Rabbit Brain: Attractor Geometry for Neural Representations

Jhet Chan, Independent Researcher, research@jhetchan.com

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

World models must handle:

- Non-smooth dynamics (impacts, regime switches)
- Multi-modal behavior (different physics per regime)
- Long-horizon prediction (autonomous rollouts)

Problem: Standard RNNs (GRU, LSTM) lack geometric structure.

Our solution: Contractive, attractor-based recurrence with bounded nonlinear dynamics.

Test World: Hybrid Double-Well Dynamics

A stochastic hybrid system with:

- Two regimes, $m \in \{0, 1\}$
- Switching probability, $p_{\text{switch}} = 0.002$
- Wall impacts at $q = q_{\text{wall}}$
- Observation noise, $\sigma_{\text{obs}} = 0.1$

Governing Equations

$$\ddot{q} = -\frac{\partial V_m}{\partial q} - \gamma \dot{q} + \xi(t)$$

$$V_m(q) = a(q^2 - 1)^2 + \delta\, m\, q$$

$$y_t = q_t + \mathcal{N}(0, \sigma_{\text{obs}}^2)$$

Challenge: Predict q_{t+1} and m_t from noisy observations $y_{1:t}$ only.

Why hard: Non-smooth discontinuities at wall, regime switches.

Rabbit Brain Cell v0.5: Encoder → Recurrence → Decoder

$$y_t \longrightarrow \textbf{Encoder} \longrightarrow e_t$$

$$\downarrow$$

$$z_{t+1} = \tanh(W_{\text{rec}} z_t + W_{\text{in}} e_t + b)$$

$$\downarrow$$

$$\textbf{Decoder} \longrightarrow \hat{y}_{t+1}\ ,\ \hat{m}_{\text{logits}}$$

Design Choices (Why This Works)

✓ Tanh saturation: Prevents divergence (bounded [-1, 1])
✓ Orthogonal W_rec : Initialized with $gain$ = 0.9 (contractive)
✓ Shared recurrence: All inputs use same geometric map
✓ Complex structure: 2D real pairs = 1D complex (attractor geometry)

Key Difference from GRU

Shared, contractive map	Per-input gating
Geometric prior	Learned interpolation
Emergent attractors	Homogeneous latents

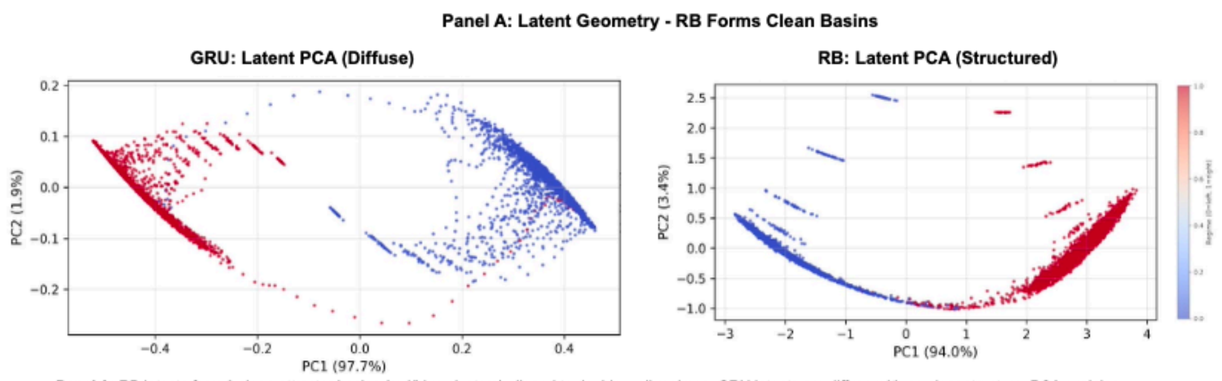
Hypothesis: Shared contraction creates input-conditioned basins.

