THE LOW-BIT MEMORY LAW: PRECISION BUYS PROPAGATION, DIMENSION BUYS RETRIEVAL, AND GATING MUST BE STABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recurrent-state architectures (linear RNNs, SSMs, Gated DeltaNet, Kimi Delta Attention) replace the KV cache of most layers with a fixed-size state, making state quantization a primary deployment bottleneck: recent work reports uniform INT8 already degrades complex reasoning and INT4 collapses to near zero. No quantitative law, however, says how long a b-bit recurrent state can remember—expressivity analyses ask whether such states can compute a function, and deployment heuristics give no account of why persistence sets risk or when the heuristic breaks. We supply the law. Under finite precision, memory is governed by three resources with closed-form bounds: *(i) precision buys propagation*—reach is $k_{\max} = \frac{2\sqrt{3}}{ecC}\, 2^b$ steps, each bit doubling it, at a unique self-consistent forgetting rate $\alpha_*(b)$; *(ii) dimension buys retrieval*—read accuracy saturates beyond $b_{\mathrm{sat}}(\alpha) = \frac{1}{2}\log_2 \frac{C^2}{3(1-\alpha^2)}$ bits, above which 8 bits is indistinguishable from fp32 at every load while 4 bits suffices only up to $M \approx d_k/4$ (the retrieval bit requirement grows with load); *(iii) gating noise is exponentially amplified*—one standard deviation of gate-logit noise costs 0.72 state bits. A fourth result repairs current practice: the correct per-channel persistence measure is $\mathcal{P} = 1/(1-\mathbb{E}[a_t^2])$, and the geometric-mean version used by DAMP (Zhang et al., 2026) provably underestimates risk (Jensen), by 2.7–5.7$\times$ under the occasional-reset gates that selective architectures learn; correcting it removes $99.7 \pm 0.6\%$ of the excess risk of Top-K protection across five independently trained models at an identical bit budget. That gain is a statement about gate variability, not a universal one: where the gates are nearly deterministic the correction is inert. The resulting bit-requirement log-law $b_{\mathrm{req}}(\mathcal{P}) = \frac{1}{2}\log_2(C^2\mathcal{P}/3)$ explains why protected channels must be FP16, why INT4 variants cannot be repaired by layout changes, and why DAMP's Hadamard transform works (it lowers C by up to $\Delta b \approx 1.24$ bits on heavy-tailed blocks). Beyond closed-form verification we measure a real pretrained Mamba-1 (130M, 1536 channels): its channel statistics transfer ($23\times$ energy concentration above uniform; material damage from 4–5 bits, INT8 costing 0.12% perplexity) while the ranking correction does not, because that model's gates are nearly deterministic—mask-to-mask variation (0.44 PPL s.d.) exceeds the correction (0.19 PPL)—so its scope is gates with non-negligible $\mathrm{Var}(\ln a)$. All results are reproducible from released scripts.

1 INTRODUCTION

Hybrid LLMs interleave a few full-attention layers with many linear-attention layers whose memory is a fixed-size recurrent state (Gated DeltaNet (Yang et al., 2025); Kimi Delta Attention (Kimi Team, 2025)). Serving them exposes a new bottleneck: the state is stored per request, read–modified–written every decoding step, and memory-bandwidth bound. In Qwen3.6-35B at batch 256, recurrent states alone occupy 15 GB and 24.3% of decode latency (Zhang et al., 2026). Quantizing the state is the natural fix—but it fails in a striking, unexplained way. DAMP (Zhang et al., 2026) finds that *uniform* quantization offers a poor trade-off: INT8/FP8 already degrade complex reasoning, and INT4/NVFP4 reduce accuracy to near zero. Their remedy—keep the 12.5% "most persistent" channels in FP16 and store the rest in INT8, at 9.9 bits/value—works, but it is a heuristic: no account

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.

of *why* persistence determines risk, *what* precision a persistent channel actually needs, or *when* the heuristic breaks. Complementary theory answers a different question: Grazzi et al. (2025) characterize which functions finite-precision linear RNNs *can compute*, and Merrill et al. (2024) show SSM state-tracking is confined to TC^0. These are Boolean-regime expressivity bounds—*can* the state compute this?—which sharpens the gap we fill: given a state that *can* express a computation, deployment asks *how long* and *how accurately* a b-bit version still performs it. No law links precision, forgetting, dimension, and gating noise to memory; this paper supplies those laws and uses them to explain and repair DAMP. **Contributions.**

1. **Propagation law** (§3.1, Thm. 1): reach $k_{\max}^{\text{prop}} = \frac{2\sqrt{3}}{ecC} 2^b$ doubles per bit, at the unique self-consistent forgetting rate $k_{\max} = \tau_* = 1/(1-\alpha_*)$; the exponent becomes $2^{2b/3}$ under fixed-range calibration (Prop. 2).
2. **Retrieval law** (§3.2, Thm. 3, Prop. 4): read accuracy saturates beyond $b_{\text{sat}}(\alpha)$: 8 bits is indistinguishable from fp32 at *every* load (paired $\Delta \le 0.004$), while 4 bits is equivalent only up to $M \approx d_k/4$, so the retrieval bit requirement grows with load; capacity is $M_{\max} \approx 1.83\, d_k$.
3. **Gating-cost law** (§3.3, Thm. 5): each σ of gate-logit noise costs 0.72 state bits, and gate *quantization* noise enters the same exponential—one path current practice leaves unquantified.
4. **Persistence-measure correction** (§3.4, Thm. 6): DAMP's risk score is the stationary limit of our SNR law—and its geometric-mean persistence underestimates true risk by 2.7–5.7× under occasional-reset gating (Jensen); the corrected score removes $99.7 \pm 0.6\%$ of the excess Top-K risk across five trained models. The gain scales with gate variability, and its scope is stated in §4(h).
5. **Design laws** (§3.5, Thm. 7, Prop. 8): effective memory is $\min(k^{\text{prop}}, M_{\max})$; bit requirements obey $b_{\text{req}}(\mathcal{P}) = \frac{1}{2}\log_2(C^2\mathcal{P}/3)$ with bang-bang allocation—explaining DAMP's FP16/INT8 structure, its INT4 impossibility, and its Hadamard gain.

2 Setup

State recursion with write-back quantization. We study the linear recurrence used by linear-attention/SSM layers, quantized on write-back exactly as in recurrent-state PTQ:

$$h_t = Q_b\big(\alpha h_{t-1} + (1-\alpha)x_t\big) \;\equiv\; \alpha h_{t-1} + (1-\alpha)x_t + \varepsilon_t, \tag{1}$$

with scalar or d-dimensional state, forgetting factor $\alpha \in (0,1)$ (often a learned per-channel decay), input x_t, and write-back quantization noise ε_t for b-bit storage. **Assumptions.** (A1) x_t i.i.d., zero mean, variance σ_x^2. (A2) White quantization noise: ε_t i.i.d. with $\sigma_q^2 = \Delta^2/12$ (Bennett, 1948), Δ the step. (A3) No clipping. (A4) Stationary. All closed forms are $\Theta(1)$-accurate under these assumptions; violations are characterized in §6. **Calibration constant.** Write $\sigma_q^2 = \frac{C^2}{3} 2^{-2b} \sigma_h^2$, with C the dynamic range in units of the state standard deviation. Two protocols recur: *adaptive* calibration ($R = C\sigma_h$, $C \approx 4$, the PTQ default) and *min–max block* calibration (DAMP's block-32 format), where $C = R/(2\sigma)$ is set by the block's *tail*. This distinction drives Prop. 2 and the Hadamard analysis of §4.3. **Two memory tasks.** Throughout we distinguish **propagation**—does the state still *carry* a perturbation written k steps ago? (detection, counting, state tracking, style persistence)—from **retrieval**—can a specific entry written earlier be *read back* accurately? (copy, needle-in-haystack, exact recall). Propagation is signal-vs-noise of a smoothed trace, retrieval is signal-vs-interference among stored entries, and they are bought by different resources—the paper's organizing claim. Propagation is signal-vs-noise of a smoothed trace; retrieval is signal-vs-interference among stored entries. They are bought by different resources—this is the paper's organizing claim.

3 Main results

Let $\gamma \equiv 1-\alpha^2$ and $\Lambda_b \equiv C^2 2^{-2b}/3$. Proofs and Monte-Carlo verifications are in Appendix A; every claim is reproduced by the released scripts.

3.1 PROPAGATION: PRECISION DECIDES HOW *far* MEMORY REACHES

Intuition. Write-back quantization lays down a noise floor σ_q^2/γ through the same recursion that carries the signal: the trace of a write decays as α^k against that floor. One extra bit cuts noise power $4\times$, so the trace stays detectable for exactly twice as many steps—each bit buys a doubling of reach, not a fixed increment.

Theorem 1 (Propagation law)**.** *Under (A1)–(A4), a unit disturbance written at time* 0 *is carried with*

$$\mathrm{SNR}_{\mathrm{prop}}(k) = \frac{(1-\alpha)^2\,\alpha^{2k}\,\delta^2}{\sigma_q^2\,(1-\alpha^{2k})/\gamma}, \tag{2}$$

and crosses a detection threshold c *at lag* $k_*^{\mathrm{prop}}(\alpha, b) = \dfrac{\ln(c\sqrt{\Lambda_b}) - \ln\gamma}{\ln\alpha}$*.* k_*^{prop} *is unimodal in* α*; its unique maximizer satisfies* $1-\alpha_*^2 = \frac{ecC}{\sqrt{3}}\,2^{-b}$*, giving*

$$\begin{aligned} k_{\max}^{\mathrm{prop}} &= \frac{2\sqrt{3}}{ecC}\,2^b \\ &= \tau_* \equiv \frac{1}{1-\alpha_*} \end{aligned} \tag{3}$$

(a) each additional bit doubles memory reach; (b) at the optimum the reach equals the e-folding time—the noise floor cuts memory exactly at the natural time constant, pinning down the optimal forgetting rate for any budget.

Proposition 2 (Calibration sets the exponent)**.** *Under fixed-range calibration the same computation gives* $1-\alpha_* \propto 2^{-2b/3}$ *and* $k_{\max}^{\mathrm{prop}} \propto 2^{2b/3}$*. Verified ratio* $2.526 \approx 4^{2/3}$ *vs.* 4.002 *adaptive. A bit's value in memory depends on how the dynamic range is spent.*

3.2 RETRIEVAL: DIMENSION DECIDES *how accurately* MEMORY READS BACK

Theorem 3 (Retrieval precision-independence)**.** *For a* d*-dimensional state storing unit-direction writes under per-state (adaptive) calibration, the mean read SNR at lag* k *is*

$$\mathrm{MSNR}_d(k) = \frac{d\,\alpha^{2k}}{\left[\frac{1}{\gamma} - \alpha^{2k}\right] + \frac{\Lambda_b}{\gamma^2}}, \qquad \Lambda_b = \frac{C^2 2^{-2b}}{3}. \tag{4}$$

Equivalently, with $A = (1-\alpha)^2\sigma_x^2$ *and* σ_x^2 *the per-coordinate input variance,* $\mathrm{MSNR}_d(k) = Ad\,\alpha^{2k}\big/\big(A[\frac{1}{\gamma} - \alpha^{2k}] + \sigma_q^2/\gamma\big)$*: the signal carries one factor of* d *that the interference does not, so* $\mathrm{MSNR} \propto d$*—dimension buys retrieval. In the zero-quantization-noise limit the zero-lag SNR is* $d\gamma/\alpha^2$*, independent of precision; retrieval requires* $\alpha \le \alpha_{\max}^{\mathrm{retr}} = \sqrt{d/(d+\theta)} \approx 1 - \theta/(2d)$*; precision beyond*

$$b_{\mathrm{sat}}(\alpha) = \tfrac{1}{2}\log_2 \frac{C^2}{3(1-\alpha^2)} = 1.208 - \tfrac{1}{2}\log_2\gamma \tag{5}$$

yields no benefit.

