DAMP: Decay-Aware Mixed-Precision Recurrent-State Quantization

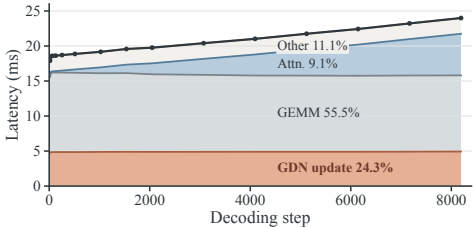
Research paper discussion – 20–30 minutes

Paper authors: Tao Zhang, Jianchao Tan, Pingwei Sun, Yanqi Yu, Zunhai Su, Zixu Jiang, Yuchen Xie, Xunliang Cai, Ziqian Zeng

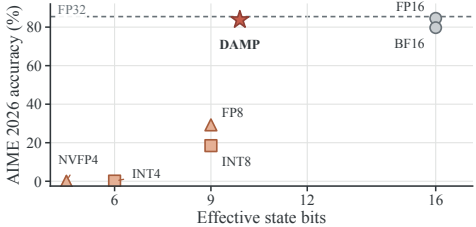
arXiv:2608.27513v2 – revised September 30, 2026
Reading date: October 13, 2026

Protect channels with large errors that persist through future decay.

Fixed-size recurrent states still cost memory and time



(a) Decode latency breakdown



(b) Accuracy–storage trade-off

- Qwen3.6, batch 256: GDN updates account for 24.3% of decode latency.
- Every step reads, updates, and requantizes the state.
- Uniform 8-bit storage damages reasoning; 4-bit is much worse.

Paper, Fig.1, PDF p.1; Sec.1, p.2. Effective bits include quantization metadata.

GDN and KDA share a decay-plus-delta recurrence

$$S_t = \underbrace{(I - \beta_t k_t k_t^\top)\Lambda_t}_{A_t} S_{t-1} + \beta_t k_t v_t^\top, \qquad o_t = S_t^\top q_t$$

- $S_t \in \mathbb{R}^{d_k \times d_v}$: key channels are rows; value dimensions are columns.
- $q_t, k_t \in \mathbb{R}^{d_k}$ are normalized; $v_t \in \mathbb{R}^{d_v}$; $\beta_t \in (0, 1)$.
- Decay the old state, partially erase the current key association, then write the new value.

The architectural difference

GDN: $\Lambda_t = \alpha_t I$ within a head.
KDA: $\Lambda_t = \mathrm{diag}(\exp(g_t))$, with channel-wise decay.

Paper, Sec.3.1, Eqs.(1)–(4), PDF p.3.

A local quantization error can survive many updates

$$S_t^q = Q(A_t S_{t-1}^q + \beta_t k_t v_t^\top), \qquad \Delta S_t = S_t^q - S_t$$
$$R_t = S_t^q - (A_t S_{t-1}^q + \beta_t k_t v_t^\top)$$

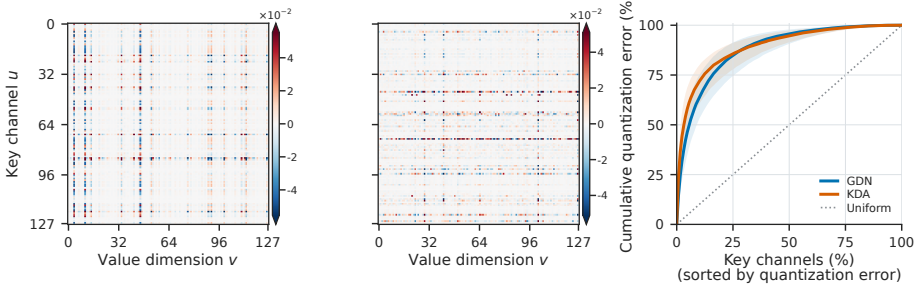
Exact state-error recurrence for a shared input trajectory

$$\Delta S_t = A_t \Delta S_{t-1} + R_t = R_t + \sum_{i=1}^{t-1} (A_t \cdots A_{i+1}) R_i$$

- R_t: newly injected error; assume $\Delta S_0 = 0$.
- A_t includes both decay and key-dependent error transformation.

Paper, Sec.3.3, Eqs.(7)–(13), PDF p.4. Shared-trajectory scope made explicit here.

