r}_{t_0}\, \kappa^{t-t_0} > R_{\max}.$$

Hence $\mathrm{acc}(\phi_t) < \alpha$ for all such t. It follows that the grace-period definition cannot certify any prefix extending beyond

$$t_0 + \left\lceil \frac{\log(R_{\max}/\bar{r}_{t_0})}{\log \kappa} \right\rceil,$$

which is exactly the claimed upper bound on $G(\rho(\mathbf{W}), \alpha)$. □

U.2 Proposition A1: Prefix-Limit Behaviour of GRACE

Proposition 7 (Prefix-limit behaviour)**.** *Let $w_t(\alpha, \tau) \propto \exp(-\alpha \max(0, g_t - \tau))$. If the threshold crossing is unique, i.e., there exists $t^\star \geq 1$ such that $g_t \leq \tau$ for $t \leq t^\star$ and $g_t > \tau$ for $t > t^\star$, then as $\alpha \to \infty$,*

$$w_t \to \begin{cases} 1/t^\star, & t \leq t^\star, \\ 0, & t > t^\star. \end{cases}$$

Proof. For $t \leq t^\star$, $\tilde{w}_t = 1$. For $t > t^\star$, $\tilde{w}_t = \exp(-\alpha(g_t - \tau)) \to 0$ as $\alpha \to \infty$ because $g_t - \tau > 0$. Normalising over the $t^\star$ surviving weights gives $w_t = 1/t^\star$ for $t \leq t^\star$. □

This result connects GRACE to prefix selection: in the large-α limit, GRACE places uniform mass on the largest contiguous norm-threshold prefix and zero mass beyond it. The prefix is not assumed to equal the accuracy-defined grace period in Definition 1; empirically, the norm-threshold prefix acts as a conservative proxy. Thus GRACE interpolates continuously between uniform averaging ($\alpha = 0$) and hard norm-threshold prefix selection ($\alpha \to \infty$).

U.3 Lemma A2: Tail Suppression Preserves Separability

Lemma 8 (Tail suppression preserves separability under bounded prefix distortion)**.** *Let $(\mathbf{x}_i, y_i)_{i=1}^N$ be labelled examples with $y_i \in \{-1, +1\}$. For each timestep $t = 1, \ldots, T$, write the observed feature as $\mathbf{h}_t^{(i)} = \mathbf{s}_t^{(i)} + \mathbf{e}_t^{(i)}$. Assume there exist a prefix length G, a separator $(\mathbf{w}, b)$, and a margin $\gamma > 0$ such that*

$$y_i\big(\mathbf{w}^\top \mathbf{s}_t^{(i)} + b\big) \geq \gamma \qquad \text{for all } i \text{ and all } t \leq G.$$

Assume also that for $t \leq G$ the distortion obeys $\|\mathbf{e}_t^{(i)}\| \leq \eta_t$ for all i. Let nonnegative, sample-dependent weights $(\lambda_t^{(i)})_{t=1}^T$ satisfy $\sum_{t=1}^T \lambda_t^{(i)} = 1$, $\lambda_t^{(i)} = 0$ for $t > G$, and

$$\|\mathbf{w}\| \sup_i \sum_{t=1}^{G} \lambda_t^{(i)} \eta_t < \gamma.$$

Then the aggregated feature

$$\bar{\mathbf{h}}^{(i)} = \sum_{t=1}^{T} \lambda_t^{(i)} \mathbf{h}_t^{(i)}$$

is linearly separable by the same separator $(\mathbf{w}, b)$, with margin at least

$$\gamma - \|\mathbf{w}\| \sup_i \sum_{t=1}^{G} \lambda_t^{(i)} \eta_t > 0.$$

Proof. Since $\lambda_t^{(i)} = 0$ for $t > G$ and $\sum_{t=1}^{T} \lambda_t^{(i)} = 1$,

$$\begin{aligned} y_i\big(\mathbf{w}^\top \bar{\mathbf{h}}^{(i)} + b\big) &= y_i \left(\mathbf{w}^\top \sum_{t=1}^{G} \lambda_t^{(i)} \big(\mathbf{s}_t^{(i)} + \mathbf{e}_t^{(i)}\big) + b \right) \\ &= \sum_{t=1}^{G} \lambda_t^{(i)}\, y_i\big(\mathbf{w}^\top \mathbf{s}_t^{(i)} + b\big) + y_i\, \mathbf{w}^\top \sum_{t=1}^{G} \lambda_t^{(i)} \mathbf{e}_t^{(i)} \\ &\geq \sum_{t=1}^{G} \lambda_t^{(i)} \gamma - \|\mathbf{w}\| \sum_{t=1}^{G} \lambda_t^{(i)} \|\mathbf{e}_t^{(i)}\| \\ &\geq \gamma - \|\mathbf{w}\| \sum_{t=1}^{G} \lambda_t^{(i)} \eta_t \\ &\geq \gamma - \|\mathbf{w}\| \sup_j \sum_{t=1}^{G} \lambda_t^{(j)} \eta_t \\ &> 0. \end{aligned}$$

Thus every aggregated feature is correctly classified by the same separator, with the stated margin. □

This lemma captures the role of tail suppression in the mechanism argument: if a prefix is separable and accumulated prefix distortion stays below the available margin, then any per-sample convex readout that excludes the tail preserves separability. GRACE approximates this behaviour by down-weighting high-growth late timesteps. The statement is written in binary margin form; the multiclass analogue follows by applying the same bound to each pairwise score gap of a linear head.

