THINK NEURON: BELIEF REVISION AS A SEQUENCE TRANSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most sequence architectures transform point representations. Probabilistic extensions usually attach uncertainty to parameters, latent variables, layer outputs, embeddings, or attention operators. We instead place the probabilistic object at the neural unit. Elemental Variational Expanse (EVE), the variational distributional neuron used here, defines an input-conditioned local posterior. Affine-EVE turns prior–posterior revision into an analytical vector-valued transition; canonical Affine-EVER uses these posterior moments with task-loss training rather than an evidence lower bound (ELBO) objective. For diagonal Gaussian prior and evidence, the transition is exactly $s_t = A_t s_{t-1} + B_t$. These affine transitions compose associatively; we call this property Variational Chain Closure (VCC). EVER is the resulting sequential architecture. Across language modeling, semantic classification, audio, biological sequences, and structured composition, EVER achieves the lowest recorded headline value on seven of eight tasks in the matched-attention generalist suite. We then use two WikiText-2 controls to isolate the source of the gain. In a frozen 6.6514M-parameter system benchmark, analytical prior–evidence precision fusion outperforms both a parameter-matched Transformer and causal cross-attention model on all five paired seeds in negative log-likelihood (NLL). In a stricter control with exactly matched architecture and parameter count, the precision-fusion model retains lower NLL: 5.0224 versus 5.0540 for a direct deterministic transition (5/5 paired seeds; paired t-test $p = 0.0170$). Brier score, accuracy, and top-5 accuracy also favor precision fusion on all five pairs. A frozen FineWeb fixed-token parameter sweep from 10M to 120M parameters provides a controlled fixed-data stress test: EVER outperforms the matched Transformer at every tested size, while Mamba-2 defines a stronger NLL reference under the evaluated configurations. A separate 100M-token validation run characterizes a quality–runtime trade-off, with Mamba-2 using $1.46\times$ the measured wall-clock time, and a five-seed core-budget control favors canonical EVER over uniform recurrent-state widening. The comparison therefore becomes an architectural diagnosis that directs higher-capacity EVE realizations toward richer state geometry rather than raw recurrent width; the exactly matched WikiText-2 controls separately isolate the transition mechanism. Same-checkpoint interventions show direct use of prior–posterior quantities, while retrained ablations reveal substantial compensation capacity. Together, these results support local belief revision as an explicit, reusable inductive bias for sequence computation and identify richer state geometry as a concrete route for extending the same VCC-compatible principle.

1 INTRODUCTION

Transformers have become a dominant approach to sequence modeling (Vaswani et al., 2017). Their probabilistic extensions can represent uncertainty in many ways, but representing a distribution and making local inference the primitive that constructs the next computational state are different choices. Bayesian neural networks place uncertainty over parameters (Blundell et al., 2015); variational autoencoders use latent variables (Kingma and Welling, 2014; Sohn et al., 2015); Variational Neural Networks (VNNs) sample stochastic layer outputs (Oleksiienko et al., 2023); probabilistic Transformers introduce distributions into attention, embeddings, or layer-wise latents (Bahuleyan

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2018; Fan et al., 2020; Hu et al., 2022; Fan et al., 2022). We ask what changes when the probabilistic object is moved to the local neural unit and its belief revision directly drives state evolution.

The starting point is the variational distributional neuron introduced as Elemental Variational Expanse (EVE) (Ruffenach, 2026). In its original formulation, a selected unit is represented by an input-conditioned local posterior $q_i(z_i \mid h_i)$, regularized against a prior and instrumented through its posterior statistics. The distinction from VNNs is therefore not merely spatial: a diagonal stochastic layer output can also factorize across units. EVE instead organizes selected units as explicit local prior–posterior objects, and EVER gives that revision a temporal computational role by using it to construct the causal transition.

We denote the resulting sequential architecture by EVER. It develops this neuron-level principle into sequence dynamics. Affine-EVE is an analytical, vector-valued specialization in which the previous causal state provides a prior, the current observation provides learned evidence, and their fusion defines a posterior. The posterior determines what should be believed after the observation; a distinct gate determines how much of that revised belief becomes the next causal state. The resulting update is exactly affine. Because affine maps remain in the same operator family under composition, successive local revisions admit an associative closure, which we call Variational Chain Closure (VCC). EVER uses this primitive and closure as its sequential core.

This yields the hierarchy

$$\boxed{\text{EVE} \;\rightarrow\; \text{Affine-EVE} \;\rightarrow\; \text{VCC} \;\rightarrow\; \text{EVER}.}$$

Our contributions are fourfold. **(1)** We formulate Affine-EVE, where local prior–posterior revision directly defines a causal state transition. **(2)** We derive its exact associative affine closure as VCC and implement an associative Triton prefix scan that reproduces the causal recurrence to floating-point precision while accelerating end-to-end execution through $T = 16384$. **(3)** We evaluate EVER across heterogeneous sequence domains with parameter-matched attention baselines, fixed-checkpoint interventions, retrained ablations, and numerical closure audits. **(4)** We strengthen attribution with matched WikiText-2 controls and separately position the resulting architecture with a 10M–120M fixed-token sweep including Transformer and Mamba-2, plus a five-seed sequential-core-budget control, topology studies, and Kullback–Leibler (KL) sensitivity analyses. Together, these experiments support a reusable distributional belief-revision transition principle, consistent gains over matched attention baselines, and direct use of explicit prior–posterior quantities in trained models. The exactly matched controls isolate the transition mechanism, while the fixed-token Mamba-2 comparison characterizes a quality–runtime trade-off and the core-budget control converts that comparison into a capacity-allocation result.

2 Related Work

Probabilistic neural computation. Stochastic neurons introduce randomness into activations or gates (Bengio, 2013); Bayes by Backprop learns weight distributions (Blundell et al., 2015); Natural-Parameter Networks propagate distributions over weights and neurons (Wang et al., 2016); and activation-level variational inference provides further local precedents (Haußmann et al., 2020). VNNs are especially close: each layer output is sampled from a Gaussian whose parameters are predicted by dedicated sublayers (Oleksiienko et al., 2023). EVE instead organizes selected units as explicit local prior–posterior inference objects; EVER then uses the resulting revision to construct a temporally causal state transition. This functional role, rather than per-unit factorization alone, is the sharper distinction.

Variational latent models and filtering. Variational autoencoders (VAEs) and conditional variational autoencoders (CVAEs) place variational distributions on example- or condition-level latent variables (Kingma and Welling, 2014; Sohn et al., 2015); T-CVAE and LlaMaVAE extend this idea to story- or sentence-level latent spaces (Wang and Wan, 2019; Zhang et al., 2023). Stochastic Recurrent Networks (STORN) and the Variational Recurrent Neural Network (VRNN) introduce stochastic latent variables into recurrent dynamics (Bayer and Osendorfer, 2014; Chung et al., 2015), while Ladder VAE uses uncertainty-weighted hierarchical fusion (Sønderby et al., 2016). Kalman filtering and learned variants provide classical precedents for combining a prior with uncertain evidence (Kalman, 1960; Krishnan et al., 2015; Karl et al., 2017; Becker et al., 2019). EVER

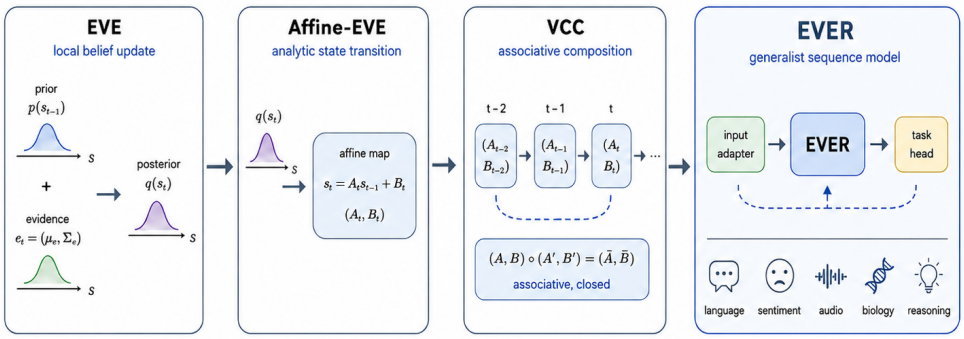


Figure 1: From EVE to EVER. EVE is the local variational belief-update primitive; Affine-EVE turns that revision into $s_t = A_t s_{t-1} + B_t$; VCC is the exact associative closure of these transitions; EVER uses the resulting dynamics as a generalist sequential core.

contributes a distinct computational organization in which local prior–posterior revision directly constructs the transition itself.

Probabilistic Transformers and sequence alternatives. Bayesian Transformers retain parameter-space uncertainty (Xue et al., 2021); Variational Attention, Bayesian Attention, DELLA, STOSA, and Distribution Transformers introduce probabilistic objects into attention or Transformer representations (Bahuleyan et al., 2018; Fan et al., 2020; Hu et al., 2022; Fan et al., 2022; Whittle et al., 2025). Separately, S4/Mamba, linear attention, RetNet, RWKV, Gated Linear Attention (GLA), Gated DeltaNet, xLSTM, Test-Time Training (TTT), Titans, and Hyena show that compact state, recurrence, gating, associative updates, test-time memory, or long convolutions can replace or complement attention (Gu et al., 2022; Gu and Dao, 2024; Dao and Gu, 2024; Katharopoulos et al., 2020; Sun et al., 2023; Peng et al., 2023; Yang et al., 2024; 2025; Beck et al., 2024; Sun et al., 2025; Behrouz et al., 2025; Poli et al., 2023). EVER differs in the source of its update: an explicit local prior–evidence revision produced by Affine-EVE.

A compact distinction is the locus of the probabilistic object. Table 1 isolates this axis from shared machinery such as Gaussian parameterization, reparameterization, or KL regularization.

Table 1: Where the probabilistic object is attached in the closest families.

Family	**Primary probabilistic locus**
Bayesian neural networks / Bayesian Transformers	Network parameters
VAE / CVAE families	Example-, sentence-, or sequence-level latent variables
Variational Neural Networks	Stochastic layer outputs
EVE	**Selected local neural units as variational states**

Extended positioning, including the full comparison table and all neighboring families, is given in Appendix A.

3 From Variational Neurons to Affine State Revision

3.1 EVE and the Affine-EVE specialization

The original EVE unit is a local variational primitive,

$$\mathcal{V}(x_t, h_{t-1}) \to \big(p_t(z \mid h_{t-1}), q_t(z \mid x_t, h_{t-1}), z_t\big),$$

with an explicit prior, amortized posterior, reparameterized latent variable, and local variational regularization (Ruffenach, 2026). At unit level, a selected neuron can be written as

$$q_i(z_i \mid h_i) = \mathcal{N}\big(\mu_i(h_i), \sigma_i^2(h_i)\big)\,,$$

which makes its posterior statistics directly observable during the forward pass. Affine-EVE preserves these prior–posterior semantics while using analytical Gaussian moments to construct a vector-valued state transition.

For observation x_t, a modality adapter produces a diagonal Gaussian evidence factor $e_t(z) = \mathcal{N}(z; \mu_{e,t}, \Sigma_{e,t})$. The previous causal state supplies the prior mean, $p_t(z \mid s_{t-1}) = \mathcal{N}(z; s_{t-1}, \Sigma_p)$. Their normalized product gives

$$\begin{aligned}
\Sigma_{q,t} &= (\Sigma_p^{-1} + \Sigma_{e,t}^{-1})^{-1}, \\
\mu_{q,t} &= \Sigma_{q,t}(\Sigma_p^{-1} s_{t-1} + \Sigma_{e,t}^{-1}\mu_{e,t}), \\
\alpha_t &= \Sigma_{q,t}\Sigma_p^{-1}, \qquad \beta_t = \Sigma_{q,t}\Sigma_{e,t}^{-1}, \qquad \alpha_t + \beta_t = 1.
\end{aligned}$$

Thus $\mu_{q,t} = \alpha_t s_{t-1} + \beta_t \mu_{e,t}$ explicitly reconciles the causal prior with current evidence.

The relative variance contraction

$$r_t = \log \frac{\Sigma_p}{\Sigma_{q,t}}$$

is injected into an adoption gate $u_t = \sigma(g_t + b + \lambda_r r_t)$, where g_t is observation-dependent and λ_r controls the influence of relative contraction. Gaussian fusion determines the posterior; u_t separately determines how much of that posterior enters the causal state:

$$\boxed{s_t = s_{t-1} + u_t(\mu_{q,t} - s_{t-1}).}$$

Substitution yields

$$K_t = u_t\beta_t, \qquad A_t = 1 - K_t, \qquad B_t = K_t\mu_{e,t}, \qquad \boxed{s_t = A_t s_{t-1} + B_t}.$$

All products involving A_t are elementwise in the diagonal Affine-EVE parameterization. The diagonal covariance is a particularly tractable VCC specialization: it parameterizes uncertainty independently in the local latent basis, yields elementwise affine coefficients, and keeps the VCC operator simple and parameter-efficient, while learned projections still mix information across features. The contraction term r_t measures uncertainty reduction only; it does not encode mean displacement, which enters through $\mu_{q,t} - s_{t-1}$. The mechanism therefore separates *inference* from *adoption*. Since $0 < \beta_t, u_t < 1$, one also has $0 < K_t < 1$ and $0 < A_t < 1$ elementwise. This specialization provides an exact and analytically transparent belief-revision core; structured, block-diagonal, or low-rank covariance is the natural higher-capacity extension when posterior cross-covariance should be represented explicitly.