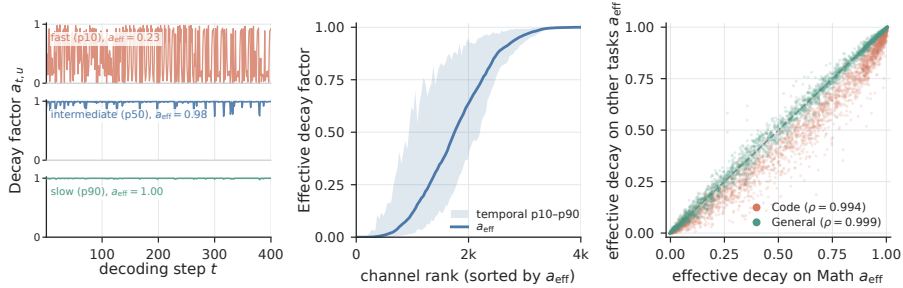
A small fraction of channels carries most local error



- Value scaling and reordering (SR) groups entries of similar magnitude.
- After SR, the top 12.5% of key channels account for about 69% (GDN) / 76% (KDA) of squared quantization error.

Paper, Fig.2(a–c), Sec.4.1, PDF pp.4–5. Author figure, panel-row crop.

Slow-decay channels remain similar across tasks



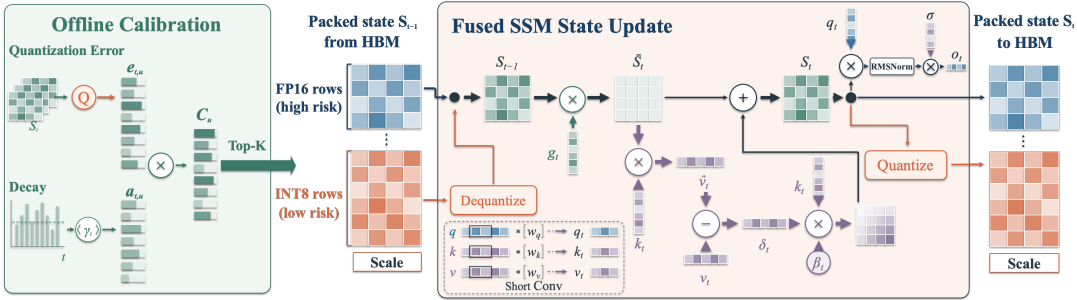
(d) Temporal decay-factor dynamics **(e) Effective decay-factor spectrum** **(f) Cross-task rank stability**

$$a_{\text{eff},u} = \exp(\mathbb{E}_t[\log a_{t,u}])$$ Near 1: weak decay, long retention

- KDA ranking correlations: Math–Code 0.994; Math–General 0.999.
- Stable rankings motivate estimating risk offline.

Paper, Fig.2(d–f), Sec.4.2, PDF pp.5–6. Author figure, panel-row crop.

DAMP turns offline risk estimates into a static cache



1. Calibrate value SR and select high-risk key channels.
2. Store selected rows in FP16; store the rest in INT8.
3. Fuse reconstruction, FP32 update, and compressed writeback.

Paper, Fig.3, PDF p.6; Sec.5.2 and Appendix A.4.

The score combines error size with future retention

$$e_{t,u} = \|[Q(S_t) - S_t]_u\|_2^2, \qquad W_{t,u} = 1 + \sum_{j=1}^{H} \prod_{r=1}^{j} a_{t+r,u}^2$$

Channel risk and budgeted selection

$$C_u = \mathbb{E}_{\mathrm{cal}}[e_{t,u} W_{t,u}], \qquad \mathcal{H}_{\ell,h} = \mathrm{TopK}(C_{\ell,h,:}, K_{\mathrm{hi}})$$

- Reference-state errors and the next $H = 256$ observed decays.
- Same $K_{\mathrm{hi}} = 16$ per head; channel identities vary by layer and head.
- C_u is a decay-based proxy; it omits the full A_t dynamics.

Paper, Sec.5.1, Eqs.(14)–(15), PDF p.6. Errors measured in original state coordinates.

Error lifetime matters even when local errors match

Teaching example: constant decay

$$W = \sum_{j=0}^{256} a^{2j}$$

Decay	a	W
Fast	0.50	1.33
Slow	0.99	49.96

Equal local error can imply very different cumulative risk.

Why decay can still matter in GDN

All channels in a head share W_t, but:

$$C_u = \mathbb{E}_t[W_t e_{t,u}]$$

Channels can incur errors at different times. Time-varying W_t can change their ranking.

Left: teaching calculation from Eq.(15), not an experiment. Right: Appendix D, PDF p.18.