U.4 Scaled-Kreiss Refinement for Smooth Activations

The main text uses Proposition 4 because it applies directly to the measured norm-ratio proxy. This appendix upgrades that proposition in operator terms. There, the conditional factor κ plays the role of a post-t_0 amplification rate; here, that rate is decomposed into a radial term and an envelope factor, first through a frozen scaled-Kreiss constant and then through a full Jacobian cocycle. The resulting corollaries replace the norm-threshold assumption of Proposition 4 with a margin-vs.-remainder criterion. The tanh remainder bound below applies directly to the bounded-activation ESN branch of our taxonomy; the modReLU and LeakyReLU cases admit analogous activation-specific remainder bounds, but we do not state them here.

For smooth activations, one can refine the main-text mechanism locally by separating linearised amplification from nonlinear remainder. The first step is a forced linear bound controlled by the scaled Kreiss constant of a frozen effective operator; the final theorem below then replaces that frozen operator by a trajectory-dependent Jacobian cocycle. Unless stated otherwise, norms in this subsection are Euclidean/operator spectral norms.

Lemma 9 (Scaled-Kreiss envelope for forced linear dynamics)**.** *Let $\mathbf{A} \in \mathbb{C}^{d\times d}$ with $\rho_* = \rho(\mathbf{A}) > 0$, and define the scaled Kreiss constant*

$$K_{\rho_*}(\mathbf{A}) := \sup_{|z|>\rho_*} (|z| - \rho_*)\, \|(z\mathbf{I} - \mathbf{A})^{-1}\| = K(\mathbf{A}/\rho_*). \tag{10}$$

Assume $K_{\rho_*}(\mathbf{A}) < \infty$. Consider the forced linear recurrence

$$\boldsymbol{\delta}_{t+1} = \mathbf{A}\boldsymbol{\delta}_t + \mathbf{q}_t, \qquad \boldsymbol{\delta}_0 = \mathbf{0}.$$

Then for every $t \geq 1$,

$$\|\boldsymbol{\delta}_t\| \leq e d\, K_{\rho_*}(\mathbf{A}) \sum_{s=0}^{t-1} \rho_*^{t-1-s}\, \|\mathbf{q}_s\|. \tag{11}$$

Proof. Let $\mathbf{B} = \mathbf{A}/\rho_*$. By variation of constants,

$$\boldsymbol{\delta}_t = \sum_{s=0}^{t-1} \mathbf{A}^{t-1-s} \mathbf{q}_s = \sum_{s=0}^{t-1} \rho_*^{t-1-s}\, \mathbf{B}^{t-1-s} \mathbf{q}_s.$$

The discrete Kreiss matrix theorem gives

$$\sup_{k\geq 0}\|\mathbf{B}^k\| \leq ed\,K(\mathbf{B}) = ed\,K_{\rho_*}(\mathbf{A})$$

[18]. Therefore

$$\|\mathbf{A}^k\| = \rho_*^k\|\mathbf{B}^k\| \leq ed\,K_{\rho_*}(\mathbf{A})\,\rho_*^k.$$

Substituting this into the variation-of-constants formula yields Equation (11). □

For smooth activations the one-step nonlinear remainder is quadratic in the distance from the linearisation point. For tanh this yields an explicit constant.

Lemma 10 (tanh linearisation remainder)**.** *Fix an input $\mathbf{x} \in \mathbb{R}^n$ and real-valued parameters $\mathbf{W} \in \mathbb{R}^{d\times d}$, $\mathbf{U} \in \mathbb{R}^{d\times n}$, and $\mathbf{b} \in \mathbb{R}^d$. Define*

$$F_{\mathbf{x}}(\mathbf{z}) = \tanh(\mathbf{W}\mathbf{z} + \mathbf{U}\mathbf{x} + \mathbf{b}),$$

with tanh applied componentwise. Let

$$J_{\mathbf{x}}(\bar{\mathbf{z}}) = D\tanh(\mathbf{W}\bar{\mathbf{z}} + \mathbf{U}\mathbf{x} + \mathbf{b})\,\mathbf{W}.$$

