VarRate

Training-Free Variable-Rate KV Cache Compression for Long-Context LLMs

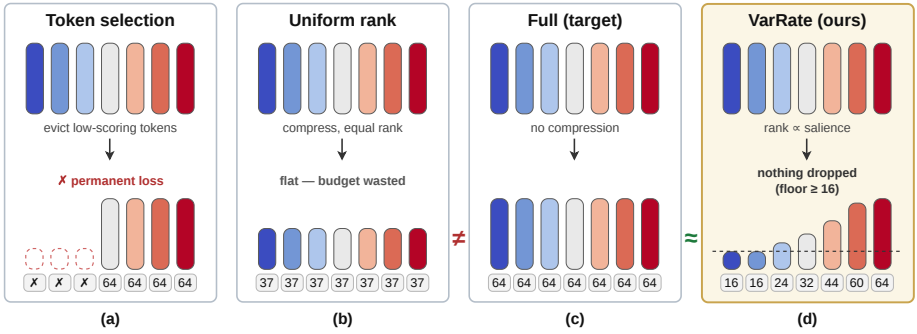
Paper authors: Shahrzad Esmat, Dhawal Shah, Ali Jannesari
Iowa State University

arXiv:2607.15498v1 — July 16, 2026
Reading presentation — September 24, 2026

Spend more rank on salient tokens; retain a representation of every token.

Independent paper discussion, not an author presentation. Main talk: 22 frames; 8 backup frames.

What happens when an importance score is wrong?

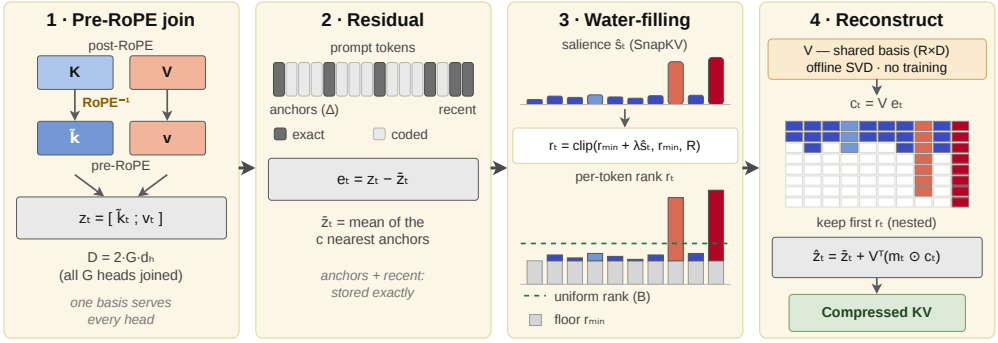


The question

Can a cheap attention score guide *fidelity* while preserving future access to every token?

Paper, Secs. 1–2; Fig. 1, PDF p. 2. Original schematic; bar heights are not experimental data.

One layer: represent, allocate, then reconstruct



The flat ablation shares the representation and basis. Only rank allocation changes.

Paper, Sec. 3, Fig. 2, PDF p. 4; Algorithm 1, p. 5. Original method schematic.

Encode residuals in a joint pre-RoPE space

$$z_t = [\tilde{k}_t; v_t] \in \mathbb{R}^D, \qquad D = 2Gd_h$$

- Undo RoPE on keys; concatenate all G KV heads and values.
- Keep strided anchors and the last w tokens exactly.
- For every other token, subtract its nearest-anchor mean:

$$\bar{z}_t = \frac{1}{c} \sum_{a \in \mathcal{N}_c(t)} z_a, \qquad e_t = z_t - \bar{z}_t.$$

Nearest means Euclidean distance in the joint KV space, not token position.

Paper, Sec. 3.1, Eqs. (1)–(2), PDF p. 3. d_h: head width; c: number of anchors.

Training-free still requires a calibrated basis

Once per model

Gather residuals on six unlabeled LongBench contexts. Use the top-R right singular vectors per layer.

$$V = [\mathbf{v}_1; \ldots; \mathbf{v}_R] \in \mathbb{R}^{R \times D}$$

- No gradients or fine-tuning.
- Default $R = 1024$.
- Llama basis: about 134 MB in half precision, all layers.
- Shared basis excluded from the per-sequence cache budget.

Deployment memory must still include the model-side basis.

Paper, Secs. 3.1 and 4.1, PDF pp. 3, 5; calibration cross-check: App. A.8, Table 10, p. 13.

The signal is cheap, but still query-dependent

$$s_t = \frac{1}{|\mathcal{W}|H} \sum_{q \in \mathcal{W}} \sum_{h=1}^{H} \alpha_{q,t}^{h}$$

- Average attention over the last w_{obs} queries and all query heads.
- Apply a length-5 moving average, then min–max normalization.
- With the question in the prompt tail, this can identify relevant content.
- Under document reuse, the question is absent when the cache is compressed.