Calibration separates fitting from configuration choice

36 prompts per model

12 MATH + 12 APPS + 12 ARC-Challenge training prompts.
Three disjoint, domain-balanced splits: fit / select / validate.

1. Generate FP32 reference trajectories: 1,024 input + 4,096 output tokens.
2. Fit candidate SR transforms and protected sets on the fitting split.
3. Choose the full mixed-precision configuration on the selection split.
4. Evaluate the frozen configuration on the validation split.

Six scoring anchors per sequence; 72 per split; 256 future decays per anchor.

Paper, Appendix A.2, Eq.(16), Algorithm 1, PDF pp.15–16.

A static layout avoids per-step channel reordering

Packed state per head

FP16 protected rows | INT8 remaining rows + FP32 group scales

- Fold fixed permutations into projections and channel-wise parameters.
- Use one CUDA kernel: load $\rightarrow$ dequantize $\rightarrow$ update/read out $\rightarrow$ quantize $\rightarrow$ write.
- Reuse state tiles in registers; recompute group scales at each write.
- FP32 arithmetic: compressed storage does not require INT8 updates.

Paper, Sec.5.2, PDF p.7; Appendix A.3–A.4, pp.16–17. Prefill writes the same packed cache.

9.9 effective bits include the group-scale overhead

Storage	Effective bits/value	Metadata
FP32	32	None
FP16 / BF16	16	None
INT8, groups of 32	$8 + 32/32 = 9$	One FP32 scale
INT4, groups of 32	$4 + 64/32 = 6$	FP32 scale + zero point

$$b_{\mathrm{DAMP}} = \frac{16}{128}\,16 + \frac{112}{128}\,9 = 9.875 \approx 9.9$$

State-storage reduction

$1 - 9.875/32 = 69.1\%$ versus FP32; 38.3% versus FP16.

Calculation from Table 1 and Appendix A.1/A.4. This is recurrent-state storage, not total model memory.

Evaluation spans two recurrent architectures and three models

Model	Recurrent + attention layers	Accuracy coverage
Qwen3.6-35B-A3B	30 GDN + 10 full attention	Six benchmarks
Kimi-Linear-48B-A3B	20 KDA + 7 MLA	Six benchmarks
Kimi-K3	69 KDA + 24 gated MLA	AIME + HMMT

- Math: AIME 2026, HMMT Feb. 2026, IMO-AnswerBench.
- General reasoning: GPQA-Diamond, MMLU-Pro. Code: LCB-v6.
- SGLang FP32-state reference; FP16/BF16, uniform 8-/4-bit, INT8+SR, and DAMP comparisons.

Paper, Sec.6.1, PDF pp.7–8. K3: 2.8T total / 104B active parameters; first two models: about 3B active.

Qwen recovers most of the accuracy lost to INT8+SR

State	Bits	AIME	HMMT	GPQA-D	LCB-v6
FP32	32	85.46	57.57	81.97	86.95
FP16	16	84.58	55.82	82.13	86.80
INT8	9.0	18.48	20.12	79.48	32.23
INT8+SR	9.0	63.98	50.09	81.89	64.51
DAMP	9.9	83.82	55.11	82.05	85.67

The sensitive tasks reveal the gain

Versus INT8+SR: +19.84 pp AIME, +21.16 pp LCB.
Versus FP32: −1.64 pp AIME, −1.28 pp LCB.

Paper, Table 1, PDF p.7; selected columns. Accuracy in %, higher is better. pp = percentage points.

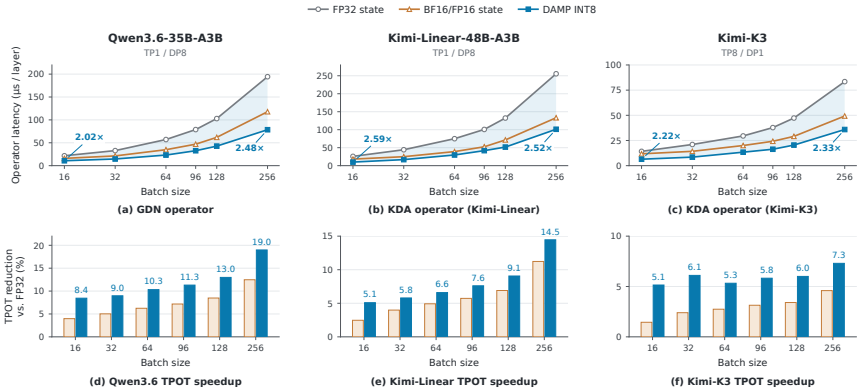
KDA results are close to FP32 at both model scales

		Kimi-Linear		Kimi-K3	
State	Bits	AIME	HMMT	AIME	HMMT
FP32	32	62.34	38.02	96.88	97.92
BF16	16	58.56	32.77	97.29	96.78
INT8+SR	9.0	58.07	32.05	93.96	91.67
DAMP	9.9	62.94	38.54	97.50	97.35

- Kimi-Linear: all six DAMP scores lie within 0.60 pp of FP32.
- K3 extends the evidence to a larger model on two math benchmarks.
- Small positive differences support near-FP32 accuracy; the tables do not provide confidence intervals.