Full derivations and the original stochastic EVE formulation are provided in Appendix B.

4 Variational Chain Closure and EVER

Let $T_t(s) = A_t s + B_t$ denote one Affine-EVE transition. Two consecutive updates compose as

$$(A_2, B_2) \circ (A_1, B_1) = (A_2 A_1,\ A_2 B_1 + B_2),$$

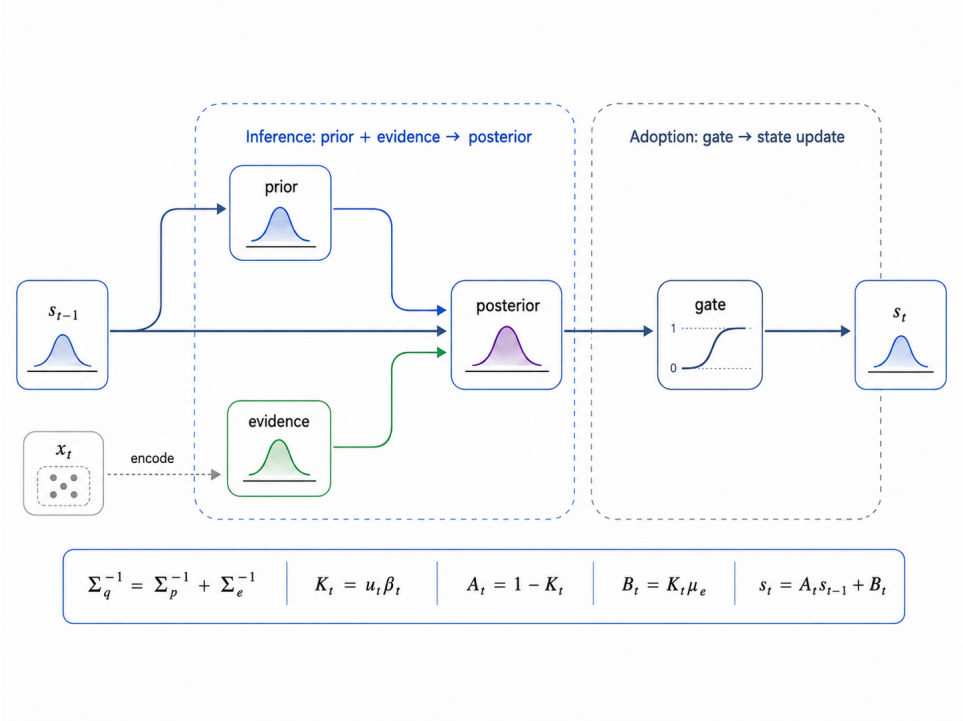


Figure 2: One Affine-EVE update. The previous state defines the prior and the observation a learned evidence factor. Their analytical fusion yields a posterior; a separate adoption gate controls how much of that inferred belief enters the causal state, producing the exact affine coefficients A_t and B_t.

and composition is associative:

$$X_3 \circ (X_2 \circ X_1) = (X_3 \circ X_2) \circ X_1.$$

The identity coefficient is $(\mathbf{1}, \mathbf{0})$ in the diagonal elementwise parameterization. Because every composition remains an affine operator of the same form, the chain of variational updates is exactly closed under composition. We call this property **Variational Chain Closure (VCC)**.

VCC provides exact mathematical associativity. We implement the same recurrence with an associative Triton prefix scan, following the general parallel-scan construction (Blelloch, 1990) and prior use of scan evaluation for linear recurrent networks (Martin and Cundy, 2018). Across $T \in \{512, 1024, 2048, 4096, 8192, 16384\}$ and five seeds, it matches the causal path to floating-point precision (maximum absolute error 5.96×10^{-7}) and is faster end-to-end at every tested length ($1.70\times$–$1.85\times$; 9.05 ms to 5.17 ms at $T = 16384$). The scan-kernel working-set scale matches the causal path; the larger end-to-end peak allocation at long context is localized to explicit BTD↔BDT layout copies, identifying kernel fusion and layout-aware execution as direct implementation optimizations. Full numerical, latency, and memory measurements are in Appendix C.

Across domains the architecture has the compact form

$$\text{modality adapter} \rightarrow \text{EVER} \rightarrow \text{task head}.$$

Across domains, the input adapter, task head, and where required the composition topology adapt to the task, while the local Affine-EVE primitive and VCC principle remain shared. Generality therefore comes from reuse of the same computational primitive across heterogeneous sequence settings.

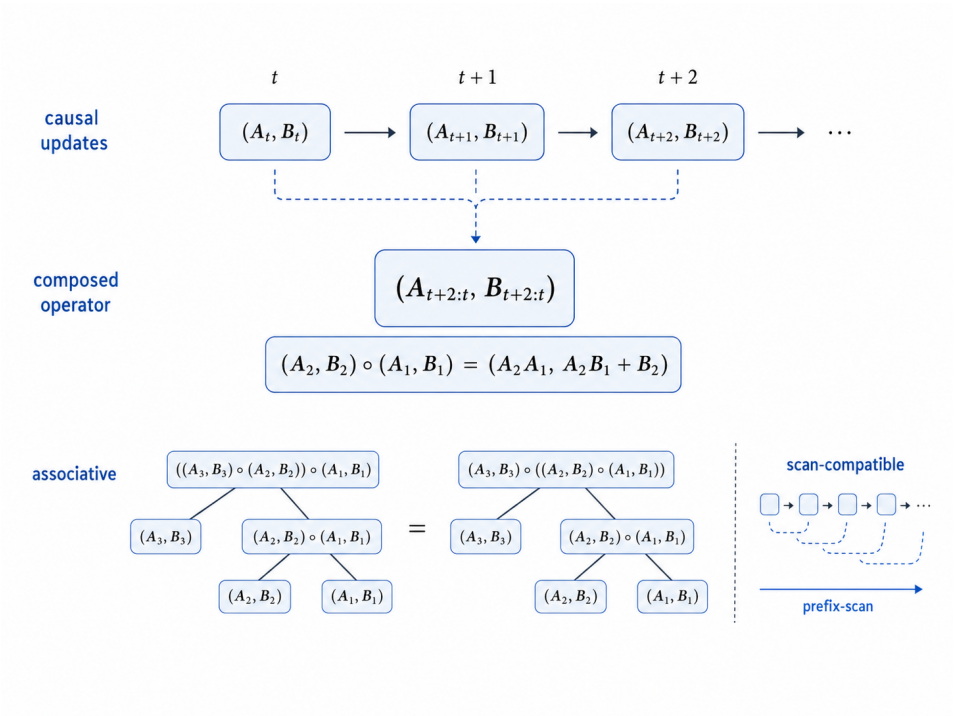


Figure 3: Variational Chain Closure. Each Affine-EVE step yields an affine operator (A_t, B_t). Exact closure and associativity allow consecutive local revisions to be represented in the same operator space and evaluated either causally or through an associative prefix scan.

5 Experiments

5.1 Protocol

We evaluate five aspects of EVER: predictive performance, cross-domain reuse, use of the prior–posterior mechanism, behavior after retraining, and numerical closure. The generalist suite compares EVER with a deterministic Transformer and a parameter-matched causal cross-attention model. A separate fixed-token language-model sweep adds Mamba-2 as a high-performance sequential reference, while the WikiText-2 controls isolate the transition mechanism by first matching complete-system parameter counts and then matching the external architecture as well. Learning rates are selected on validation, comparison seeds are paired, and terminal tests are frozen before evaluation in the most heavily audited suites. Canonical Affine-EVER uses the task loss alone; BBH-hard additionally uses $\beta_{\mathrm{KL}} = 10^{-4}$. Full protocol details and the control rationale are in Appendix D.

5.2 Generalist results

Table 2 reports the headline metric for each domain. For standard benchmarks, "EVER" denotes canonical Affine-EVER; for Logical Arguments and BBH-hard it denotes the task-specific composition variants described below.

Table 2: Headline results across heterogeneous sequence tasks. Lower is better for NLL and mean squared error (MSE).

Task	Domain	Metric	EVER	Transformer	Cross-Attn.
Tiny Shakespeare	language	NLL	**1.8948**	1.9299	1.9174
Penn Treebank	language	NLL	**5.0176**	5.0242	5.0477
WikiText-2	language	NLL	5.1871	5.2057	**5.1659**
SST-2	sentiment	NLL	**0.4564**	0.6581	0.6122
FSDD	audio	NLL	**0.6878**	0.8958	0.8009
Splice Junction	biology	NLL	**0.1216**	0.1556	0.1308
Logical Arguments	composition	MSE	**0.1537**	0.1690	0.1615
BBH-hard postfix/product	composition	NLL	**0.3884**	0.8494	0.4980

EVER records the lowest headline value on seven of the eight rows. On Penn Treebank, NLL is closely matched to the Transformer, while EVER improves accuracy and Brier score on all ten paired seeds and significantly outperforms cross-attention in NLL. On SST-2, EVER has lower NLL than both baselines under paired tests; on FSDD, the frozen confirmatory suite shows a clear NLL advantage. WikiText-2 provides complementary metric behavior: cross-attention records the lowest mean NLL in the original generalist suite, while EVER records higher mean accuracy together with better Brier score and expected calibration error (ECE). This complementary pattern motivates the more tightly matched WikiText-2 controls below. Full per-task statistics are reported in Appendix E.

5.3 Matched controls and fixed-token parameter sweep

We use two WikiText-2 controls to move from system-level comparison to mechanism-level attribution. The first is a frozen system benchmark with exactly 6,651,400 stored trainable parameters for each model: analytical prior–evidence precision fusion, a deterministic Transformer, and causal cross-attention. Precision fusion obtains TEST NLL 5.02237 ± 0.00785, compared with 5.24359 ± 0.02243 for the Transformer and 5.27432 ± 0.12693 for cross-attention. It wins all 5/5 paired seeds against both baselines. The mean paired NLL differences are -0.22121 and -0.25195; the Holm-adjusted primary-endpoint values are 1.11×10^{-5} and 0.00932. Brier score, accuracy, top-5 accuracy, and area under the risk–coverage curve (AURC) also favor precision fusion on all five pairs. This first control establishes a strong system-level result at the same stored parameter budget.

The second control also matches the external architecture. Both models use the same embedding and readout, local projection, affine recurrence, gate, and exactly 6,651,400 trainable parameters. The difference is local: the precision-fusion model derives its transition coefficient from Gaussian precision fusion, whereas the deterministic control learns the corresponding coefficient directly. Learning-rate tuning is symmetric on full WikiText-2, and TEST is materialized only after all ten checkpoints pass the global freeze. The two controls therefore answer complementary questions: the first measures system-level performance, and the second isolates the contribution of precision fusion under matched external architecture.

Table 3: Exactly architecture- and parameter-matched WikiText-2 control (5 paired seeds).

	EVE precision fusion	Deterministic matched control
Parameters	6,651,400	6,651,400
Selected learning rate	0.009	0.006
TEST NLL ↓	**5.02237 ± 0.00785**	5.05401 ± 0.01371
Perplexity (PPL) ↓	**151.77**	156.66

The precision-fusion model has lower NLL on all five paired seeds. The mean paired difference is -0.03164, with a pre-specified paired t-test of $p = 0.0170$. Secondary metrics show the same pattern on all five pairs: PPL (-4.886, $p = 0.0173$), Brier score (-0.001665, $p = 0.0285$), accuracy ($+0.001066$, $p = 0.0127$), and top-5 accuracy ($+0.002571$, $p = 0.0155$). AURC favors precision fusion on 5/5 pairs and ECE on 4/5 pairs. Because the precision-fusion forward path uses analytical coefficients without stochastic sampling, this control isolates the *Gaussian precision-fusion parameterization of the transition* relative to learning the deterministic coefficient directly.

A separate frozen FineWeb fixed-token parameter sweep evaluates approximately 10M, 30M, 60M, and 120M trainable parameters under the same 100M-token budget, with two terminal seeds per size. EVER outperforms the matched Transformer on both seeds at every tested size. Mean NLL is 4.8268/4.9511/5.2039/5.4286 for EVER and 6.1164/5.7356/5.7822/6.0089 for the Transformer. At 120M, the fixed budget falls below one token per parameter, so the experiment specifically probes parameter scaling under a controlled data budget. In this setting, Mamba-2 provides a high-performance sequential reference and reaches lower NLL, including 3.8994 versus 5.4286 for EVER at 120M. Read together with the exactly matched WikiText-2 controls, the two studies separate system-level scaling behavior from transition-level mechanism isolation.