Then for all $\mathbf{z}, \bar{\mathbf{z}} \in \mathbb{R}^d$,

$$\|F_{\mathbf{x}}(\mathbf{z}) - F_{\mathbf{x}}(\bar{\mathbf{z}}) - J_{\mathbf{x}}(\bar{\mathbf{z}})(\mathbf{z}-\bar{\mathbf{z}})\| \leq \frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2\,\|\mathbf{z}-\bar{\mathbf{z}}\|^2, \tag{12}$$

where

$$L_{\tanh} := \sup_{u\in\mathbb{R}}|\tanh''(u)| = \frac{4}{3\sqrt{3}}.$$

Proof. Write $\mathbf{a}(\mathbf{z}) = \mathbf{W}\mathbf{z} + \mathbf{U}\mathbf{x} + \mathbf{b}$. Since tanh acts componentwise,

$$J_{\mathbf{x}}(\mathbf{z}) = D\tanh(\mathbf{a}(\mathbf{z}))\,\mathbf{W}.$$

For any $\mathbf{z}, \bar{\mathbf{z}}$, the operator norm of the diagonal difference obeys

$$\begin{aligned}\|D\tanh(\mathbf{a}(\mathbf{z})) - D\tanh(\mathbf{a}(\bar{\mathbf{z}}))\| &\leq \max_j|\tanh'(a_j(\mathbf{z})) - \tanh'(a_j(\bar{\mathbf{z}}))| \\ &\leq L_{\tanh}\,\|\mathbf{a}(\mathbf{z}) - \mathbf{a}(\bar{\mathbf{z}})\|_\infty \\ &\leq L_{\tanh}\,\|\mathbf{W}\|\,\|\mathbf{z}-\bar{\mathbf{z}}\|.\end{aligned}$$

Multiplying by $\|\mathbf{W}\|$ gives the Jacobian Lipschitz bound

$$\|J_{\mathbf{x}}(\mathbf{z}) - J_{\mathbf{x}}(\bar{\mathbf{z}})\| \leq L_{\tanh}\,\|\mathbf{W}\|^2\,\|\mathbf{z}-\bar{\mathbf{z}}\|.$$

The integral remainder formula then yields

$$\begin{aligned}&\|F_{\mathbf{x}}(\mathbf{z}) - F_{\mathbf{x}}(\bar{\mathbf{z}}) - J_{\mathbf{x}}(\bar{\mathbf{z}})(\mathbf{z}-\bar{\mathbf{z}})\| \\ &\quad\leq \int_0^1 \|J_{\mathbf{x}}(\bar{\mathbf{z}} + \theta(\mathbf{z}-\bar{\mathbf{z}})) - J_{\mathbf{x}}(\bar{\mathbf{z}})\|\,\|\mathbf{z}-\bar{\mathbf{z}}\|\,d\theta \\ &\quad\leq \int_0^1 \theta\,L_{\tanh}\,\|\mathbf{W}\|^2\,\|\mathbf{z}-\bar{\mathbf{z}}\|^2\,d\theta = \frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2\,\|\mathbf{z}-\bar{\mathbf{z}}\|^2.\end{aligned}$$

□

The previous two lemmas combine into a simple margin-vs.-remainder criterion for any frozen linearisation.

Proposition 11 (Frozen-linear margin criterion)**.** *Let $(\mathbf{x}_i, y_i)_{i=1}^N$ with $y_i \in \{-1, +1\}$, and suppose that at a fixed timestep t the true features admit a decomposition*

$$\mathbf{z}_t^{(i)} = \tilde{\mathbf{z}}_t^{(i)} + \mathbf{e}_t^{(i)}.$$

Assume there exists a separator $(\mathbf{w}_t, b_t)$ and a margin $m_t > 0$ such that

$$y_i\big(\mathbf{w}_t^\top\tilde{\mathbf{z}}_t^{(i)} + b_t\big) \geq m_t \qquad \text{for all } i.$$

If

$$\max_i \|\mathbf{e}_t^{(i)}\| < \frac{m_t}{\|\mathbf{w}_t\|},$$

then the same separator classifies the true features correctly, with margin at least

$$m_t - \|\mathbf{w}_t\| \max_i \|\mathbf{e}_t^{(i)}\| > 0.$$

Moreover, as an idealized autonomous/frozen-linearisation case, suppose the remainders satisfy a frozen linear recurrence

$$\mathbf{e}_{s+1}^{(i)} = J_* \mathbf{e}_s^{(i)} + \mathbf{r}_s^{(i)}, \qquad \mathbf{e}_0^{(i)} = \mathbf{0},$$

for some fixed operator $J_* \in \mathbb{R}^{d\times d}$ with $\rho_* = \rho(J_*) > 0$, and assume $\|\mathbf{r}_s^{(i)}\| \le R_s$ for all i, s. Then

$$\max_i \|\mathbf{e}_t^{(i)}\| \le ed\, K_{\rho_*}(J_*) \sum_{s=0}^{t-1} \rho_*^{t-1-s} R_s. \tag{13}$$

In this idealized case, if J_* is the local Jacobian used in the linearisation and each $\mathbf{r}_s^{(i)}$ is the corresponding one-step Taylor remainder at a reference point $\bar{\mathbf{z}}_s^{(i)}$, then Lemma 10 gives

$$R_s \le \frac{L_{\tanh}}{2} \|\mathbf{W}\|^2 D_s^2, \qquad D_s := \max_i \|\mathbf{z}_s^{(i)} - \bar{\mathbf{z}}_s^{(i)}\|.$$

