Kalman Delta Networks

Uncertainty-aware Associative Memory

Paper: Ngoc Bui, Tinglin Huang, Rex Ying
Yale University

arXiv:2609.07816v1 • 7 September 2026
Independent reading presentation • 22 September 2026

Track confidence in a recurrent memory to decide how strongly to edit it.

22 main slides for a 20–30 minute discussion; 4 backup slides.

https://arxiv.org/abs/2609.07816v1

A fixed memory must decide what to overwrite

The online decision

Each token edits a compressed prefix before future queries are known.

$$S_t = D_t S_{t-1} + \beta_t k_t \big(v_t - (D_t S_{t-1})^\top k_t\big)^\top$$

- $S_t \in \mathbb{R}^{d_k \times d_v}$ stores a key–value map; read $o_t = S_t^\top q_t$.
- A large residual can mean stale memory or an unreliable observation.
- A token-predicted gate has no explicit covariance state recording prior support.

Paper, Secs. 1–2, PDF pp. 1–4. "Fixed gain" does not mean a constant gate across tokens.

Separate the memory transition from the write gain

Model	Transition D_t	Gain κ_t
DeltaNet	I	$\beta_t k_t$
Gated DeltaNet	$\alpha_t I$	$\beta_t k_t$
KDA	$\mathrm{diag}(\alpha_t)$	$\beta_t k_t$
Isotropic KDN	$\mathrm{diag}(\alpha_t)$	covariance-derived $\beta_t k_t$
Diagonal KDN	$\mathrm{diag}(\alpha_t)$	covariance-derived vector

What changes in KDN?

Propagate uncertainty with the transition; derive the residual gain from uncertainty and observation noise.

Paper, Sec. 3.3, Eqs. (18)–(20), PDF pp. 7–8; Sec. 4, pp. 8–11.

Treat each key–value pair as a noisy measurement

$$\widetilde{S}_t = D_t \widetilde{S}_{t-1} + W_t, \qquad W_t \sim \mathcal{N}_{\mathrm{col}}(0, \Omega_t)$$
$$v_t = \widetilde{S}_t^\top k_t + e_t, \qquad e_t \sim \mathcal{N}(0, r_t I_{d_v})$$

- $\widetilde{S}_t$: latent map; S_t: its estimated mean.
- $P_t \in \mathbb{R}^{d_k \times d_k}$: key-space covariance shared by independent value columns.
- Ω_t: uncertainty from drift; r_t: reliability of the current value.

Scope of the model

Linear–Gaussian dynamics with conditionally independent noises. These are modeling assumptions, not measured laws of language.

Paper, Sec. 3.1, Eqs. (7), (9)–(11), PDF pp. 4–6.

The Kalman gain balances uncertainty and reliability

$$\widehat{S}_t = D_t S_{t-1}, \qquad \widehat{P}_t = D_t P_{t-1} D_t^\top + \Omega_t$$

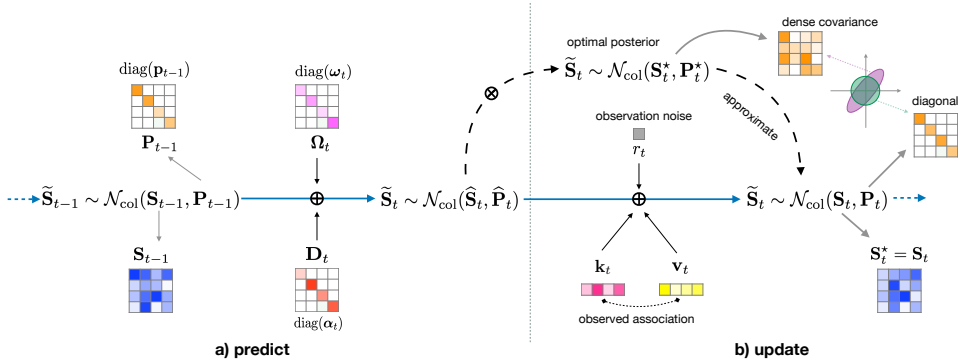
$$\kappa_t = \frac{\widehat{P}_t k_t}{r_t + k_t^\top \widehat{P}_t k_t}, \qquad S_t = \widehat{S}_t + \kappa_t \underbrace{(v_t - \widehat{S}_t^\top k_t)^\top}_{\text{innovation}}$$

$$P_t = (I - \kappa_t k_t^\top)\widehat{P}_t, \qquad o_t = S_t^\top q_t$$

- Effective same-key write: $\beta_t^{\text{eff}} = z_t/(r_t + z_t)$, where $z_t = k_t^\top \widehat{P}_t k_t$.
- Larger directional uncertainty permits a stronger correction.

