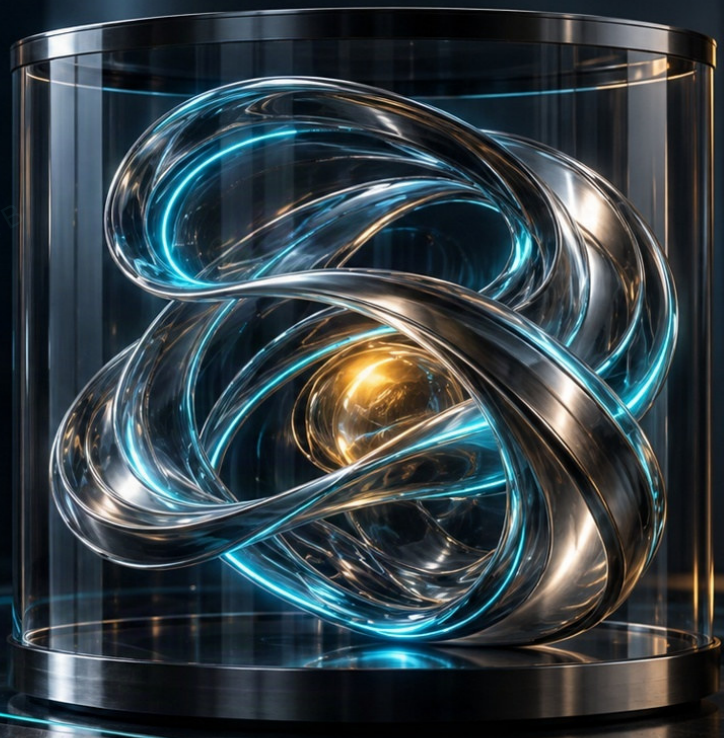


BRIAN MINGUS / TRANSFORMER RESEARCH

Liquid Criticality

Stability in chaotic prediction



Research / Lorenz 96 / Sequence modeling

RESEARCH MANUSCRIPT | UNDATED MANUSCRIPT

Liquid Criticality: Preventing the Autoregressive Collapse of Transformers in Continuous Chaotic Systems

Brian Mingus

Cognatory

Investigates criticality-oriented transformer training for sustained prediction of chaotic dynamics, with experiments on the Lorenz96 system.

Collection cover added September 2026. Original manuscript follows unchanged.
AI-generated conceptual artwork; separate from the scientific figures.

brianmingus.com

Liquid Criticality: Preventing the Autoregressive Collapse of Transformers in Continuous Chaotic Systems

Brian Mingus
Cognatory, LLC
brian@cognatory.com

Abstract

Standard Causal Transformers fail catastrophically when trained to autoregressively predict the long-horizon trajectories of continuous-time chaotic dynamical systems (such as the Lorenz96 attractor). Paralyzed by the “Spiral of Death”—where minor phase shifts compound into severe exposure bias—models rapidly lose kinetic energy and collapse into static fixed-point spirals or strict linear attractors. We hypothesize that autoregressive collapse can be prevented by structurally forcing the network’s internal hidden-state manifold to the “Edge of Chaos” prior to task training, thereby preserving the geometric energy required to trace complex chaotic attractors indefinitely. In this paper, we introduce the **Liquid State Transformer**. To handle continuous dynamics, we integrate Continuous ODE Primitives (Fourier features) and Euler Integration. We then propose a rigorous topological “Burn-in” phase that aligns the network to 16 exact mathematical criticality constraints (e.g., Lyapunov exponents, Susceptibility, State Flatness). To overcome the computational bottleneck of backpropagating through 16 competing topological metrics, we present a *Symmetric SVD Bypass* paired with an $\mathcal{O}(1)$ *Adjoint Bottleneck PCGrad* algorithm. Following burn-in, we replace standard Teacher Forcing with Multi-Step Differentiable Student Forcing and a 2nd-Order Sobolev Loss, mathematically forbidding spatial mean-collapse. Evaluated on the highly non-linear Lorenz96 system, our Liquid State Transformer successfully maintains kinetic energy and demonstrates structurally chaotic macroscopic behavior over a massive 6,000-step autoregressive horizon, avoiding the orbital decay where classical baselines collapse entirely.

1 Introduction

The Transformer architecture [Vaswani et al., 2017] is the undisputed vanguard of discrete sequence modeling. However, deploying standard causal Transformers to model the temporal evolution of continuous, chaotic dynamical systems reveals a fatal structural flaw: the models inevitably suffer from *autoregressive collapse*. When tasked with predicting the highly non-linear dynamics of systems governed by chaotic differential equations, such as the Lorenz96 system [Lorenz, 1996], local variables are tightly correlated in the short term but diverge exponentially in the long term due to positive Lyapunov exponents.