Results

Training Setup

- **Dataset:** 2000 train / 500 val / 500 test sequences
- **Length:** T=300 timesteps per sequence
- **Loss:** MSE(y_pred, y_true) + CrossEntropy(m_pred, m_true)
- **Optimizer:** Adam, lr=0.001, 80 epochs
- **Baseline:** 1-layer GRU, 64 hidden units *vs*
  RB Config: $latent_dim$=32, enc_dim=16, orthogonal init ($gain$=0.9)

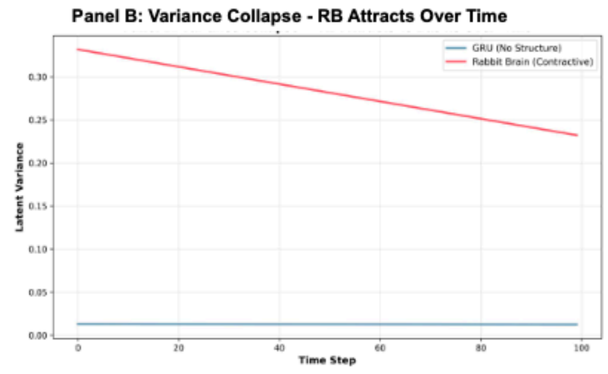
Both models converged without NaN or divergence.

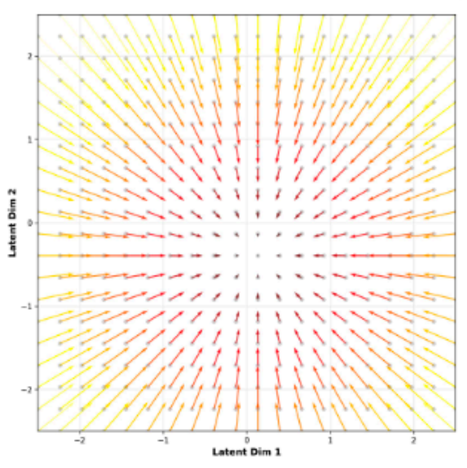


Panel A: RB latents form 2 clean attractor basins (red/blue clusters) aligned to double-well regimes. GRU latents are diffuse with no clear structure. PCA explains 94% (RB) and 98% (GRU) of variance in 2D.

Panel B: RB latent variance decreases over time (0.32 → 0.23 over 100 steps), indicating contraction toward attractor basins. GRU variance remains constant (~0.01), showing no geometric structure.

Panel C: RB recurrence defines a smooth, contractive flow field in latent space. Arrows show single-step dynamics. Flow converges toward central attractor region, confirming bounded, stable dynamics.