Paper, Proposition 3.1, PDF p. 7; Appendix B, pp. 21–22. MMSE optimality is conditional on the assumed model.

A measurement creates correlations between key channels



- Prediction can stay diagonal; the exact posterior is generally dense.
- Reverse-KL projection restores a diagonal state and preserves the one-step mean.

Paper, original Fig. 1, PDF p. 5. Exactness refers to the retained predictive prior.

Diagonal KDN projects precision, not covariance entries

$$\widehat{p}_t = \alpha_t^2 \odot p_{t-1} + \omega_t, \qquad \kappa_t = \frac{\widehat{p}_t \odot k_t}{r_t + \sum_i \widehat{p}_{t,i} k_{t,i}^2}$$

$$q_t = \arg\min_{q \in \mathcal{Q}_{\mathrm{col}}} \mathrm{KL}\big(q \,\|\, \mathcal{N}_{\mathrm{col}}(S_t^\star, P_t^\star)\big)$$

$$S_t = S_t^\star, \qquad p_{t,i} = \frac{1}{(P_t^{\star -1})_{ii}} = \left(\widehat{p}_{t,i}^{-1} + \frac{k_{t,i}^2}{r_t}\right)^{-1}$$

- The approximation retains d_k uncertainty scalars per head.
- One-step mean preservation does not recover the full exact filtering trajectory.

Paper, Sec. 4.1, Eqs. (23)–(30), PDF pp. 8–9; proof in Appendix A.2, p. 21. Here $\mu = 1$.

Information scaling protects subsequent writes

$$u_t = \frac{\mu}{r_t} k_t \odot k_t, \qquad p_t = \frac{\widehat{p}_t}{1 + u_t \odot \widehat{p}_t}$$

- Compute the current gain from $\widehat{p}_t$ *before* this update.
- Larger μ lowers retained variance and weakens later overlapping writes.
- $\mu = 1$ gives the original variational solution; main runs use $\mu = d_k$.

A balanced-key example

For m active coordinates with $k_i^2 = 1/m$ and variance p:

$$V_{\text{exact}} = \frac{rp}{r + p}, \qquad V_{\text{proj}}(\mu) = \frac{mrp}{mr + \mu p}.$$

$\mu = m$ matches this direction, not the full covariance.

Paper, Eq. (34), PDF p. 9; Appendix B, Eqs. (60)–(65), p. 22.

A nonlinear variance recurrence admits an associative scan

$$p_{t,i} = \frac{\alpha_{t,i}^2 p_{t-1,i} + \omega_{t,i}}{u_{t,i}\alpha_{t,i}^2 p_{t-1,i} + 1 + u_{t,i}\omega_{t,i}}$$

$$M_{t,i} = \begin{bmatrix} \alpha_{t,i}^2 & \omega_{t,i} \\ u_{t,i}\alpha_{t,i}^2 & 1 + u_{t,i}\omega_{t,i} \end{bmatrix}, \quad \begin{bmatrix} n_{t,i} \\ d_{t,i} \end{bmatrix} = M_{t,i} \begin{bmatrix} n_{t-1,i} \\ d_{t-1,i} \end{bmatrix}$$

Möbius representation

$p_{t,i} = n_{t,i}/d_{t,i}$. Composition becomes 2×2 matrix multiplication: associative, but order-sensitive.

- First scan uncertainty to obtain gains; then scan the affine memory recurrence.
- The associative formulation permits $O(\log T)$ parallel depth.

Paper, Eqs. (32)–(39), PDF pp. 9–11.

Isotropic KDN trades directionality for a scalar state

$$a_t = \frac{1}{d_k}\sum_i \alpha_{t,i}^2, \qquad \widehat{b}_t = a_t b_{t-1} + \omega_t, \qquad P_t = b_t I$$

$$\beta_t = \frac{\widehat{b}_t}{r_t + \widehat{b}_t \|k_t\|^2}, \qquad b_t = \frac{\widehat{b}_t}{1 + \mu \widehat{b}_t \|k_t\|^2/(d_k r_t)}$$

- Prediction: trace projection, preserving the average eigenvalue.
- Posterior: reverse-KL projection, yielding a harmonic mean when $\mu = 1$.
- The gain stays parallel to k_t, but depends on accumulated uncertainty.

State cost per head

Extra uncertainty: $O(1)$ isotropic or $O(d_k)$ diagonal. Both still store the $O(d_k d_v)$ associative memory.

