NEXUS-1: A Modular Foundation Model Architecture — Design and Empirical Validation

Ronisky Lodenyo
Openwist AI Research
research@openwist.com · openwist.com/research/project-cortex
July 2026, revised November 2026

Abstract

We present NEXUS-1 (Neural EXponential Unified System v1), a modular foundation model architecture that replaces quadratic Transformer self-attention with a Mamba state-space sequence model, liquid time-constant (LTC) units, hierarchical predictive coding and a three-tier memory system. Three experiments are reported. EXP-002 shows that a modern Hopfield episodic memory achieves above 90% recall at capacities more than 160 times the classical bound, consistent with O(exp(D/2)) theory. EXP-004b evaluates the SPINAL fast-path module: SpinalLNN achieves the highest reward on continuous Pendulum-v1 (-1213 versus -1514 for a GRU and -1677 for an MLP) with matched parameter counts and three random seeds, and meets the target of under 2 ms CPU inference, while ranking last on discrete CartPole-v1. This task dependence supports the two-path design. EXP-005 evaluates the SOMA consolidation engine: combined elastic weight consolidation (EWC) regularisation and surprise-weighted experience replay retain 98.1% of Task-A (MNIST) accuracy after training on Task-B (Fashion-MNIST), against 28.4% with no protection. Replay alone accounts for nearly all of the protective effect (97.4% retention), with EWC adding a small but consistent margin. Two experiments remain.

1. Introduction

Self-attention in the Transformer [1] has a cost that grows quadratically with sequence length, which limits its use on long-horizon tasks. NEXUS-1 replaces each Transformer component with a linear-time or constant-time alternative: state-space sequence modelling [5], continuous-time recurrent units [4, 6], predictive coding [2, 10], associative memory [7] and continual learning [3, 12, 13].

Contributions. (1) The complete NEXUS-1 architecture. (2) Empirical validation of O(exp(D/2)) Hopfield capacity. (3) An analysis of the task dependence of LTC units, which supports the two-path design. (4) Empirical validation of EWC plus replay for continual learning, with a decomposition of the contribution of each mechanism.

2. Related work

2.1 State-space models

Mamba [5] adds input-dependent selectivity to structured state-space models [8] and matches Transformer quality at O(n) cost.

2.2 Associative memory

Ramsauer et al. [7] derive O(exp(D/2)) capacity for modern Hopfield networks with one-step softmax retrieval, which is equivalent to dot-product attention over a fixed external memory bank.

2.3 Liquid time-constant networks

Hasani et al. [6] introduce LTC networks. The input-dependent time constant tau(x,h) produces adaptive temporal dynamics that have been shown to work well for sensorimotor control.

2.4 Predictive coding and continual learning

Predictive coding [10] and free-energy formulations [2] describe layered models that iteratively minimise prediction error. EWC [12] protects important weights using Fisher information, and Tadros et al. [13] show that interleaved replay reduces forgetting by 40–80% in artificial networks. The consolidation engine in NEXUS-1 combines both.

---

3. NEXUS-1 architecture

NEXUS-1 processes all input through seven functional components. Every modality is projected into a shared latent space, after which two processing paths run in parallel: a fast path and a deliberative path. A routing gate decides which path's output to use, and an offline consolidation process periodically reorganises long-term memory without interrupting the live system.

HERMES maps text, image, audio and sensor input into the same R^D space, so that no modality is privileged. SPINAL is the fast path: a small stack of continuous-time recurrent units producing sub-2 ms responses for time-critical control. CORTEX is the deliberative path: stacked predictive-coding layers that settle iteratively toward a low-error interpretation of the input. MNEMON supplies CORTEX with episodic and semantic context retrieved from memory. THALAMUS routes between the two paths using an uncertainty signal. SOMA consolidates experience offline, moving episodic memory into the semantic store. The component names are identifiers only.

Component	Function	Architecture	Cost
HERMES	Multimodal encoding	Per-modality encoders	O(n)
SPINAL	Fast path	2x LTC units	O(1)
CORTEX	Deliberative path	L x RPSS layers	O(n)
MNEMON	Episodic and semantic memory	Hopfield + GNN	O(Nn)
SOMA	Offline consolidation	Replay + EWC	Offline
THALAMUS	Routing between paths	Uncertainty-based gate	O(1)

Table 1. NEXUS-1 components, their function, architecture and computational cost.