Paper, Tables 1–2, PDF pp.7–8. Accuracy in %, higher is better.

Kernel speedups translate into smaller whole-model gains



Up to 2.59× kernel speedup; batch-256 TPOT reductions: 19.0% / 14.5% / 7.3%.

Paper, Fig.4, p.8; Sec.6.3, p.9. ShareGPT: 256 input / 64 output tokens. Qwen and Kimi-Linear: TP1/DP8; K3: TP8/DP1. The 2.59× point is Kimi-Linear at batch 16.

Smaller states also improve multi-turn cache reuse

State	Mean TTFT (s) ↓	P95 TTFT (s) ↓	Cache hit (%) ↑
FP32	38.427	72.471	47.20
BF16	35.646	69.500	56.96
DAMP	30.464	61.559	65.29

- K3, fixed total memory budget for recurrent states and KV cache.
- 256 four-turn WildChat conversations; prefix caching enabled.
- Versus FP32: 20.7% lower mean TTFT, +18.09 pp cache hits.

Paper, Table 3, PDF p.9; Appendix E, p.19. TTFT = time to first token; mean token-based cache hit rate.

At the same budget, error plus retention selects better rows

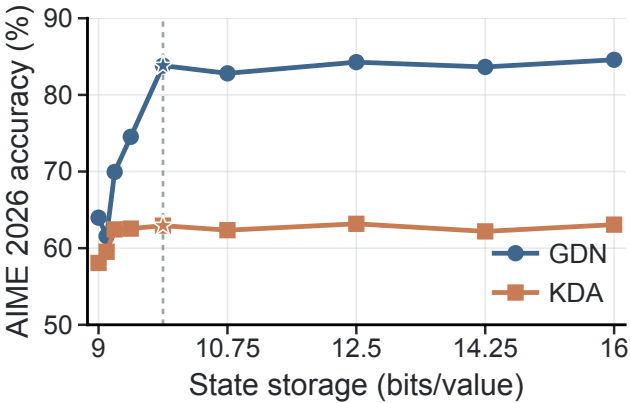
Selector	AIME	GPQA-D	LCB-v6
Random	59.06	62.12	58.68
State energy	60.36	63.38	59.67
Error-only	61.40	63.85	60.34
Retention-only	61.22	63.69	60.69
DAMP: $\mathbb{E}[eW]$	62.94	64.56	61.02

Matched comparison on Kimi-Linear (KDA)

Same INT8+SR quantizer and 9.9-bit budget.
DAMP improves AIME by 1.54 pp over error-only selection.

Paper, Table 4, PDF p.9. Accuracy in %, higher is better. GDN comparison in backup slides.

Protecting more than 16 rows gives limited extra benefit



Paper, Fig.5, Sec.6.4, PDF p.9. AIME 2026 accuracy; author figure.

- More FP16 rows increase effective state bits.
- Most recovery occurs by $K_{\text{hi}} = 16$.
- The curves are not strictly monotonic.

16 rows is a practical quality–storage choice.

Long-context accuracy and calibration splits are stable

RULER: 4K–128K, compressed state throughout prefill and decode

Largest absolute gap from FP32, in percentage points:

	Qwen	Kimi-Linear
DAMP	0.04	0.02
INT8+SR	0.13	0.15

Two disjoint, domain-balanced calibration subsets

Top-16 overlap: 92.0% KDA / 93.2% GDN.
Full-ranking Spearman: 0.990 / 0.986; random overlap expectation: 12.5%.

Paper, Appendix B, Table 6, PDF p.17; Appendix C.1, Table 7, p.18. Long-input RULER and difficult reasoning generation test different workloads.

The approximation and deployment scope remain important

- Risk proxy: only decay is modeled; key-dependent mixing in A_t and dynamics beyond the 256-step horizon are omitted.
- Budget constraint: every head receives 16 FP16 rows; no global reallocation across heads.
- 4-bit gap: INT4/NVFP4 variants do not recover FP32 accuracy.
- Reproduction details: GPU model and the numerical multi-turn memory budget are not specified; SR candidate construction is only sketched.