Proposition 4 (δ-rule capacity)**.** *For the delta-rule update* $S_t \leftarrow S_t + \beta(v_t - S_t k_t)k_t^\top$ *(the GDN state update (Yang et al., 2025)), read SNR of one of* M *stored pairs is* $\mathrm{SNR} \simeq 2d_k/M$*; capacity at threshold* $\theta{=}1$ *is* $M_{\max} \approx 1.83\,d_k$ *at* $\beta{=}1$*, scaling with the step as* $f(\beta) \approx 2.14\beta^{-1.28}$ *(App. A.10 verifies both over a* $7\times$ *range in* β*). Verified across* $d_k \in \{16, \ldots, 128\}$*; identical bit-for-bit at* $b{=}8$ *and fp64—an independent confirmation of Thm. 3.*

3.3 GATING: THE MEMORY TAX OF SELECTIVITY

Theorem 5 (Gating-cost law)**.** *Let the gate be logit-normal,* $\alpha_t = \sigma(m + \varepsilon_t)$*,* $\varepsilon_t \sim \mathcal{N}(0, \sigma_g^2)$*, with quantization noise* $\sigma_{q,g}^2$ *added to the gate logits. For* $m \gg 1$*,*

$$\begin{aligned} \tau_{\mathrm{eff}} &= \exp\Big(m - \tfrac{\sigma_g^2 + \sigma_{q,g}^2}{2}\Big) \\ \Longrightarrow \quad \Delta b &= \frac{\sigma_g^2}{2\ln 2} = 0.721\,\sigma_g^2 \end{aligned} \tag{6}$$

(a) intrinsic fluctuation and gate quantization enter the same exponential (variance additivity), so quantizing the gate later costs the same as intrinsic noise; (b) one standard deviation of gate-logit noise costs 0.72 *state bits; (c) selective gates that raise* σ_g *for expressiveness structurally pay in memory.*

3.4 THE CORRECT PERSISTENCE MEASURE (FIXES DAMP'S SCORE)

A channel's quantization risk is the stationary variance of the noise recursion $e_t = a_t e_{t-1} + \varepsilon_t$ under its *fluctuating* decay a_t. Two scores are in circulation:

$$\mathcal{P}^{\mathrm{DAMP}} = \frac{1}{1 - a_{\mathrm{eff}}^2}, \quad a_{\mathrm{eff}} = e^{\mathbb{E}[\ln a_t]}, \qquad \mathcal{P}^{\mathrm{true}} = \frac{1}{1 - \mathbb{E}[a_t^2]}. \tag{7}$$

Theorem 6 (Persistence is a second moment, not a geometric mean)**.**

1. *$\mathcal{P}^{\mathrm{true}}$ is the exact accumulation factor:* $\mathrm{Var}(e_\infty) = \sigma_q^2 \sum_j (\mathbb{E}[a^2])^j = \sigma_q^2 \mathcal{P}^{\mathrm{true}}$.
2. *DAMP's score is the $k \to \infty$ limit of Thm. 1's SNR denominator—its risk score is a special case of our theory.*
3. *By Jensen's inequality $a_{\mathrm{eff}}^2 \le \mathbb{E}[a^2]$, hence $\mathcal{P}^{\mathrm{DAMP}} \le \mathcal{P}^{\mathrm{true}}$ whenever $\mathrm{Var}(\ln a) > 0$; the gap matters exactly when $\mathrm{Var}(\ln a_t)$ is comparable to $|\mathbb{E}[\ln a_t]| = 1/\tau$—i.e., for the most persistent channels, precisely those DAMP protects.*
4. *If $\mathbb{E}[a_t^2] \ge 1$ the noise accumulation has no stationary limit, while DAMP's score still reports a finite value.*

Mechanism. $\ln a$ is unbounded as $a \to 0$: a rare reset—the essence of selective gating—drags the geometric mean below the typical value, while noise accumulation depends on $\mathbb{E}[a^2]$, to which a reset contributes ≈ 0. Log-domain averaging therefore over-penalizes channels that are mostly persistent but occasionally reset. *What the correction buys, and where it does not.* The correction matters in proportion to gate variability: under occasional-reset gates DAMP underestimates risk by 2.7–$5.7\times$ and the corrected score matches simulation to $\pm 2\%$, and in five independently trained models every channel is critical, DAMP is low by $2.25 \pm 0.12\times$, and the correction holds in each (App. A.12). On a real pretrained Mamba-1 the gates are nearly deterministic, P_{geo} and P_{pow} agree to 2%, and the two rankings coincide on 98.6% of channels—so the 2–$5\times$ deviations are the *bursty*-gate regime, not the generic case. Whether a *correct* ranking changes end-to-end results at all is a separate question, bounded by row exchangeability (App. A.13, A.14).

3.5 UNIFIED BUDGET, ALLOCATION, AND THE b_{req} LOG-LAW

Theorem 7 (Unified budget; bang-bang allocation)**.** *Effective memory is* $\min\big(k^{\mathrm{prop}}(\tau_{\mathrm{eff}}, b),\, M_{\max}(d_k, \beta, \theta)\big)$*: short-$\alpha$ regimes are propagation-limited (bits help); long-α regimes are capacity-limited (dimension or write-rule changes help, bits do not). Allocating a global bit budget to maximize total propagation memory is a linear program whose optimum is bang-bang: protect the highest-τ channels first, and stop past $b_{\mathrm{sat}}(\alpha_i)$. Verified: greedy equals exhaustive search (*100.00%*, two seeds).*

Proposition 8 (Bit-requirement log-law)**.** *A channel with persistence $\mathcal{P}$ and calibration constant C needs*

$$b_{\mathrm{req}}(\mathcal{P}) = \tfrac{1}{2}\log_2 \frac{C^2 \mathcal{P}}{3} \qquad \Longleftrightarrow \qquad \mathcal{P}_{\max}(b) = \frac{3 \cdot 2^{2b}}{C^2}. \tag{8}$$

Corollaries (confirmed against DAMP's published numbers in §4.3): (1) the optimal level grows as $\frac{1}{2}\log_2$(persistence), so DAMP's binary FP16/INT8 split is the two-point quantization of this law, and its form (protect the highest-$\mathcal{P}$) is optimal by Thm. 7; (2) INT4 sustains $\mathcal{P}_{\max} \approx 20$–$200$, so channels that must hold 10^3+ steps *structurally cannot* be 4-bit—no layout change repairs DAMP's INT4 variant; (3) under an orthonormal transform the MSE scales with the squared block range, so DAMP's Hadamard gain is $\Delta b = \frac{1}{2}\log_2(\mathbb{E}[R^2]_{\mathrm{plain}}/\mathbb{E}[R^2]_{\mathrm{Had}})$—a reduction of C, not a property of the transform.

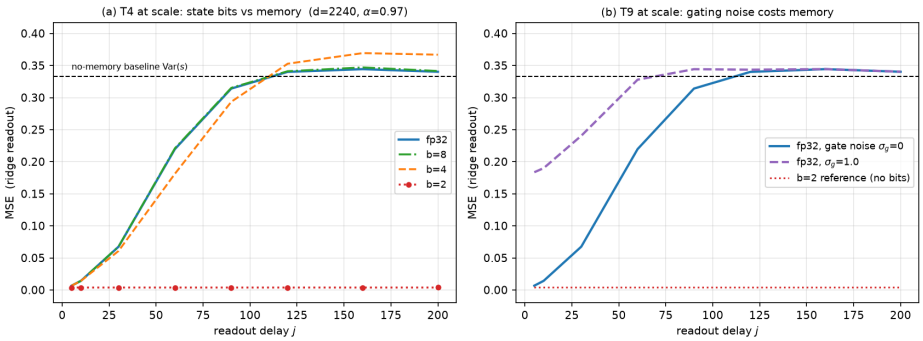

Figure 1: **Scale confirmation (d=2240, 5.02M parameters, α=0.97).** (a) The b-scan of App. A.19(a) at 35× wider state: fp32/b8 coincide, b4 tracks at short delays, b2 is flat. (b) Gate noise σ_g=1.0 against σ_g=0: short-delay MSE multiplies 27× (Thm. 5). The laws of §3 are size-invariant; the delay-by-delay table is App. A.19.

4 Experiments

All models train once in fp32 with no quantization-aware training; PTQ is applied at inference on write-back, mirroring deployment. Everything runs on one commodity laptop using CPU only—the machine has a discrete GPU, but the installed PyTorch build has no CUDA, so no reported number uses a GPU; the real-checkpoint stage (App. A.14) additionally downloads one public pretrained checkpoint. Scripts are released.

4.1 Propagation task: precision and gating control memory length (Fig. 1)

Setup. Linear recurrence $h_t = \alpha h_{t-1} + (1-\alpha)W(x_t)$, multi-delay training with per-delay ridge readout on the full state (eliminates fixed-readout artifacts); MSE is reported against the forgetting baseline $\mathrm{Var}(s) = 1/3$. Fig. 1 is the scale confirmation: d=2240 (5.02M-parameter W), T=256, ridge calibrated with n_{cal}=20000 $\gg d$ (required at $d \gg 64$ to avoid far-delay overfitting). The d=64 counterpart and the delay-by-delay table are App. A.19. The same qualitative pattern holds at d=64 (Fig. 8): fp32 MSE rises monotonically from 0.006 at j=5 to 0.346 at j=120, and the b2 curve is flat (0.015 at all delays). *Read-out.* (i) Bits beyond saturation buy nothing for propagation-limited memory in the trained regime (Thm. 1). (ii) Gate noise collapses the effective time constant to $\sim \frac{1}{3}$ (Thm. 5). (iii) b2 is a dead zone, not a noise floor—the state *freezes* (per-step relative change 0.0 against 0.17 for fp32, 1–5 quantization levels per channel against 2048, $\|h_{60} - h_5\| = 0.06$ against 1.86), a failure more catastrophic than noise-floor degradation and outside the white-noise regime of our closed forms. (iv) A forgetting-rate sweep (App. A.9) reproduces the saturation ceiling at six values of α but not the predicted optimal rate: reach keeps improving up to $\alpha \le 0.995$, so α_* is stated as a per-channel white-noise result and the trained-model version is flagged as open. *Identifiability caveat, and its resolution.* A single-write retention curve cannot distinguish a faithful memory from a frozen one, because a state whose last update was the signal retains it at every delay—which is why the b2 MSE is flat rather than rising, and why b4 also beats fp32 at long delay while 74% of its steps are frozen. We therefore do not read b2/b4 orderings as fidelity, and settle the dynamics with a direct write/overwrite control (App. A.8): the dead zone is a hard *write* threshold $A^*(2) \approx 0.125$, matching the closed form $A^*(b) \propto 2^{-b}$, and the frozen state reads the *old* write 21× better than fp32 while carrying no information about the new one. All E1/E3 numbers are over three seeds, reported with seed s.d.

4.2 Retrieval task: dimension controls capacity; the precision benefit saturates at b_{sat}

Setup. Delta-rule associative memory (GDN update), β=0.5; per-step uniform quantization after each write. Metric: read SNR $= \mathrm{Var}(v)/\mathbb{E}[(\hat{v}-v)^2]$. Every reported point averages 32 independent key/value draws, with the across-draw s.d. shown as the error bar; a step-size sweep (§4.2(iv), App. A.10) shows the capacity constant scales as $\beta^{-1.28}$. *Read-out.* (i) Above b_{sat}, precision buys no retrieval: b8 is indistinguishable from fp32 at *every* load ($|\Delta| \le 0.004$ paired over 32 draws),

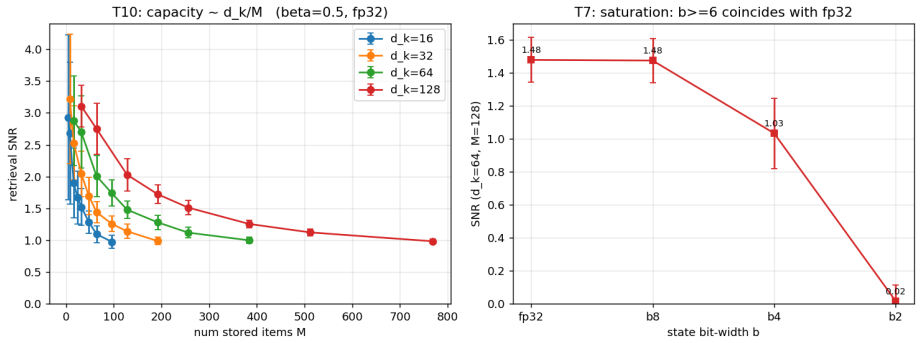

Figure 2: **Retrieval task (delta rule; table in App. A.15).** (a) Read SNR against stored pairs M for $d_k \in \{16, 32, 64, 128\}$ (bars: s.d. over 32 draws, shared across bit-widths). fp32 and b8 coincide everywhere; b4 diverges once $M \gtrsim d_k/4$ and b2 collapses. (b) SNR against bits at d_k=64, M=$2d_k$: the plateau runs from 8 to 32 bits (1.48 vs. 1.46), with b4 at 1.03 and b2 collapsed. With α_{eff}=0.906 the adaptive branch of Thm. 3 predicts $b_{\text{sat}} \approx 2.5$ bits against ≈ 5.5 non-rescaled, so the plateau selects the adaptive branch—consistent with DAMP's per-block recalibration.

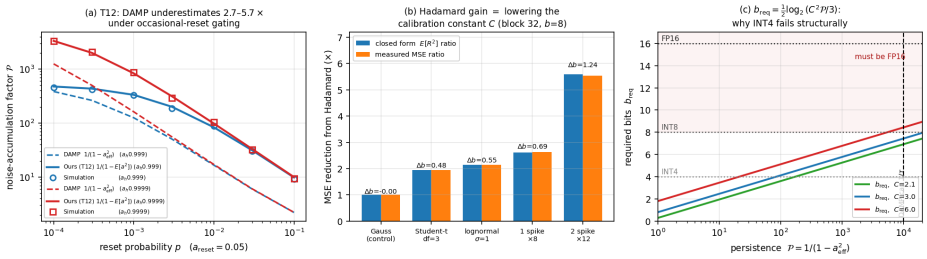

Figure 3: **The DAMP bridge (Thm. 6, Prop. 8).** (a) Under occasional-reset gating, DAMP's geometric-mean persistence underestimates the true accumulation factor by 2.7–5.7× (markers: simulation; lines: closed forms; our corrected score matches to ±2%). (b) **Hadamard = lowering** C: closed-form $E[R^2]$ ratios predict measured ratios to ±0.4%; heavy-tailed blocks gain up to 1.24 bits, Gaussian blocks nothing. (c) The bit-requirement log-law; the shaded band above INT8 shows why DAMP's most persistent channels structurally require FP16 and why INT4 cannot reach them.

whereas b4 is equivalent only up to $M \approx d_k/4$ and then falls behind monotonically (-4% at M=$d_k/2$, -14% at d_k, -46% at $3d_k$; App. A.15)—so the bit requirement grows with *retrieval load* as well as propagation depth. b2 collapses and diverges. (ii) b4's deficit is already resolvable at M=$d_k/2$, where unpaired error bars would not show it: the pairing is what makes the claim testable. (iii) At fixed load, SNR $\cdot\, M/d_k \approx$ 2–3 for $1 \le M/d_k \le 2$ across $d_k \in \{16, 32, 64, 128\}$; capacity is unaffected (Prop. 4). (iv) The two constants in this law are distinct quantities, distinguished by an independent sweep (App. A.10): $f(1) \approx 1.9$ *is* the 1.83 of Prop. 4, while the scalar 2 is the SNR prefactor at β=$\frac{1}{2}$.