The recorded implementation places Mamba-2 and canonical EVER at distinct points on a quality–runtime frontier. In a separate 100M-token FineWeb validation run, Mamba-2 reaches NLL 3.9628 in 536.2 s, while canonical EVER reaches 4.7881 in 366.8 s, so the lower-NLL reference uses $1.46\times$ the measured wall-clock time ($+46.2\%$). These measurements characterize the recorded implementation and hardware. A separate 30M terminal suite on FineWeb, C4-clean, and C4-noclean reproduces the NLL ordering. At 120M total parameters, the audit assigns 44.53M parameters to the EVER sequential core and 103.13M to Mamba-2. A five-seed validation control then widens only the canonical EVER recurrent state until its audited core reaches 103.21M parameters. Canonical EVER is favored on all five pairs, with 5.1557 ± 0.0458 versus 5.3790 ± 0.0044 for core-matched EVER, while Mamba-2 obtains 3.9185 ± 0.0035. This turns the comparison into a capacity-allocation result: the paired outcome identifies state geometry, structured covariance, richer cross-channel interactions, and alternative EVE transition parameterizations as more informative higher-capacity design axes than uniform recurrent width. Full details are in Appendix H.

5.4 Mechanistic validation

Mechanistic validation complements predictive performance by intervening directly on the same trained Penn Treebank (PTB) checkpoint. The mechanistic suite is distinct from the headline PTB benchmark, which is why its full-model NLL (5.0127) differs slightly from the headline mean (5.0176).

Table 4: Same-checkpoint interventions on Penn Treebank.

Intervention	**NLL** $\downarrow$
Full model	5.0127
Relation-neutral gate	5.0534
Constant variance	6.8951
Variance permutation	6.1514
Posterior–prior identity $q = p$	7.2019

The intervention pattern shows direct use of the prior–posterior quantities exposed by Affine-EVE. Targeted perturbations increase NLL from 5.0127, while a fixed-variance model retrained from scratch reaches 4.9902. These two results reveal complementary properties: dynamic variance is causally active in the trained canonical checkpoint, while retraining shows that the surrounding architecture can redistribute predictive function when that pathway is constrained. The mechanism therefore supplies an explicit inductive organization without making the system brittle to a single representational route. Full ablations are in Appendix F.

5.5 Structured composition

Structured composition tests the same primitive in richer task topologies. On Logical Arguments, EVER with a hierarchical hourglass composition topology is evaluated on a frozen terminal holdout with ten seeds. It obtains MSE 0.1537, compared with 0.1690 for the Transformer and 0.1615 for cross-attention. The preregistered EVER–cross-attention comparison gives $\Delta = -0.007855$, 95% confidence interval (CI) $[-0.013196, -0.002515]$, $p = 0.008832$.

A separate development control compares the frozen hierarchical hourglass configuration with a nearly parameter-matched flat topology built from the same VCC primitive (245,881 versus 245,713

parameters). Averaged over ten seeds and three development folds, the flat topology obtains MSE 0.1394 versus 0.1496 for the hierarchical topology. This shows that the Affine-EVE/VCC primitive remains effective under both composition organizations.

On BBH-hard postfix/product, the task-specific EVER configuration reaches 89.4% accuracy, versus approximately 83.9% for both comparators, and NLL 0.3884 versus 0.8494/0.4980. This task-specific configuration uses a small KL coefficient, and the accompanying five-seed development sweep includes $\beta_{\mathrm{KL}} = 0$, where predictive performance remains strong across the tested range. The result therefore characterizes the complete BBH system, while the exactly matched WikiText-2 study provides the architecture-level transition attribution.

6 Discussion

The experiments support local analytical belief revision as a reusable sequential inductive structure across heterogeneous domains. Same-checkpoint interventions confirm mechanism use, while retraining reveals compensatory representational flexibility.

Canonical Affine-EVER uses analytical Gaussian fusion with the task loss, unlike original EVE's stochastic reparameterization and ELBO. Here, "variational" therefore refers to the prior–evidence–posterior organization rather than to ELBO regularization. The posterior-mean refinement in Appendix H, with $u_t = 1$, remains affine and VCC-compatible, making the adoption gate a design choice rather than a requirement.

EVER and attention are complementary: attention selects information across positions, whereas Affine-EVE determines how strongly current evidence revises the causal state.

In the 10M–120M fixed-token study, EVER outperforms the matched Transformer at every tested size, while Mamba-2 provides the stronger NLL reference and a distinct measured quality–runtime trade-off. The five-seed core-budget control shows that matching Mamba-2's sequential-core size by widening EVER favors the smaller canonical state on all five pairs, directing capacity toward richer state structure rather than raw width. Diagonal covariance provides a tractable VCC specialization with learned feature mixing; structured or low-rank covariance and richer cross-channel interactions are direct higher-capacity extensions.

VCC's Triton scan matches the causal recurrence numerically and accelerates end-to-end execution over $T = 512$–16384. Its long-context memory difference is localized to layout conversions, identifying kernel fusion and layout-aware execution as concrete systems optimizations. Details are in Appendices H and I.

7 Conclusion

EVE moves the variational object to the local neural unit. Affine-EVE turns prior–posterior revision into an analytical causal transition; VCC makes these transitions exactly closed and associatively composable. EVER uses the resulting dynamics as a generalist sequential core. Across language, sentiment, audio, biological sequences, and structured composition, EVER obtains the lowest recorded headline value on seven of eight comparisons in the matched-attention suite. The WikiText-2 controls provide system-level and exactly architecture-matched evidence for precision fusion. In the fixed-token 10M–120M sweep, EVER outperforms the matched Transformer at every tested size, while Mamba-2 supplies the stronger NLL reference and the paired core-budget control identifies parameter organization, rather than uniform recurrent width, as the informative next capacity axis. The diagonal Affine-EVE specialization provides an exact, tractable VCC core, with structured state geometry offering a direct route to richer local interactions; the current scan already delivers exact parallel evaluation and measured speedup, with its additional peak allocation localized to layout copies that are directly targetable by kernel fusion. Same-checkpoint interventions establish active use of prior–posterior quantities, and retraining demonstrates compensation capacity. The combined evidence supports local analytical belief revision as a reusable transition rule and defines concrete architectural and systems routes for higher-capacity VCC-compatible sequence models.

AI USE STATEMENT

Generative AI tools were used as general-purpose digital assistants for literature retrieval and organization, translation and language editing, manuscript formatting, and the drafting and iterative refinement of portions of the paper. Their role was comparable to that of other computational tools supporting the research and writing workflow; scientific judgment, validation, and responsibility remained with the authors. All AI-assisted text and retrieved material were checked by the authors, and the mathematical statements, citations, numerical results, and scientific claims were verified against the underlying sources and recorded results. The authors take full responsibility for the final content of this work.

REFERENCES

Bahuleyan, H., Mou, L., Vechtomova, O., & Poupart, P. (2018). Variational Attention for Sequence-to-Sequence Models. COLING 2018.

Bayer, J., & Osendorfer, C. (2014). Learning Stochastic Recurrent Networks. arXiv:1411.7610.

Blelloch, G. E. (1990). Prefix Sums and Their Applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended Long Short-Term Memory. NeurIPS 2024.

Becker, P., Pandya, H., Gebhardt, G., Zhao, C., Taylor, C. J., & Neumann, G. (2019). Recurrent Kalman Networks: Factorized Inference in High-Dimensional Deep Feature Spaces. ICML 2019, PMLR 97:544–552.

Behrouz, A., Zhong, P., & Mirrokni, V. (2025). Titans: Learning to Memorize at Test Time. NeurIPS 2025.

Bengio, Y. (2013). Estimating or Propagating Gradients Through Stochastic Neurons. arXiv:1305.2982.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight Uncertainty in Neural Network. ICML 2015, PMLR 37:1613–1622.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., & Bengio, Y. (2015). A Recurrent Latent Variable Model for Sequential Data. NeurIPS 2015.

Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. ICML 2024, PMLR 235:10041–10071.

Fan, X., Zhang, S., Chen, B., & Zhou, M. (2020). Bayesian Attention Modules. NeurIPS 2020.

Fan, Z., Liu, Z., Wang, A., Nazari, Z., Zheng, L., Peng, H., & Yu, P. S. (2022). Sequential Recommendation via Stochastic Self-Attention. WWW 2022.

Gu, A., Goel, K., & Ré, C. (2022). Efficiently Modeling Long Sequences with Structured State Spaces. ICLR 2022.

Gu, A., & Dao, T. (2024). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Conference on Language Modeling (COLM) 2024.

Haußmann, M., Hamprecht, F. A., & Kandemir, M. (2020). Sampling-Free Variational Inference of Bayesian Neural Networks by Variance Backpropagation. UAI 2019 / PMLR 115.

Hu, J., Yi, X., Li, W., Sun, M., & Xie, X. (2022). Fuse It More Deeply! A Variational Transformer with Layer-Wise Latent Variable Inference for Text Generation. NAACL 2022.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, 82(1), 35–45.

Karl, M., Soelch, M., Bayer, J., & van der Smagt, P. (2017). Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. ICLR 2017.

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. ICML 2020, PMLR 119:5156–5165.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. ICLR 2014. arXiv:1312.6114.

Krishnan, R. G., Shalit, U., & Sontag, D. (2015). Deep Kalman Filters. arXiv:1511.05121.

Lahoti, A., Li, K. Y., Chen, B., Wang, C., Bick, A., Kolter, J. Z., Dao, T., & Gu, A. (2026). Mamba-3: Improved Sequence Modeling using State Space Principles. ICLR 2026.

Martin, E., & Cundy, C. (2018). Parallelizing Linear Recurrent Neural Nets Over Sequence Length. ICLR 2018.

Oleksiienko, I., Tran, D. T., & Iosifidis, A. (2023). Variational Neural Networks. Procedia Computer Science, 222, 104–113. doi:10.1016/j.procs.2023.08.148.

Peng, B., et al. (2023). RWKV: Reinventing RNNs for the Transformer Era. Findings of EMNLP 2023, 14048–14077.

Poli, M., et al. (2023). Hyena Hierarchy: Towards Larger Convolutional Language Models. ICML 2023, PMLR 202:28043–28078.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. Nature Neuroscience, 2, 79–87.

Ruffenach, Y. (2026). Variational Distributional Neuron. arXiv:2602.18250.

Sohn, K., Lee, H., & Yan, X. (2015). Learning Structured Output Representation Using Deep Conditional Generative Models. NeurIPS 2015, 3483–3491.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). Ladder Variational Autoencoders. NeurIPS 2016.

Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., & Wei, F. (2023). Retentive Network: A Successor to Transformer for Large Language Models. arXiv:2307.08621.

Sun, Y., et al. (2025). Learning to (Learn at Test Time): RNNs with Expressive Hidden States. ICML 2025, PMLR 267:57503–57522.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. NeurIPS 2017.

Wang, H., Shi, X., & Yeung, D.-Y. (2016). Natural-Parameter Networks: A Class of Probabilistic Neural Networks. NeurIPS 2016.

Wang, T., & Wan, X. (2019). T-CVAE: Transformer-Based Conditioned Variational Autoencoder for Story Completion. IJCAI 2019, 5233–5239. doi:10.24963/ijcai.2019/727.

Whittle, G., Ziomek, J., Rawling, J., & Osborne, M. A. (2025). Distribution Transformers: Fast Approximate Bayesian Inference With On-The-Fly Prior Adaptation. arXiv:2502.02463.

Xue, B., Yu, J., Xu, J., Liu, S., Hu, S., Ye, Z., Geng, M., Liu, X., & Meng, H. (2021). Bayesian Transformer Language Models for Speech Recognition. ICASSP 2021, 7378–7382. doi:10.1109/ICASSP39728.2021.9414046.

Yang, S., Wang, B., Shen, Y., Panda, R., & Kim, Y. (2024). Gated Linear Attention Transformers with Hardware-Efficient Training. ICML 2024, PMLR 235:56501–56523.

Yang, S., Kautz, J., & Hatamizadeh, A. (2025). Gated Delta Networks: Improving Mamba2 with Delta Rule. ICLR 2025.

Zhang, Y., Carvalho, D. S., Pratt-Hartmann, I., & Freitas, A. (2023). LlaMaVAE: Guiding Large Language Model Generation via Continuous Latent Sentence Spaces. arXiv:2312.13208.

A Extended Related Work

The relevant literature is organized from local computation to full architectures: (i) probabilistic or variational neural units, (ii) sequential and structured variational models, (iii) Bayesian filtering and Kalman-type updates, (iv) probabilistic or distributional Transformers, and then (v) attention alternatives based on a state, memory, or non-attentional operator. This organization follows the hierarchy EVE $\rightarrow$ Affine-EVE $\rightarrow$ VCC $\rightarrow$ EVER and distinguishes the precedents for each ingredient from their organization in the proposed architecture.

A.1 Probabilistic, Stochastic, and Variational Computation at the Neuron Level

Stochastic units and Bayesian neural networks have a long history. Stochastic neurons explicitly introduce randomness into activation or gating (Bengio, 2013); Bayes by Backprop learns a distribution over weights (Blundell et al., 2015); Natural-Parameter Networks (NPN) (Wang et al., 2016) go further by representing both weights and neurons with exponential-family distributions and propagating these distributions from layer to layer. Haußmann et al. (2020) further show that latent variables can be associated with activations in a variational inference procedure. These works provide established precedents for neural stochasticity, distributional activation representations, and variational quantities at the local level.