Hence timestep t remains separable whenever

$$\frac{L_{\tanh}}{2} \|\mathbf{W}\|^2\, ed\, K_{\rho_*}(J_*) \sum_{s=0}^{t-1} \rho_*^{t-1-s} D_s^2 < \frac{m_t}{\|\mathbf{w}_t\|}. \tag{14}$$

Proof. For the first claim,

$$\begin{aligned} y_i\big(\mathbf{w}_t^\top \mathbf{z}_t^{(i)} + b_t\big) &= y_i\big(\mathbf{w}_t^\top \tilde{\mathbf{z}}_t^{(i)} + b_t\big) + y_i\, \mathbf{w}_t^\top \mathbf{e}_t^{(i)} \\ &\ge m_t - \|\mathbf{w}_t\|\, \|\mathbf{e}_t^{(i)}\| \\ &\ge m_t - \|\mathbf{w}_t\| \max_j \|\mathbf{e}_t^{(j)}\| \\ &> 0. \end{aligned}$$

Thus the same separator still classifies the true features, with the stated margin.

For the bound Equation (13), apply Lemma 9 to the recurrence for each i and then take the maximum over i. The tanh specialisation follows by substituting the Taylor bound from Lemma 10. The final separability criterion is then exactly the first claim applied with the resulting upper bound on $\max_i \|\mathbf{e}_t^{(i)}\|$. □

The frozen operator above is a convenient expository model. The rigorous trajectory-dependent treatment is the nonautonomous Jacobian-cocycle argument below, where the local Jacobian may vary with both sample and timestep.

Theorem 12 (Trajectory-dependent margin criterion)**.** *Fix real-valued tanh dynamics*

$$F_{\mathbf{x}}(\mathbf{z}) = \tanh(\mathbf{W}\mathbf{z} + \mathbf{U}\mathbf{x} + \mathbf{b}).$$

For each sample i, let $(\bar{\mathbf{z}}_s^{(i)})_{s=0}^t$ be a reference path and define

$$J_s^{(i)} := D \tanh(\mathbf{W}\bar{\mathbf{z}}_s^{(i)} + \mathbf{U}\mathbf{x}_i + \mathbf{b})\, \mathbf{W},$$

and the Jacobian cocycle

$$\Phi_{t:s}^{(i)} := \begin{cases} J_{t-1}^{(i)} \cdots J_s^{(i)}, & s < t, \\ \mathbf{I}, & s = t. \end{cases}$$

Let the true trajectory satisfy

$$\mathbf{z}_{s+1}^{(i)} = F_{\mathbf{x}_i}(\mathbf{z}_s^{(i)}),$$

and let the linearised surrogate start from the same initial condition $\tilde{\mathbf{z}}_0^{(i)} = \mathbf{z}_0^{(i)}$ *and evolve by*

$$\tilde{\mathbf{z}}_{s+1}^{(i)} = F_{\mathbf{x}_i}(\bar{\mathbf{z}}_s^{(i)}) + J_s^{(i)}(\tilde{\mathbf{z}}_s^{(i)} - \bar{\mathbf{z}}_s^{(i)}).$$

Assume there exist deterministic numbers $M_{t,s}$ *such that*

$$\max_i \|\Phi_{t:s}^{(i)}\| \le M_{t,s} \qquad \textit{for all } 0 \le s \le t.$$

If the surrogate features at time t *admit a separator* $(\mathbf{w}_t, b_t)$ *with margin* $m_t > 0$*,*

$$y_i\big(\mathbf{w}_t^\top \tilde{\mathbf{z}}_t^{(i)} + b_t\big) \ge m_t \qquad \textit{for all } i,$$

then the true features are correctly classified whenever

$$\frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2 \sum_{s=0}^{t-1} M_{t,s+1} D_s^2 < \frac{m_t}{\|\mathbf{w}_t\|}, \qquad D_s := \max_i \|\mathbf{z}_s^{(i)} - \bar{\mathbf{z}}_s^{(i)}\|. \tag{15}$$

In particular, if there exist $\rho_* > 0$ *and* C_t *such that*

$$M_{t,s+1} \le C_t\, \rho_*^{t-1-s} \qquad \textit{for all } s < t,$$

then the sufficient condition

$$\frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2\, C_t \sum_{s=0}^{t-1} \rho_*^{t-1-s} D_s^2 < \frac{m_t}{\|\mathbf{w}_t\|}, \tag{16}$$

also guarantees separability of the true timestep-t *features.*

Proof. Let $\mathbf{e}_s^{(i)} = \mathbf{z}_s^{(i)} - \tilde{\mathbf{z}}_s^{(i)}$. Subtracting the surrogate recurrence from the true recurrence gives

$$\mathbf{e}_{s+1}^{(i)} = J_s^{(i)} \mathbf{e}_s^{(i)} + \mathbf{r}_s^{(i)},$$

where

$$\mathbf{r}_s^{(i)} := F_{\mathbf{x}_i}(\mathbf{z}_s^{(i)}) - F_{\mathbf{x}_i}(\bar{\mathbf{z}}_s^{(i)}) - J_s^{(i)}(\mathbf{z}_s^{(i)} - \bar{\mathbf{z}}_s^{(i)}).$$