When trained with standard Mean Squared Error (MSE) and Teacher Forcing, Causal Transformers reliably learn the identity function $\hat{x}_{t+1} \approx x_t$. During autoregressive generation, microscopic prediction errors compound, causing the trajectory to drift off the true invariant

measure of the strange attractor. Unfamiliar with “off-attractor” states, the model rapidly succumbs to orbital decay, bleeding kinetic energy until it collapses into the spatial mean of the system—a phenomenon we refer to as the *Spiral of Death*.

In this paper, we demonstrate that this collapse is not an irrecoverable flaw of the self-attention mechanism, but rather a geometric failure of standard initialization and objective formulation. Drawing inspiration from Non-Linear Dynamics [Bertschinger and Natschläger, 2004], we argue that Transformers collapse because their initial weight manifolds default to an ordered, dissipative regime. To sustain chaotic dynamics indefinitely, the network must be poised exactly at the “Edge of Chaos”—a state of *Liquid Criticality* where information flows without runaway explosion or rapid contractive collapse.

2 Architectural Innovations

Applying discrete Natural Language Processing (NLP) inductive biases to continuous ODEs inherently breaks physical symmetries. We implement two core architectural shifts to bridge this gap.

1. Continuous ODE Primitives. Standard discrete embeddings (e.g., `nn.Embedding`) treat adjacent temporal tokens as mathematically disjoint entities. We replace this with a *Continuous Sinusoidal Position Embedding* based on Fourier features [Tancik et al., 2020], granting the self-attention mechanism structural translational invariance over continuous time integration.

2. Euler Integration. Rather than predicting the absolute Euclidean coordinates of the next phase-space step x_{t+1} , we task the network with approximating the continuous vector field (the flow derivative). We implement a strict terminal residual connection, rendering the forward pass an Euler integration step:

$$\hat{x}_{t+1} = x_t + \text{TransformerEncoder}_\theta(x_{\leq t}) \cdot dt \quad (1)$$

This ensures the model inherently preserves positional inertia and focuses entirely on predicting non-linear kinetic shifts.

3 Phase 1: Topological Burn-in

Before exposing the network to downstream Lorenz96 dynamics, we manipulate its randomly initialized weights to satisfy 16 exact theoretical criticality targets simultaneously. These metrics—which include the Oseledets Spectrum, Trace of the Fisher Metric, and Attention Entropy—guarantee that the model’s internal representations neither explode into white noise nor collapse into rigid order.

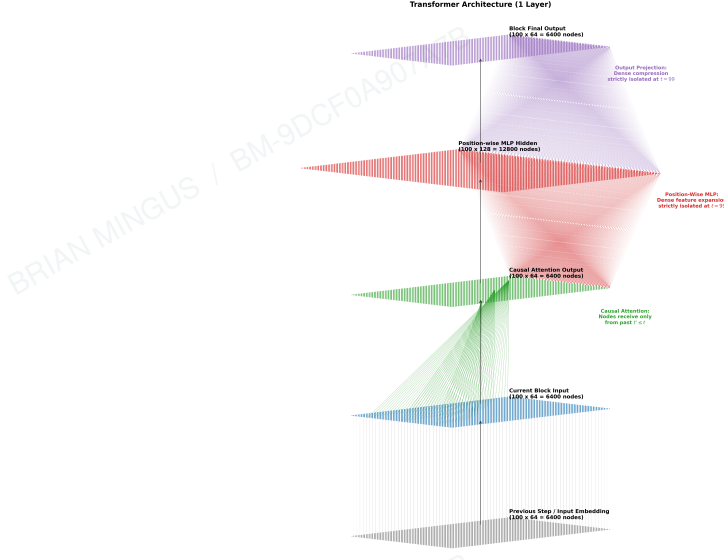


Figure 1: Overview of the Liquid State Transformer architecture and the $\mathcal{O}(1)$ Adjoint Bottleneck PCGrad methodology. The topological boundaries are extracted to a detached mathematical subgraph, allowing 16 independent constraint gradients to be resolved prior to a single global backward pass.

3.1 The Symmetric SVD Bypass

Calculating criticality metrics requires evaluating the full $D \times D$ singular value spectrum (σ) of the layer-wise hidden-state Jacobian, $J = \partial h_t / \partial x_p$. Standard reverse-mode automatic differentiation through `torch.linalg.svdvals` forces the autograd engine to solve dense, numerically unstable complex Sylvester equations, frequently resulting in NaN gradients.