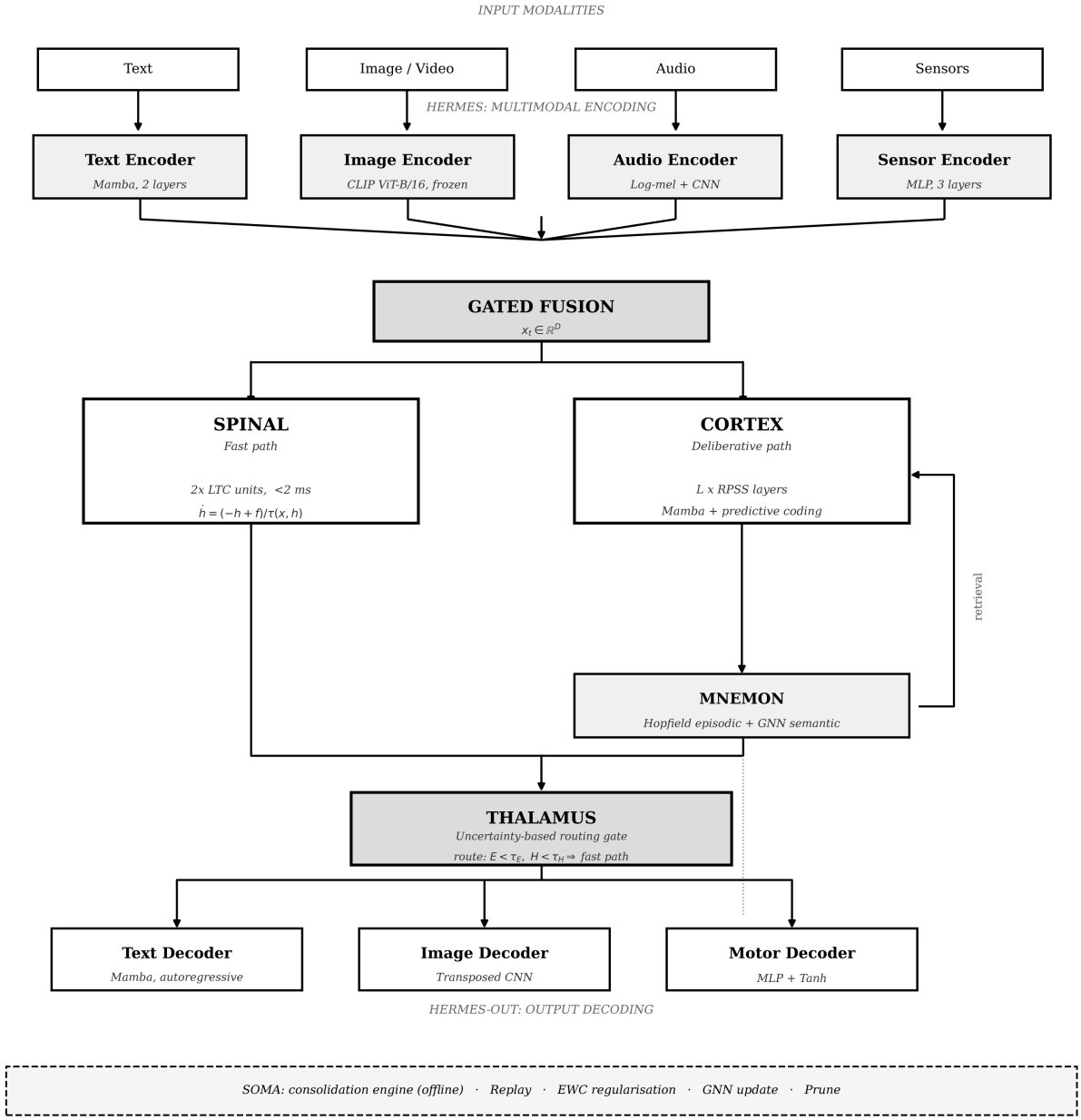


Figure 1. NEXUS-1 system architecture. Input modalities are encoded by HERMES into a shared latent x_t in R^D. SPINAL (fast path) and CORTEX (deliberative path) process this latent in parallel. CORTEX queries MNEMON for episodic and semantic context (retrieval loop, right). THALAMUS routes between the two outputs and passes the result to the appropriate decoder. SOMA consolidates memory offline between sessions.