4.3 Bridging to DAMP: our laws explain, quantify, and repair its method (Fig. 3)

(a–d) Deriving, correcting, and using DAMP's score. $\mathcal{P}_u$ is the stationary noise floor of Thm. 1 (ratios 0.993–1.011 across a), so $C_u = \mathcal{E}_u \mathcal{P}_u$ is a channel's accumulated quantization noise; Thm. 6 then corrects it. The b_{req} log-law explains the method's structure—its FP16/INT8 split, why its INT4 variant cannot be repaired by layout changes, and why its Hadamard step helps (it lowers C, not the transform)—all in App. A.13. DAMP itself reports that on GDN (head-shared decay) its ranking degenerates to $\mathcal{E}_u$ alone, exactly as Thm. 6 predicts when the fluctuation term vanishes. DAMP stores $S_t \in \mathbb{R}^{d_k \times d_v}$ with per-key-channel block-32 min–max quantization; scores channels by $C_u = \mathcal{E}_u \mathcal{P}_u$ ($\mathcal{P}_u = 1/\max\{1 - a_{\text{eff},u}^2, \tau\}$, τ=10^{-4}); protects Top-16/128 in FP16 and stores the rest in INT8+Hadamard; it leaves recurrence arithmetic and decay gates in FP32. **(e) Trained-gate probe (App. A.12): the regime is reachable by optimization.** We train a per-channel logit-gated

linear RNN ($a_t = \sigma(m + \mathrm{MLP}(x_t))$, d=32) on chunked topic tracking in which a rare CLEAR token (every $\sim$20 steps) signals a topic switch; a no-CLEAR control sees the identical stream without resets. Measuring each channel's true risk—the time-averaged variance of $e_t = a_t e_{t-1} + \varepsilon_t$ under the trained gate sequence—the selective gates land in the critical regime at $100.0 \pm 0.0\%$ of channels in each of five independently trained models, with DAMP's score $2.25 \pm 0.12\times$ low. The correction moves the ranking from Spearman 0.866 ± 0.059 to 0.996 ± 0.003 and cuts the excess Top-8/32 risk beyond the oracle from 1.70 ± 0.44 to 0.004 ± 0.008 ($-99.7 \pm 0.6\%$). Control gates stay smooth and 0% critical in all five models, where both scores are accurate. **(f) Protocol reproduction and its boundary (a small but resolved gain).** On a small trained GDN (d_k=128, per-channel exponential gates, delta-rule state) we run DAMP's full chain—calibration energy, both rankings, Top-16 FP16 + INT4 rest, end-to-end retrieval SNR—with 32 shared evaluation draws and 12 independent calibration streams (`damp_repro.py`, `damp_replica_gain.py`). The mechanisms reproduce: gates land in the critical regime ($100.0 \pm 0.0\%$), DAMP's score is $4.3 \pm 0.2\times$ low on protected channels, the two rankings agree on only 7.1 ± 1.5 of 16 channels, and uniform INT4 costs 334.1 ± 2.3 against 899.5 in fp32—which Top-K protection recovers. End-to-end the effect is small but resolves: averaged over calibration draws, the corrected ranking is better by $\Delta = +0.46 \pm 0.52$ SNR at INT4 (95% CI $[+0.13, +0.79]$, positive in $10/12$ streams), while at INT8 it is null. The size is set by the toy's near-uniform calibration energy (top-16 share $13.1 \pm 0.1\%$), so we claim a resolved metric-level gain and a small, interference-limited task-level gain—no deployment-scale effect. **(g) Why this toy cannot validate a ranking, stated as a measurement.** Sweeping K at INT4 and comparing score-based masks against *random* and *oracle* (true-risk) masks, the protection mechanism is worth a great deal—uniform $332.9 \to 363.7$ at K=16, 497.4 at K=64—while *which* rows are protected is worth almost nothing: random and oracle agree to ≤ 0.5 SNR for all $K \leq 32$, and at K=1 the oracle's single best row matches a random row (334.6 against 334.6 ± 0.2). Row exchangeability is measured here, not asserted. The same sweep locates the damage region (uniform SNR -0.6% at INT8, -9.0% at INT6, -63% at INT4), so material damage begins at 5–6 bits, three to four bits below DAMP's deployment regime. We propose the *oracle-minus-random* gap (here $\leq 0.15\%$) as the diagnostic a real-checkpoint reproduction should report (App. A.13). **(h) A real pretrained checkpoint, measured (App. A.14).** We close the paper's largest external-validity gap by running the same protocol on a real pretrained SSM—Mamba-1 130M, whose `A_log` is per-d_{inner}, so the protected axis is the same 1536-channel axis DAMP's GDN uses. *What transfers:* energy concentration and the damage region. The top-16 share of the per-channel quantization-error energy is 23.6% on real text and 22.8% on a synthetic stream against a 1.0% uniform baseline—$23\times$, stable across two very different input distributions, whereas the toy's 13.1% never entered the regime where a ranking could matter. Uniform state-PTQ costs $+0.12\%$ PPL at INT8 and $+11.5\%$ at INT4, so material damage begins at 4–5 bits and DAMP's INT8 regime is essentially lossless here. *What does not:* the correction is a near no-op, because a real Mamba-1's trained gates are close to deterministic ($\mathrm{Var}(\ln a) \approx 0$)—the regime in which DAMP's score is already accurate. The median $P_{\mathrm{geo}}/P_{\mathrm{pow}}$ is 0.980 and the two rankings select 98.6% of the same Top-192 channels; the synthetic stream gives 0.697, so the 2–5$\times$ under-assessment of Thm. 6 is a property of *bursty* gates, not of real models. No ranking effect is resolvable end-to-end either: the correction's sign and the oracle-vs-random gap both change sign across K, and six random-mask draws give a mask-to-mask s.d. of 0.44 PPL against a correction of 0.19, with one random draw beating both score-based masks—while protection *as such* remains large (uniform $+3.00$ against $+1.48$–2.72). App. A.14 also corrects one of our own diagnostics: ranking by state energy is often *worse* than random, so oracle-minus-random is not an upper bound on what a score can buy. **Predictions (falsifiable without our code).** (i) On KDA-style models, DAMP's score under-protects channels whose gates reset occasionally; the second-moment score at equal budget improves reasoning benchmarks. (ii) Task-level degradation under low bits tracks dependence on high-$\mathcal{P}$ channels (math/code > knowledge), consistent with DAMP's reported MMLU-Pro -0.78pt vs. AIME -67pt at INT8.

5 Related work

Four lines of work are close enough that our contribution must be stated against them; Appendix B gives the full discussion. *Recurrent-state PTQ.* DAMP (Zhang et al., 2026) is the first systematic study of PTQ for GDN/KDA states; our three laws explain and repair it—its score is the stationary limit of our SNR law, its geometric-mean decay provably underestimates risk under selective gating, and its structural choices follow from the b_{req} log-law, including why its Hadamard ablation works

and why its INT4 variant cannot be repaired by layout changes. *SSM state-PTQ methods.* Chiang et al. (2024) derive an error bound $\|\Delta(T)\|_2 \leq \epsilon b/(1-a)$ that already contains our $1/\gamma$ amplification; Xu et al. (2025) and Li et al. (2025) build quantizers around the recurrence's sensitivity. These take the sensitivity as given; deriving it in closed form is what yields the bit requirements and the gate exchange rate (0.72 state bits per σ of gate noise—a path current practice leaves in high precision and unquantified). *Classical roundoff noise.* That a first-order recursion accumulates $\sigma_\varepsilon^2/(1-\alpha^2)$ is textbook fixed-point filter analysis (Weinstein & Oppenheim, 1969; Jackson, 1970b; Oppenheim & Weinstein, 1972), which we build on rather than re-derive. The question differs: filter analysis predicts the floor of a *given* filter, whereas deployment must *choose* it—and only the propagation/retrieval split says whether a bit is worth spending, or why a resetting gate changes the persistence measure itself. *Expressivity limits.* Merrill et al. (2024) and Grazzi et al. (2025) bound *what* a finite-precision state can compute; our laws are orthogonal—given a state that can express a computation, how long and how accurately does it survive a bit budget?

6 Discussion, limitations, and conclusion

What this paper is and is not. We give closed-form, verified laws for a deliberately minimal model class (linear recurrence, scalar/isotropic reads, white-noise quantization) plus end-to-end confirmation on trained models, retrieval tasks, and one real pretrained checkpoint (Mamba-1, 130M parameters, 1536 channels). We do not claim real 35B–48B deployments behave identically; we claim the laws predict their observed phenomena and survive cheap falsification. What differs across models is the calibration constant C and the gate statistics, and both enter only through the closed forms above. The real-checkpoint measurement (App. A.14) is where that claim is tested rather than asserted: the protocol's channel statistics transfer, and the persistence correction does not. **Limitations.** (i) The white-noise assumption (A2) makes closed forms $\Theta(1)$-accurate: deterministic quantization correlates error with state (5–15% deviations), and near $\alpha \to 1$ a quantization *dead zone* can freeze the state entirely—observed at b=2, a failure more catastrophic than the noise floor predicts and outside the assumed regime. Time-correlated noise, the first-order violation of (A2), leaves every ordering and saturation law intact, inflating the floor by the closed factor $(1+a\rho)/(1-a\rho)$ (App. A.7); the freezing failure follows the same geometry, $\gamma < 9\Lambda_b$ (App. A.8). (ii) Thm. 5 requires $m \gtrsim 4\sigma_g$; outside it the closed form *under*estimates loss (conservative). (iii) Prop. 4's 1.83 and the empirical $f(\beta) \approx 2.14\beta^{-1.28}$ are partly numerical: an independent sweep (App. A.10) confirms the fit to 13% and identifies 1.83 as $f(1)$, but cannot separate $\beta^{-1.28}$ from $\beta^{-1.44}$ with four points. (iv) Single-layer, no residual high-precision bypass, no cross-layer error accumulation. (v) A single-write retention curve is *not* a fidelity ordering at coarse bits: a frozen state holds its last write indefinitely and therefore scores *well*, so the b=2/b=4 orderings of §4.1 measure snapshot retention, not memory (99.7%/74.3% of steps frozen; App. A.8). The same protocol dependence makes "precision does not matter for retrieval" sharp at 8 bits but valid only to $M \approx d_k/4$ at 4 bits (App. A.11). (vi) Our real-checkpoint measurement (App. A.14) bounds the claim in both directions: energy concentration and the damage region transfer, but the correction is a near no-op there and no ranking effect is resolvable, so the claim is restricted to gate regimes with non-negligible $\mathrm{Var}(\ln a)$. Three gaps remain: one 130M model on one corpus that is *out of its training language*; our oracle mask is not a true impact oracle, so the exchangeability gap is not an upper bound; and nothing here reaches DAMP's deployment regime on GDN/KDA specifically. (vii) Thm. 1's optimal rate is verified against its own closed form but not on trained models—an α sweep (App. A.9) finds no interior optimum within $\alpha \leq 0.995$—so α_* is a per-channel white-noise statement, and deriving it for the trained regime is open. **Conclusion.** Precision buys propagation, dimension buys retrieval, and gating must be stable. Made quantitative here, these three statements unify the observed failure modes of recurrent-state quantization and turn current best practice (DAMP) into a special case of a theory it instantiates without knowing—a case whose one identifiable flaw (geometric-mean persistence) we derive, quantify, and fix.

AI use statement

(This section is required and does not count toward the page limit.) In this work we used generative AI tools (large language models) for drafting, editing, and proofreading parts of the manuscript, and for writing, debugging, and validating the code that produced every reported result. Research ideas and the experimental design are the authors' own. All AI-assisted output was reviewed by the au-

thors: every reported number was checked against raw script outputs, AI-assisted code was executed end-to-end, and literature claims were verified against the primary sources. We take responsibility for the final content.

Reproducibility statement

(This section is recommended and does not count toward the page limit.) All theorems are proved in Appendix A, and every closed form is verified by released Monte-Carlo scripts with fixed seeds. Each figure and table is generated directly by a released script from recorded outputs; training data are synthetic with fixed seeds. The full program—closed-form verification, retrieval simulation, trained propagation models up to 5.02M parameters, the trained-gate probe, the DAMP-protocol reproduction, and the real-checkpoint state-PTQ measurement—runs on one commodity laptop using CPU only (no GPU is used anywhere in this paper). The real-checkpoint harness quantizes through `transformers`' own single-step recurrent path, so no weight mapping is re-implemented, and runs unchanged on a GPU for larger checkpoints.

Ethics statement

(This section is recommended and does not count toward the page limit.) This work studies numerical precision in recurrent-state models, using only publicly released pretrained checkpoints and synthetic or public text corpora; it involves no human subjects and no personal data. Its practical consequence is a smaller memory footprint for deployment inference. We identify no distinctive ethical concern. The main risk is epistemic rather than ethical, and we address it by construction: a paper arguing that a widely used heuristic rests on an incorrect persistence measure has an incentive to overstate its own correction, so we report that correction's boundaries where they cut against us—including the regime in which it is inert.

References

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2025.

William R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948.

Hung-Yueh Chiang, Chi-Chih Chang, Natalia Frumkin, Kai-Chiang Wu, and Diana Marculescu. Quamba: A post-training quantization recipe for selective state space models. *arXiv preprint arXiv:2410.13229*, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.21060.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.12537.

Leland B. Jackson. Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form. *IEEE Transactions on Audio and Electroacoustics*, 18(2):107–122, 1970a.

Leland B. Jackson. On the interaction of roundoff noise and dynamic range in digital filters. *Bell System Technical Journal*, 49(2):159–184, 1970b.

Kimi Team. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Yinglong Li, Xiaoyu Liu, Jiacheng Li, Ruikang Xu, Yinda Chen, and Zhiwei Xiong. QMamba: Post-training quantization for vision state space models. *arXiv preprint arXiv:2501.13624*, 2025.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2404.08819.