Variational Neural Networks (VNNs) (Oleksiienko et al., 2023) are an especially close precedent. VNNs place uncertainty at the layer-output level by sampling each layer output from a Gaussian distribution whose mean and variance are predicted by dedicated sublayers. EVE shares the idea that internal representations can be stochastic and distributional, but it places the variational object at the level of the individual unit. In the original EVE formulation (Ruffenach, 2026), each selected unit is represented by an input-conditioned local posterior

$$q_i(z_i \mid h_i) = \mathcal{N}\big(\mu_i(h_i), \sigma_i^2(h_i)\big)\,,$$

regularized against a prior and directly instrumented through its posterior statistics during the forward pass. The distinction is therefore more specific than stochastic hidden activity. When a VNN layer uses a diagonal Gaussian output, its stochastic factorization can also be written per unit, so granularity alone is not decisive. The sharper distinction is functional and temporal: EVE explicitly organizes a local prior–posterior relation, and EVER uses that relation to construct causal transition coefficients over time.

EVE operates at a similar level of granularity but with a different organization. In its original formulation, an EVE unit is explicitly constructed as a local inference primitive: it carries a prior, an amortized posterior, a reparameterized latent variable, and local variational regularization. The EVE formulation is therefore characterized more precisely as a unit that locally organizes prior–posterior revision and gives this revision computational semantics. EVER extends this organization by using the resulting local belief revision as the direct source of a causal sequential transition.

The foundational distinction of the EVE lineage can therefore be stated more precisely: neuron as distribution is a precedent; EVE targets neuron as local belief-update primitive.

A.2 Sequential Variational Models and Structured Variational Inference

Canonical variational autoencoders (VAEs) (Kingma and Welling, 2014) place the variational distribution on a latent variable that serves as an example-level bottleneck between inference and generative networks. Conditional VAEs (CVAEs) (Sohn et al., 2015) extend this construction by conditioning the generative latent-variable model on observed context. In Transformer-based language generation, T-CVAE (Wang and Wan, 2019) uses a latent variable to model coherent story-plot distributions, while LlaMaVAE (Zhang et al., 2023) explicitly operates in a continuous sentence-level latent space. These models are important precedents for Gaussian latent variables, reparameterized sampling, and KL-regularized variational objectives. Their probabilistic object is an example-, sentence-, or sequence-level latent representation. EVE places the probabilistic object at local-unit level, creating a distinct locus, granularity, conditioning pattern, and computational role while drawing on the same established Gaussian, reparameterization, and KL machinery.

Sequential latent-variable models provide an essential precedent. Stochastic Recurrent Networks (STORN) (Bayer and Osendorfer, 2014) introduces stochastic latent variables into recurrent dynamics, while the Variational Recurrent Neural Network (VRNN) (Chung et al., 2015) defines, at each step, a latent variable with prior and posterior conditioned on the recurrent state and observation. These models therefore establish direct precedents for the use of variational latent variables, prior and posterior distributions, and temporal dependencies.

Ladder Variational Autoencoders (Sønderby et al., 2016) provide a particularly close precedent from the standpoint of information fusion: the model is hierarchical, with bottom-up information and the top-down prior combined using precision-related weighting. EVER builds on this established uncertainty-weighted fusion principle by assigning it a local transition role inside sequential computation.

The difference from STORN, VRNN, and structured variational models concerns the role assigned to inference. In these approaches, the variational latent variable participates in recurrent or hierarchical dynamics that remain a distinct architectural object. In Affine-EVE, prior-posterior revision directly provides the quantities that construct the causal-state transition; VCC then closes these transitions under composition, and EVER uses this revision as the source of its sequential dynamics.

A.3 Bayesian Filtering, Kalman Methods, and Learned Updates

Bayesian filtering is a central precedent for any architecture that confronts a prior belief with an uncertain observation. The Kalman filter (Kalman, 1960) combines a predicted state and a measurement according to their uncertainties to produce a posterior estimate that is then propagated over time. Covariance weighting and the prior $\rightarrow$ observation $\rightarrow$ posterior update are classical principles, and the role of β_t in Affine-EVE is conceptually related to this uncertainty weighting. Affine-EVE specializes the idea through a learned evidence factor embedded directly in the neural transition, with local analytical fusion supplying the quantities used by the recurrence.

Deep Kalman Filters (Krishnan et al., 2015) and Deep Variational Bayes Filters (DVBF) (Karl et al., 2017) combine deep learning, latent variables, and probabilistic state models; Recurrent Kalman Networks (RKN) (Becker et al., 2019), in turn, learn a high-dimensional latent representation in which Kalman-type updates can be computed efficiently with an explicit representation of uncertainty. These works are therefore closer neighbors of EVER than purely deterministic state-space models (SSMs) for the "prior + observation + posterior" component.

EVER organizes these precedents in four linked steps. First, the EVE neural primitive supplies a local variational state. Second, the observation becomes a learned evidence factor that is fused with the prior. Third, inference and adoption are separated: the posterior q_t defines the revised belief, and the gate u_t controls how much of that revision becomes the causal state, yielding the effective coefficient $K_t = u_t\beta_t$. Fourth, Affine-EVE turns this adopted revision into an affine operator, and VCC provides associative closure under composition. Their integration defines the EVER construction.

Predictive coding (Rao and Ballard, 1999) provides an additional conceptual parallel: a prediction is confronted with the observation, and the resulting error revises representations. EVER realizes belief revision through local probabilistic prior–evidence fusion, followed by gated adoption and an affine transition.

A.4 Variational Transformers and Distributional Representations

Bayesian Transformer language models (Xue et al., 2021) provide a parameter-space counterpart inside the Transformer family: variational inference is used to estimate posterior distributions over parameters in multi-head self-attention, feed-forward, and embedding modules. This retains the Bayesian-neural-network locus of uncertainty at the model-parameter level, while EVE places the variational object directly inside selected local hidden units.

Several works make Transformer components stochastic or variational. Variational Attention (Bahuleyan et al., 2018) treats alignment as a latent variable; Bayesian Attention Modules (Fan et al., 2020) constructs stochastic attention distributions with a data-dependent prior; DELLA (Hu et al., 2022) distributes variational latent variables across Transformer layers. In these architectures, the

variational mechanism modifies or accompanies a Transformer component. In EVER, EVE defines the local primitive and Affine-EVE makes the revision produced by this primitive the sequential transition itself.

Distributional representations form another line of work. STOSA (Fan et al., 2022) represents items as Gaussian distributions and constructs self-attention from Wasserstein distances. Distribution Transformers (Whittle et al., 2025) directly learns transformations of distributions, notably from priors to observation-conditioned posteriors, while retaining a Transformer architecture to perform these transformations. The similarity to EVER is genuine at the level of probabilistic objects; the difference is that, in EVER, the local prior–posterior relation directly constructs the next causal state and therefore serves as the sequential transition itself.

A distribution describes where a belief lies; EVE locally formalizes its revision, Affine-EVE turns that revision into a transition, and VCC makes these transitions composable.

A.5 Attention Alternatives Based on State, Memory, or Non-Attentional Operators

The original Transformer replaced recurrence with self-attention. Since then, several families have explored alternative sequential cores: state-space models, linearized attention rewritten as recurrence, retention, delta rules, modernized recurrent memories, test-time state learning, or implicit long convolutions. These works are important conceptual comparators for EVER at the level of sequential composition, even when their local primitive is very different.

A.5.1 State-Space Models, Mamba, and State Space Duality

State-space models show that a compact causal state can replace explicit attention interaction as the primary support for sequential integration. S4 (Gu et al., 2022) introduces structured state dynamics suited to long sequences. Mamba (Gu and Dao, 2024) is the most important architectural neighbor of EVER: it makes several parameters of the SSM dynamics input-dependent in order to select, according to content, what should be propagated, forgotten, or exposed. State Space Duality, underlying Mamba-2 (Dao and Gu, 2024), establishes close theoretical links between certain SSM classes and attention variants; Mamba-3 (Lahoti et al., 2026) continues this line by enriching recurrence from state-space principles.

Mamba and EVER can therefore exhibit externally similar forms of recurrence without sharing the same origin. Mamba starts from state-space dynamics made selective by the input. EVER starts from the EVE variational neuron: context defines a prior, the observation produces distributional evidence, their combination defines a posterior, and quantities derived from this revision construct the state transition.

The difference is therefore more fundamental than a choice of parameterization for A_t and B_t. In Mamba, the coefficients belong to the selective SSM mechanism. In Affine-EVE, transition coefficients are derived from explicit local inference: relative uncertainties determine the contribution of the evidence, prior-posterior contraction enters the adoption gate, and the adopted state then provides the basis for the next prior. Mamba makes state dynamics selective; EVER constructs state dynamics from the revision produced by a variational neural primitive.

A.5.2 Linear Attention, Retention, and Delta Rules

A second family starts directly from attention or associative memory and derives a recurrent execution. Linear Transformers (Katharopoulos et al., 2020) show that kernel self-attention can be reformulated iteratively; RetNet (Sun et al., 2023) has parallel, recurrent, and chunkwise-recurrent forms; RWKV (Peng et al., 2023) combines a mechanism related to linear attention with recurrent execution; Gated Linear Attention (Yang et al., 2024) uses data-dependent gates and a matrix state. Gated DeltaNet (Yang et al., 2025) combines gating and a delta rule to control memory erasure and perform targeted corrections.

The similarity to EVER lies in the presence of a recurrently updated state or memory and, for several of these models, an algebra compatible with efficient scans. These families provide established precedents for gating, recurrence, and associativity. EVER is distinguished by the source of its

correction: an explicit interaction between a causal prior and distributional evidence within the Affine-EVE primitive.

A.5.3 Recurrent Memories and Test-Time Updates

xLSTM (Beck et al., 2024) modernizes the LSTM family through exponential gates and new memory structures, including a matrix memory. Test-Time Training (TTT) (Sun et al., 2025) offers another perspective: the hidden state itself becomes a learned model, and its update rule corresponds to a self-supervised learning step executed along the sequence. Titans (Behrouz et al., 2025) introduces a long-term neural memory learned at test time and combines it with attention. These works establish persistent state update as a broad modern design principle. EVER contributes a specific form of that principle: its state carries the mean of the next local prior and is revised through explicit prior–evidence inference in which uncertainty directly affects the amount adopted.

A.5.4 Convolutional Alternatives

Attention alternatives also include non-recurrent operators. Hyena (Poli et al., 2023), for example, is a subquadratic attention replacement based on implicitly parameterized long convolutions and data-controlled gating. EVER occupies a complementary mechanistic position, with a core built from causal state revision derived from the EVE primitive.

A.6 Exact Positioning of EVE, Affine-EVE, VCC, and EVER

The literature provides precedents for neural distributions, stochasticity, variational inference, uncertainty-weighted fusion, sequential priors and posteriors, Bayesian filtering, causal state, gating, affine recurrence, associativity, and attention replacement. EVER's contribution lies in organizing these elements around the EVE primitive so that local belief revision directly becomes the sequential transition.

The hierarchy is as follows: EVE denotes the local variational-inference neural primitive; Affine-EVE preserves its prior-posterior semantics and analytically turns the revision into a causal transition; VCC denotes exclusively the associative affine closure of these transitions; EVER denotes the sequential architecture built from these elements. Affine-EVER's immediate distinction therefore lies in a state transition derived from local belief revision, while the foundational difference traces back to the nature of the EVE primitive from which this transition is derived.

A compact way to separate the closest probabilistic families is to ask where the distribution is attached and what role it plays in the computation:

Bayesian neural networks / Bayesian Transformers → network parameters
VAE / CVAE → example- or sequence-level latent variable
Variational Neural Networks → stochastic layer outputs
EVE → local variational unit states

This distinction isolates the probabilistic locus from the mathematical tools used to implement it. Gaussian parameterizations, reparameterized sampling, and KL regularization are shared precedents; the distinguishing question is which object becomes probabilistic and whether that object directly constitutes the local computation.

The following table summarizes the closest similarities and the architectural distinction that positions EVER relative to each family.

Table 5: Positioning of EVER relative to neighboring model families.