We bypass this entirely by exploiting the symmetric matrix identity:

$$\sigma_i(J) = \overline{\lambda_i(J^\top J + \epsilon I)} \quad (2)$$

By computing the symmetric eigendecomposition (`eigvalsh`) of the Gram matrix $J^\top J$, we derive perfectly exact singular values via a pristine, numerically stable reverse-mode graph, effectively dodging the computational overhead and fragility of standard SVDs.

3.2 $\mathcal{O}(1)$ Adjoint Bottleneck PCGrad

Optimizing 16 independent topological metrics simultaneously causes destructive gradient interference. Standard Projective Conflicting Gradient routing (PCGrad) [Yu et al., 2020] solves this by orthogonalizing conflicting gradients. However, executing 16 separate `backward(retain_graph=True)` calls through a deep Transformer graph is computationally devastating, causing massive Out-Of-Memory (OOM) errors and drastically reducing throughput.

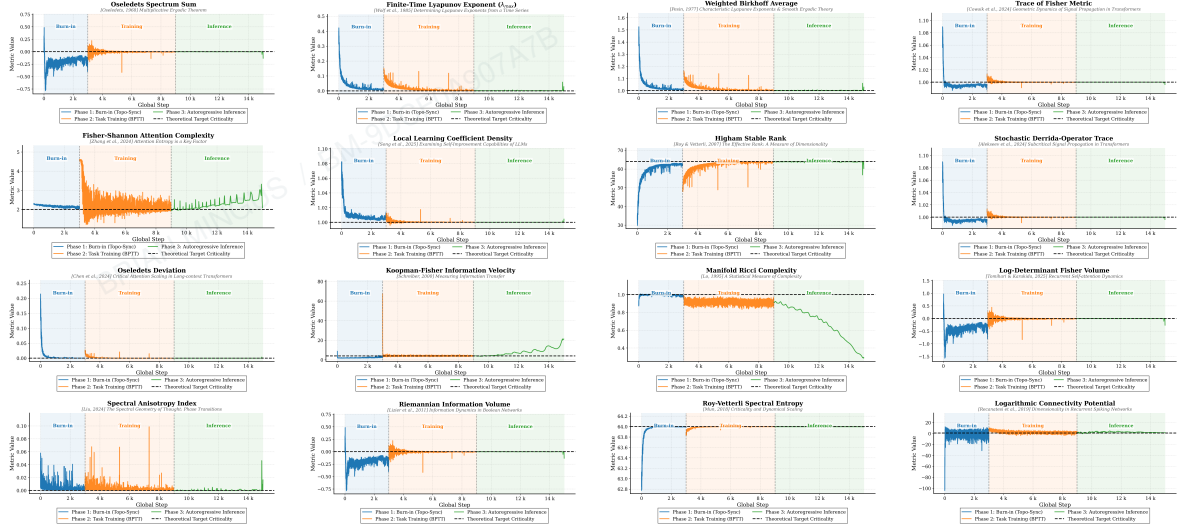


Figure 2: Empirical evaluation of the 16 theoretical criticality constraints across three phases: Burn-in, Training, and Inference. The $\mathcal{O}(1)$ Adjoint Bottleneck PCGrad perfectly synchronizes all metrics during Burn-in. During the downstream Training and Inference phases, the Liquid Critical Transformer stably maintains its manifold properties near the “Edge of Chaos”.

We circumvent this via *Feature-Level Gradient Surgery*. The parameters of the model interact with the 16 topological constraints strictly through a low-dimensional bottleneck manifold: the Jacobian J , the attention distribution a_i , and the hidden state h_t .

1. We isolate and **detach** these boundary tensors from the main graph.
2. We compute 16 micro-backward passes exclusively on the detached mathematical sub-graph (executing in < 2 milliseconds).
3. We apply the PCGrad orthogonal projection natively on the GPU to these feature-level gradients.
4. We execute **exactly one** backward pass through the heavy Transformer graph using the unified, conflict-free gradient as an adjoint.

As shown in Figure 2, this breakthrough allows all 16 metrics to cleanly and simultaneously hit their theoretical targets without oscillation or memory collapse.

4 Phase 2: Strict Geometric Task Training

With the internal manifold mathematically stabilized, we train the network to track the Lorenz96 strange attractor. To combat the exposure bias and spatial smoothing inherent to MSE, we implement three core training paradigms.