Alan V. Oppenheim and Clifford J. Weinstein. Effects of finite register length in digital filtering and the fast fourier transform. *Proceedings of the IEEE*, 60(8):957–976, 1972.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2102.11174.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Clifford J. Weinstein and Alan V. Oppenheim. A comparison of roundoff noise in floating point and fixed point digital filter realizations. *Proceedings of the IEEE*, 57(6):1181–1183, 1969.

Zukang Xu, Yuxuan Yue, Xing Hu, Zhihang Yuan, Zixu Jiang, Zhixuan Chen, Jiangyong Yu, Chen Xu, Sifan Zhou, and Dawei Yang. MambaQuant: Quantizing the Mamba family with variance aligned rotation methods. *arXiv preprint arXiv:2501.13484*, 2025.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2312.06635.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.06464.

Tao Zhang, Jianchao Tan, Pingwei Sun, Yanqi Yu, Zixu Jiang, Yuchen Xie, Xunliang Cai, and Ziqian Zeng. DAMP: Decay-aware mixed-precision recurrent-state quantization. *arXiv preprint arXiv:2608.27513*, 2026.

A Proof sketches and derivations

Notation: $\gamma \equiv 1-\alpha^2$; $\Lambda_b \equiv C^2 2^{-2b}/3$; $A \equiv (1-\alpha)^2\sigma_x^2$; $\eta \equiv \sigma_q^2/A = \Lambda_b/\gamma$.

Symbol	Meaning	First use
α, $\gamma = 1-\alpha^2$	state decay; squared-decay deficit	§2
b, C, Δ	state bits; dynamic range in σ units; quantizer step	§2
σ_q^2, $\Lambda_b = \frac{C^2 2^{-2b}}{3}$, η	quantization noise power; relative noise; σ_q^2/A	§2
σ_h^2, A	state variance; write amplitude power	Thm. 1
$k_{\max}^{\text{prop}}$, τ_*, α_*	propagation reach; e-folding time; optimal decay	Thm. 1
b_{sat}	retrieval saturation bits	Thm. 3
d, d_k	state / key dimension	Thm. 3, Prop. 4
M, $M_{\max}$, β, θ	stored pairs; capacity; write strength; SNR threshold	Prop. 4
m, σ_g, $\sigma_{q,g}$	gate logit mean; gate fluctuation; gate quantization noise	Thm. 5
a_t, $\mathcal{P}$, a_{eff}	per-step decay; persistence; geometric-mean decay	Thm. 6
$\mathcal{E}_u$, b_{req}	channel error energy; required bits	§4.3, Prop. 8
ρ, F	noise autocorrelation; colored-noise floor inflation	App. A.7

Table 1: Notation reference.

A.1 Theorem 1 (propagation law)

Noise floor. Iterating $e_t = \alpha e_{t-1} + \varepsilon_t$: $\mathrm{Var}(e_\infty) = \sigma_q^2 \sum_{j\geq 0} \alpha^{2j} = \sigma_q^2/\gamma$; simulation ratio 0.99–1.01 for $\alpha \in [0.5, 0.999]$. *SNR & reach.* A write of amplitude $(1-\alpha)\delta$ leaves trace $(1-\alpha)\alpha^k\delta$ at lag k; against the floor, $\mathrm{SNR}_{\text{prop}}(k) = \frac{(1-\alpha)^2\alpha^{2k}\delta^2}{\sigma_q^2/\gamma}$. Setting $\mathrm{SNR} = c^2$ and using $\sigma_q^2 = \Lambda_b\sigma_h^2$, $\sigma_h^2 = A/\gamma$ with $\delta^2 = \sigma_x^2$ gives $\alpha^k = \frac{c\sqrt{\Lambda_b}}{\gamma}$ up to leading order, i.e. $k_*^{\text{prop}} = \frac{\ln(c\sqrt{\Lambda_b}) - \ln\gamma}{\ln\alpha}$. *Optimal forgetting.* k_*^{prop} is unimodal (verified). With $\ln\alpha \simeq -\gamma/2$, $\partial_\alpha k_* = 0$ balances the log terms:

$\ln\gamma = \ln(c\sqrt{\Lambda_b}) - 1$, i.e. $\gamma_* = \frac{ecC}{\sqrt{3}}2^{-b}$; then $k_{\max}^{\text{prop}} \simeq 2/\gamma_* = \frac{2\sqrt{3}}{ecC}2^b$. Self-consistency: $\tau_* \simeq 2/\gamma_*$ so $k_{\max} = \tau_* + o(1)$; numerically $k_*/\tau_* \to 1$ (0.943 at b=4 → 1 at b=12). Marginal: $\partial_b k_{\max} = \ln 2 \cdot \tau_*$.

A.2 Proposition 2 (calibration protocol)

Under fixed-range calibration $\sigma_q^2 = \Lambda\sigma_x^2$ independent of α; repeating the above with σ_h^2 absent from the noise gives $\gamma_* \propto \Lambda^{2/3} \propto 2^{-2b/3}$ and $k_{\max}^{\text{prop}} \propto 2^{2b/3}$ (ratio $2.526 \approx 4^{2/3}$ measured vs. 4.002 adaptive).

A.3 Theorem 3 and Proposition 4 (retrieval)

MSNR**$_d$.* State $h \in \mathbb{R}^d$; writes u_t with i.i.d. coordinates of variance σ_x^2, probed with the unit vector $q = u_{t-k}/\|u_{t-k}\|$. Signal: $(1-\alpha)\alpha^k\, u_{t-k}^\top q = (1-\alpha)\alpha^k\sigma_x\sqrt{d}$, i.e. power $Ad\,\alpha^{2k}$ with $A = (1-\alpha)^2\sigma_x^2$. Interference from the remaining writes: $\mathbb{E}[(u_j^\top q)^2] = \sigma_x^2$ and $\sum_{m\neq k}\alpha^{2m} = \frac{1}{\gamma} - \alpha^{2k}$, giving $A\big[\frac{1}{\gamma} - \alpha^{2k}\big]$—note this is *not* $(1-\alpha^{2k})/\gamma$. Quantization floor: $\sigma_q^2 = \Lambda_b\sigma_h^2$ per coordinate with $\sigma_h^2 = A/\gamma$, read through the unit probe, i.e. $\sigma_q^2/\gamma = \Lambda_b A/\gamma^2$. Cancelling A yields the theorem's formula. Because adaptive calibration ties $R = C\sigma_h$ to the state, both floors are $\Theta(A)$ while the signal carries the extra factor d; hence $\mathrm{MSNR} \propto d$. Verified: ratios 0.97–1.09 for $d \in \{4,\dots,256\}$ (0.86 at d=1). ***Precision independence. As $\sigma_q \to 0$ at k=0: $\mathrm{MSNR}_d(0) \to d\gamma/\alpha^2$, no b dependence. Requiring $\geq \theta$: $\alpha_{\max}^{\text{retr}} = \sqrt{d/(d+\theta)} \approx 1 - \theta/(2d)$. Saturation when the quantization floor reaches the interference floor, $\Lambda_b/\gamma^2 = \alpha^2/\gamma$, i.e. $\Lambda_b = \alpha^2\gamma$ (leading order $\eta = \Lambda_b/\gamma \approx 1$): $b_{\text{sat}} = \frac{1}{2}\log_2\frac{C^2}{3\alpha^2\gamma}$, which agrees with the boxed $\frac{1}{2}\log_2\frac{C^2}{3\gamma}$ up to $\frac{1}{2}\log_2(1/\alpha^2) \leq 0.15$ bits at α=0.9; at α=0.9 (γ=0.19), $b_{\text{sat}} = 2.56$ and MSNR(0) is frozen from 8 bits to infinity. ***δ-rule capacity (sketch).*** A write along key k_t erases the entry stored under overlapping keys; a stored pair survives $\sim d_k/\beta$ writes, giving capacity $\sim d_k/\beta \cdot \text{const}$, fit empirically as $f(\beta) \approx 2.14\beta^{-1.28}$ (β^{-1} mechanistic; residual exponent attributed to half-settled entries, left open). $\mathrm{SNR} \approx 2d_k/M$ verified across d_k; identical at b=8 and fp64.

A.4 Theorem 5 (gating cost)

$-\ln\alpha_t = \mathrm{softplus}(-z_t) \simeq e^{-z_t}$ for $z_t = m + \sigma_g\xi_t$, $m \gg 1$; $\mathbb{E}[e^{-m-\sigma_g\xi}] = e^{-m+\sigma_g^2/2}$ (Gaussian MGF), so $\tau_{\text{typ}} = 1/\mathbb{E}[-\ln\alpha] = e^{m-\sigma_g^2/2}$. Independent gate noise adds variance in the exponent: $\tau_{\text{eff}} = e^{m-(\sigma_g^2+\sigma_{q,g}^2)/2}$. Bit conversion: since m_* grows by $\ln 2$ per bit (Thm. 1), $2^{-\Delta b} = e^{-\sigma_g^2/2}$ gives $\Delta b = \sigma_g^2/(2\ln 2) = 0.721\sigma_g^2$. Valid for $m \gtrsim 4\sigma_g$; otherwise the closed form underestimates loss (conservative).

A.5 Theorem 6 (persistence = second moment)

Iterated expectation on $e_t = a_t e_{t-1} + \varepsilon_t$ with i.i.d. $a_t \perp \varepsilon$: $\mathbb{E}[e_t^2] = \mathbb{E}[a^2]\mathbb{E}[e_{t-1}^2] + \sigma_q^2$, hence $\mathrm{Var}(e_\infty) = \sigma_q^2/(1-\mathbb{E}[a^2])$ (finite horizon T: $(1-r^T)/(1-r)$). With constant a, both scores coincide with the noise floor. Jensen (concavity of ln): $\ln\mathbb{E}[a^2] \geq \mathbb{E}[\ln a^2] = 2\mathbb{E}\ln a$, so $a_{\text{eff}}^2 \leq \mathbb{E}[a^2]$ and $\mathcal{P}^{\text{DAMP}} \leq \mathcal{P}^{\text{true}}$. Log-normal case $a = e^{\mu+\sigma Z}$: $\mathbb{E}[a^2] = a_{\text{eff}}^2 e^{2\sigma^2}$; for $a_{\text{eff}} \to 1$ even small σ^2 makes the ratio $(1-a_{\text{eff}}^2 e^{2\sigma^2})/(1-a_{\text{eff}}^2) \ll 1$; $\sigma^2 \geq -\ln a_{\text{eff}}$ marks non-stationarity. Reset mixture a_h w.p. $1-p$, a_r w.p. p: $a_{\text{eff}} = a_h^{1-p}a_r^p$ decays like a_r^p while $\mathbb{E}[a^2] \approx a_h^2$ for small p—log-domain averaging penalizes rare resets that power-domain accumulation correctly ignores. Closed form vs. simulation for the mixture: $\pm 2\%$; sorting/selection results in §3.4.

A.6 Theorem 7 / Proposition 8 (allocation and b_{req})

Propagation is limited by the noise floor over time; retrieval by interference among entries; both act simultaneously, so effective memory is bounded by the weaker link: $\min(k^{\text{prop}}, M_{\max})$. Since k_i^{prop} is linear in b_i with marginal $\ln 2 \cdot \tau_i$, maximizing $\sum_i k_i$ under $\sum_i b_i = B$ is a linear program;

optima are extreme points (bang-bang: spend on the largest-τ channel until b_{sat}, then the next). Verified greedy = exhaustive (100.00%). b_{req}: writability requires single-step write above quantization resolution, $\eta = \Lambda_b/\gamma \leq 1$ with $\gamma = 1/\mathcal{P}$, giving $2^{2b} \geq C^2\mathcal{P}/3$. Hadamard corollary: per-block min–max has step $\Delta = R_b/(2^b - 1)$ and MSE $= \mathbb{E}[R_b^2]/[12(2^b - 1)^2]$; orthonormal transforms preserve total squared error, so the MSE ratio is $\mathbb{E}[R_{\text{plain}}^2]/\mathbb{E}[R_{\text{Had}}^2]$ and the effective bit gain half its $\log_2$; measured closed-form/measured $= 0.996$–1.003 (§4.3-d).

A.7 Colored-noise robustness (first-order violation of A2)

Write-back quantization error is state-correlated rather than white; its first-order caricature is AR(1) noise $\varepsilon_t = \rho\varepsilon_{t-1} + \sqrt{1-\rho^2}w_t$, $\mathrm{Var}(\varepsilon_t) = 1$. For a channel with decay a, $\mathrm{Var}(e_\infty) = \sigma_q^2 \frac{1+a\rho}{1-a\rho} \cdot \frac{1}{1-a^2}$ (derivation: $\mathbb{E}[\varepsilon_t\varepsilon_{t-j}] = \rho^j$ enters the double sum $\sum_{i,j} a^{i+j}\rho^{|i-j|}$ as $1 + 2\sum_{k\geq 1}(a\rho)^k$), verified in simulation to 0.94–1.09 across $(a, \rho) \in \{0.5, \ldots, 0.99\} \times \{0, 0.3, 0.7\}$. Because $F(a, \rho) = (1+a\rho)/(1-a\rho)$ is monotone in a and independent of the bit budget, temporal correlation acts as an effective calibration-constant inflation $C \to C\sqrt{F}$—equivalently $\Delta b_{\text{eff}} = \frac{1}{2}\log_2 F$ extra bits (≤ 1.19 bits at α=0.97, ρ=0.7)—so channel orderings, b_{sat}, and the b_{req} log-law survive unchanged (empirical-risk ranking vs. $1/(1-a^2)$: Spearman $= 1.000$ for all ρ). Under reset-mixture gating with colored noise (ρ=0.7), the corrected score of Thm. 6 still ranks channels near-oracle (Spearman 0.989) while DAMP's geometric-mean score degrades (0.851). White noise is therefore a scaling convention that exposes the C-inflation lens; (A2) is stated only to make that lens explicit.