Paper, Sec. 4.2, Eq. (40), PDF p. 11; Appendix C, pp. 23–26.

The practical kernel uses two chunkwise stages

1. **Gain computation:** reduce token-local Möbius matrices, obtain exclusive chunk carries, replay chunks to emit κ_t.
2. **Memory computation:** use compact WY for the asymmetric delta rule:

$$S_t = (I - \kappa_t k_t^\top) D_t S_{t-1} + \kappa_t v_t^\top.$$

- Gain chunks: $C_g = 256$; memory chunks: $C_m \in \{32, 64\}$.
- The implemented short gain carry pass is serial across chunks.
- FP32 homogeneous rescaling and log-decay differences improve stability.

Paper, Appendix D, PDF pp. 26–27. Scan compatibility is not a claim of zero serial work in the implementation.

Evaluation controls data and capacity at two scales

	Smaller run	Larger run
Parameters / tokens	750M / 50B	1.3B / 100B
Layers / hidden width	16 / 2048	18 / 2304
KDN heads / key width	16 / 128	18 / 128
Training data	FineWeb-Edu, 4K sequences	
Hybrid setting	—	9 recurrent + 9 SWA layers

- Recurrent non-embedding parameters matched within 0.03%; Mamba-3 states matched too.
- Evaluate perplexity, six-task zero-shot accuracy, synthetic and real retrieval.
- Main KDN runs: $\mu = d_k = 128$, $r_{\min} = 0.01$, learned initial variance per head.

Paper, Sec. 5.1, PDF pp. 11–13. Hybrid and attention-only quality baselines use 2K sliding windows.

At 750M, both variants improve recurrent-model quality

Model (750M / 50B)	Wiki. PPL ↓	LMB. PPL ↓	Avg. % ↑
KDA	18.85	15.06	53.87
Mamba-3 MIMO	18.99	15.67	54.39
Isotropic KDN	**18.42**	14.68	54.41
Diagonal KDN	18.64	**14.15**	**54.97**

- Diagonal mean accuracy: +1.10 percentage points over KDA.
- Against the strongest non-KDN mean (Mamba-3 MIMO): +0.58 points.
- Individual tasks differ: Mamba-3 MIMO still leads HellaSwag and ARC-c.

Paper, selected rows of Table 1, PDF p. 12. Avg.: unweighted mean over six zero-shot tasks.

At 1.3B, average gains are small and hybrid results vary

Model (1.3B / 100B)	Wiki. PPL ↓	LMB. PPL ↓	Avg. % ↑
KDA	15.40	10.09	60.28
Isotropic KDN	15.30	9.98	60.35
Diagonal KDN	**15.04**	**9.75**	**60.45**
KDA + SWA	15.08	10.81	59.94
Isotropic KDN + SWA	**14.98**	10.90	59.38
Diagonal KDN + SWA	15.10	**10.41**	**60.11**

Diagonal's recurrent mean gain over KDA is only +0.17 points.

- No reported multi-seed uncertainty establishes statistical significance.
- In hybrids, neither KDN variant dominates every aggregate metric.

Paper, selected rows of Table 1, PDF p. 12. Bold values compare only within each displayed block.

Diagonal KDN shows stronger synthetic retrieval gains

Recurrent model	750M / 50B	1.3B / 100B
KDA	62.11	64.31
GDN-2	52.54	71.00
Isotropic KDN	62.36	68.66
Diagonal KDN	**70.33**	**76.51**

- Values: **our equal-weight recomputation** of Table 2's 14 cells (%).
- At 1.3B: +12.20 points over KDA; +5.51 over the strongest baseline, GDN-2.
- Not every cell wins: S-NIAH-2 / 8K is 25.8 vs GDN-2's 28.0.
- Modified RULER sampling, 500 matched examples/cell; not a full-suite score.

Paper, Table 2 and caption, PDF p. 13. Differences use unrounded means. Contexts span 1K–8K; training is 4K.

Local attention changes the real-retrieval ranking

Model (1.3B / 100B)	Recurrent	Hybrid + SWA
KDA	33.76	45.09
Mamba-3 MIMO	32.23	44.76
GDN-2	32.09	44.24
Isotropic KDN	33.40	**45.94**
Diagonal KDN	**34.86**	45.55

- Six-task mean accuracy (%); all inputs limited to 2K tokens.
- Diagonal wins recurrent average; Isotropic wins hybrid average.
- Hybridization improves Diagonal by 10.69 points, larger than the mixer-only gap.

Paper, Table 3, PDF p. 14. Tasks: SWDE, SQuAD, FDA, TriviaQA, NQ, DROP. These are short-input retrieval results.