3.1 Key equations

Each component is governed by an equation from signal-processing or dynamical-systems theory. The Mamba recurrence gives linear-time sequence modelling; the LTC equation gives input-adaptive temporal dynamics; the RPSS error term implements top-down prediction; the Hopfield update implements one-step associative retrieval; and the EWC penalty protects previously learned weights during continual training.

```
Mamba: h_t = A_bar*h_(t-1) + B_bar*u_t
LTC: dh/dt = (-h + f(x,h)) / tau(x,h)
RPSS: eps^l = Mamba(x) - W_pred*z^(l+1)
Hopfield: h* = X*softmax(beta*X^T*q), beta = sqrt(D)
EWC: L = L_B + (lambda/2)*SUM F_i*(theta_i - theta*_i)^2
```

3.2 Image encoder update (v1.1)

The original patch CNN trained from scratch is replaced by a frozen CLIP ViT-B/16 [14] or DINOv2-B [15] backbone with a single learned linear projection (768 to D). The backbone adds no GPU training cost because its weights are frozen; only the projection layer is trained. The downstream Mamba-based CORTEX module is unaffected, since it operates on whichever projected latent it receives. The change gives the model a stronger starting visual representation, learned from large-scale pretraining. The trade-off is a larger model footprint, roughly 86M frozen parameters for ViT-B/16.

4. MNEMON: memory architecture

MNEMON has three tiers: working memory (a circular buffer, K=512, O(1)); episodic memory (modern Hopfield, O(exp(D/2))); and semantic memory (a GNN trained during SOMA consolidation). Surprise-weighted retrieval biases recall toward events with high prediction error: $\text{score}_i = \text{beta} * q^T * p_i + \text{alpha} * \text{surprise}_i$.

5. EXP-002: Hopfield capacity

5.1 Setup

D in {64, 128, 256}, N in {5, 10, ..., 800}, noise in {0.05, 0.10, 0.20}, 5 trials per condition. A recall is correct when cosine similarity exceeds 0.99.

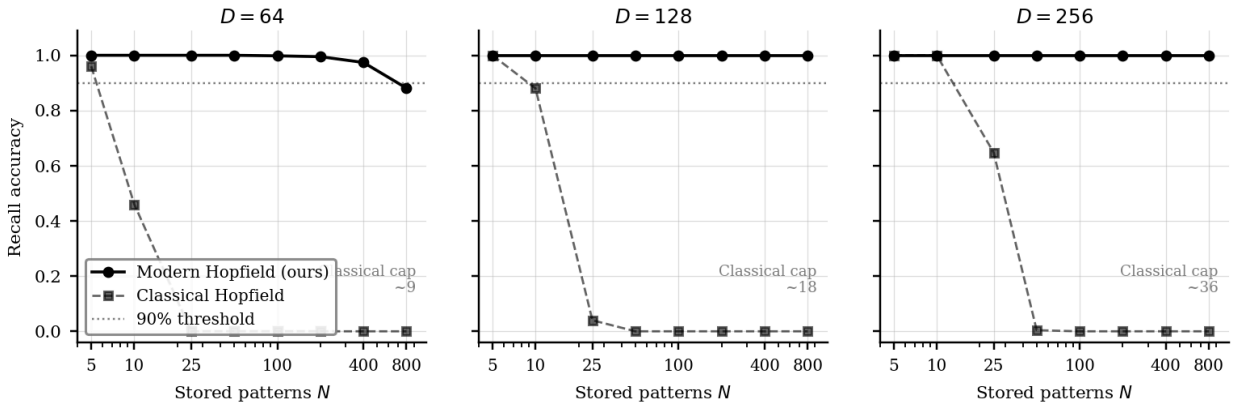


Figure 2. Recall accuracy against N at noise 0.10. The modern Hopfield memory (solid) maintains near-perfect recall. The classical memory (dashed) collapses by N=25 in every dimension tested.

D	Classical capacity	Modern capacity	Ratio
64	N ≤ 5	N = 400	80x
128	N ≤ 5	N > 800	>160x
256	N ≤ 10	N > 800	>80x

Table 2. Capacity at noise 0.10. For D ≥ 128 the ceiling was not found.