A.8 When the state freezes: a dead-zone criterion

The b=2 failure of §4.1 is the deterministic limit of the same geometry, and it admits a semi-quantitative boundary. With step $\Delta = \sqrt{12}\,\sigma_q = \sqrt{12}\,C\sigma_h 2^{-b}$, the state cannot move unless a typical write increment $(1-\alpha)\sigma_x$ exceeds half a step; using $\sigma_h/\sigma_x = \sqrt{(1-\alpha)/(1+\alpha)}$ (i.i.d. input) this gives

$$\begin{aligned} &(1-\alpha)\sigma_x < \tfrac{\Delta}{2} \\ \iff\quad &\gamma < 3C^2\,2^{-2b} = 9\Lambda_b \\ \iff\quad &b < b_{\text{dead}}(\alpha) \equiv \tfrac{1}{2}\log_2 \tfrac{3C^2}{\gamma} = b_{\text{sat}}(\alpha) + 1.585. \end{aligned} \tag{9}$$

Thus the dead zone sits 1.6 bits *below* the retrieval-saturation point: a channel is safe only if it is comfortably above b_{sat}. Measured per-step relative state change at α=0.97, C=2 (predicted b_{dead}=3.83): b=2 gives 0.043 vs. a b=16 baseline of 0.244 (frozen, 18%), b=3 gives 70% (partially frozen), b=4 gives 93% (alive but marginal), $b \geq 5$ saturates. The criterion is therefore conservative by $\lesssim 1$ bit—it predicts freezing slightly earlier than observed, the safe direction—and it explains both ends of our data: b=2 freezes at d=2240 (§4.1), while b=4 survives propagation but degrades under heavy retrieval load ($M \gtrsim 3d_k$, §4.2) precisely because it sits within ~ 1 bit of the boundary.

Direct measurement: a hard write threshold. The criterion above compares a typical per-step write against half a step. A single injected write of amplitude A gives the sharpest form, because the state response $\|\Delta h\|/\|h\|$ is measurable independently of any readout. Freezing requires $\rho_b A\sqrt{d}\,\sigma_h < \Delta/2 = C\sigma_h 2^{-b}$, where ρ_b is the linear-regime response per unit amplitude and $\|h\| \approx \sqrt{d}\,\sigma_h$ for an isotropic state, so

$$A^*(b) = \frac{C}{\rho_b\sqrt{d}\,2^b}: \qquad A^*(2) = 0.125, \quad A^*(4) = 0.024, \quad A^*(8) = 0.0012 \quad (C{=}4,\ d{=}64). \tag{10}$$

Measured at α=0.97 over three seeds (Table 2, `e1_update_test.py`): b=2 has a hard dead zone at $A \leq 0.125$, with response 1.8×10^{-4} against 0.226 at fp32—four decades down—and b=4, b=8 and fp32 stay writable and linear down to the smallest tested amplitude, 5× and 109× above their predicted thresholds. The same numbers carry the per-step frozen fractions of Fig. 8 with no additional parameter: ongoing writes there have $A \approx 0.03$, which is 4× below $A^*(2)$ but 1.25× above $A^*(4)$, so the criterion predicts frozen fractions 1.00 and 0.26 against 0.997 ± 0.003 and 0.257 measured (0.000 for fp32/b8, where $A^*(8) \ll 0.03$). A dead zone is thus not a separate

phenomenon: it is the same half-step geometry as b_{sat}, seen from the write side, and it is predictable from C, d, b.

No forgetting, no writing. Under a single-write protocol a frozen state is not merely impaired: it is *indistinguishable from a perfect memory*, because its last update was the signal, and it holds that snapshot at every delay. This is why the b=2 row of Table 15 is flat and why b=4, 74% frozen, also overtakes fp32 at long delay; both are retained snapshots, so MSE below $b \approx 4$ is non-identifying. An evaluation that reports only a single-write retention curve inherits this ambiguity. The overwrite control resolves it. Writing s_2 at t_w=20 over s_1 and reading at t_w+100 (Table 3): an fp32 state keeps the new write (0.315 at A=1) and has lost the old one (0.331, the no-information level $\mathrm{Var}(s)$=1/3); at A=0.125 the b2 state is the exact mirror image—*no* information about s_2 (0.333, indistinguishable from $\mathrm{Var}(s)$), and near-perfect information about s_1 (0.016, 21× better than fp32 at the same delay, whose state by then carries neither write). A state that cannot forget is a state that cannot be written to. Retention-by-freezing and retention-by-decay are different objects that one retention curve conflates—which is exactly the distinction Thm. 6 builds into the persistence measure, and a direct argument that repairing DAMP's score is necessary rather than cosmetic. *A*

b	frozen steps	ρ_b (A=1)	resp. at A=0.125	$A^*(b)$ pred.	MSE at j=60 / j=120	
					this protocol	*overwrite*
fp32	0.000	1.700	0.226	0.0012	0.213 / 0.325	0.315
8	0.000	1.700	0.226	0.0012	0.213 / 0.325	0.315
4	0.743 ± 0.007	1.327	0.181	0.024	0.095 / 0.253	0.234
2	0.997 ± 0.003	0.998	0.0002	0.125	0.015 / 0.015	0.174

Table 2: Write dead zone at d=64, α=0.97, C=4 (three seeds; std $\leq$ 0.003 where shown). Prediction $A^*(b) = C/(\rho_b\sqrt{d}2^b)$ matches the measured hard threshold at b=2 (0.1253 predicted, frozen at $A \leq 0.125$), and places the b=4/b=8 thresholds below the tested range. Frozen-step fraction is the fraction of steps $t \geq 1$ with $\Delta h = 0$. The last column is the *new*-write read at A=1 under the overwrite protocol—a frozen state scores well on single-write retention and badly the moment the content changes.

b	read at t_w+100	A=1	A=0.5	A=0.25	A=0.125
fp32	s_2 (new)	0.315	0.327	0.330	0.331
fp32	s_1 (old)	0.331	0.331	0.331	0.331
4	s_2 (new)	0.234	0.301	0.328	0.336
4	s_1 (old)	0.295	0.270	0.258	0.256
2	s_2 (new)	0.174	0.264	0.335	0.333
2	s_1 (old)	0.181	0.096	0.030	0.016

Table 3: Overwrite control (three seeds, std $\leq$ 0.003): s_1 written at t=0, s_2 at t_w=20, read at lag 100; no-information level is $\mathrm{Var}(s)$=1/3. The b2 rows invert: as A decreases it loses s_2 entirely (0.333, no information at A=0.125) while its hold on s_1 *improves* (0.030 → 0.016), because the state stops accepting input at all. fp32 shows the opposite monotonicity on both targets.

corrected claim about retrieval precision-independence. An earlier version stated that fp32/b8/b4 "agree to three digits" for $M \lesssim d_k$, on the basis of unpaired error bars. Testing on paired differences (shared key/value draws) shows the claim holds exactly for b8 at every load ($|\Delta| \leq 0.004$) but *not* for b4, whose deficit is already resolvable at M=$d_k/2$ and reaches -59% at M=$6d_k$ (App. A.11). The corrected statement is load-dependent, and it strengthens rather than weakens the resource-allocation thesis: retrieval load, not only propagation depth, sets a bit requirement.

A.9 Testing $\alpha_*(b)$ on trained models: what holds, and what does not

The second half of Thm. 1 predicts an interior optimum: a tighter budget should force faster forgetting, with α_* set by $1 - \alpha_*^2 = \frac{ecC}{\sqrt{3}}2^{-b}$. We tested it directly (`e3_alpha_b_grid.py`, `e3b_fixed_readout.py`): d=64, T=128, three seeds, six budgets $b \in \{32, 8, 6, 5, 4, 2\}$ and six forgetting rates $\alpha \in \{0.80, \ldots, 0.995\}$, reach measured over 31 lags under two readouts—one refit per delay on the quantized states, and one fitted only on fp32 states and then applied unchanged, so that the decoder cannot absorb quantization error. Table 4 reports $\log_{10}$ MSE at lag 60. *What*

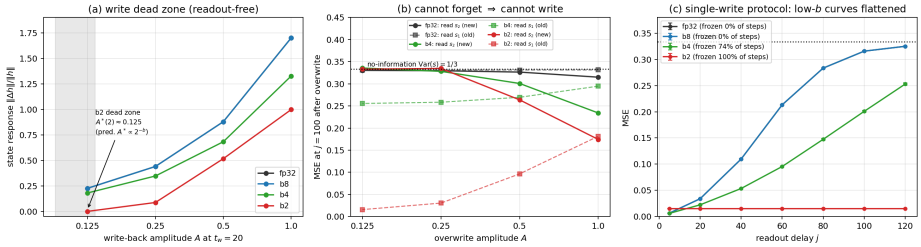

Figure 4: **A measured write dead zone and the "cannot forget ⇒ cannot write" dual** (d=64, α=0.97, three seeds; `e1_update_test.py`). (a) State response to a single write of amplitude A at t_w=20, readout-free: b=2 collapses to 1.8×10^{-4} at $A \leq 0.125$ against a predicted threshold $A^*(2)$=0.125; b=4 and fp32/b8 remain writable and linear. (b) Overwrite control at j=100: the same frozen state has *no* information about the new write s_2 and near-perfect information about the old write s_1—retention-by-freezing, not memory. (c) Single-write MSE(j) with seed error bars, flattened by (partial) freezing, which is what makes this protocol non-identifying below $b \approx 4$.

holds. (i) Saturation is reproduced across the entire α sweep: b8 and fp32 coincide at every α, and b6 stays within 10% of them–an independent confirmation of b_{sat} from a second protocol, at six forgetting rates rather than one. (ii) At every budget, reach increases monotonically with α over the tested range; the top of the range is at or near the best. The two readouts agree on both points. *What does not.* The predicted turnover is not observed. At b=6 the closed form puts α_*=0.95; measured MSE at lag 60 there is 0.309, against 0.063 at α=0.995—longer memory keeps helping well past the predicted optimum. Two candidate explanations were tested and both are *refuted*: the quantizer is not a systematic gain contraction ($\kappa = \mathbb{E}\langle \hat{h}, h\rangle / \mathbb{E}\|h\|^2 = 1.000$ with $\geq 92\%$ of the error in the non-compensable residual, at every $b \geq 4$; `_shrinkage_diag.py`), and a matched readout does not recover the missing factor by averaging quantization noise over d coordinates, because the quantizer step is set by the *per-coordinate* state scale and does not shrink with d. *Where the law must be read with care.* Two limits are now visible. First, states at $b \leq 4$ partially freeze, and *freezing is itself α-dependent* (larger α means smaller increments against a fixed step): b4 beats fp32 at lag 60 by 12.5× at α=0.99, which App. A.8 identifies as snapshot retention, not as precision. An α sweep at coarse bits is non-identifying for exactly the reason a single-write curve is, so only the unfrozen rows ($b \geq 6$) test α_* cleanly. Second, the closed form for k_*^{prop} has no floor other than quantization: it grows without bound in b and turns negative as soon as $c\sqrt{\Lambda_b} > \gamma$, so it cannot by itself produce the saturation ceiling observed in §4.1. Adding the task/gate floor Λ_0, $k_*^{\text{prop}} = \left[\ln\gamma - \frac{1}{2}\ln\left(c(\Lambda_b + \Lambda_0)\right)\right] / |\ln\alpha|$, restores both the ceiling and a turnover, and is the form we recommend for deployment estimates. We therefore state α_* as a white-noise, per-channel result, not as a trained-model prediction, and we flag the trained-model version of this law as open.

b	α=0.80	0.90	0.95	0.97	0.99	0.995	$\alpha_*(b)$
fp32	−0.48	−0.48	−0.51	−0.67	−1.08	−1.20	1.000
b8	−0.48	−0.48	−0.51	−0.67	−1.08	−1.20	0.988
b6	−0.48	−0.48	−0.51	−0.67	−1.06	−1.15	0.950
b5	−0.48	−0.48	−0.51	−0.66	−1.16	−1.40	0.897
b4	−0.48	−0.48	−0.54	−1.02	−2.18	−2.29	0.780
b2	−0.47	−1.58	−1.97	−1.82	−1.42	−1.33	—

Table 4: $\log_{10}$ MSE at lag 60 vs. forgetting rate α and budget b (d=64, T=128, three seeds, readout refit per delay). No column is a minimum in α: reach keeps improving to the top of the tested range, so the closed-form $\alpha_*(b)$ of Thm. 1 (last column) is not the empirical optimum. The fp32-fitted-readout control agrees with every entry for $b \geq 5$ to ≤ 0.02 dex and deviates only in the rows App. A.8 marks non-identifying (b=4 at $\alpha \geq 0.97$, up to 0.9 dex; b=2, up to 1.6 dex)—the size of that gap measures how much of a coarse state's apparent retention is decoder adaptation rather than channel memory. Rows $b \leq 4$ are additionally confounded in α, because a larger α shrinks the per-step increment against a fixed step and raises the frozen-step fraction; the clean comparison is $b \geq 6$, where the states do not freeze and the monotone-in-α trend persists.