By Lemma 10,

$$\|\mathbf{r}_s^{(i)}\| \le \frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2\, \|\mathbf{z}_s^{(i)} - \bar{\mathbf{z}}_s^{(i)}\|^2 \le \frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2 D_s^2.$$

The variation-of-constants formula for time-varying linear recurrences yields

$$\mathbf{e}_t^{(i)} = \sum_{s=0}^{t-1} \Phi_{t:s+1}^{(i)} \mathbf{r}_s^{(i)}.$$

Therefore

$$\max_i \|\mathbf{e}_t^{(i)}\| \le \frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2 \sum_{s=0}^{t-1} M_{t,s+1} D_s^2.$$

If Equation (15) holds, then $\max_i \|\mathbf{e}_t^{(i)}\| < m_t/\|\mathbf{w}_t\|$, so Proposition 11 implies that the same separator correctly classifies the true features. The geometric specialisation Equation (16) follows immediately from the bound $M_{t,s+1} \le C_t \rho_*^{t-1-s}$. □

The remaining step is to package the cocycle envelope into a single finite-horizon quantity.

Definition 13 (Finite-horizon cocycle envelope)**.** *Under the hypotheses of Theorem 12, fix* $\lambda > 0$ *and define*

$$\mathcal{R}_t(\lambda) := \max_i \sup_{(\mathbf{q}_0,\dots,\mathbf{q}_{t-1}) \not\equiv \mathbf{0}} \frac{\Big\| \sum_{s=0}^{t-1} \Phi_{t:s+1}^{(i)} \mathbf{q}_s \Big\|}{\sum_{s=0}^{t-1} \lambda^{\,t-1-s} \|\mathbf{q}_s\|}. \tag{17}$$

This is a weighted input-output norm of the nonautonomous linearised system, and serves as the finite-horizon analogue of the frozen scaled-Kreiss constant.

Lemma 14 (Cocycle envelope controls Jacobian products)**.** *For every $0 \le s < t$,*

$$\max_i \|\Phi^{(i)}_{t:s+1}\| \le \mathcal{R}_t(\lambda)\, \lambda^{\,t-1-s}. \tag{18}$$

Moreover, in the autonomous frozen case $J_u^{(i)} \equiv J_*$ for all i, u, with $\rho_* = \rho(J_*)$, one has

$$\mathcal{R}_t(\rho_*) \le ed\, K_{\rho_*}(J_*). \tag{19}$$

Proof. For Equation (18), fix $s < t$, an index i, and a vector $\mathbf{v} \neq \mathbf{0}$. In Equation (17), choose $\mathbf{q}_u = \mathbf{0}$ for $u \neq s$ and $\mathbf{q}_s = \mathbf{v}$. Then

$$\frac{\|\Phi^{(i)}_{t:s+1}\mathbf{v}\|}{\lambda^{\,t-1-s}\|\mathbf{v}\|} \le \mathcal{R}_t(\lambda).$$

Taking the supremum over $\mathbf{v} \neq \mathbf{0}$ and then the maximum over i proves Equation (18).

For Equation (19), the output of the frozen linear system

$$\boldsymbol{\delta}_{u+1} = J_*\boldsymbol{\delta}_u + \mathbf{q}_u, \qquad \boldsymbol{\delta}_0 = \mathbf{0},$$

at time t is

$$\boldsymbol{\delta}_t = \sum_{s=0}^{t-1} J_*^{\,t-1-s}\mathbf{q}_s.$$

Lemma 9 yields

$$\|\boldsymbol{\delta}_t\| \le ed\, K_{\rho_*}(J_*) \sum_{s=0}^{t-1} \rho_*^{\,t-1-s}\|\mathbf{q}_s\|.$$

Taking the supremum over all nonzero forcing sequences gives Equation (19). □

Corollary 15 (Cocycle-envelope margin criterion)**.** *Under the hypotheses of Theorem 12, if for some $\lambda > 0$,*

$$\frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2\,\mathcal{R}_t(\lambda) \sum_{s=0}^{t-1} \lambda^{\,t-1-s} D_s^2 < \frac{m_t}{\|\mathbf{w}_t\|}, \tag{20}$$

then the true timestep-t features are correctly classified by the same separator.