Family	Relevant similarity	Key difference from EVER
Bayesian neural networks / Bayesian Transformers	Distribution over network parameters; uncertainty is represented in weight space.	EVE can keep parameters deterministic while making selected local hidden units variational.
NPN	Distributions over both weights and neurons propagated through the network.	Distributional neurons are a precedent; EVE specifically organizes selected units as local prior-posterior revisions with directly observable posterior statistics.
VAE / CVAE / T-CVAE / LlaMaVAE	Gaussian latent variables and variational objectives at example, sentence, or sequence/story level.	EVE places the posterior at local-unit level rather than using a global or sequence-level latent bottleneck.
VNN	Gaussian stochastic intermediate representations sampled at layer-output level.	In VNN, the layer output is stochastic; in EVE, the selected neuron itself is the local probabilistic computational object.
Stochastic neurons	Randomness in activations or gating.	EVE gives the local stochastic state explicit prior-posterior semantics and instruments its posterior statistics.
STORN / VRNN / Ladder VAE	Variational latent variables; prior/posterior; precision-based fusion or weighting.	EVER uses local inference to construct an Affine-EVE transition and VCC closure; Ladder VAE provides a hierarchical precision-fusion precedent.
Kalman / Deep Kalman / DVBF / RKN	Prior + uncertain observation $\rightarrow$ posterior; covariance weighting; temporal propagation.	EVE is the local primitive; learned evidence factor; separation of inference/adoption through u_t; Affine-EVE transition followed by VCC closure.
Variational Attention / DELLA / STOSA / Distribution Transformers	Latent variables, distributions, or uncertainty integrated into Transformers.	In EVER, EVE revision itself supplies the sequential transition.
S4 / Mamba / Mamba-2 / Mamba-3	Compact causal state; input-dependent dynamics; recurrent/scan execution.	State-space primitive and SSM selectivity versus a transition derived from prior-evidence revision produced by EVE.
Linear Attention / RetNet / RWKV / GLA / Gated DeltaNet	Recurrent state or memory; gating; retention/delta updates; associativity for some variants.	Update derived from key-value/retention/delta memory versus an update derived from local prior-posterior inference.
xLSTM / TTT / Titans	Persistent memory or actively updated hidden state.	Gated memory or test-time learning versus a state interpreted as a causal belief and revised probabilistically.
Hyena	Replacement of self-attention by a subquadratic operator.	Implicit long convolutions + gating, contrasted with EVER's local prior–posterior primitive and VCC closure.
EVER	Causal state and observation-dependent update.	EVE $\rightarrow$ prior-posterior revision $\rightarrow$ Affine-EVE $\rightarrow$ causal transition $\rightarrow$ VCC; EVER is the resulting architecture.

The resulting positioning is precise: EVE provides a local variational revision primitive; Affine-EVE turns this revision into a state transition; VCC describes its associative affine closure; EVER constructs a sequential core from this chain. Mamba makes state-space dynamics selective, while EVER constructs its dynamics from the revision produced by a variational neural primitive.

The literature map includes both experimentally trained comparators and conceptual neighbors. The trained baselines are defined in the experimental protocol, while the broader families establish the architectural context of EVE, Affine-EVE, VCC, and EVER.

B FULL EVE AND AFFINE-EVE DERIVATIONS

B.1 THE EVE VARIATIONAL NEURON

EVE begins by modifying the elementary computational unit. A deterministic unit applies a transformation to its input and produces an activation. The variational neuron adds local inference: before taking the current observation into account, it defines a prior distribution; after the observation, it constructs a posterior distribution conditioned on that information.

$\mathcal{V}(x_t, h_{t-1}) \to \big(p_t(z \mid h_{t-1}), q_t(z \mid x_t, h_{t-1}), z_t\big)$

In its original formulation (Ruffenach, 2026), EVE is a local VAE-type variational unit: it has an explicit prior, an amortized posterior, local reparameterization of the latent variable, and local variational regularization. Its activation results from inference between what was expected before the observation and what is inferred after it.

The variational neuron represents a local transformation of belief.

B.2 FROM THE EVE NEURON TO THE AFFINE-EVE PRIMITIVE

Affine-EVER uses a local variational primitive derived from the EVE neuron and applied to a vector state. This analytical, vector-valued specialization preserves the semantics of prior-posterior revision while adapting it to sequential state dynamics.

In this specialization, the transition directly uses the moments of the prior and posterior obtained by analytical Gaussian fusion. The latent sample z_t retains its role in the original variational formulation, whereas the canonical Affine-EVER benchmarks use analytical moments and the task loss. The evidence lower bound (ELBO) characterizes the full variational objective of the original EVE neuron; Affine-EVE isolates the computational component of local inference.

Variational information then becomes a transition rule: the posterior defines the direction of revision, while the relative uncertainties of the prior and evidence determine how much of that revision should actually modify the causal state.

EVE variational neuron → prior-posterior update → Affine-EVE primitive → state transition → Variational Chain Closure

EVE provides the local primitive; Affine-EVE derives the transition from it; VCC organizes its closure and composition over time; EVER denotes the resulting architecture.

B.3 LOCAL EVIDENCE

Let $x_1, \ldots, x_T$ be a sequence of observations. A modality-specific adapter transforms each x_t into a local representation from which EVE produces the parameters of a diagonal Gaussian evidence factor:

$$e_t(z) = \mathcal{N}(z; \mu_{e,t}, \Sigma_{e,t}), \qquad \Sigma_{e,t} > 0.$$

In the implementations studied, the mean is bounded by a nonlinearity and the variance is made strictly positive by a softplus-type transformation.

e_t denotes a local evidence factor, or learned observational factor. It is the Gaussian factor with which the prior is fused. In the Affine-EVER benchmarks, e_t is operationally defined as a learned evidence factor; it represents what the current observation proposes independently of its adoption by the causal state.

B.4 CAUSAL PRIOR

The previous causal state s_{t-1} provides the prior mean:

$$p_t(z \mid s_{t-1}) = \mathcal{N}(z; s_{t-1}, \Sigma_p),$$

where Σ_p is a learned diagonal variance shared across positions in the Affine-EVER core. Context therefore dynamically determines the prior mean through s_{t-1}, while its uncertainty scale is parameterized by this learned vector. This static-in-time covariance parameterization is computationally consequential: $\Sigma_{q,t}$, α_t, β_t, and the contraction term can be computed from the current evidence and the shared Σ_p before temporal composition. If instead the posterior covariance were propagated as the next prior covariance, these coefficients would become sequentially history-dependent and the present precomputed prefix-scan execution would no longer apply unchanged.

The causal state s_{t-1} thus acquires a simple probabilistic interpretation: it represents the belief available immediately before observation x_t and becomes the basis of the next local prior.

B.5 Precision-Space Fusion

The local posterior is operationally defined as the normalized product of the prior and the learned evidence factor:

$$q_t(z) \propto p_t\left(z \mid s_{t-1}\right)\, e_t(z).$$

For diagonal Gaussians, this fusion is analytical:

$$\Sigma_{q,t} = \left(\Sigma_p^{-1} + \Sigma_{e,t}^{-1}\right)^{-1},$$

and

$$\mu_{q,t} = \Sigma_{q,t}\left(\Sigma_p^{-1} s_{t-1} + \Sigma_{e,t}^{-1}\mu_{e,t}\right).$$

Defining

$$\alpha_t = \Sigma_{q,t}\Sigma_p^{-1},\ \beta_t = \Sigma_{q,t}\Sigma_{e,t}^{-1},$$

we obtain elementwise

$$\alpha_t + \beta_t = 1$$

and

$$\boxed{\mu_{q,t} = \alpha_t s_{t-1} + \beta_t \mu_{e,t}.}$$

The posterior is therefore an explicit compromise between the previous causal state and the information provided by the observation.

The posterior reconciles memory and evidence.

B.6 Prior-Posterior Relation

Variance contraction is represented by

$$\rho_t = \log \frac{\Sigma_{q,t}}{\Sigma_p}.$$

For the Gaussian fusion considered here,

$$\Sigma_{q,t} \le \Sigma_p,$$

and therefore $\rho_t \le 0$. The following positive quantity is also used

$$r_t = -\rho_t = \log \frac{\Sigma_p}{\Sigma_{q,t}},$$

which measures how much the observation has reduced the uncertainty of the prior.

This quantity is explicitly injected into the transition gate. In Affine-EVER, the prior-posterior relation refers specifically to relative variance contraction, separately from other possible measures such as posterior-mean displacement or Kullback-Leibler (KL) divergence.

B.7 From Local Inference to State Transition

The resulting posterior serves as the revision target. Affine-EVE learns how much of this new belief should become the new computational state.

A gate $u_t \in (0, 1)$ is constructed from local information and the relative prior-posterior variance contraction:

$$u_t = \sigma\left(g_t + b + \lambda_r r_t\right),$$

where g_t denotes a term learned from the observation and λ_r controls the influence of the relative contraction r_t.

The transition is then

$$\boxed{s_t = s_{t-1} + u_t\left(\mu_{q,t} - s_{t-1}\right).}$$

This equation captures the interpretation of the EVER core. In the fast execution path, the code directly computes α_t, β_t, r_t and then the affine coefficients A_t and B_t; the implemented transition is algebraically equivalent to this partial update toward the local posterior.

When u_t is close to zero, the state mainly retains its prior belief. When u_t increases, it moves toward the posterior constructed from the new observation.

Substituting $\mu_{q,t}$ by its expression:

$$s_t = s_{t-1} + u_t\beta_t\left(\mu_{e,t} - s_{t-1}\right).$$

The effective revision coefficient is therefore

$$K_t = u_t\beta_t.$$

This form makes the interpretation immediate: β_t, determined by the relative uncertainties of the prior and the evidence, sets the evidence contribution implied by relative uncertainty; u_t, modulated in particular by the prior-posterior contraction r_t, controls the portion actually adopted; their product $K_t = u_t\beta_t$ is the revision actually applied.

The causal state s_t is distinguished from its normalized representation h_t = LayerNorm(s_t). The next-step prior is centered on s_t before normalization, whereas h_t is the representation exposed to the readout or next module. This distinction directly matches the notebook implementation.

Affine-EVE addresses two questions in succession:

"What is the posterior?"

then:

"How much of this posterior should become the next computational state?"

The separation between local inference and adoption of the revision, together with its exact affine form, is detailed in Figure 4.

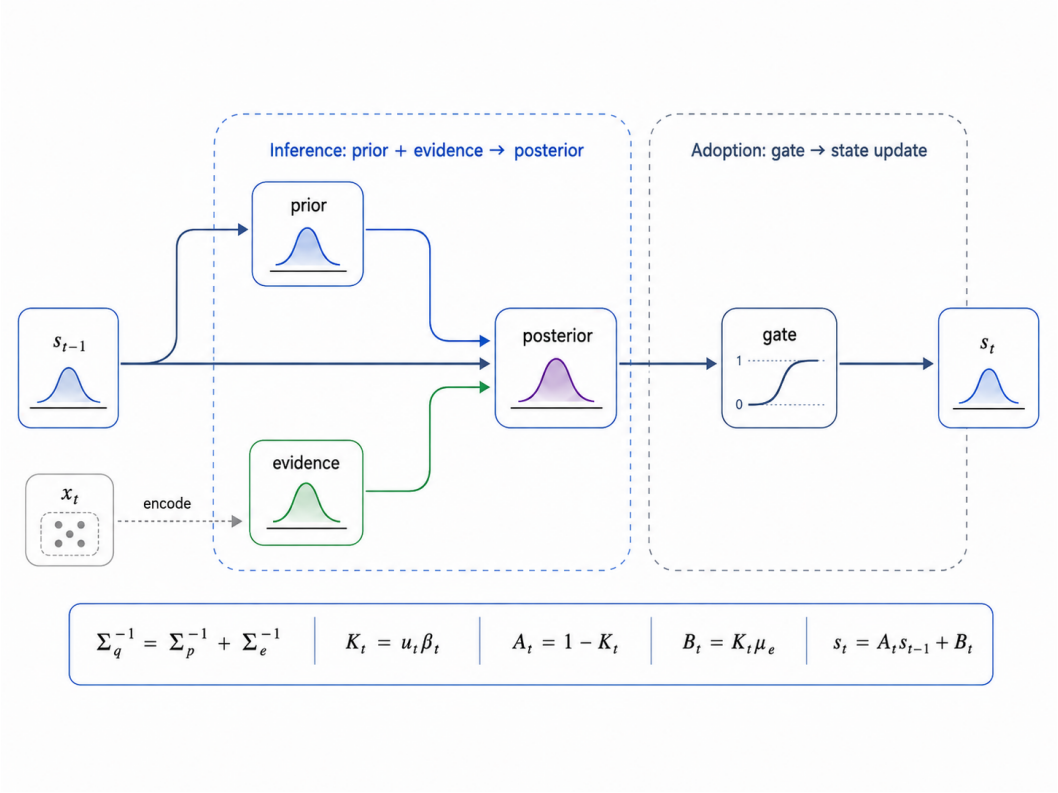


Figure 4: One Affine-EVE update. The previous causal state s_{t-1} defines the local prior, while the current observation x_t is encoded into a learned evidence factor. Precision-space fusion yields the posterior, with $\Sigma_q^{-1} = \Sigma_p^{-1} + \Sigma_e^{-1}$. A distinct adoption gate then controls how much of this inferred belief is incorporated into the computational state. The effective update coefficient $K_t = u_t\beta_t$ induces $A_t = 1 - K_t$ and $B_t = K_t\mu_e$, giving the exact affine transition $s_t = A_t s_{t-1} + B_t$. This explicit separation between inference—what the observation implies—and adoption—how much of that inference becomes the next causal state—is the central computational bias implemented by Affine-EVE.

B.8 Additional Prior-Posterior Properties

The EVE primitive has several distinctive properties when it jointly maintains a probabilistic state and its revision.