A.10 THE CAPACITY CONSTANT $f(\beta) = M_{\max}/d_k$, MEASURED

Prop. 4 reports two constants that are numerically close and had been easy to conflate: the SNR prefactor 2 and the capacity constant 1.83. They are different quantities. Sweeping the load out to $32\,d_k$ at d_k=64 with 32 draws per point (`verify_retrieval.py --fracs, _e6_capacity_fit.py`) and locating the crossing of SNR at threshold θ=1 by log–log interpolation gives Table 5. Three points resolve inside the swept range and agree with the closed fit $f(\beta) = 2.14\beta^{-1.28}$ to within 13%, over a 7× range in β and 1.5 decades in $M_{\max}$. The fourth, β=0.1 (predicted $40.8\,d_k$), lies beyond the sweep: its SNR is still 1.01 at M=$32d_k$, so it is consistent but unresolved. A free two-parameter fit to the three resolved points returns $1.96\beta^{-1.44}$—the sweep cannot discriminate the exponent, but it does establish that the β^{-1} mechanism carries the scaling, that $f(1) = 1.9$ is the 1.83 of Prop. 4, and that the SNR prefactor 2 is a separate number evaluated at β=$\frac{1}{2}$. This also explains why a load scan that stops at $6\,d_k$ looks "saturated" at small β (Fig. 2(a)): the capacity point has simply moved out of frame.

β	1	0.50	0.25	0.10
$M_{\max}/d_k$ measured	1.9	5.8	13.7	> 32
$2.14\,\beta^{-1.28}$	2.1	5.2	12.6	40.8
deviation	−13%	+12%	+9%	(unresolved)

Table 5: Capacity constant vs. step size (d_k=64, 32 draws, threshold SNR= θ=1, load swept to M=$32d_k$). The closed fit of Prop. 4 is confirmed to 13% on the three resolvable points. $f(1) = 1.9$ coincides with the $1.83\,d_k$ capacity quoted there.

A.11 PAIRED COMPARISONS AND UNCERTAINTY QUANTIFICATION

A claim that two configurations "agree" is a claim about a *difference*, so it has to be tested on the difference. In §4.2 every bit-width sees the same 32 key/value/query draws; the comparison is therefore paired even though the across-draw spread of the SNR itself is large (at M=16, fp32 is 2.88 ± 0.70). Reporting two independent standard errors hides the effect the claim is about. Table 6 gives the paired difference at d_k=64, β=0.5, n_{q}=4000: *Where each experiment sits on this*

M	16	32	64	96	128	192	256	384
Δ(b8)	−0.001	−0.000	−0.000	−0.003	−0.004	−0.001	−0.002	−0.00
paired s.d.	0.016	0.013	0.010	0.009	0.006	0.006	0.005	0.00
Δ(b4)	+0.018	−0.111	−0.285	−0.362	−0.445	−0.594	−0.632	−0.59
paired s.d.	0.307	0.242	0.299	0.219	0.212	0.263	0.283	0.24
95% CI of Δ(b4), % of fp32	[−3, +4]	[−7, −1]	[−20, −9]	[−25, −16]	[−35, −25]	[−54, −39]	[−66, −47]	[−68, −
b4 equivalent to fp32 (±5%)	yes	**no**	no	no	no	no	no	no

Table 6: Paired test of "precision does not matter above b_{sat}" (32 shared draws; `retrieval_paired_test.py`). $\Delta(b)$ =SNR$_b$−SNR$_{\mathrm{fp32}}$ on identical draws. b8 is equivalent to fp32 at every load—the CI of Δ(b8) stays inside ±0.4% of the fp32 mean and its paired s.d. ($\leq$ 0.016) is $\sim 40\times$ smaller than the unpaired one. b4 is equivalent only at M=$d_k/4$; from M=$d_k/2$ the CI excludes zero and the deficit grows monotonically to −59% at M=$6d_k$. The load dependence of the b4 requirement is only resolvable under pairing.

spectrum. §4.1 and App. A.8 report three *training* seeds per configuration with seed-level s.d. as error bars (Fig. 4c is one such case; the seed spread there is ≤ 0.003 on MSE, i.e. smaller than every effect we claim). §4.3-b and -c report closed-form-vs-measured ratios rather than a significance test, because given the inputs that comparison is deterministic; the honest statement there is agreement to ±0.4% and ±2%, not a p-value. The trained-gate probe (§4.3) is the weakest statistical claim in the paper: one model per condition, with N=120k *tokens* providing token-level averaging, not model-level replication. We state this rather than dress it up, and we flag model-level replication of the gate probe as the highest-value remaining experiment alongside real-checkpoint channel statistics (Limitations).

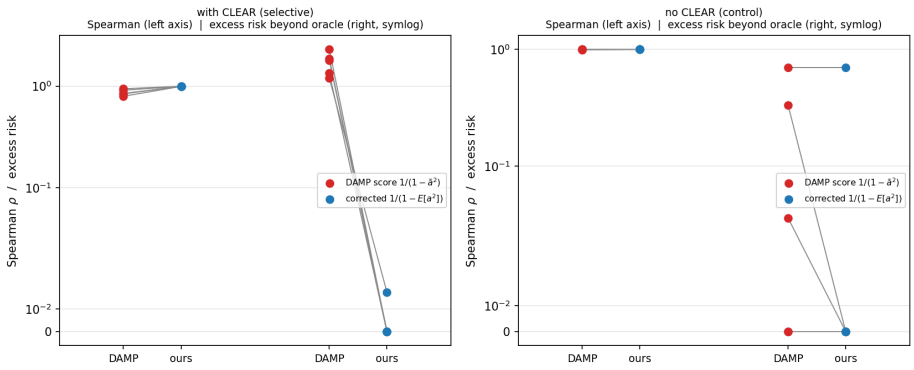

Figure 5: **Model-level replication of the gate probe** (five independent trainings per condition; grey segments pair the two scores within one model). Left pair: Spearman of score against true risk; right pair: excess risk beyond the oracle (symlog). The DAMP→corrected gain in the selective condition is consistent in every model; in the control the excess risk is at the floor (0–0.7) in both, which is why that column carries no relative claim.

A.12 The trained-gate probe at model level

§4.3 originally reported one trained model and one evaluation stream, so its percentages were token-level averages and its Spearman had no interval. We retrain five models per condition, each with independent initialisation, training stream and evaluation stream (`e7_gate_probe_seeds.py`), and report model-level mean ± s.d.; Spearman additionally carries a 4000-draw bootstrap over the 32 channels (Table 7, Fig. 5).

	with CLEAR (selective)	no CLEAR (control)
channels in critical regime	$100.0 \pm 0.0\%$	$0.0 \pm 0.0\%$
DAMP underestimate	$2.25 \pm 0.12\times$	$1.05 \pm 0.01\times$
Spearman, DAMP score	0.866 ± 0.059	0.994 ± 0.005
Spearman, corrected	0.996 ± 0.003	0.998 ± 0.002
excess Top-8/32 risk, DAMP	1.70 ± 0.44	0.22 ± 0.30
excess Top-8/32 risk, corrected	0.004 ± 0.008	0.14 ± 0.31
Top-8 overlap, DAMP vs corrected	5.4 ± 0.5	7.6 ± 0.5

Table 7: Trained-gate probe over five independently trained models per condition (d=32; mean ± s.d.). The selective condition reproduces in *every* model: all channels critical, DAMP low by 2.13–2.45×, ranking correlation moving from 0.795–0.941 to 0.993–1.000, and excess risk beyond the oracle reduced by 98.5–100%. In the control the excess risk is itself ~ 0.2 (comparable to its own spread), so its *ratio* is undefined; we report absolute values there and make no relative claim.

A.13 Row exchangeability and the damage region

Values moved out of §4.3(f) for space: with 32 shared draws and 12 calibration streams, uniform INT4 costs 334.1 ± 2.3 against 899.5 in fp32, and Top-K protection recovers it to 365.4 ± 2.5 (DAMP's ranking) and 365.5 ± 2.4 (ours); the corrected-vs-DAMP difference is $\Delta = +0.46 \pm 0.52$ SNR at INT4 (95% CI $[+0.13, +0.79]$). The single-stream values quoted in an earlier version (366.0 against 365.7) are superseded by these resampled ones.

Details of the DAMP bridge (§4.3(a–d)). (a) We derive DAMP's score. $\mathcal{P}_u$ is the stationary noise floor of Thm. 1 (ratios 0.993–1.011 across a); $C_u = \mathcal{E}_u \mathcal{P}_u$ is the channel's accumulated quantization noise.

(c) b_{req} explains its structure and INT4 failure. INT4 sustains $\mathcal{P}_{\max} \approx$ 20–200; DAMP caps $\mathcal{P}$ at 10^4, so the most persistent channels need 6.9–8.4 bits—beyond INT8, hence FP16. Its reported limitation "INT4 variants cannot recover FP32" is thus a hard constraint of the log-law; exits are lowering C (rotation) or lowering $\mathcal{P}$ (reparameterizing decay).

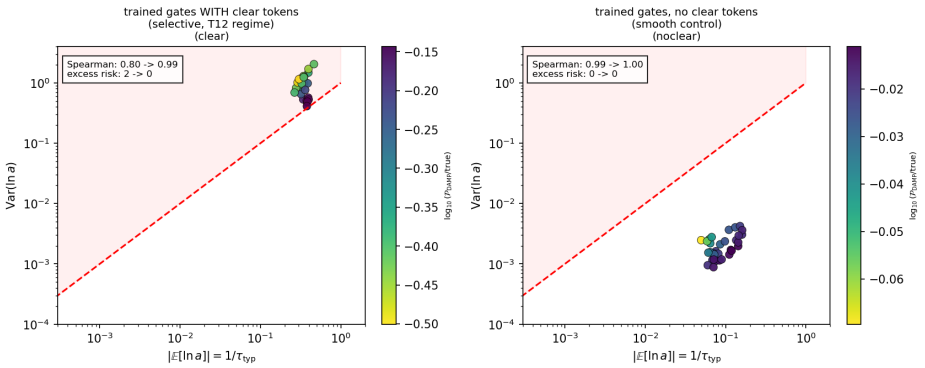

Figure 6: **Trained-gate probe (Thm. 6, §4.3).** Left: gates trained under selective pressure (rare CLEAR tokens) sit in the critical regime $\mathrm{Var}(\ln a) > |\mathbb{E}[\ln a]|$ (above the dashed divergence line)—100% of channels—where DAMP's score is low by $\sim 2.3\times$ and the corrected score matches true risk (color: $\log_{10} \mathcal{P}_{\mathrm{DAMP}}/\mathrm{true}$). Right: the no-CLEAR control learns smooth gates (0% in the critical regime), where both scores are accurate. Spearman and excess-risk reductions annotated in each panel.

(d) Hadamard gain = lowering C. Measured on five block distributions (block 32, b=8): the closed-form MSE ratio $\mathbb{E}[R^2]_{\mathrm{plain}}/\mathbb{E}[R^2]_{\mathrm{Had}}$ matches measurement to $\pm 0.4\%$; heavy-tailed blocks gain Δb=0.47–1.24 bits (2-spike blocks: C=3.05 $\rightarrow$ 1.77); Gaussian blocks gain nothing—the gain is in the tails, not the transform.

How each protocol-reproduction claim is quantified. *How each claim here is quantified.* Calibration energy, the protected-channel underestimate, and the ranking overlap are means over 12 independently drawn calibration streams; SNR values are means over 32 shared evaluation draws with the across-draw s.d.; the ranking comparison is a paired difference on the shared draws. Two measurements decide whether a scoring fix can matter for a given toy, and neither needs retraining (`e5_damage_region.py`). (i) *Exchangeability*: protect K rows in fp32 and quantize the rest, choosing the mask by DAMP's score, by the corrected score, by the true per-row risk (oracle), or uniformly at random, and sweep K at INT4 (Table 8). (ii) *Damage region*: uniform quantization for $b \in \{8, \ldots, 2\}$. The damage curve completes the picture: uniform SNR against fp32's 899.5 is

K	uniform	random	DAMP	corrected	oracle	oracle−random
1	332.9	334.6 ± 0.2	334.6	334.9	334.6	+0.0
2	332.9	336.6 ± 0.2	336.8	336.4	336.8	+0.1
4	332.9	340.3 ± 0.4	340.1	340.3	340.1	−0.2
8	332.9	347.6 ± 0.1	347.7	347.9	347.8	+0.2
16	332.9	363.4 ± 0.3	363.7	364.7	363.8	+0.4
32	332.9	399.6 ± 0.4	400.2	401.6	400.1	+0.5
64	332.9	495.2 ± 1.8	497.4	498.7	495.9	+0.7

Table 8: Protection budget sweep at INT4 (d_k=128, 8 shared draws; random masks averaged over 3 draws each). Raising the budget buys +164 SNR from K=0 to K=64; choosing *which* rows to protect buys at most +0.7 (oracle over random) and $\leq$ +1.6 (corrected over DAMP). At K=1 the oracle's single best row is indistinguishable from a random row. Row exchangeability is thus measured rather than assumed, and it is what bounds the ranking gain in §4.3-f.

893.9 at INT8 (−0.6%), 877.6 at INT7 (−2.4%), 818.2 at INT6 (−9.0%), 641.0 at INT5 (−28.7%), 332.9 at INT4 (−63%), 94.7 at INT3 and 9.3 at INT2. The toy's INT8 regime is therefore barely damaged, and the toy enters a materially damaged regime only at 5–6 bits. For calibration of the same quantity on the model side: per-row quantization energy E_u has top-16 share 13.2% against the 12.5% exchangeable baseline (Gini 0.020), DAMP's score $E_u \times \mathcal{P}_{\mathrm{geo}}$ reaches 14.6%, and the empirical per-row risk reaches 17.2% (Gini 0.122)—all close to exchangeable, which is exactly why the sweep above is flat.