Proof. By Lemma 14,

$$M_{t,s+1} := \mathcal{R}_t(\lambda)\lambda^{\,t-1-s}$$

satisfies the hypothesis of Theorem 12. Substituting this choice into Equation (15) gives Equation (20). □

Theorem 16 (Lifted resolvent representation of the cocycle envelope)**.** *Let $\mathcal{X}_t = (\mathbb{R}^d)^{t+1}$, and equip the input space $(\mathbb{R}^d)^t$ with the block norm*

$$\|(\mathbf{u}_0, \dots, \mathbf{u}_{t-1})\|_{\ell_{1,2}} := \sum_{s=0}^{t-1} \|\mathbf{u}_s\|,$$

and for each sample i define the lifted operator $\mathcal{T}^{(\lambda)}_{i,t} : \mathcal{X}_t \to \mathcal{X}_t$ by

$$(\mathcal{T}^{(\lambda)}_{i,t}\mathbf{x})_0 = \mathbf{0}, \qquad (\mathcal{T}^{(\lambda)}_{i,t}\mathbf{x})_{s+1} = \lambda^{-1} J_s^{(i)}\mathbf{x}_s \quad (0 \le s \le t-1).$$

Let $\mathcal{B}_t : (\mathbb{R}^d)^t \to \mathcal{X}_t$ inject the input blocks by

$$\mathcal{B}_t(\mathbf{u}_0, \dots, \mathbf{u}_{t-1}) = (\mathbf{0}, \mathbf{u}_0, \dots, \mathbf{u}_{t-1}),$$

and let $\Pi_t : \mathcal{X}_t \to \mathbb{R}^d$ extract the last block, $\Pi_t(\mathbf{x}_0, \dots, \mathbf{x}_t) = \mathbf{x}_t$. Then

$$\mathcal{R}_t(\lambda) = \max_i \left\| \Pi_t (I - \mathcal{T}^{(\lambda)}_{i,t})^{-1} \mathcal{B}_t \right\|_{\ell_{1,2}\to 2}. \tag{21}$$

Moreover, for every $0 < r < 1$,

$$\mathcal{R}_t(\lambda) \le \frac{r}{1-r} \max_i \sup_{|z|=r} \left\| \Pi_t (zI - \mathcal{T}^{(\lambda)}_{i,t})^{-1} \mathcal{B}_t \right\|_{\ell_{1,2}\to 2}. \tag{22}$$

For the block $\ell_{1,2}$ norm used here, Remark 17 immediately below turns this lifted resolvent object into an exact finite-horizon block maximum, which we evaluate numerically in Figure 19.

Proof. Fix i. Since $\mathcal{T}_{i,t}^{(\lambda)}$ is strictly lower block triangular, it is nilpotent, hence

$$(I - \mathcal{T}_{i,t}^{(\lambda)})^{-1} = \sum_{k=0}^{t} (\mathcal{T}_{i,t}^{(\lambda)})^k.$$

For $\mathbf{u} = (\mathbf{u}_0, \dots, \mathbf{u}_{t-1})$, the last block of $(I - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t\mathbf{u}$ is therefore

$$\Pi_t(I - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t\mathbf{u} = \sum_{s=0}^{t-1} \lambda^{-(t-1-s)}\Phi_{t:s+1}^{(i)}\mathbf{u}_s,$$

with the convention $\Phi_{t:t}^{(i)} = \mathbf{I}$. Now substitute $\mathbf{u}_s = \lambda^{t-1-s}\mathbf{q}_s$ in Equation (17). This turns the denominator into $\sum_{s=0}^{t-1} \|\mathbf{u}_s\| = \|\mathbf{u}\|_{\ell_{1,2}}$ and yields Equation (21).

For Equation (22), use the Laurent expansion of the nilpotent resolvent. For $0 < r < 1$, the contour encloses the pole at $z = 0$ but excludes $z = 1$. Since

$$\Pi_t(zI - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t = \sum_{k=0}^{t} z^{-k-1}\Pi_t(\mathcal{T}_{i,t}^{(\lambda)})^k\mathcal{B}_t$$

and $(1-z)^{-1} = \sum_{m\geq 0} z^m$ on $|z| < 1$, the residue at $z = 0$ is $\Pi_t(I - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t$. Hence

$$\Pi_t(I - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t = \frac{1}{2\pi i}\oint_{|z|=r} \frac{\Pi_t(zI - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t}{1-z}\, dz.$$

Taking operator norms, using $|1 - z| \geq 1 - r$ on $|z| = r$, and the contour length $2\pi r$, gives

$$\left\|\Pi_t(I - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t\right\|_{\ell_{1,2}\to 2} \leq \frac{r}{1-r}\sup_{|z|=r}\left\|\Pi_t(zI - \mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t\right\|_{\ell_{1,2}\to 2}.$$

Taking the maximum over i proves Equation (22). □

Remark 17 (Exact evaluation under the block $\ell_{1,2}$ norm). For the block norm used in Theorem 16,

$$\mathcal{R}_t(\lambda) = \max_i \max_{0\leq s<t} \lambda^{-(t-1-s)}\|\Phi_{t:s+1}^{(i)}\|. \tag{23}$$

Proof. For fixed i, write

$$A_i(\mathbf{u}_0, \dots, \mathbf{u}_{t-1}) := \sum_{s=0}^{t-1} \lambda^{-(t-1-s)}\Phi_{t:s+1}^{(i)}\mathbf{u}_s.$$

Then

$$\|A_i(\mathbf{u})\| \leq \sum_{s=0}^{t-1} \lambda^{-(t-1-s)}\|\Phi_{t:s+1}^{(i)}\|\,\|\mathbf{u}_s\| \leq \Big(\max_{0\leq s<t} \lambda^{-(t-1-s)}\|\Phi_{t:s+1}^{(i)}\|\Big)\|\mathbf{u}\|_{\ell_{1,2}}.$$

Hence $\|A_i\|_{\ell_{1,2}\to 2}$ is at most the displayed maximum. Equality is attained by choosing a single nonzero block $\mathbf{u}_s$ aligned with a top singular vector of the maximizing matrix. Taking the maximum over i and using Equation (21) gives Equation (23). □

For fixed finite horizon, the cocycle envelope is therefore numerically accessible: one can either evaluate the exact block maximum in Equation (23) or upper-bound it via the corrected lifted resolvent contour in Equation (22).