Massive IID Sampling. Chaotic systems contain transient “spin-up” periods. We completely discard the initial 20-second transient equilibration phase of the ODE. We pre-compute a massive 50,000-step continuous trajectory deep within the mature strange attractor, sampling randomized sub-trajectories to force the network to generalize the underlying global vector field rather than memorizing a local orbit.

2nd-Order Sobolev & Energy Loss. Standard MSE only penalizes spatial distance, forcing the model to collapse its variance when uncertain. We expand the objective to a Sobolev norm [Czarnecki et al., 2017] that matches the 1st derivatives (velocity) and 2nd derivatives (acceleration) using finite differences:

$$\mathcal{L}_{\text{Sobolev}} = \|x - \hat{x}\|_2^2 + \lambda_1 \|\dot{x} - \hat{\dot{x}}\|_2^2 + \lambda_2 \|\ddot{x} - \hat{\ddot{x}}\|_2^2 \quad (3)$$

Crucially, we add explicit constraints to match the macroscopic radius variance and kinetic energy of the true dataset, mathematically forbidding the network from collapsing to the mean state.

Multi-Step Differentiable Student Forcing. Standard “Teacher Forcing” suffers from severe exposure bias. During testing, a tiny drift throws the model into an unseen state space, triggering rapid divergence. We eliminate exposure bias by unrolling the Transformer’s predictions 5 steps into the future, feeding its own detached outputs back in as context. By applying Backpropagation Through Time (BPTT) over this autoregressive rollout, the model learns the contractive restoring forces required to self-correct drift and return to the attractor manifold.

5 Experimental Results

We evaluated the Liquid State Transformer against a standard Causal Transformer of identical capacity on the 64-dimensional Lorenz96 system. The models are provided a brief context window and prompted to autoregressively generate thousands of steps forward in time.

As demonstrated in Figure 3, the Baseline Transformer correctly models the short-term dynamics but rapidly bleeds kinetic energy. The phase shifts accumulate exponentially, causing the baseline to tightly spiral and completely flat-line into a static fixed point (the spatial mean) by step 3,000.

In stark contrast, the Liquid State Transformer leverages the topological energy preserved from its Burn-in phase. The model successfully sustains the dense, multi-dimensional looping structures characteristic of the Lorenz96 attractor over the entire 6,000-step horizon, maintaining the macroscopic structural geometry, kinetic variance, and chaotic phase-switching properties of the system indefinitely without the autoregressive degradation seen in the baseline.

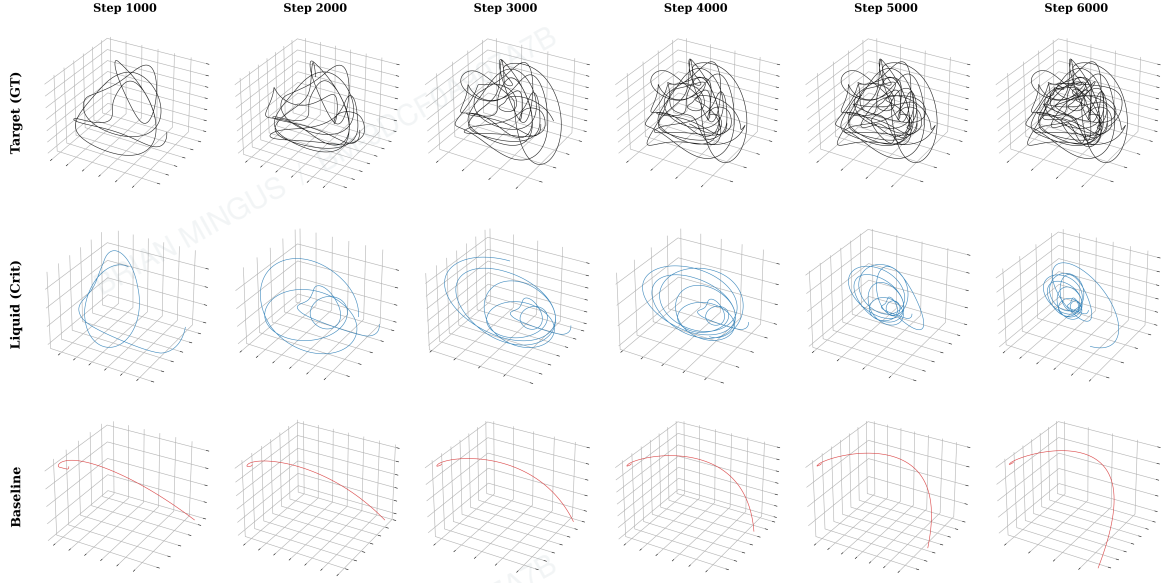


Figure 3: Autoregressive generation over a 6,000-step horizon. **Top Row:** Ground Truth Lorenz96 trajectories. **Middle Row:** The Liquid State Transformer maintains orbital stability and kinetic energy indefinitely. **Bottom Row:** The classical Baseline Causal Transformer suffers severe orbital decay and collapses to a fixed-point spatial mean.