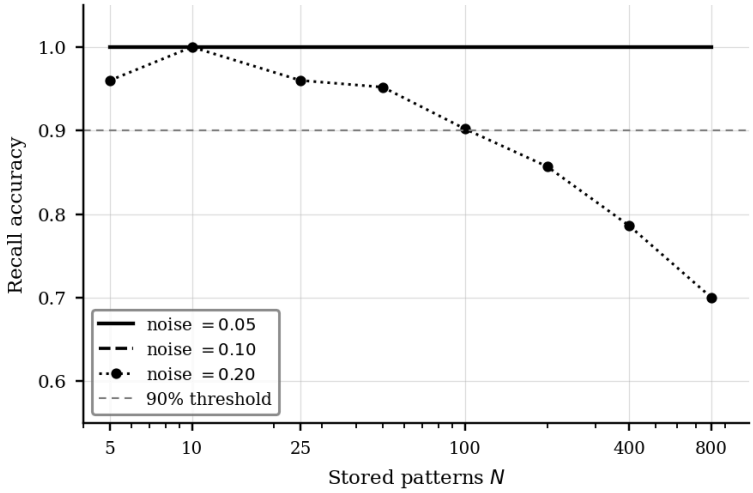


Figure 3. Noise robustness at D=128. At noise ≤ 0.10, recall stays at 1.000. At noise 0.20 it degrades gradually from N≈100.

N	noise 0.05	noise 0.10	noise 0.20
5	1.000	1.000	0.960
10	1.000	1.000	1.000
25	1.000	1.000	0.960
100	1.000	1.000	0.902
200	1.000	1.000	0.857
400	1.000	1.000	0.786
800	1.000	1.000	0.700

Table 3. Modern Hopfield recall at D=128. Hypothesis EXP-002: confirmed.

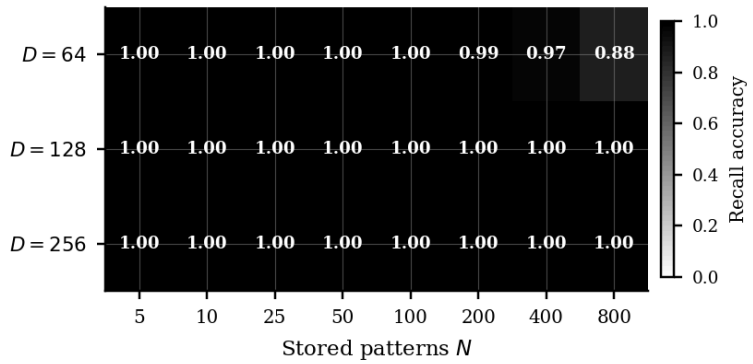


Figure 4. Recall accuracy at noise 0.10. D ≥ 128 stays at 1.00 across all N. D=64 drops only at N=800 (0.88).

6. EXP-004b: SPINAL control validation

6.1 EXP-004 failure and redesign

The initial EXP-004 was inconclusive: SpinalLNN had 9.5 times more parameters than the MLP; Pendulum was discretised into 11 bins, which removed the LTC advantage; and REINFORCE had too little signal in 800 episodes.

Flaw	EXP-004	EXP-004b
Parameters	9.5x mismatch	All within ±8% of 800
Pendulum action	11 discrete bins	Continuous Gaussian
Algorithm	REINFORCE	Actor-critic with baseline
Episodes / seeds	800 / 1	500 + 2000 / 3

Table 4. Corrections from EXP-004 to EXP-004b.