Corollary 18 (Finite-horizon lifted-resolvent margin criterion)**.** *Under the hypotheses of Theorem 12, if for some* $\lambda > 0$ *and* $0 < r < 1$*,*

$$\frac{L_{\tanh}}{2}\,\|\mathbf{W}\|^2\,\frac{r}{1-r}\,\max_i \sup_{|z|=r}\left\|\Pi_t(zI-\mathcal{T}_{i,t}^{(\lambda)})^{-1}\mathcal{B}_t\right\|_{\ell_{1,2}\to 2}\sum_{s=0}^{t-1}\lambda^{\,t-1-s}D_s^2 < \frac{m_t}{\|\mathbf{w}_t\|}, \tag{24}$$

then the true timestep-t *features are correctly classified by the same separator.*

Proof. Combine Equation (22) with Equation (20). □

To show that this appendix quantity is not merely formal, we numerically evaluate the finite-horizon cocycle envelope on the actual real-ESN/tanh cross-system sweep. Across the full set of 27 runs (3 seeds $\times$ 9 spectral radii), we form the true Jacobian cocycle along 128 held-out MNIST test trajectories and compute the exact $t = T = 40$ block maximum from Equation (23). Figure 19 reports both the raw envelope $\widehat{\mathcal{R}}_{40}(1)$ and the scaled envelope $\widehat{\mathcal{R}}_{40}(\rho(\mathbf{W}))$. The raw envelope grows sharply with spectral radius and correlates strongly with the measured early-window advantage ($r = 0.87$) and lower terminal accuracy ($r = -0.88$). After removing the per-radius mean, the residual correlation between $\log_{10}\widehat{\mathcal{R}}_{40}(1)$ and the early-window gap remains positive ($r = 0.79$), indicating that the envelope explains more than spectral radius alone. By contrast, the scaled envelope stays $O(1)$ and decreases from about 2.13 at $\rho = 1.0$ to about 1.55 at $\rho = 1.6$, quantifying how tanh saturation attenuates the Jacobian cocycle relative to the nominal radial factor even as the raw finite-horizon amplification increases.

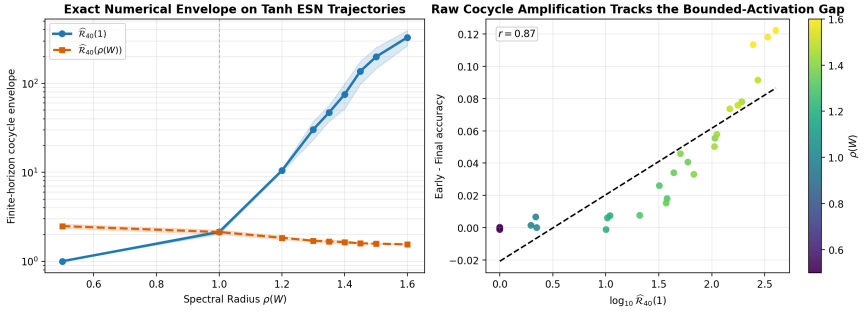


Figure 19: **Concrete numerical estimates of the cocycle envelope on the tanh ESN branch.** *Left:* the raw finite-horizon envelope $\widehat{\mathcal{R}}_{40}(1)$ grows across the real-ESN sweep, while the scaled envelope $\widehat{\mathcal{R}}_{40}(\rho(\mathbf{W}))$ remains $O(1)$ and gently decreases. Shading shows ± 1 std across three seeds. *Right:* $\log_{10}\widehat{\mathcal{R}}_{40}(1)$ correlates with the Early$-$Final gap across all 27 tanh runs ($r = 0.87$), with positive within-radius residual correlation ($r = 0.79$).

Proposition 11 is the frozen-Jacobian special case of Corollary 15, with $\mathcal{R}_t(\rho_*) \le ed\,K_{\rho_*}(J_*)$. A full nonlinear pseudospectral theorem would require computable lifted-resolvent or output-weighted pseudospectral bounds for $\mathcal{T}_{i,t}^{(\lambda)}$; here we evaluate $\mathcal{R}_t(\lambda)$ directly on the bounded tanh branch and use the simpler frozen-radial proxy for the main complex-modReLU sweep.