B.8.1 A Natural Origin for Revision

Two distinct quantities should not be conflated. The contraction statistic r_t measures relative variance reduction and is blind to posterior-mean displacement. Consequently, two observations with the same prior/evidence variances but different mean displacement contribute the same contraction term to the adoption gate. The direction and geometric size of the proposed revision are instead carried by $\mu_{q,t} - s_{t-1}$. Affine-EVE therefore combines uncertainty-based weighting and adoption with mean displacement, but r_t alone does not distinguish confirmation of a nearby prior from revision of a distant prior.

B.8.2 Directional Confidence

The weight

$$\beta_t = \Sigma_{q,t}\Sigma_{e,t}^{-1}$$

depends directly on the relative uncertainty of the evidence and prior.

A precise observation can therefore move the state more than an uncertain observation.

The mechanism combines:

evidence value,

and

relative credibility of the evidence.

B.8.3 Separation Between Inference and Adoption

Probabilistic fusion determines μ_q, but the gate u_t still controls adoption of this posterior.

This separation distinguishes two questions:

What should be believed after the observation?

and

How much should this new belief modify the computational state?

This separation is a design degree of freedom of the canonical Affine-EVER parameterization, not a requirement for VCC. Setting $u_t = 1$ gives direct posterior-mean recurrence, $s_t = \mu_{q,t}$, which remains affine and associatively composable; Appendix H evaluates this special case in a sparse-evidence refinement.

C VCC Algebra and Numerical Closure Audits

A variational chain denotes the causal succession of local updates in which each observation defines a posterior relative to the current prior, the adopted revision produces the next causal state, and this state in turn provides the mean of the next prior. VCC denotes the closure of this chain enabled by the exact affine form of each transition.

$p_t,\ e_t \to q_t \to s_t \to p_{t+1},\ e_{t+1} \to q_{t+1} \to s_{t+1} \to \cdots$

The affine form of each transition is the mathematical object that closes this chain under composition:

$T_t(s) = A_t\, s + B_t.$

"Variational" describes the origin and semantics of the prior-posterior updates; "affine" describes the exact mathematical property of closure under composition.

C.1 Affine Form of the Update

The previous equation can be rearranged:

$$s_t = (1 - u_t\beta_t)\, s_{t-1} + u_t\beta_t\mu_{e,t}.$$

Setting

$$A_t = 1 - u_t\beta_t$$

and

$$B_t = u_t\beta_t\mu_{e,t},$$

which gives

$$\boxed{s_t = A_t s_{t-1} + B_t.}$$

Each local analytical belief-revision operation therefore produces an affine transformation of the causal state s_t. Moreover, strict positivity of the prior and evidence variances implies $0 < \beta_t < 1$ elementwise. Since the sigmoid gate satisfies $0 < u_t < 1$, the effective coefficient obeys $0 < K_t = u_t\beta_t < 1$ and hence $0 < A_t = 1 - K_t < 1$. Products of the homogeneous coefficients therefore cannot multiplicatively amplify dependence on the initial state.

C.2 Affine Closure

Consider two transitions:

$$s_1 = A_1 s_0 + B_1, \qquad s_2 = A_2 s_1 + B_2.$$

Then

$$s_2 = A_2 A_1 s_0 + A_2 B_1 + B_2.$$

We can therefore define the composition

$$(A_2, B_2) \circ (A_1, B_1) = (A_2 A_1, A_2 B_1 + B_2)\,.$$

This operation is associative:

$$X_3 \circ (X_2 \circ X_1) = (X_3 \circ X_2) \circ X_1.$$

All products involving A_t are elementwise in the diagonal Affine-EVE parameterization. Accordingly, the identity operator is represented by the all-ones coefficient vector and the all-zeros offset vector:

$$(A, B) = (\mathbf{1}, \mathbf{0})\,.$$

The affine operators induced by the variational updates are therefore exactly closed under composition in the same operator space.

Variational Chain Closure (VCC) denotes the affine closure under composition of this chain of variational updates.

local inference $\rightarrow (A_t, B_t) \rightarrow$ associative composition $\rightarrow$ causal state.

Affine composition, its associative closure, and its compatibility with prefix-scan-style evaluation are summarized in the main-text Figure 3.

C.3 Causal and Associative Execution

Associativity of affine composition permits exact prefix evaluation through an associative scan (Blelloch, 1990). We implement this execution path in Triton in addition to the causal reference recurrence, following the same general idea used to parallelize linear recurrent networks over sequence length (Martin and Cundy, 2018). Both compute the same Affine-EVE semantics; they differ only in how temporal composition is scheduled.

Table 6 reports an NVIDIA A100 benchmark with five independent seeds at each context length. The associative implementation matches the causal recurrence to floating-point precision and is faster end-to-end at every tested length.

Table 6: VCC causal versus associative prefix-scan execution on NVIDIA A100. Values are means over five seeds. The end-to-end associative path includes the current layout conversions.

T	Causal (ms)	Scan kernel (ms)	Scan end-to-end (ms)	Speedup
512	0.297	0.041	0.175	1.696×
1024	0.571	0.067	0.324	1.759×
2048	1.210	0.127	0.679	1.786×
4096	2.637	0.247	1.524	1.731×
8192	4.635	0.492	2.511	1.846×
16384	9.051	0.978	5.167	1.752×

Across the complete benchmark, the maximum absolute discrepancy is 5.96×10^{-7} and the maximum relative discrepancy is 1.61×10^{-7}. The current end-to-end scan path explicitly materializes batch–time–dimension (BTD) to batch–dimension–time (BDT) and BDT-to-BTD contiguous layout conversions. At $T = 16384$, measured peak extra memory is 1.63 GiB for the end-to-end associative path versus 0.41 GiB for the causal path. Crucially, the scan-kernel working-set scale itself matches the causal path, localizing the additional allocation to explicit layout copies rather than to the affine scan algebra. This is a favorable systems property: layout-aware integration or fused transpose/scan stages can target the copies directly without changing the recurrence, VCC semantics, or associative operator. The measured memory difference therefore identifies a concrete kernel-integration opportunity while preserving the exactness and speed benefits of the parallel path.

The recurrence defines the semantics; associativity changes how the same recurrence can be executed.

D Experimental Protocol and Hyperparameters

The experiments use a common architecture organized according to the general decomposition

modality adapter → EVER → task head.

For symbolic sequences, the adapter is primarily an embedding. For continuous signals, it becomes a feature projection. The head may produce a next-token distribution, a class, or a continuous value.

The Affine-EVER core remains based on the same chain:

evidence → prior → posterior → gate → (A, B) → closure.

Structured-composition tasks use task-specific composition topologies around the same Affine-EVE primitive derived from the variational neuron. These tasks serve as additional tests of sequential composition rather than as a separate architectural contribution; the demonstrated generality concerns the reused variational primitive.

D.1 Full Experimental Protocol

D.1.1 Research Questions

The protocol examines five research questions (RQs).

RQ1 — Performance. Can a transition based on prior-posterior revision compete with deterministic attention mechanisms?

RQ2 — Generality. Does the same primitive remain useful across heterogeneous modalities and objectives?

RQ3 — Mechanism. Does the trained model actually use prior-posterior information?

RQ4 — Compensation. To what extent can a variant with a simplified variational mechanism recover performance after retraining?

RQ5 — Closure. Do the affine form and fused execution paths faithfully reproduce the mathematically defined causal recursion?

D.1.2 BASELINES

The main comparators are:

1. **EVER;**
2. **Parameter-matched deterministic Transformer;**
3. **Parameter-matched causal Cross-Attention.**

In the main standard benchmarks, recorded active parameter counts are matched exactly; some task-specific composition variants differ by only a few parameters. The separate WikiText-2 system-level control uses a different accounting convention: it matches the stored trainable-parameter total and therefore includes explicitly reported filler parameters in the two attention comparators (Appendix H). The stricter WikiText-2 architecture-matched control uses natural active parameters on both sides. This distinction is kept explicit rather than treating stored filler as active model capacity.

D.1.3 SELECTION AND FAIRNESS

The models have their own learning rates, selected exclusively on validation. Grids are extended whenever an optimum lies on a boundary. Seeds in the main comparisons are paired. Checkpoints are selected using the validation metric defined before terminal evaluation.

The most heavily audited suites explicitly lock the terminal test until the protocol is frozen. The notebooks also record model sizes, fingerprints, splits, hyperparameters, closure errors, and, depending on the experiment, system metrics.

Canonical Affine-EVER benchmarks use the task loss alone. The EVER variant used for BBH-hard adds a small KL coefficient, $\beta_{\mathrm{KL}} = 10^{-4}$, to the cross-entropy, while the baselines use cross-entropy alone. The terminal BBH-hard row therefore evaluates the complete task-specific systems under their selected objectives. A separate development sweep includes $\beta_{\mathrm{KL}} = 0$ and shows strong predictive behavior across the tested KL range, establishing the regularizer as an optional task-specific control rather than a structural requirement of Affine-EVE. Transition-level architectural attribution is supplied separately by the exactly matched WikiText-2 control. The general Affine-EVER transition remains defined by its analytical prior–posterior structure.

E FULL BENCHMARK STATISTICS

The available confirmatory experiments report mean performance by task and domain. For language and probabilistic classification tasks, NLL is used as a common metric where appropriate; for Logical Arguments, the preregistered primary metric is MSE.

Table 7: Generalist benchmark results across tasks and domains.

Task	Domain	Metric	EVER	Transformer	Cross-Attn.
Tiny Shakespeare	language	NLL ↓	**1.8948**	1.9299	1.9174
Penn Treebank	language	NLL ↓	**5.0176**	5.0242	5.0477
WikiText-2	language	NLL ↓	5.1871	5.2057	**5.1659**
SST-2	sentiment	NLL ↓	**0.4564**	0.6581	0.6122
FSDD	audio	NLL ↓	**0.6878**	0.8958	0.8009
Splice Junction	biology	NLL ↓	**0.1216**	0.1556	0.1308
Logical Arguments	composition	MSE ↓	**0.1537**	0.1690	0.1615
BBH-hard postfix/product	composition	NLL ↓	**0.3884**	0.8494	0.4980

Model scope. For the standard benchmarks, EVER refers to canonical Affine-EVER; for Logical Arguments and BBH-hard, it refers to the task-specific EVER composition variants described in Section G.

Three trends emerge from the results.

The gains span several domains: language, sentiment, audio, biological sequences, and structured composition.

Performance depends on the domain and metric. On WikiText-2, cross-attention obtains the best mean NLL, while EVER has slightly higher mean accuracy as well as better Brier score and expected calibration error (ECE). On Tiny Shakespeare, EVER obtains the best NLL and ECE, while the best accuracy and Brier scores belong to other configurations. The performance profile is therefore multidimensional.

Several comparisons show particularly clear paired statistical evidence. On PTB, the NLL difference between EVER and the Transformer is small and statistically compatible with zero, while EVER improves accuracy and Brier score on all ten seeds and significantly outperforms cross-attention in NLL. On SST-2, EVER has lower NLL than both baselines with significant paired t-tests. On FSDD, the frozen confirmatory suite retains a clear EVER advantage in NLL over both comparators. On the Logical Arguments terminal holdout, the preregistered comparison against cross-attention satisfies the predefined criteria for MSE and NLL.

These results show that the variational revision primitive is competitive across heterogeneous domains and particularly advantageous in several regimes.

F Mechanistic Validation and Retrained Ablations

Mechanistic validation combines two complementary protocols: interventions applied to an already trained model and modified architectures that are then retrained. The first tests whether the trained model uses the prior-posterior relation; the second tests how well alternatives compensate after retraining.

The two protocols probe complementary properties of the mechanism.

F.1 Interventions at the Same Checkpoint

On Penn Treebank, interventions are applied to the same EVER checkpoint, directly measuring the causal role of variational components in the learned computation.

The full model obtains a mean NLL of

$$5.0127.$$

This 5.0127 value is the full-model reference mean for the dedicated PTB mechanistic-analysis protocol; the main PTB benchmark comparison reports 5.0176 in Table 7. The two values summarize distinct evaluation protocols.

A relation-neutral gate gives

$$5.0534.$$

A constant-variance intervention gives

$$6.8951,$$

A temporally permuted-variance intervention gives

$$6.1514,$$

and a posterior–prior identity intervention,

$$q = p,$$

gives

$$7.2019.$$

These interventions show that the learned model causally uses the variational quantities provided to it. Alignment of the variances with the corresponding observation directly affects computation at the trained checkpoint.

F.2 Retrained Ablations

Retraining then measures how well the architectural variants compensate for the change.

Under the same PTB protocol, the values are approximately:

Table 8: Retrained ablations under the Penn Treebank protocol.

Retrained variant	**NLL ↓**
Full EVER	5.0127
Direct deterministic transition	5.0540
Relation-neutral transition	5.0112
Gaussian transition with simplified inference	5.0074
Constant-variance retraining	**4.9902**

Several simplified variants recover comparable or slightly improved NLL after retraining. In particular, the constant-variance result shows that the surrounding architecture can compensate when dynamic, data-dependent variance is removed, while the same-checkpoint interventions show that the canonical trained model actively uses that variance pathway. This combination demonstrates both mechanism use and representational flexibility.

This result quantifies how well the variants compensate after retraining.

Same-checkpoint interventions establish causal use of the variational quantities by the trained EVER solution.