6.2 Pendulum-v1 (continuous control)

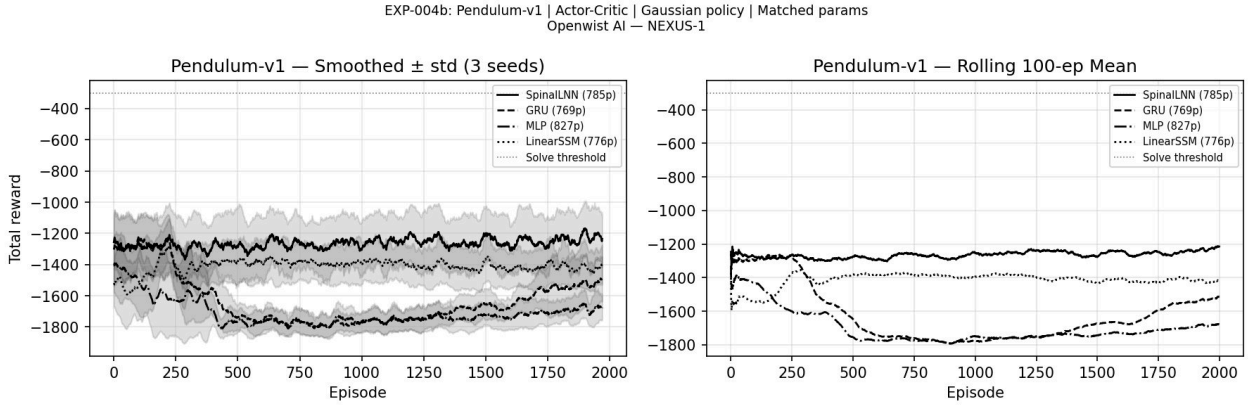


Figure 5. Pendulum-v1, 2000 episodes, 3 seeds. SpinalLNN (solid) leads from about episode 250 and holds the lead. GRU and MLP plateau much lower.

Model	Params	final_100	± std	Rank
SpinalLNN	785	-1213.1	47.9	1
LinearSSM	776	-1415.4	32.1	2
GRU	769	-1514.1	120.3	3
MLP	827	-1677.0	22.1	4

Table 5. Pendulum (higher is better). SpinalLNN leads the GRU by 301 and the MLP by 463.

6.3 CartPole-v1 (discrete control)

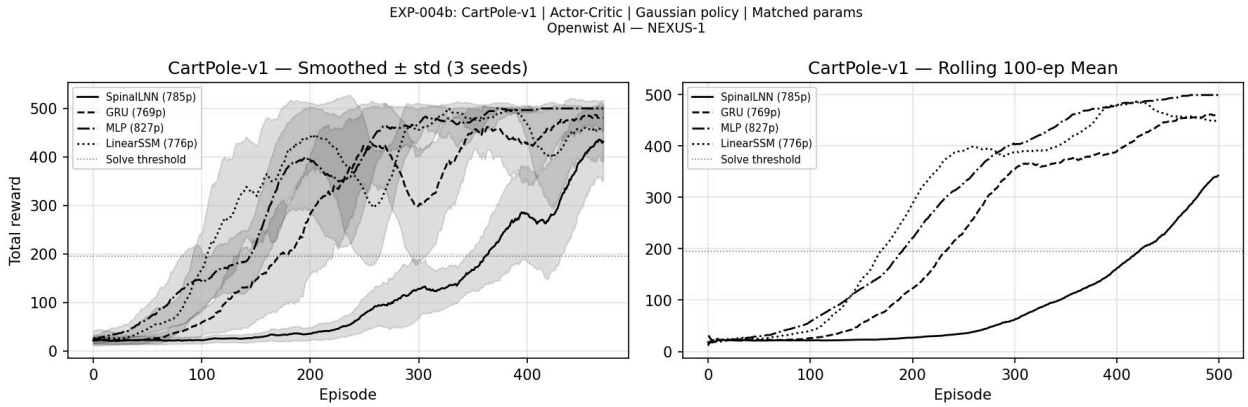


Figure 6. CartPole-v1, 500 episodes, 3 seeds. The MLP converges fastest and scores highest. SpinalLNN finishes last, an expected result (Section 6.4).

Model	final_100	± std	Rank
MLP	498.9	0.7	1
GRU	462.5	19.1	2
LinearSSM	447.3	19.6	3
SpinalLNN	343.6	15.2	4

Table 6. CartPole (higher is better). The MLP wins on discrete control.

6.4 Task-dependent ODE advantage

Pendulum (continuous). The adaptive tau(x,h) lets h(t) track the pendulum's dynamics. A smooth h(t) gives smooth torque, which matches the physics. The GRU and MLP produce piecewise outputs that fit continuous dynamics less well.