A.14 A REAL PRETRAINED CHECKPOINT, MEASURED

The synthetic Zipf stream's fp32 perplexity is 116,192, which is why that stream's *absolute* PPL is not interpretable (it is out of distribution for the model) and only its paired comparisons are used. Every measurement in §4.3 is synthetic or a small trained model. This appendix reports the same protocol on a real pretrained SSM, which is the largest external-validity gap the paper had (`ereal_kit/`). The model is Mamba-1 130M (Pile-trained, 24 layers, 129M parameters), obtained from ModelScope; its `A_log` is shaped $(d_{\text{inner}}, d_{\text{state}})$ with d_{inner}=1536, so the channel axis on which DAMP's GDN protects rows is reproduced at the same granularity. Implementation notes: we do not re-implement the recurrence or transcribe weights. We load the checkpoint with `transformers` and run it one token at a time with a cache, so the model takes its own single-step path and the state lands in `cache.layers[i].recurrent_states[0]`; rewriting it between steps is exactly the write-back quantization of §4.3. Stepwise and full-sequence logits agree to 1.4×10^{-6}, and a self-test asserts that clamping the state actually changes the output, so a misplaced hook aborts rather than producing a plausible-looking null. Two streams are used: a synthetic Zipf token stream, and real natural text (Chinese Wikipedia, 521 MB). The latter is *out of the model's training language*—it is the only real text corpus with data files available on the permitted hub—so it should be read as a genuine-text, distribution-shifted probe, not as an in-distribution evaluation. The synthetic stream is the one on which PPL is not interpretable in absolute terms (it is ∼124k); all comparisons within a stream are paired on the same tokens. *Channel statistics* are collected

quantity	synthetic stream	real natural text	synthetic toy (§4.3)
SSM layers instrumented	24/24	24/24	—
channels (protected axis)	1536	1536	128
channels in critical regime	100.0%	99.9%	100%
top-16 share of E_u	22.8%	23.6%	13.1%
uniform baseline for top-16	1.0%	1.0%	12.5%
median $\text{Gini}(E_u)$	0.741	—	—
median $P_{\text{geo}}/P_{\text{pow}}$	0.697	**0.980**	1/4.3
Top-192 mask overlap (DAMP vs. corrected)	93.1%	**98.6%**	7.1/16 disagree

Table 9: Channel statistics on a real pretrained Mamba-1 130M. Energy concentration transfers and is stable across two very different streams (23× the uniform baseline in both), so the ranking problem is real at deployment granularity. But the score *defect* does not transfer: on real text the two persistence estimates agree to 2% and the rankings select almost the same channels, because a trained Mamba-1's gates are nearly deterministic. The toy's 13.1% vs. a 12.5% baseline is why its end-to-end ranking effect was tiny.

in Table 9. Energy concentration transfers and is stable: the top-16 share of E_u is 22.8% on the synthetic stream and 23.6% on real text, against a 1.0% uniform baseline (23×), so the concentration is a property of the trained model rather than of the input stream. The score *defect* does not transfer: the median $P_{\text{geo}}/P_{\text{pow}}$ is 0.980 on real text against 0.697 on the synthetic stream, and the two rankings agree on 98.6% (real text) and 93.1% (synthetic) of the Top-192 channels. The mechanism is visible in the same table: a trained Mamba-1's gates are nearly deterministic, so the geometric and power means of a coincide, and the correction has almost nothing to repair—which is exactly the smooth-gate regime in which §4.1's verification (i) already finds DAMP accurate. The bursty gates that make the correction matter are a property of the task, not of synthetic-ness. *Damage region.* Table 10 locates where state PTQ starts to cost anything. On real text (fp32 PPL 26.00), INT8 and INT6 are within 1%, INT5 costs 2.7%, and INT4 costs 11.5%—material damage begins at 4–5 bits, consistent with $b_{\text{sat}} \approx 3.2$ bits, and DAMP's INT8 deployment regime is essentially lossless for this model. The same measurement on the synthetic stream is uninformative (PPL ∼124k, quantization changes it by 0.5%, and a random mask beats the energy-based oracle) because the model is far out of distribution there; we report it only to show that the real-text arm is the one carrying information. *Can a ranking matter here?* Top-192 protection recovers roughly half the INT4 damage (uniform +3.00 versus +1.48 to +2.22 for the four masks), so protection as such is a large effect. Which channels are protected is not resolvable, and a resampled control settles the question quantitatively. Table 12 sweeps K at INT4: *both* the corrected-minus-DAMP difference and the oracle-minus-random gap change sign across K. Redrawing the random mask six times at K=192 (last row of Table 11) gives a mask-to-mask s.d. of 0.44 PPL and a range of 1.22 PPL, while the correction is 0.19 and the oracle-vs-random gap 0.04; one of the six random draws (27.502) is better than

bits	fp32	uniform	cost	relative
INT8	26.00	26.03	+0.030	+0.12%
INT6	26.00	26.23	+0.225	+0.87%
INT5	26.00	26.71	+0.706	+2.71%
INT4	26.00	29.00	+3.000	+11.54%

Table 10: Uniform state-PTQ on Mamba-1 130M reading real natural text. This is the deployment-relevant number the paper previously had only for synthetic models.

both score-based masks (27.678 and 27.484). The variation induced by *which* random subset is protected therefore exceeds every effect we are trying to measure, so we report no resolvable end-to-end ranking effect on this model. What is resolvable is the *number* of protected channels, not their identity. *A correction to our own diagnostic.* The "oracle" in this appendix ranks channels

random mask #	0	1	2	3	4	5
PPL at K=192, INT4	28.038	27.502	28.121	28.622	28.717	28.079

Table 11: Six independently drawn random Top-192 masks, same tokens, same model (fp32 26.00; no protection 29.00; DAMP 27.678; corrected 27.484). Mean 28.177, s.d. 0.442, range $[27.502, 28.717]$. The spread between two random draws exceeds the DAMP-vs-corrected difference, which is the quantitative reason the ranking question is unanswerable at this scale.

K	uniform	DAMP	corrected	random	oracle
1	29.00	28.208	28.377	28.451	28.835
16	29.00	28.030	28.163	28.661	28.194
192	29.00	27.678	27.484	27.988	28.222

Table 12: Protection-budget sweep at INT4 on real text (lower PPL is better; fp32 is 26.00). Protection as such is worth a great deal (29.00 → 27.48), but the ordering of the four masks is not stable: corrected−DAMP is $+0.169, +0.133, -0.194$ and oracle−random is $+0.383, -0.467, +0.234$. Table 11 shows why—the single random mask per cell here has a mask-to-mask s.d. of 0.44 PPL, larger than any of these differences.

by state energy. The real-model data show it is frequently *worse than random* (28.835 vs. 28.451 at K=1), and the reason is instructive: block min-max error depends on a block's dynamic range rather than its magnitude, so a high-energy channel is not necessarily a high-*error* channel. The E_u-based scores beat it (DAMP's top-1 row gives 28.208), which supports the paper's choice of quantization-error energy as the ranking axis over raw state energy. But it also means oracle-minus-random, as defined here, is *not* an upper bound on what a score can buy, and the exchangeability conclusion of App. A.13 should be read with that caveat. A true impact oracle requires per-channel ablation against the task metric.

A.15 Retrieval task: paired comparison table

The load-resolved table quoted in §4.2.

A.16 Closed-form verification details

Numerical checks quoted in §3, moved here for space (the corresponding scripts are released). *Verification.* $k_*(b)/k_*(b-2)$ converges $4.697 \to 4.153 \to 4.037 \to 4.009 \to 4.002$ (ratio 4 expected); $k_*/\tau_* \to 1$; α_* closed-form vs. numerical argmax agrees to three decimals. Marginal gain $\partial k_{\max}/\partial b = \ln 2 \cdot \tau_*$—one bit is worth one e-folding time of the channel it is spent on. *Verification.* Closed form vs. simulation: ratios 0.97–1.09 over $d \in \{4, 16, 64, 256\}$, $b \in \{6, 8, 12\}$, $k \in \{0, 5\}$; at d=1 the ratio degrades to 0.86, where one-dimensional writes are ± 1 and the Gaussian-direction approximation behind the interference term is worst. $\mathrm{MSNR} \propto d$; saturation at b_{sat} is observed both as a plateau (MSNR(0) frozen from 8 to ∞ bits at α=0.9) and at the half-power point ($b \approx 3$ measured vs. b_{sat}=2.56 predicted); delta-rule associative memory agrees to three significant digits across fp32/b8/b4 and diverges at b2 (§4.2). *Verification.* Closed form vs. simulation: ratios 1.000–1.093 across the valid regime ($m \gtrsim 4\sigma_g$); variance additivity confirmed with quantized gates; end-to-end on trained models: gate noise σ_g=1.0 multiplies short-delay propagation MSE

b	M=16	32	64	96	128	192	256	384
fp32	2.88	2.70	2.01	1.74	1.48	1.28	1.12	1.00
8	2.88	2.70	2.01	1.74	1.48	1.28	1.12	1.00
4	2.90	2.59	1.72	1.38	1.03	0.69	0.49	0.41
2	0.75	0.13	0.05	0.00	0.02	0.00	0.00	0.00
b4/fp32	1.01	0.96	0.86	0.79	0.70	0.54	0.44	0.41
Δ(b4)	$+0.02$	-0.11	-0.29	-0.36	-0.45	-0.59	-0.63	-0.59
b4 equiv. ($\pm5\%$)	yes	**no**	no	no	no	no	no	no

Table 13: Retrieval SNR vs. number of stored pairs M (d_k=64); each entry averages 32 independent key/value draws. b2 collapses *and diverges* (MSE $\to 10^{54}$: 2-bit states break the contraction of $(I-\beta kk^\top)$), reproducing DAMP's INT4-collapse at minimal scale. Because all bit-widths share the same draws, the right comparison is the *paired* difference Δ(b)=SNR$_b-$SNR$_{\text{fp32}}$: b8 is equivalent to fp32 at every load ($|\Delta| \le 0.004$, paired s.d. ≤ 0.016), whereas b4 is equivalent only at $M{=}d_k/4$ and then degrades monotonically with load (-4% at $d_k/2$, -14% at d_k, -46% at $3d_k$; paired s.d. 0.21–0.31, so the drift is resolved well before it is visible in the unpaired error bars). Retrieval therefore has a load-dependent bit requirement, not merely a propagation-dependent one.

31× at d=64 and 27× at d=2240 (§4.1). The logit parameterization covers the log-/exp-space gates used by gated linear attention (Yang et al., 2024): for $m \gg 1$ only the softplus tail matters, so the law is parameterization-agnostic. *Verification.* (i) Smooth sigmoid gates (m=6): DAMP accurate within 0.6% for $\sigma_g \le 1$ (its valid regime), 6–16% off for σ_g=1.5–2. (ii) Occasional-reset mixtures: DAMP underestimates by 2.66–5.73×; the corrected closed form matches simulation to $\pm2\%$. (iii) Trained gates (§4.3, App. A.12) confirm the regime is reachable by optimization, in five independently trained models per condition: all channels critical (uncertainty 0), DAMP low by $2.25 \pm 0.12\times$, and the correction holding in every model. (iv) Sorting 128 channels and protecting Top-16 (DAMP's protocol): Spearman of score vs. true risk $0.954 \to 0.991$; residual risk beyond the oracle drops 83%. Whether such a ranking can matter at all is bounded by row exchangeability, which App. A.13 measures directly with an oracle-vs-random control. (v) The scope of the correction is bounded by gate variability, and a real pretrained model supplies the counter-case: on a real Mamba-1 130M, trained gates are nearly deterministic, P_{geo} and P_{pow} agree to 2%, and the corrected and original rankings coincide on 98.6% of channels (App. A.14). The 2–5× deviations above are the *bursty*-gate regime, not the generic case.

A.17 Dimension buys retrieval only if the range is recalibrated

Moved here from §3.2 for space; referenced in Limitations (ii). The d in Thm. 3 is tied to the calibration protocol: adaptive calibration gives $\sigma_h \propto 1/\sqrt{d}$ for unit-norm writes, so the quantization floor shrinks with d exactly as fast as the interference does and b_{sat} is dimension-free. If the range is instead not rescaled with the state (a range reused across widths), the same derivation gives $\text{MSNR}_d^{\text{nr}}(k) = d\,\alpha^{2k}/\big([\frac{1}{\gamma} - \alpha^{2k}] + d\,\Lambda_b/\gamma^2\big)$, hence $b_{\text{sat}}^{\text{nr}} = b_{\text{sat}} + \frac{1}{2}\log_2 d$ and $\text{MSNR}_d^{\text{nr}} \to \alpha^{2k}\gamma^2/\Lambda_b$ as $d \to \infty$: *dimension stops buying retrieval*. Simulated at α=0.9, b=6, d=256: 59.5 (adaptive, $\propto d$) against 19.0 (fixed, ceiling 27.7); the measured saturation point moves by ≤ 1 bit across $d \in \{1, \dots, 64\}$ under adaptive calibration versus $+4$ bits under fixed range (App. A.18). DAMP recalibrates per block, so the adaptive branch—and with it the boxed b_{sat}—is the one that governs deployment.