Retrained ablations show that alternative parameterizations can learn a comparable or better mapping. The fixed-variance contrast is particularly informative: the same-checkpoint intervention shifts NLL from 5.0127 to 6.8951, while a fixed-variance model retrained from scratch reaches 4.9902. Read together, these results show active use of learned variance in the trained EVER solution and strong compensation after re-optimization. Affine-EVE therefore provides an explicit and controllable inductive structure for prior–posterior revision while preserving representational flexibility around that structure.

G Structured Composition Details

These tasks extend the evaluation to sequential composition and assess how the same Affine-EVE primitive operates when successive pieces of information must be integrated and composed under task-specific topologies.

G.1 Logical Arguments

EVER with a hierarchical hourglass composition topology is evaluated on a frozen terminal holdout with ten seeds. The terminal means are:

Table 9: Logical Arguments results on the frozen terminal holdout.

Model	**MSE ↓**	**Brier ↓**	**NLL ↓**	**Accuracy ↑**
EVER (hierarchical composition)	**0.1537**	**0.7684**	**1.5452**	**0.2941**
Transformer	0.1690	0.8452	2.0710	0.2588
Cross-Attention	0.1615	0.8077	1.6296	0.1647

The preregistered EVER versus Cross-Attention comparison gives, for MSE, $\Delta = -0.007855$ with 95% CI $[-0.013196, -0.002515]$ and $p = 0.008832$. For NLL, $\Delta = -0.084449$ with 95% CI $[-0.149785, -0.019113]$ and $p = 0.016924$.

The two confirmatory criteria defined before the terminal set was opened are therefore satisfied.

G.2 BBH-Hard Postfix/Product

The task-specific EVER variant used for BBH-hard is evaluated on three seeds with exactly the same parameter budget as both comparators.

It reaches:

$$89.4\%$$

accuracy, compared with approximately

$$83.9\%$$

for the Transformer and cross-attention.

The mean NLL decreases from

$$0.8494$$

for the Transformer and

$$0.4980$$

for cross-attention to

$$0.3884$$

for the BBH-hard EVER configuration.

Across the three terminal seeds, these results provide a complementary test showing that the variational primitive remains effective in a richer composition topology.

H Matched Controls, Fixed-Token Parameter Sweeps, and Robustness

This appendix collects matched controls, frozen fixed-token parameter sweeps, robustness studies, and complementary development analyses that extend the main results. Terminal and development datasets remain separated throughout the experimental pipeline.

H.1 Frozen WikiText-2 System-Level Parameter-Matched Control

Before the stricter causal control, the parameter-matched WikiText-2 system benchmark evaluates EVE with analytical prior–evidence precision fusion against two conventional system-level comparators under an exactly matched stored trainable-parameter budget. All three models have 6,651,400 trainable parameters and use five paired production seeds. Learning rates are selected before a global freeze, and TEST is opened only after the 15 production checkpoints are frozen. The resulting means are:

Table 10: Frozen parameter-matched WikiText-2 system comparison (five paired seeds).

Metric	EVE: analytical precision fusion	Transformer	Cross-Attn.
TEST NLL ↓	**5.02237**	5.24359	5.27432
PPL ↓	**151.77**	189.39	196.54
Brier ↓	**0.89439**	0.91019	0.91153
Accuracy ↑	**0.23309**	0.21442	0.21270
Top-5 ↑	**0.44041**	0.41974	0.41847
AURC ↓	**0.58820**	0.62549	0.62894

For the primary endpoint, EVE with analytical prior–evidence precision fusion beats the Transformer on 5/5 pairs with mean paired $\Delta\mathrm{NLL} = -0.22121$ (paired t-test $p = 5.54 \times 10^{-6}$; Holm-adjusted $p = 1.11 \times 10^{-5}$), and beats cross-attention on 5/5 pairs with mean paired $\Delta\mathrm{NLL} = -0.25195$ ($p = 0.00932$; Holm-adjusted $p = 0.00932$). Brier, accuracy, top-5 accuracy, and AURC are favorable on all five pairs against both baselines. These secondary metrics provide convergent support for the primary NLL result.

This experiment provides system-level evidence under an exactly matched stored trainable-parameter budget. EVE with analytical prior–evidence precision fusion uses all 6,651,400 parameters as natural model parameters; the Transformer and cross-attention comparators use 6,534,494 natural parameters plus 116,906 filler parameters (1.76%) to close the same budget. The subsequent architecture-matched WikiText-2 causal control tightens attribution by matching the external architecture and using natural active parameters on both sides.

H.2 Exactly Matched WikiText-2 Causal Control

The architecture-matched WikiText-2 causal control compares EVE with analytical prior–evidence precision fusion with a deterministic architecture-matched control under exact external architecture and parameter matching. Both models have 6,651,400 trainable parameters, natural active parameters throughout, the same embedding and readout budgets, the same local projection, affine recurrence, and gate, and all stored parameters active on the prediction path. Fifteen common tensors are identical at initialization. The six learning rates $\{0.003, 0.0045, 0.006, 0.0075, 0.009, 0.0135\}$ are evaluated symmetrically on full WikiText-2 with the same tuning seed. The selected learning rates are interior to the tested grid.

Table 11: Symmetric WikiText-2 learning-rate tuning before TEST.

Learning rate	EVE precision-fusion val NLL	Deterministic control val NLL
0.0030	5.12947	5.13930
0.0045	5.09827	5.12134
0.0060	5.09088	**5.10197**
0.0075	5.10526	5.13021
0.0090	**5.07570**	5.12835
0.0135	5.08400	5.18215

All ten production checkpoints are loadable and fingerprint-verified before the global freeze; TEST is materialized only afterwards. Seed-level TEST NLL is:

Table 12: Seed-level TEST NLL in the exactly matched WikiText-2 control.

Seed	EVE precision fusion	Deterministic matched control	Δ fusion–control
2601	5.02278	5.04659	−0.02381
2602	5.03253	5.03734	−0.00481
2603	5.01482	5.05029	−0.03547
2604	5.02724	5.07101	−0.04376
2605	5.01449	5.06481	−0.05032

The mean NLL difference is -0.03164 (5/5 favorable pairs; paired t-test $p = 0.0170$). Mean PPL decreases from 156.66 to 151.77. Secondary paired differences are also favorable for Brier (-0.001665, 5/5, $p = 0.0285$), accuracy ($+0.001066$, 5/5, $p = 0.0127$), and top-5 accuracy ($+0.002571$, 5/5, $p = 0.0155$). ECE is favorable on 4/5 pairs and AURC on 5/5 pairs, adding directional agreement across calibration and risk-oriented metrics.

All five paired NLL differences have the same favorable sign. With five pairs, the exact two-sided sign statistic has a discrete resolution of $2/2^5 = 0.0625$, while the pre-specified paired t-test yields $p = 0.0170$.

The causal object isolated by this experiment is the local transition parameterization. EVE with analytical prior–evidence precision fusion computes its coefficient through Gaussian precision fusion, while the deterministic architecture-matched control learns the corresponding deterministic coefficient directly and then uses the same relation, gate, and affine recurrence. Production uses cross-entropy, with KL retained as a diagnostic. This design directly attributes the observed performance difference to the *precision-fusion parameterization of the transition*.

H.3 Frozen FineWeb Fixed-Token Parameter Sweep with Mamba-2

The terminal parameter sweep uses the same 100M-token budget for each model and size and two production seeds. It is designed as a fixed-data architecture comparison, not as a compute-optimal scaling-law experiment. As parameter count increases, tokens per parameter decrease; at 120M parameters the budget is below one training token per parameter. The table reports mean TEST NLL.

Table 13: Frozen FineWeb fixed-token parameter sweep. Two terminal seeds per size; lower NLL is better.

Target size	EVER	Transformer	Mamba-2
10M	4.826762	6.116402	**4.149514**
30M	4.951101	5.735620	**3.986431**
60M	5.203944	5.782173	**3.933532**
120M	5.428550	6.008929	**3.899391**

EVER outperforms the matched Transformer on both terminal seeds at all four sizes. Mamba-2 provides the strongest NLL reference in the same suite, giving a consistent three-model ordering across the full 10M–120M range. The increasing NLL of EVER with parameter count under this fixed token budget is therefore interpreted as an undertraining/optimization outcome of this protocol, not as evidence of an intrinsic adverse capacity-scaling law.

A separate 30M natural-web terminal campaign uses approximately 50M training tokens and two seeds on each corpus:

Table 14: Natural-web terminal replication at approximately 30M parameters.

Corpus	EVER	Transformer	Mamba-2
FineWeb	5.488906	6.207971	**4.325568**
C4-clean	5.545144	6.274279	**4.289469**
C4-no-clean	4.906554	6.048472	**3.455624**

The ordering again holds on both seeds of every corpus: EVER consistently outperforms the matched Transformer in these terminal language-model suites, while Mamba-2 establishes the strongest NLL reference under the evaluated configurations.

H.3.1 Parameter-allocation audit

Total trainable parameters are closely matched, and the allocation audit reveals distinct sequential-core budget profiles across architectures.

Table 15: Sequential-core parameter fraction in the frozen fixed-token sweep audit, reported as a bookkeeping measure of relative core allocation.

Target	EVER core fraction	Mamba-2 core fraction
10M	11.2%	54.0%
30M	21.8%	69.3%
60M	29.3%	79.5%
120M	37.1%	85.9%

At 120M, the audit assigns 44.53M parameters to the EVER sequential core and 103.13M to the Mamba-2 core. Mamba-2 therefore allocates approximately 2.3 times as many trainable parameters to modules classified as sequential core at nearly identical total trainable count. This bookkeeping ratio provides architectural context for the *total-parameter-matched* comparison. We therefore evaluate a direct sequential-core-budget control.

H.3.2 Sequential-core-budget control

To isolate raw sequential-core allocation, we construct an additional EVER configuration that preserves the canonical gated recurrence and embedding width ($E = 3328$) while increasing only recurrent state width from $D = 6672$ to $D = 15480$. Under the same audit definition, this raises the EVER sequential core from 44.53M to 103.21M parameters, within 0.075% of the 103.13M Mamba-2 sequential core. This construction is a *core-budget* match rather than a total-parameter match: widening D also enlarges the projected readout, giving the widened EVER configuration approximately 208.0M stored trainable parameters.

All configurations are trained for 100M tokens. Family-specific learning rates are selected on development data, and the full comparison uses five paired seeds and a held-out validation window.

Table 16: Five-seed sequential-core-budget control. Lower validation NLL is better.

Configuration	Audited core params	Selected learning rate	Validation NLL
Canonical EVER	44.53M	6×10^{-4}	$\mathbf{5.1557 \pm 0.0458}$
Core-matched EVER	103.21M	6×10^{-5}	5.3790 ± 0.0044
Mamba-2	103.13M	2×10^{-3}	$\mathbf{3.9185 \pm 0.0035}$

Canonical EVER is favored on all five paired seeds; the mean paired difference for the widened configuration is $+0.2233$ NLL. Because the embedding width and recurrence form are held fixed while only D is widened, this control isolates raw state width as a capacity axis. The smaller canonical state remains better, showing that additional recurrent parameters are productive only when their organization changes, not when they are added uniformly along the same diagonal state dimension. Training remains numerically stable. The recorded gradient norms also reveal persistent global gradient clipping in the widened configuration, whereas canonical EVER progressively exits the clipped regime during training, providing a concrete optimization signature for this allocation choice. This result narrows the higher-capacity design space toward state geometry, covariance structure, cross-channel interaction, and transition parameterization rather than uniform width alone.

The audited Mamba implementation is Mamba-2 from `mamba-ssm 2.3.2.post1` with `causal-conv1d 1.7.0`; the loaded implementation sources are hashed, and the source-hashed runtime version is recorded with the experiment artifacts.

H.3.3 Controlled context corruption

On C4-clean, a controlled context-corruption diagnostic measures both absolute quality and sensitivity to increasing corruption. Mean NLL over two seeds is:

Table 17: Controlled context corruption on C4-clean.

Corruption	EVER	Transformer	Mamba-2
0%	5.5298	6.2545	**4.2789**
5%	5.7051	6.3719	**4.6202**
10%	5.8851	6.4913	**4.9509**
20%	6.2429	6.7315	**5.5731**
30%	6.6047	6.9762	**6.1439**

From clean context to 30% corruption, the absolute NLL increases are approximately 1.075 for EVER, 0.722 for the Transformer, and 1.865 for Mamba-2. Because the clean NLLs differ substantially, these absolute deltas are not normalized stability scores. Relative to each model's clean NLL, the increases are approximately 19.4%, 11.5%, and 43.6%, respectively. Under both absolute and relative views, the Transformer degrades least, Mamba-2 most, and EVER lies between them in this diagnostic.