Paper, Sec.5.1; Appendices A.2 and F. Interpretation and reproduction gaps identified in this reading.

Allocate precision to errors that are large and long-lived

1. Mechanism: combine local quantization error with future decay retention.
2. Implementation: static FP16/INT8 rows and a fused FP32 update.
3. Evidence: near-FP32 task accuracy at 9.9 effective bits; 69.1% less state storage and up to 19.0% lower TPOT.

Discussion

At the same memory budget, which helps more: richer transition-aware scores or allocating different budgets to different heads?

Tao Zhang et al., DAMP: Decay-Aware Mixed-Precision Recurrent-State Quantization. arXiv:2608.27513v2, September 30, 2026.

Backup: the transformed state preserves the recurrence

$$T_t = P_K S_t D_c^{-1} P_V^\top, \quad q'_t = P_K q_t, \quad k'_t = P_K k_t, \quad v'_t = P_V D_c^{-1} v_t$$
$$A'_t = P_K A_t P_K^\top, \qquad T_t = A'_t T_{t-1} + \beta_t k'_t {v'_t}^\top$$
$$D'_c = P_V D_c P_V^\top, \quad o'_t = D'_c T_t^\top q'_t = P_V o_t, \quad W'_O = W_O P_V^\top$$

- P_K, P_V: fixed key/value permutations; D_c: positive diagonal scaling.
- Restore value scales before normalization and gating; associated parameters follow the same value order.
- With the initial state transformed, the unquantized block is equivalent.

Paper, Appendix A.3, Eqs.(17)–(18), PDF p.16.

Backup: choose SR using both low- and high-precision error

$$J(c) = \mathbb{E}_{\text{tune}}\left[\sum_{u \notin \mathcal{H}_c} W_{t,u} e_{t,u}^{(8,c)} + \sum_{u \in \mathcal{H}_c} W_{t,u} e_{t,u}^{(16,c)}\right]$$

- c: candidate value scaling/reordering; $\mathcal{H}_c$: its fitted FP16 set.
- Restore original state coordinates before measuring either error.
- Freeze $\mathcal{H}_c$ during selection; choose the candidate minimizing $J(c)$.
- Sample two 1,024-update windows, starting at updates 0 and 2,048; anchors follow updates 256, 512, and 768 in each window.

Paper, Appendix A.2, Eq.(16), Algorithm 1, PDF pp.15–16. Reported calibration time: about 7 / 9 / 78.75 min for Qwen / Kimi-Linear / K3.

Backup: accuracy uses repeated stochastic generations

Benchmark	Qwen / Kimi-Linear	Kimi-K3
AIME 2026	128	16
HMMT February 2026	64	8
IMO-AnswerBench	16	–
GPQA-Diamond	32	–
MMLU-Pro	16	–
LiveCodeBench-v6	8	–

- Qwen: temperature 0.6, top-p 0.95, top-k 20; max. output 65,536.
- Kimi-Linear: temperature 1.0, top-p 1.0; max. output 262,144.
- K3: temperature 1.0, top-p 0.95; max. output 262,144.

Paper, Appendix A.5, Table 5, PDF p.17. Table entries are generations per problem, not pass@k metrics.

Backup: GDN benefits from informed channel selection

Selector	AIME	GPQA-D	LCB-v6
Random	66.59	80.98	66.88
State energy	82.19	81.37	85.17
DAMP	83.82	82.05	85.67

- Same 9.9-bit budget and INT8+SR quantizer on Qwen3.6.
- Retention-only cannot distinguish rows within a GDN head.
- DAMP can reweight errors over time; this table does not include an error-only baseline.

Paper, Appendix D, Table 8, PDF pp.18–19. Accuracy in %, higher is better.

Backup: RULER remains close across 4K–128K

State	4K	8K	16K	32K	64K	128K
Qwen3.6-35B-A3B						
FP32	96.96	96.94	96.65	96.56	96.32	96.13
INT8+SR	96.93	96.89	96.60	96.43	96.28	96.07
DAMP	97.00	96.94	96.65	96.56	96.32	96.10
Kimi-Linear-48B-A3B						
FP32	95.27	94.44	93.53	92.37	91.44	88.02
INT8+SR	95.27	94.35	93.41	92.24	91.36	87.87
DAMP	95.27	94.42	93.53	92.36	91.44	88.02

All formats are used throughout prefill and decoding.

Paper, Appendix B, Table 6, PDF p.17; selected rows. Macro accuracy in %, higher is better.