CartPole (discrete). The task needs only a decision boundary, for which the MLP's stateless direct mapping is the right inductive bias. SpinalLNN's Euler step adds latency that slows convergence on immediate reactive switching. LinearSSM ranking second on Pendulum suggests that any temporal state helps on continuous control. The 202-point gap between SpinalLNN and LinearSSM isolates the LTC dynamics specifically.

Task	Winner	SpinalLNN	Mechanism
Continuous	SpinalLNN	1 of 4	ODE dynamics match the physics
Discrete	MLP	4 of 4	Pattern matching is sufficient

Table 7. Task dependence. The ODE advantage is specific to the domain.

Criterion	Result	Status
SpinalLNN > MLP, Pendulum	-1213 vs -1677	Pass (+463)
SpinalLNN > GRU, Pendulum	-1213 vs -1514	Pass (+301)
Variance (3 seeds)	std = 47.9	Pass
Inference latency	0.7 ms CPU	Pass (< 2 ms)
CartPole, all models solve	all > 343	Pass

Table 8. EXP-004b criteria. Hypothesis: supported.

7. EXP-005: SOMA continual-learning validation

7.1 Setup

A shared CNN (2 convolutional layers, a 64-dimensional latent and a linear classifier head) is trained on full MNIST (60,000 images, 3 epochs), which sets the Task-A baseline. A surprise-weighted replay buffer stores raw exemplars during training (5% of each batch, capacity 2,000), and EWC Fisher information is computed over 100 batches of Task-A data using the implementation in core/soma/ewc.py. From the identical post-Task-A checkpoint, four training runs then proceed on Fashion-MNIST (60,000 images, 3 epochs) under four conditions: NONE (no protection), EWC_ONLY, REPLAY_ONLY, and FULL_SOMA (both combined). Three random seeds are used throughout.

The replay buffer stores raw pixel exemplars rather than post-encoder latents. The encoder is shared across both tasks and keeps updating during Task-B training, so latents cached under the Task-A encoder would drift out of sync with

the evolving feature space. Raw exemplars are re-encoded at every replay step, which keeps the replay signal valid. This matches raw-exemplar replay in established continual-learning methods such as iCaRL and GEM.

Parameter	Value
Seeds	42, 7, 123
Epochs (Task A / Task B)	3 / 3
EWC lambda	400.0, 100 Fisher batches
Replay capacity / batch	2,000 / 64
Pass: FULL_SOMA retention	> 80%
Pass: NONE retention	< 40%

Table 9. EXP-005 configuration.

7.2 Results

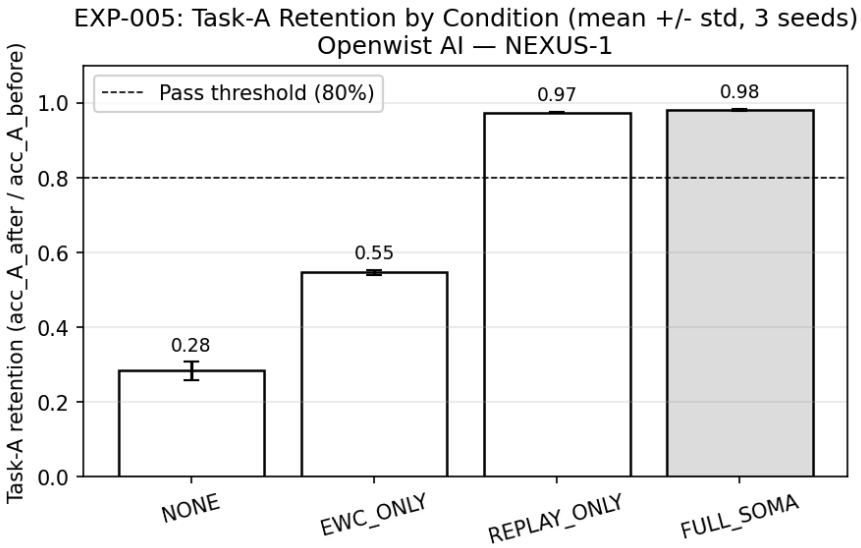


Figure 7. Task-A retention by condition, mean ± std across 3 seeds. FULL_SOMA and REPLAY_ONLY both clear the 80% threshold by a wide margin; NONE falls to 28%.