A.18 Correction log: where the dimension sits, and why it matters

Superseded point estimates. An earlier version of the gate probe quoted one model and one N=120k-token stream: Spearman $0.802 \to 0.989$, excess Top-8/32 risk reduced by 97%, DAMP low by 2.3×. Those numbers lie inside the per-model ranges now reported in App. A.12 ($0.866 \pm 0.059 \to 0.996 \pm 0.003$, $-99.7 \pm 0.6\%$, $2.25 \pm 0.12\times$) and are retained here for traceability. The protocol reproduction likewise superseded a single-stream comparison; see App. A.13. The retrieval law depends on where the state dimension d is placed, and an earlier version of this paper placed it inconsistently: the interference carried a factor $1/d$ (unit-norm writes) while the quantization floor was written with $\sigma_h^2 = A/\gamma$ (per-coordinate unit variance). Taken literally, that combination multiplies the quantization term by d and predicts $b_{\text{sat}} = b_{\text{sat}}(1) + \frac{1}{2}\log_2 d$. We record the resolution

because it is cheaply checkable and because the corrected statement carries an additional result. *(i) Under adaptive calibration the dimension-free b_{sat} is the right one.* With a unit-norm probe the signal, not the interference, is what carries the factor d: the probe reads the full key norm, so the signal power is $Ad\,\alpha^{2k}$ while both floors stay $\Theta(A)$. Simulating the half-power point at $\alpha{=}0.9$ gives $b \approx 3$ for every $d \in \{1, 4, 16, 64\}$, against a closed form of 2.56 for all d; the inconsistent placement would have required $2.56 \to 5.56$ over the same range. *(ii) The discarded prediction is a real regime, not an artifact.* If the quantization range is *not* recalibrated with the state—a range fixed once and reused across widths—σ_q^2 stops shrinking with d, the theorem becomes $\mathrm{MSNR}_d^{\mathrm{nr}}(k) = d\,\alpha^{2k} \big/ \big([\frac{1}{\gamma} - \alpha^{2k}] + d\,\Lambda_b/\gamma^2\big)$, and both predictions return: $b_{\mathrm{sat}}^{\mathrm{nr}} = b_{\mathrm{sat}} + \frac{1}{2}\log_2 d$, and $\mathrm{MSNR}_d^{\mathrm{nr}} \to \alpha^{2k}\gamma^2/\Lambda_b$, i.e. dimension stops buying retrieval. Table 14 summarizes both branches. *Dead-zone guard.* At coarse

	d				
	1	4	16	64	256
MSNR, adaptive ($b{=}6$)	0.23	0.93	3.72	14.9	59.5
MSNR, non-rescaled range ($b{=}6$)	0.23	0.91	3.31	9.7	19.0
b_{sat} measured (adaptive)	4	3	3	3	—
b_{sat} measured (non-rescaled)	4	4	5	6	—
b_{sat} closed form (adaptive)	2.56, independent of d				
b_{sat} closed form (non-rescaled)	2.56	3.56	4.56	5.56	—

Table 14: Dimension and calibration at $\alpha{=}0.9$, $C{=}4$ (`verify_bsat_regimes.py`, $T{=}6000$, unit-norm writes). Adaptive calibration is linear in d and dimension-free in b_{sat}; the non-rescaled range saturates at the ceiling $\gamma^2/\Lambda_b = 27.7$ and shifts b_{sat} by $\approx \frac{1}{2}\log_2 d$ (measured $+4$ bits from $d{=}1$ to $d{=}64$ at the 98%-saturation criterion, $+2$ bits at the half-power criterion—the gap is integer-bit resolution plus the dead zone that removes $b \le 4$ from the non-rescaled sweep at $d{=}64$).

bits the state stops moving, the read variance vanishes and the SNR estimator degenerates to $+\infty$; those bits are detected explicitly (per-step state change on fewer than 5% of steps) and reported as dead zone rather than as saturation, which is what makes the non-rescaled $d{=}64$ sweep start at $b{=}5$. *A typographical correction in Thm. 1.* An earlier version wrote the propagation SNR with numerator $(1-\alpha)^{2k}\delta^2$; the correct numerator is $(1-\alpha)^2\alpha^{2k}\delta^2$, the square of the trace $(1-\alpha)\alpha^k\delta$ derived in App. A.1. Only the corrected form reproduces the stated $k_*^{\mathrm{prop}} = \big[\ln(c\sqrt{\Lambda_b}) - \ln\gamma\big]/\ln\alpha$, and it is the form implemented in the released scripts.

A.19 Propagation task: additional figure and table

These are the $d{=}64$ propagation figure and the $d{=}2240$ table quoted in §4.1.

B Extended related work

This appendix carries the full version of the related-work discussion summarized in §5.

Recurrent-state PTQ. DAMP (Zhang et al., 2026) is the first systematic study of PTQ for GDN/KDA recurrent states, finding poor uniform-quantization trade-offs, error energy concentrated in few channels, and stable relative decay strength across prompts; protecting 12.5% of channels in FP16 at 9.9 bits/value nearly matches FP32 while cutting state storage 69.1%. We show its score function is the stationary limit of our SNR law (Thm. 1/6), that inserting the geometric-mean decay into that second-moment quantity provably underestimates risk under selective gating (Thm. 6), and that its structural decisions follow from our b_{req} log-law (Prop. 8), including an explanation of its Hadamard ablation and the impossibility of its INT4 variant. We also point out a blind spot shared by current practice: decay gates are left in high precision, yet Thm. 5 gives the exchange rate (0.72 state bits per σ of gate noise) that marks the gate path as a prime—and currently unquantified—target.

PTQ for the Mamba family and SSM states. Beyond GDN/KDA, post-training quantization of selective-SSM states is an active line: Chiang et al. (2024) derive an error bound $\|\Delta(T)\|_2 \le \epsilon b/(1-a)$ that already exhibits the $1/(1-a)$ amplification which our variance floor σ_q^2/γ makes precise, and use a Hadamard transform to quantize outlier-heavy output activations; Xu et al. (2025)

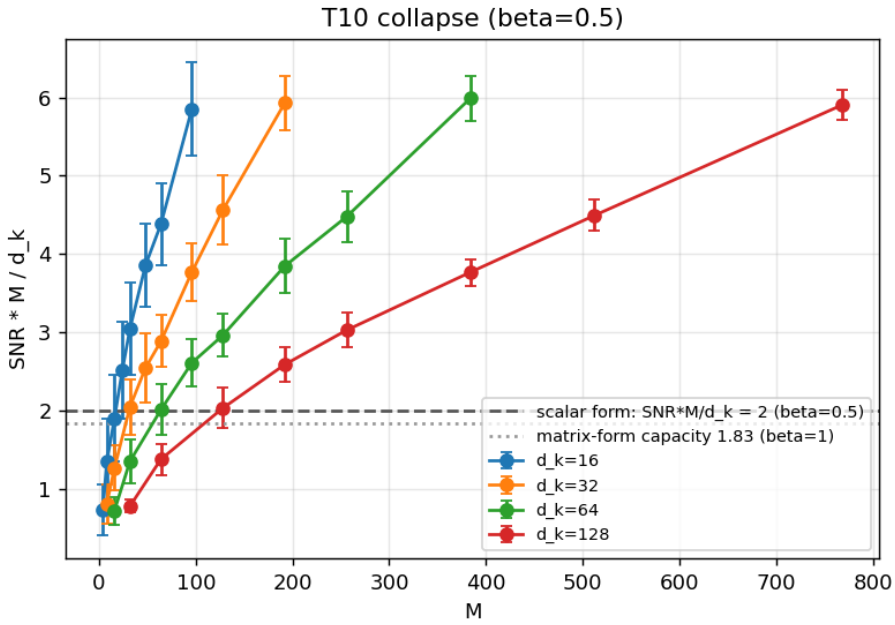

Figure 7: **Retrieval scaling collapse (Prop. 4; $\beta{=}\frac{1}{2}$).** SNR times M/d_k is consistent across d_k at fixed load ($\approx$ 2–3 for $1 \leq M/d_k \leq 2$ against the scalar-form 2), demonstrating capacity $\propto d_k$ and SNR $\propto d_k/M$ in the interference-limited regime; past $\approx 3d_k$ the measured SNR decays more slowly than $1/M$, so the product drifts upward. The two reference lines are deliberately distinct: 2 is the SNR prefactor measured at $\beta{=}\frac{1}{2}$, while 1.83 is the *capacity* constant $M_{\max}/d_k$ at $\beta{=}1$ (App. A.10); they are not two estimates of one number. The b2 divergence breaks the collapse.

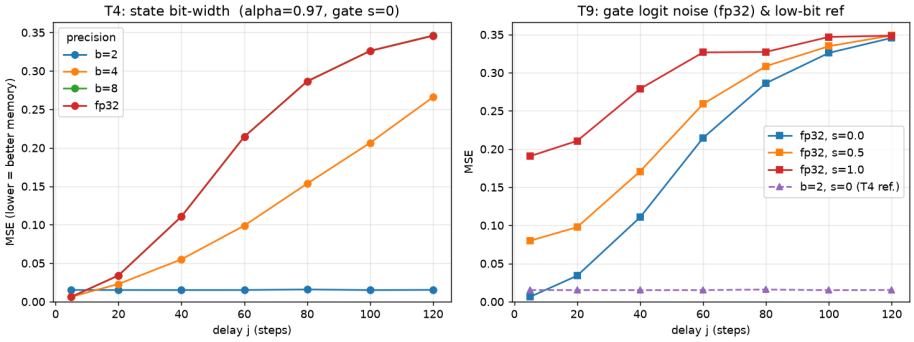

Figure 8: **Propagation task at $d{=}64$ ($\alpha{=}0.97$; ridge readout).** (a) State-bit scan: fp32 and b8 agree to $\leq 1\%$ at every delay (saturation past $b_{\text{sat}} \approx 3.2$ bits); b4 matches them at short delay and is *better* at long delay, but App. A.8 shows this is snapshot retention (74% of its steps are frozen), not fidelity. The b2 curve is flat at MSE ≈ 0.015—memory dynamics destroyed by coarse quantization, not merely degraded; that flatness is precisely why a single-write curve cannot score it, and Fig. 4 supplies the write/overwrite control. (b) Gating noise (Thm. 5): fp32 with gate-logit noise $\sigma_g{=}1.0$ behaves at delay $j{=}5$ like $\sigma_g{=}0$ does near the no-memory ceiling; the b2 curve is shown as reference. Dashed line: no-memory baseline $\text{Var}(s) = 1/3$.

prove that Hadamard does *not* equalize channel variance and replace it with KLT- and smooth-fused rotations; Li et al. (2025) design quantizers for the vision-SSM state from its distributional pathologies (long-tailed $\bar{A}$, time-varying hidden-state ranges). These are *method* papers: they take the sensitivity of the recurrence as given and build a quantizer around it. We take the complementary step—deriving the sensitivity itself in closed form (noise floor, b_{sat}, b_{req}, gate exchange rate)—which is what lets us say *why* the Hadamard step helps (it lowers the calibration constant C, worth $\Delta b \leq 1.24$ bits), how many bits a persistent channel actually needs, and where a persistence-based ranking is biased (Thm. 6).

b	σ_g	j=5	j=30	j=60	j=120	j=200
fp32	0	0.007	0.067	0.220	0.340	0.340
8	0	0.007	0.068	0.221	0.341	0.341
4	0	0.007	0.061	0.181	0.353	0.367
2	0	0.004†	0.004	0.004	0.004	0.004
fp32	1.0	0.184	0.240	0.328	0.343	0.340

Table 15: Propagation MSE vs. readout delay j at scale (d=2240, 5.02M parameters, α=0.97); the no-memory baseline is $\mathrm{Var}(s)$=1/3. fp32 and b8 coincide to $\sim 1\%$ (saturation past $b_{\mathrm{sat}} \approx 3.2$ bits); gate noise σ_g=1.0 multiplies short-delay MSE 27×. †The b2 row is flat at *all* delays because the state freezes (App. A.8): the entry is the retained first-write snapshot, not a fidelity ordering, and is not comparable to the rows above.

Roundoff noise in recursive filters. That the noise of a first-order recursion accumulates to $\sigma_\varepsilon^2 \sum_n |h(n)|^2 = \sigma_\varepsilon^2/(1-\alpha^2)$—our σ_q^2/γ—and that this floor trades against dynamic range is classical fixed-point filter analysis (Weinstein & Oppenheim, 1969; Jackson, 1970b;a; Oppenheim & Weinstein, 1972). We build on that floor rather than re-derive it; the questions differ. Filter analysis *predicts* the floor of a given filter, whereas deployment must *choose* the filter: given a budget b, which forgetting rate maximizes reach and how far does it reach (Thm. 1: $\alpha_*(b)$ and $k_{\max} = \tau_*$)—an optimization the classical results do not pose. Nor do they separate the two memory tasks: the same recursion is noise-floor limited for propagation, where bits buy reach, but interference limited for retrieval, where dimension buys accuracy and bits stop helping past b_{sat} (Thm. 3); only that split answers whether a given bit is worth spending. Finally, the quantities that dominate risk in modern architectures—token-dependent learned gates and their own quantization (Thm. 5), and second-moment persistence when gates reset (Thm. 6)—have no counterpart when filter coefficients are fixed and known.

Finite-precision expressivity and state-tracking limits. Merrill et al. (2024) prove SSM state-tracking power is confined to TC^0—Mamba-style states cannot provably track permutation composition, chess notation, or long-narrative entities—and confirm the failure empirically; Grazzi et al. (2025) characterize which finite-precision linear updates realize which regular languages. Both bound *what a state can compute*. Our laws answer the orthogonal, deployment-relevant question: *given a state that can express a computation, how long and how accurately does it survive under a bit budget?* Expressivity is about function class; memory laws are about fidelity within that class.

Linear recurrence, gating, and memory-as-design lineages. Gated DeltaNet (Yang et al., 2025), Kimi Delta Attention (Kimi Team, 2025), and the decay-based linear-recurrence line before them (Sun et al., 2023; Dao & Gu, 2024) define the state updates we analyze; gated linear attention (Yang et al., 2024) introduced the per-channel log-space gating our Thm. 5 analyzes; delta-rule associative memories date to Schlag et al. (2021). On the memory-design side, Behrouz et al. (2025) treat the state as a neural *long-term memory* whose write strength is modulated by surprise—a philosophy in which gating noise and persistence are first-class quantities, and for which our Thm. 5/6 provide the first quantitative precision accounting. Quantization at large (weights, activations, KV cache) is well studied, but the theories do not transfer: KV-cache entries are written once and read independently, so their errors stay independent across positions and per-entry PTQ bounds apply; a recurrent state is re-written at every step, so its error is re-injected and accumulates as σ_q^2/γ—the phenomenon our noise-floor analysis formalizes.