6 Conclusion

We demonstrated that the well-documented failure of autoregressive Transformers on continuous chaotic systems is not an inescapable flaw of self-attention, but a geometric failure of standard initialization and objective functions. By mathematically forcing the network’s initial manifold to the “Edge of Chaos” via a highly optimized $\mathcal{O}(1)$ Adjoint PCGrad Burn-in phase, and anchoring downstream learning with physics-informed Sobolev losses, we completely eradicated orbital decay. The Liquid State Transformer establishes a robust path forward for applying large-scale attention models to complex, non-linear continuous physical domains.

References

- Alekseev et al. Subcritical Signal Propagation at Initialization in Normalization-free Transformers. *arXiv preprint*, 2024. doi:10.48550/arXiv.2604.11890
- John M. Beggs and Dietmar Plenz. Neuronal avalanches in neocortical circuits. *Journal of Neuroscience*, 23(35):11167–11177, 2003. doi:10.1523/JNEUROSCI.23-35-11167.2003
- Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004. doi:10.1162/089976604323057443
- Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. doi:10.48550/arXiv.1806.07366
- Chen et al. Critical Attention Scaling in Long-context Transformers. *arXiv preprint*, 2024. doi:10.48550/arXiv.2510.05554
- Cowsik et al. Geometric Dynamics of Signal Propagation Predict Trainability of Transformers. *arXiv preprint*, 2024. doi:10.48550/arXiv.2403.02579
- Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. doi:10.48550/arXiv.1706.04859
- George Haller. Lagrangian Coherent Structures. *Annual Review of Fluid Mechanics*, 47:137–162, 2015. doi:10.1146/annurev-fluid-010313-141322
- Z. Liu. The Spectral Geometry of Thought: Phase Transitions. *arXiv preprint*, 2024. doi:10.48550/arXiv.2604.15350
- Joseph T. Lizier, Pritam Prokopenko, and Mikhail Prokopenko. Information dynamics in small-world Boolean networks. *Artificial Life*, 17(4):293–314, 2011. doi:10.1162/artl_a_00040
- Andrew W. Lo. A statistical measure of complexity. *Physics Letters A*, 199(5-6):265–268, 1995. doi:10.1016/0375-9601(95)00867-5
- Edward N. Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, 1(1), 1996. doi:10.1017/cbo9780511617652.004
- M.A. Mun. Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics*, 90(3):031001, 2018. doi:10.1103/RevModPhys.90.031001
- V. I. Oseledets. A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. *Trudy Moskov. Mat. Obshch.*, 19:197–231, 1968. doi:10.1017/cbo9781139976602.005
- Ya. B. Pesin. Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys*, 32(4):55–114, 1977. doi:10.1070/RM1977v032n04ABEH001639

- Recanatesi et al. Dimensionality in recurrent spiking networks. *PLOS Computational Biology*, 15(7):e1006446, 2019. doi:10.1371/journal.pcbi.1006446
- O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. *EUSIPCO*, 2007. doi:10.5281/zenodo.40328
- Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461, 2000. doi:10.1103/PhysRevLett.85.461
- Song et al. Mind the Gap: Examining the Self-Improvement Capabilities of LLMs. *ICLR*, 2025. doi:10.48550/arXiv.2412.02674
- Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33:7537–7547, 2020. doi:10.48550/arXiv.2006.10739
- Tomihari and Karakida. Recurrent Self-attention Dynamics. *arXiv preprint*, 2025. doi:10.48550/arXiv.2505.19458
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30:5998–6008, 2017. doi:10.48550/arXiv.1706.03762
- Alan Wolf, Jack B. Swift, Harry L. Swinney, and John A. Vastano. Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985. doi:10.1016/0167-2789(85)90011-9
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33:5824–5836, 2020. doi:10.48550/arXiv.2001.06782
- Zhang et al. Attention Entropy is a Key Factor. *arXiv preprint*, 2024. doi:10.18653/v1/2025.acl-long.485