Condition	Retention	acc_A_after	acc_B
NONE	0.284 ± 0.024	0.279	0.886 ± 0.005
EWC_ONLY	0.546 ± 0.007	0.538	0.882 ± 0.001
REPLAY_ONLY	0.974 ± 0.002	0.960	0.885 ± 0.004
FULL_SOMA	0.981 ± 0.003	0.966	0.880 ± 0.006

Table 10. EXP-005 summary across 3 seeds. Task-A baseline accuracy: 0.985 (mean).

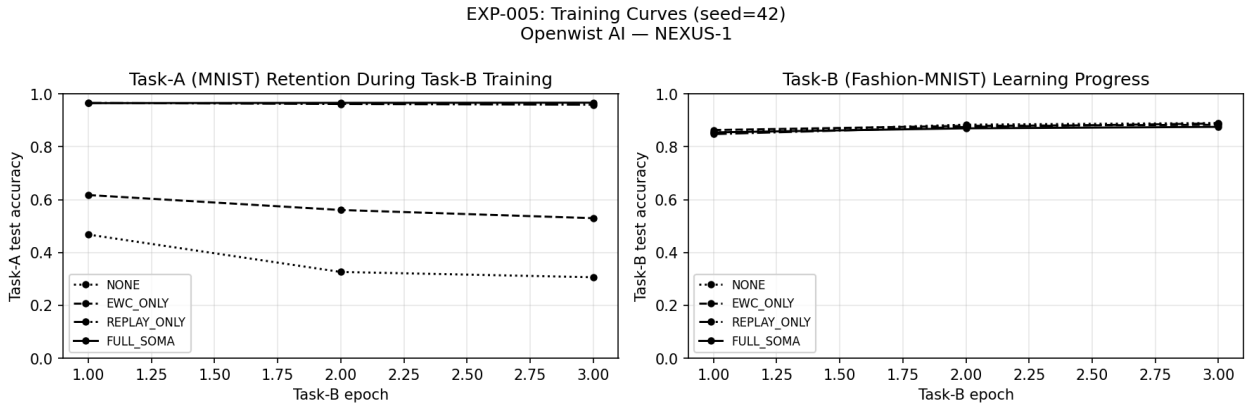


Figure 8. Training curves, seed 42. Left: Task-A accuracy across Task-B epochs. NONE and EWC_ONLY decay steadily; REPLAY_ONLY and FULL_SOMA stay flat near the ceiling. Right: Task-B learning progress. All four conditions converge to within 0.6 points of each other.

7.3 Mechanism decomposition: replay dominates

The four-condition design isolates the contribution of each mechanism. REPLAY_ONLY alone reaches 97.4% retention, within 0.7 percentage points of the full combined system. EWC_ONLY alone reaches 54.6%, which is real but partial protection. In this benchmark, surprise-weighted replay accounts for nearly all of SOMA's protective effect.

FULL_SOMA's margin over REPLAY_ONLY is small (+0.7 percentage points on average) but positive in all three seeds (+0.8, +0.6, +0.6) with no overlap. It is a small but consistent effect. The training curves show why EWC alone is insufficient here: Task-A accuracy under EWC_ONLY keeps falling across all three Task-B epochs (62% to 53%), whereas REPLAY_ONLY and FULL_SOMA hold flat from the first epoch. Task-B accuracy confirms that none of the protective mechanisms costs new-task learning: all four conditions land within 0.6 percentage points of each other on Fashion-MNIST.

Finding	Evidence
NONE confirms catastrophic forgetting	28.4% retention; collapse visible within 1 epoch
Replay is the primary protective mechanism	REPLAY_ONLY 97.4% vs EWC_ONLY 54.6%
EWC adds a small, consistent margin on top of replay	FULL_SOMA +0.7 pp over REPLAY_ONLY, positive in 3 of 3 seeds
Protection does not impair new-task learning	All conditions within 0.6 pp on Task-B

Table 11. EXP-005 mechanism decomposition. Hypothesis: supported.