Broader Impact (Supplementary)

This work studies when terminal representations degrade while usable prefixes remain. Better readout timing could improve reliability in iterative and physical computing, including analog reservoirs and photonic circuits. A potential downside is deploying trajectory-aware stopping heuristics without calibration in safety-relevant settings. For example, a photonic inference accelerator could stop early and appear confident even though its trajectory has left the calibrated regime, masking silent degradation. These methods are diagnostics, not deployment guarantees; reliable use requires checking the proxy against task accuracy in the target operating conditions.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction frame the paper as a discovery paper about endpoint-readout failure, an accuracy-defined grace period, practical trajectory-aware decoders, and a cross-system taxonomy. These claims are supported by experiments in Sections 3 through 6.4, Appendix O, and theory in Section 4.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 explicitly discusses limitations: fixed dynamical substrates, frozen-feature rather than end-to-end high-accuracy CIFAR-10, single-laptop-GPU scaling limits, controlled trained-system gain sweeps spanning LeakyReLU and LSTM, three-seed Tiny ImageNet horizon comparisons with frozen ConvNeXt features, checkpoint-selection sensitivity, and the looseness of the theoretical grace period bound.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Propositions 4 and 6, together with Lemma 8, are stated with explicit assumptions and include complete proofs in the main text or Appendix U.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The manuscript specifies the experimental settings that affect the main claims: system parameters (d, n, T, modReLU bias, spectral radius values, magnitude clipping), probe details (`RidgeClassifierCV`), dataset splits (MNIST/FashionMNIST, 5K train / 5K test; frozen-feature CIFAR-10, 10K train / full 10K test; row-sequential CIFAR-10 LSTM, 10K train / 2K test), the use of ImageNet-pretrained ResNet-18 weights for the frozen-feature CIFAR-10 check, GRACE hyperparameters (τ, α), fixed defaults, and random seed counts; see Section 3, Appendix A, and Appendices O and P.

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

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

   Answer: [No]

   Justification: The original conference submission included supplementary code and saved results. This preprint provides the full experimental protocols and uses publicly available benchmarks and pretrained backbones, but does not yet link a public code archive for the extended experiments.

6. **Experimental setting/details**

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

   Answer: [Yes]

Justification: All experimental details are in Section 3 (main paper) and Appendices A, B, O and P (appendix), including data splits, early-window sizes, fixed defaults, the reference-radius audit, probe type, GRACE hyperparameters, system construction, the trained LeakyReLU setup (MNIST split, terminal cross-entropy, Adam, learning rate, weight decay, schedule, batch size, epochs, and gradient clipping), and the trained row-sequential CIFAR-10 LSTM setup plus timestep-audit probe protocol.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main reservoir results report mean $\pm$ standard deviation across 10 seeds; the trained LeakyReLU study reports mean across 5 seeds and the sequential CIFAR-10 LSTM check reports mean $\pm$ standard deviation across 3 seeds because each operating point requires full end-to-end training. Paired Wilcoxon signed-rank tests with Holm-Bonferroni correction are reported for key comparisons, and Cohen's d is interpreted alongside raw variance and a rank-based Cliff's δ check. The modern recursive panel reports sample standard deviations across three paired seeds without significance tests. See Section 3, Appendix C, and Appendices O and P.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix A reports the compute worker type, GPU memory, representative per-experiment runtimes from saved logs/results, an overall compute estimate for the reported runs, the one-time ResNet-18 feature extraction cost, a brief disclosure that exploratory pilot runs required additional compute, and the limitation that the small compute footprint does not establish scaling behavior. Section 6.4 and Appendix S reports the additional Tiny ImageNet model scale, RTX 5080 Laptop GPU runtime, and allocated/reserved memory.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This work is foundational research on dynamical systems for machine learning. It uses standard, publicly available benchmark datasets (MNIST, FashionMNIST, CIFAR-10, Tiny ImageNet-200) and publicly available pretrained ResNet-18 and ConvNeXt-Tiny weights only as a frozen feature extractor, and raises no ethical concerns.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper includes an unnumbered supplementary Broader Impact discussion at the end of the supplementary material that discusses both potential benefits (more reliable readout in physical and iterative systems) and a concrete downside risk (miscalibrated deployment in safety-relevant settings).

11. **Safeguards**

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

    Answer: [N/A]

Justification: This paper does not release pre-trained models or scraped datasets. The work analyses fixed dynamical systems on standard benchmarks and uses torchvision's ImageNet-pretrained ResNet-18 and ConvNeXt-Tiny weights only as a frozen feature extractor for one CIFAR-10 check.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Appendices A and P lists the benchmark sources and usage terms for MNIST, FashionMNIST, CIFAR-10, Tiny ImageNet-200, and the torchvision ImageNet-pretrained ResNet-18 and ConvNeXt-Tiny weights, while the manuscript cites the original datasets, scikit-learn, and prior methods.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper studies a readout regime and introduces a diagnostic readout variant (GRACE), but does not release a new dataset, benchmark, or model asset as part of the submission.

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

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

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

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were not used as an important, original, or non-standard component of the core methodology. Any LLM use was limited to writing, editing, formatting, or submission-preparation assistance and did not impact the core methodology, scientific rigor, or originality of the research.