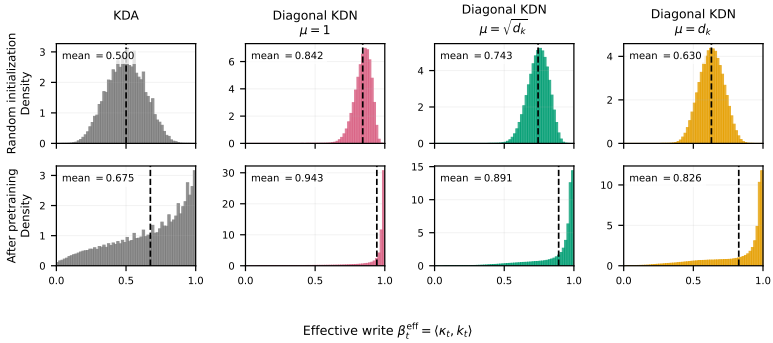
The best information scale depends on the metric

μ	Wiki. PPL ↓	LMB. PPL ↓	Avg. % ↑
1	18.69	15.51	54.12
$\sqrt{d_k}$	18.93	15.10	54.20
d_k	18.64	14.15	**54.97**
$4d_k$	18.97	**13.91**	54.92
Learned	**18.51**	14.68	54.62

- Diagonal KDN, 750M / 50B; learned scale initialized at d_k.
- Larger fixed scales improve LAMBADA PPL monotonically in this sweep.
- d_k is a practical compromise, not a universal optimum.

Paper, Table 4, PDF p. 14; $d_k = 128$.

Larger scales reduce the observed effective-write strength



- Top: initialization. Bottom: trained. Dashed lines: pooled means.
- Trained Diagonal means: $0.943 \rightarrow 0.891 \rightarrow 0.826$ as μ increases.
- This supports reduced overwrite, not calibrated factual confidence.

Paper, original Fig. 2, PDF p. 10. Panels have different density-axis ranges. Fixed-weight probe: Appendix B, Table 6, p. 23.

Learning both noise terms does not win every metric

Setting ($\mu = 1$)	Wiki. PPL ↓	LMB. PPL ↓	Avg. % ↑
Learn r_t, ω_t	18.69	**15.51**	**54.12**
Fix $r_t = 1$	**18.46**	15.72	53.77
Fix both at 1	18.64	15.84	53.87

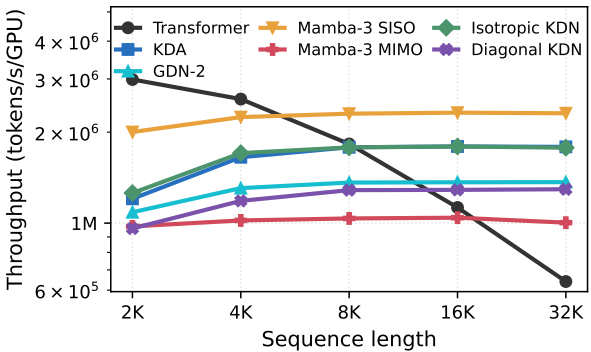
- Fixing observation noise improves WikiText, but lowers average accuracy.
- The evidence supports mixed, metric-dependent effects.
- This ablation uses $\mu = 1$, not the main setting $\mu = d_k$.

Missing comparison

The paper does not report the full noise-factorial ablation at the main information scale.

Paper, Table 5, PDF p. 15.

Isotropic stays close to KDA; Diagonal costs more



- One H200; mixer layers only.
- Batch 4, width 2048; median of 30 runs.
- BF16 setup; Diagonal uses FP32/TF32x3.
- Mamba-3 SISO leads at long lengths.

Full attention here; 2K SWA in the quality tables.

Paper, original Fig. 4 and throughput paragraph, PDF p. 15. Beyond 4K: runtime scaling probes, not quality validation. No end-to-end speedup is established.

The evidence does not imply universally better memory

- **Approximation:** diagonal/scalar states discard correlations; $\mu \neq 1$ adds a correction beyond the original variational posterior.
- **Statistics:** modest 1.3B accuracy gains without multi-seed uncertainty.
- **Architecture dependence:** at S-NIAH-1 / 8K, 1.3B Diagonal scores 99.8 recurrent vs 28.0 hybrid.
- **Efficiency:** extra gain scan and precision cost; no measured end-to-end decode benefit.
- **Interpretation:** learned covariance is not demonstrated to be calibrated factual confidence.