8. Discussion

8.1 Validated findings

Finding	Experiment	Key result	Status
Hopfield capacity O(exp(D/2))	EXP-002	1.000 recall at N=800, D=128	Confirmed
Classical bound 0.14D is insufficient	EXP-002	Ceiling not found for D ≥ 128	Confirmed
LTC ODE beats GRU/MLP on continuous control	EXP-004b	-1213 vs -1677	Confirmed
LTC ODE trails MLP on discrete control	EXP-004b	343 vs 499	Confirmed
SpinalLNN under 2 ms latency	EXP-004b	0.7 ms CPU	Confirmed
Recurrence helps on continuous control	EXP-004b	LinearSSM ranks 2nd on Pendulum	Confirmed
SOMA prevents catastrophic forgetting	EXP-005	98.1% vs 28.4% retention	Confirmed
Replay is the dominant protective mechanism	EXP-005	REPLAY_ONLY 97.4% alone	Confirmed
EWC adds a small consistent margin over replay	EXP-005	+0.7 pp, positive in 3 of 3 seeds	Confirmed

Table 12. All validated findings to date.

8.2 Architectural implications

EXP-004b supports the two-path design: SPINAL handles continuous control tasks (vehicle steering, humanoid joints, drone control) where ODE dynamics match physical systems, and discrete decisions are routed to CORTEX through THALAMUS. EXP-005 supports SOMA's design, with the refinement that replay capacity is the main lever for retention in this regime. EWC's role may grow in settings with constrained replay buffers or much longer task sequences, which Section 8.3 identifies as an open question. The CLIP/DINOv2 image-encoder update provides strong visual features at no extra GPU training cost.

8.3 Limitations

EXP-002 uses random patterns; real encoder outputs may reduce effective capacity. EXP-004b does not reach the Pendulum solve threshold; PPO or SAC may clarify full-solve potential. EXP-005 uses a two-task sequence with a replay buffer holding 2,000 of 60,000 training images; longer task sequences or smaller buffers may show a larger EWC contribution than observed here. Each experiment tests one component on its own, and the full architecture has not yet been evaluated end to end. Two experiments remain.

9. Remaining experiments

ID	Hypothesis	Status
EXP-002	Hopfield O(exp(D/2))	Done
EXP-004b	LTC wins continuous control	Done
EXP-005	SOMA retains over 80% of Task-A	Done
EXP-001	Mamba perplexity comparable to a Transformer	GPU needed
EXP-003	RPSS MSE below Mamba	After EXP-001

Table 13. Roadmap. Three of five experiments are complete.

10. Conclusion

NEXUS-1 replaces Transformer self-attention with state-space sequence modelling, continuous-time recurrent units, predictive coding, associative memory and continual learning. EXP-002 shows that the modern Hopfield memory exceeds 160 times the classical capacity bound, with perfect recall for D ≥ 128. EXP-004b shows that SpinalLNN wins continuous control at under 2 ms latency while ranking last on discrete control, which supports the two-path design. EXP-005 shows that SOMA retains 98.1% of Task-A performance after Task-B training, against 28.4% unprotected, with surprise-weighted replay as the dominant mechanism. Two experiments remain. Project page: openwist.com/research/project-cortex. Contact: research@openwist.com.

References

[1] Vaswani et al. (2017). Attention Is All You Need. NeurIPS.

[2] Friston, K. (2010). The free-energy principle. Nature Reviews Neuroscience 11(2).

[3] McClelland, McNaughton & O'Reilly (1995). Complementary learning systems. Psychological Review.

[4] Kandel et al. (2013). Principles of Neural Science, 5th ed.

[5] Gu & Dao (2023). Mamba: Linear-Time Sequence Modeling. arXiv:2312.00752.

[6] Hasani et al. (2022). Closed-form continuous-time neural networks. Nature Machine Intelligence 4.

[7] Ramsauer et al. (2021). Hopfield Networks is All You Need. ICLR 2021.

[8] Gu et al. (2022). Efficiently Modeling Long Sequences with Structured State Spaces. ICLR 2022.

[9] Hopfield (1982). Neural networks with collective computational properties. PNAS 79(8).

[10] Rao & Ballard (1999). Predictive coding in the visual cortex. Nature Neuroscience.

[11] Williams (1992). REINFORCE. Machine Learning 8.

[12] Kirkpatrick et al. (2017). Overcoming catastrophic forgetting in neural networks. PNAS 114(13).

[13] Tadros et al. (2022). Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. Nature Communications 13.

[14] Radford et al. (2021). CLIP. ICML 2021.

[15] Oquab et al. (2023). DINOv2. TMLR 2024.