H.4 Sparse-Evidence Posterior-Mean Recurrence

A development study evaluates a refinement of EVER that combines a 12.5%-fanin local evidence projection with direct posterior-mean recurrence. Relative to the baseline EVER recurrence,

$$A_t = 1 - u_t(1 - \alpha_t), \qquad B_t = u_t \beta_t \mu_{e,t},$$

the refinement uses

$$A_t = \alpha_t, \qquad B_t = \beta_t \mu_{e,t},$$

so that the pre-normalization recurrent update is the analytical posterior mean, $s_t = \alpha_t s_{t-1} + \beta_t \mu_{e,t} = \mu_{q,t}$. The recurrence change is exactly direct posterior adoption ($u_t = 1$); it remains affine and VCC-compatible. Thus the refinement removes the baseline partial-adoption gate from the predictive recurrence while restricting the local evidence projection to 12.5% of the input fan-in. Its development performance tests the gate as an optional design degree of freedom rather than a necessary component of belief revision. The natural-web study uses a complete five-seed validation block, preserving the terminal split for subsequent evaluation.

At approximately 30M stored trainable parameters and five seeds:

Table 18: Five-seed 30M natural-web development comparison.

Corpus	Baseline EVER recurrence	EVER: sparse evidence + posterior mean	Transformer	Mamba-2
FineWeb	5.2896	4.8716	6.3260	**4.3028**
C4-clean	5.3480	4.8804	6.3286	**4.2547**
C4-no-clean	4.7675	4.2853	6.1105	**3.4026**

This refinement of EVER outperforms the baseline EVER recurrence and the Transformer on all 15 paired seed–corpus comparisons. On the FineWeb parameter sweep, complete validation points are available at 10M (three seeds) and 30M (five seeds): performance is closely matched at 10M (4.7389 versus 4.7338), while at 30M the refinement improves NLL to 4.6776 from 4.7881.

The stored trainable parameter count is exactly matched between the baseline EVER recurrence and the refinement. A structural prediction-path audit estimates active fractions of 90.27%, 80.98%, 74.42%, and 67.55% at nominal 10M, 30M, 60M, and 120M sizes, respectively, quantifying the parameter economy introduced by the local-input mask.

Fixed-token wall-clock measurements characterize the recorded implementation and hardware under an identical token budget. In the 30M natural-web validation block with the same fixed token budget, the refinement Pareto-dominates the baseline EVER recurrence on all three corpora: FineWeb improves from NLL 5.2896 in 182.7 s to 4.8716 in 156.7 s (-14.3% wall-clock); C4-clean from 5.3480 in 181.4 s to 4.8804 in 157.8 s (-13.0%); and C4-no-clean from 4.7675 in 179.7 s to 4.2853 in 156.3 s (-13.1%). A separate 100M-token FineWeb validation run shows the same direction: the baseline EVER recurrence records NLL 4.7881 in 366.8 s and the refinement 4.6776 in

339.1 s, while the Transformer reaches 5.8442 in 324.7 s and Mamba-2 3.9628 in 536.2 s. Relative to canonical EVER, Mamba-2 therefore lowers NLL by 17.2% in this run while requiring $1.46\times$ the measured wall-clock time ($+46.2\%$). Thus the refinement simultaneously improves quality and observed wall-clock time relative to the baseline EVER recurrence across the complete 30M validation block. These measurements characterize *fixed-token wall-clock efficiency* under the recorded implementation and hardware. These development results identify sparse-evidence posterior-mean recurrence as a promising refinement of canonical EVER. This refinement is evaluated on development data only; the terminal split is not used.

H.5 BBH-Hard KL Sensitivity

The published BBH-hard composition uses $\beta_{\mathrm{KL}} = 10^{-4}$. A separate five-seed development sweep holds architecture and learning rate fixed and evaluates $\beta_{\mathrm{KL}} \in \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$, directly mapping predictive behavior and internal KL across the regularization range.

Table 19: BBH-hard KL sensitivity on development data (five seeds per coefficient).

β_{KL}	NLL ↓	MSE ↓	Accuracy ↑	Internal KL
0	0.3305	0.07072	92.07%	1.1608
10^{-5}	**0.2915**	0.06894	91.80%	1.1381
10^{-4}	0.3152	**0.06856**	**92.33%**	1.0888
10^{-3}	0.3011	0.07012	91.73%	0.8735

Predictive metrics remain strong across the tested KL range: the 10^{-4} setting gives the best mean MSE and accuracy, 10^{-5} gives the best mean NLL, and the unregularized $\beta_{\mathrm{KL}} = 0$ setting remains strong. Internal prior–posterior KL decreases systematically as regularization strengthens. The sweep therefore shows two useful properties at once: predictive robustness across regularization strengths and direct control of the internal KL statistic. The strong $\beta_{\mathrm{KL}} = 0$ result also shows that KL is a tunable task-specific regularizer rather than a prerequisite of the Affine-EVE primitive. The terminal BBH-hard row reports the complete task-specific system, while the exactly matched WikiText-2 controls provide architecture-level attribution. No KL regularization is used or tuned in the FineWeb or natural-web parameter sweeps.

H.6 Logical Arguments Topology Control

A development topology study compares the frozen hierarchical hourglass composition configuration with a flat composition topology built from the same Affine-EVE/VCC primitive over three development folds and ten seeds (60 total runs). This directly evaluates how the shared primitive behaves across two surrounding composition organizations.

Table 20: Logical Arguments topology control on development folds.

	Flat composition topology	Hierarchical hourglass topology
MSE ↓	**0.139444**	0.149593
NLL ↓	**1.375932**	1.487629
Accuracy ↑	**38%**	32%
Trainable parameters	245,713	245,881

The parameter gap is 168 parameters (0.0683%) with fully functional parameterization on both sides. Averaging the three development folds by seed, the flat composition topology reaches MSE 0.139444 versus 0.149593 for the hierarchical hourglass topology, while numerical closure checks reproduce the reference recurrence to floating-point tolerance. Together, these results show that the Affine-EVE/VCC primitive remains effective across distinct composition organizations.

I COMPUTATIONAL AND EXTENDED DISCUSSION

The affine equation gives VCC exact associative closure, and the same algebra is evaluated through both causal and associative execution paths. A Triton prefix scan reproduces the causal recurrence to floating-point precision while accelerating end-to-end execution over the tested $T = 512$–16384 range. The measured memory increase is localized in the recorded implementation to explicit layout conversions, which makes kernel fusion and layout-aware scan execution a concrete systems optimization target.

System benchmarks show that Affine-EVER spans multiple latency regimes, with several configurations achieving competitive or advantageous wall-clock behavior and richer composition variants providing additional task-specific structure.

Two properties should be distinguished:

Mathematical property:

$$(A, B)$$

admits exact associative composition.

System property:

actual speedup depends on the kernel, hardware, state size, and topology used.

The first is inherent to VCC; the second depends on implementation.

This distinction precisely attributes associativity to VCC and runtime performance to implementation, hardware, state size, and topology.

I.1 EXTENDED DISCUSSION

I.1.1 ARCHITECTURAL POSITIONING OF EVER

In the canonical benchmarks, Affine-EVER is an alternative sequential core to the self-attention-based mechanisms used by the Transformer and cross-attention comparators. EVER is a generalist sequential architecture based on a variational primitive.

EVE can also be integrated into a Transformer. EVER explores the regime in which the Affine-EVE primitive, derived from the variational neuron, itself becomes the sequence-level dynamics: EVE provides the variational unit, Affine-EVE extracts an analytical transition from it, VCC organizes its composition over time, and EVER denotes the resulting architecture.

I.1.2 VARIATIONAL STRUCTURE AND TRAINING OBJECTIVE

The term "variational" refers here to the local distributional inference structure inherited from EVE: a prior and evidence define a posterior, and the relation between the two explicitly enters the computation. For canonical Affine-EVER this terminology does not imply optimization of an ELBO or sampled variational inference during the forward pass.

Two levels are distinguished:

variational structure of the computation

and

variational regularization of the objective.

The original EVE neuron explicitly included a local ELBO and stochastic reparameterization. In Affine-EVER, the posterior is obtained analytically through Gaussian fusion and its moments are used directly; the canonical benchmarks use the task loss to train this transition. The EVER variant used for BBH-hard adds a small KL coefficient ($\beta_{\mathrm{KL}} = 10^{-4}$).

This organization clearly separates the variational structure of the transition from regularization of the objective. Canonical Affine-EVER benchmarks evaluate the former with the task loss, while the

EVER variant used for BBH-hard explicitly adds the KL term specific to this composition configuration.

I.1.3 Computational Bias and Compensation Capacity

The PTB results reveal two complementary properties. Same-checkpoint interventions show that the trained canonical solution actively uses dynamic variance. After retraining, the fixed-variance variant reaches comparable or lower NLL, showing that the surrounding architecture can redistribute predictive function when that pathway is constrained.

EVE therefore makes the belief-revision computation explicit while preserving representational flexibility around it. The combination of causal mechanism use and compensation capacity is a robustness property of the architecture rather than dependence on a single irreplaceable pathway.

I.1.4 Complementarity Between Attention and EVER

In canonical Affine-EVER architectures, EVER serves as the sequential core through causal dynamics that maintain and revise a state as evidence arrives. Each observation provides evidence, which is fused with the prior carried by the previous state, and the adopted revision becomes the next causal state. Attention organizes interactions between positions; EVER organizes the sequential integration of evidence.

Attention essentially asks:

Which other states are relevant to the current state?

EVER asks:

To what extent should the current evidence revise the causal state?

The former primarily organizes interactions between positions; the latter organizes sequential evidence integration.

The two operations are conceptually complementary. Combining them is a natural direction, pairing attentional selection with variational revision.

I.2 Design Implications and Extensions

The experiments yield several concrete design implications.

Scale. The frozen terminal study covers approximately 10M–120M total trainable parameters under a fixed 100M-token budget. Holding data constant isolates parameter scaling under a common training budget, including the data-limited 120M regime. Size-appropriate token budgets provide the complementary axis for subsequent scaling studies.

Sequential baselines. The cross-domain generalist table uses matched attention comparators to preserve parameter and domain alignment, while the separate language-model sweeps add Mamba-2 as a high-performance sequential reference. EVER consistently improves over the matched Transformer across the tested sizes, and the Mamba-2 comparison maps a distinct quality–runtime operating point. The five-seed sequential-core-budget control then shows that matching Mamba-2's audited core count by uniformly widening the canonical EVER state favors the smaller canonical configuration on all five pairs. The comparison therefore becomes a capacity-allocation experiment: the informative variable is how capacity is organized across state, representation, gating, covariance, and projections, not simply how many parameters are assigned to the recurrent core. Finer gating granularity, structured or low-rank covariance, richer cross-channel interactions, and matrix-valued state organizations are direct higher-capacity realizations of this design space. Preserving affine operator closure would retain the central VCC property while enriching state interactions.

Cross-domain structure. The generalist WikiText-2 suite and the exactly architecture-matched causal control provide complementary views: the former compares complete systems across several metrics, while the latter isolates the precision-fusion transition against a deterministic direct-transition counterpart. Together they separate system-level behavior from mechanism-level attribution.

Mechanism and compensation. The PTB same-checkpoint interventions establish causal use of prior–posterior quantities by the trained checkpoint, while retrained ablations show that the architecture can redistribute predictive function across alternative representations, including a slightly better fixed-variance PTB solution. This combination demonstrates active mechanism use together with compensation capacity.

Covariance and uncertainty semantics. The generalist core uses diagonal covariances as a deliberately tractable specialization: the local update is analytical and parameter-efficient, the affine coefficients remain elementwise, and learned projections still mix information across channels. This provides a clean exact VCC baseline in which uncertainty weighting is explicit at local-unit level. Structured, block-diagonal, or low-rank covariance then defines a direct higher-capacity extension for explicit posterior cross-correlation while preserving the prior–evidence–posterior organization. Dedicated calibration studies can further connect internal variance and relation quantities to operational uncertainty signals.

Composition topologies and regularization. Structured-composition variants reuse the Affine-EVE primitive with task-specific organization. Development studies show effective behavior under both flat and hierarchical hourglass composition topologies. The BBH KL sweep remains strong from $\beta_{\mathrm{KL}} = 0$ to 10^{-3} while internal KL changes systematically, showing that the regularizer can tune the internal relation without being structurally required by the primitive.

Sparse-evidence posterior-mean refinement. This EVER refinement combines a 12.5%-fanin local evidence projection with direct posterior-mean recurrence, corresponding to direct posterior adoption in the recurrence. Complete development blocks show consistent quality and wall-clock gains over the baseline EVER recurrence at 30M, together with substantial active-parameter economy from the local-input mask. The result identifies the adoption gate as a genuine design degree of freedom while preserving strong development performance and VCC compatibility. The terminal split remains untouched.

Temporal parallelization. VCC associativity is realized by an associative Triton prefix scan, following standard parallel-scan ideas (Blelloch, 1990; Martin and Cundy, 2018), that reproduces the causal recurrence to floating-point precision and yields approximately $1.70\times$–$1.85\times$ end-to-end speedup over the tested $T = 512$–16384 range. The scan-kernel working-set scale matches the causal path; the measured peak-memory difference is localized to explicit BTD↔BDT layout conversions in the recorded end-to-end implementation. This separation is useful because kernel fusion and layout-aware scan execution can target the copies directly while leaving the exact VCC operator unchanged.

Compute matching. Recorded wall-clock results provide implementation-level efficiency measurements under fixed token budgets. Dedicated fixed-time compute matching offers a complementary systems benchmark for future scaling studies.