Paper, Secs. 4–5, Tables 1–2, PDF pp. 8–15; Appendix D, pp. 26–27. Calibration and generalization cautions are this reading's analysis.

Carry uncertainty forward, then let evidence control edits

1. **Concept:** interpret delta writes as Kalman innovations in a drifting associative map.
2. **Algorithm:** compress covariance to recover a Möbius gain scan and an affine memory scan.
3. **Evidence:** improved recurrent-model aggregates; the strongest gains appear in selected synthetic retrieval settings.

Open question

Can richer covariance and rotating transitions preserve both useful correlations and efficient scans?

Paper, Sec. 7, PDF p. 16. Rotation is proposed future work.

Ngoc Bui, Tinglin Huang, Rex Ying. *Kalman Delta Networks: Uncertainty-aware Associative Memory*. arXiv:2609.07816v1, 2026.

`https://arxiv.org/abs/2609.07816v1`

Backup: reverse KL explains the diagonal update

Let $\Lambda^\star = (P^\star)^{-1}$ and $q = \mathcal{N}_{\mathrm{col}}(S, \mathrm{diag}(p))$.

$$\mathrm{KL}(q\|p^\star) = \frac{1}{2}\,\mathrm{tr}\big((S - S^\star)^\top \Lambda^\star (S - S^\star)\big) + \frac{d_v}{2}\sum_i \big(\Lambda^\star_{ii} p_i - \log p_i\big) + C.$$

- Positive definite $\Lambda^\star$ gives $S = S^\star$.
- $\partial\,\mathrm{KL}\,/\partial p_i = 0$ gives $p_i = 1/\Lambda^\star_{ii}$.
- $\Lambda^\star = \widehat{P}^{-1} + kk^\top/r$ yields the elementwise recurrence.

Which posterior is exact?

The one formed from the already approximated predictive prior. The mean statement is local to this step.

Paper, Appendix A.2, Eqs. (54)–(57), PDF p. 21.

Backup: protection has a causal ordering

For a token-local intervention μ_t, hold the predictive prior fixed:

$$\frac{\partial \kappa_t}{\partial \mu_t} = 0, \qquad \frac{\partial S_t}{\partial \mu_t} = 0, \qquad \frac{\partial p_{t,i}}{\partial \mu_t} = -\frac{k_{t,i}^2}{r_t} p_{t,i}^2 \leq 0.$$

With next-token parameters fixed, this implies

$$\frac{\partial \beta_{t+1}^{\text{eff}}}{\partial \mu_t} \leq 0.$$

μ	1	$\sqrt{d_k}$	d_k	$4d_k$
Trained-at-scale mean	.943	.891	.826	.805
Fixed-weight probe	.886	.873	.862	.857

Paper, Appendix B, Eqs. (64)–(65), Table 6, PDF pp. 22–23. Probe protocols differ; compare within each row. Fixed weights do not fix downstream hidden states.

Backup: compact WY handles asymmetric write directions

Within a chunk, let $\gamma_r = \prod_{j \le r} \alpha_j$ and

$$\overline{k}_r = \gamma_r \odot k_r, \qquad \overline{\kappa}_r = \kappa_r \oslash \gamma_r.$$

Stack rows into $\overline{K}$, $\overline{K}^{(\kappa)}$, and V. Then

$$T = \mathrm{tril}\big(\overline{K}(\overline{K}^{(\kappa)})^{\top}, -1\big), \quad A = (I + T)^{-1}, \quad R = AV - A\overline{K}S^{\mathrm{in}}.$$

- A unit-lower-triangular solve computes chunk-local factors in parallel.
- Write direction κ_t and erase direction k_t need not be parallel.
- Include the query–write diagonal because output follows the current write.
- Chunk memory carries remain; decay ratios use log differences for stability.

Paper, Appendix D.2, PDF p. 27. R here stacks chunk residuals, not the observation-noise covariance.

Backup: training and evaluation details that affect scope

- AdamW: peak LR 4×10^{-4}, $\beta = (0.9, 0.95)$, weight decay 0.1; gradient clipping 1.0.
- Cosine schedule; 1% token-budget warm-up; global batch 0.5M tokens.
- Learned positive $c_{0,h}$ per head replaces the theoretical unit initial variance.
- RULER: 25-word increments and deterministic random essay windows; matched keys, values, and depths.
- Real retrieval: cloze formatting following JRT, inputs limited to 2K tokens.
- No training reproduction or independent author-code audit in this reading.

Paper, Sec. 5.1 and Table 2 caption, PDF pp. 11–13; Table 3, p. 14. Sources read: author TeX and all substantive PDF sections/appendices, v1.