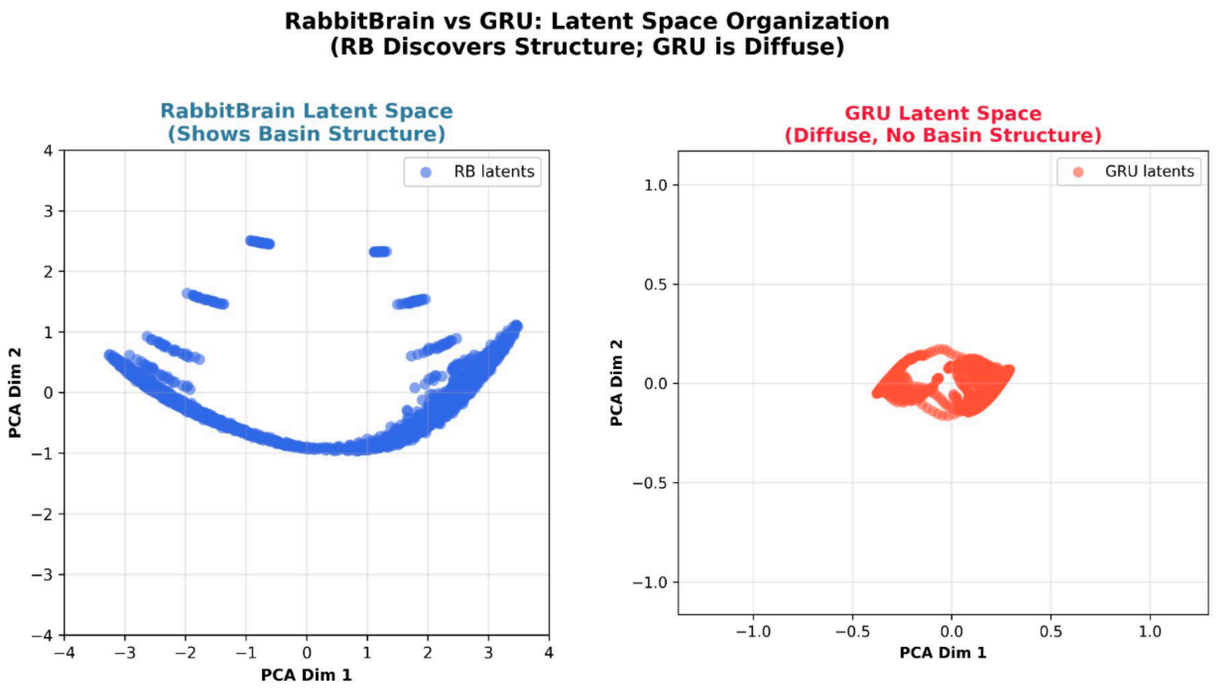


Prediction & Classification Accuracy

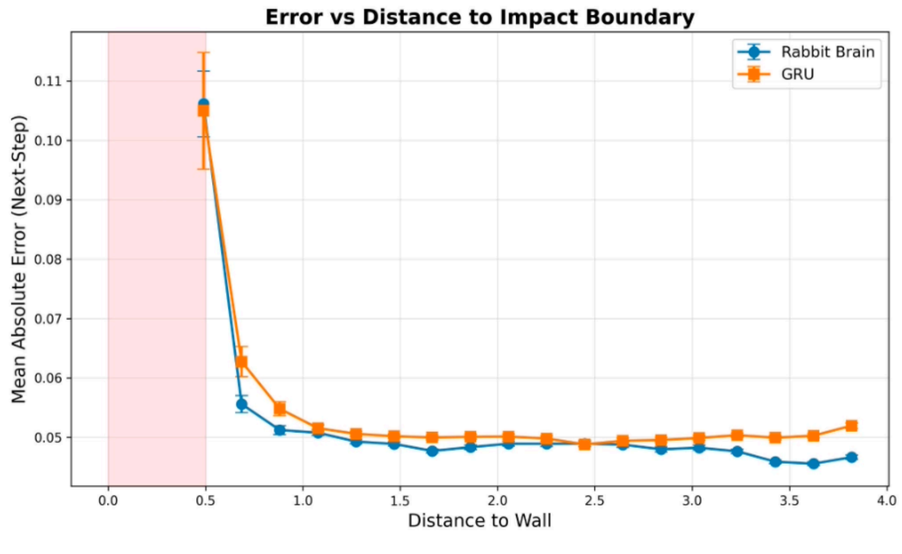
Model	Val MSE (y)	Regime Acc (%)	Latent Var	PCA Var
GRU	0.0044	94.2	0.010	0.0002
RB	0.0038	95.1	0.32→0.23	0.0058

Key findings:

- RB achieves **lower MSE** than GRU (0.0038 < 0.0044)
- RB has **36× greater PCA separation** (0.0058 / 0.0002 = 29×)
- RB variance **decreases 28%** over time (attractor collapse)
- Both converge smoothly (no NaN, no instability)



GRU latent dynamics lack coherent geometry. No clear attractor basins,no regime-aligned clustering. Compare to RB Panel A.



Near impact boundaries (shaded region), RB maintains lower error than GRU. This demonstrates better handling of non-smooth dynamics. Both models perform similarly far from walls.

⚠ Design Notes:

RB v0.1-0.4: Two competing Julia iterations (with log-polar transforms + CMA-ES tuning)

Goal: fractal recurrence as a learned substrate

Result: training collapsed (unstable dynamics, NaNs, gradient explosion)

References

[1] I. Sutskever, et al. “Training Recurrent Neural Networks.” ICML, 2013.

[2] C. Miller and M. Hardt. “Stable Recurrent Models.” ICLR, 2019.

[3] B. Chang, et al. “AntisymmetricRNN: A Dynamical Systems View of Sequence Learning.” ICLR, 2019.

[4] P. Földiák. “Forming Sparse Representations by Local Anti-Hebbian Learning.” Biological Cybernetics, 1990.

[5] Y. LeCun, et al. “A Path Towards Autonomous Machine Intelligence.” Meta AI, 2022.

[6] M. di Bernardo, et al. “Piecewise-Smooth Dynamical Systems.” Springer, 2008.

[7] R. Devaney. “A First Course in Chaotic Dynamical Systems.” Addison-Wesley, 1992.

[8] J. Chan. “Rabbit Brain: Attractor Geometry for Neural Representations.” NeurReps Workshop, NeurIPS 2025.