Design choice

Use the same imperfect signal to coarsen low-scoring tokens instead of deleting their entire KV entries.

Paper, Sec. 3.2, Eq. (4), PDF p. 4. Signal inherited from SnapKV.

Water-filling redistributes a fixed coefficient budget

L: prompt length; κ: retained fraction; $\mathcal{E}, \mathcal{C}$: exact/coded sets.

$$B = D(\kappa L - |\mathcal{E}|), \qquad r_t = \mathrm{clip}(r_{\min} + \lambda \hat{s}_t, r_{\min}, R)$$

$$\sum_{t \in \mathcal{C}} r_t = B, \qquad |\mathcal{C}| r_{\min} \le B \le |\mathcal{C}| R.$$

- Start at the floor; assign extra rank in proportion to salience.
- Cap tokens at R and redistribute the remaining budget.

Teaching example: $B = 20$, floor 2, cap 8

Scores $(0, \frac{1}{3}, \frac{2}{3}, 1)$ give ranks $(2, 4, 6, 8)$ with $\lambda = 6$.
Uniform allocation would use $(5, 5, 5, 5)$.

Paper, Eqs. (3), (5), PDF pp. 3–4. Feasibility inequality and toy example: reader derivation.

Nested reconstruction gives a local error guarantee

$$\mathbf{c}_t = Ve_t, \qquad \hat{z}_t = \bar{z}_t + \sum_{j \le r_t} c_{t,j}\mathbf{v}_j$$

- Retain only the leading r_t coefficients; split back into K and V.
- Reapply each token's RoPE to its reconstructed key.
- Orthogonal directions: extra rank cannot increase the same token's squared reconstruction error.

What this does not establish

No guarantee of monotone task accuracy or globally optimal attention distortion. Discarded coefficients are still lost.

Paper, Sec. 3.3, Eq. (6), PDF p. 4. Guarantee boundary: reader interpretation.

Keep the evaluation protocols separate

- Main: Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct; 16 LongBench tasks, official metrics, greedy decoding.
- Default: $\kappa = .20$, $r_{\min} = 16$, $R = 1024$, $\Delta = 16$, $c = 4$, $w = w_{\text{obs}} = 64$.
- Main setup states 200 examples/task; paired comparisons over 16 tasks with 20,000 bootstrap resamples.
- Reuse: compress without the question; passage / MuSiQue / Qasper. Standard three-task ablations use GovReport instead.

Budget and fairness

KIVI-2 is near 20%; KIVI-4 is 31%. KVzip receives the same 64-token tail pin as VarRate.

Paper, Secs. 4.1–4.3, PDF pp. 5–7. AVG-16 averages heterogeneous task scores.

At 20%, the strongest result is the flat ablation

Method	Llama AVG-16	Qwen AVG-16
Full	46.01	48.79
VarRate	45.69	48.00
flat	43.47	30.50
SnapKV	45.79	46.25
PyramidKV	46.02	44.90
KIVI-2	44.82	46.82

- Relative to Full: -0.32 / -0.79 score points.
- Relative to flat: $+2.22$ / $+17.50$ at matched storage.

Selected rows from paper Table 1, PDF p. 6. Higher is better. KIVI-2 uses about 19%, not exactly 20%.

A nominal lead is not always statistically supported

VarRate minus baseline	Llama 95% CI	Qwen 95% CI
flat	$[+0.72, +4.05]$	$[+10.77, +24.69]$
SnapKV	$[-1.18, +0.93]$	$[+0.49, +3.59]$
KIVI-2	$[-0.73, +2.76]$	$[-0.30, +3.11]$

- Flat is significantly worse on both models.
- Selection-family wins are significant mainly on Qwen.
- Neither model establishes a win over KIVI-2.

An interval crossing zero is a non-significant difference, not proof of equivalence.

Paper, App. A.13, selected rows from Table 15, PDF p. 16; paired bootstrap over 16 tasks.

Reuse hurts selection more in the tested protocol

Query-agnostic, three-task mean	Llama	Qwen
Full control	59.91	58.21
VarRate	56.21	51.79
SnapKV	45.75	41.35
Ada-KV	47.48	40.28
KVzip, identical tail pin	55.34	56.42

- From query-aware to reuse: VarRate loses 3.53 / 5.50 points.
- SnapKV loses 12.96 / 14.66; Full barely changes.
- KVzip is stronger on Qwen reuse.

Paper, Sec. 4.3; Table 2, PDF p. 7. Tasks: passage / MuSiQue / Qasper. Scenario drops are not gaps to Full.

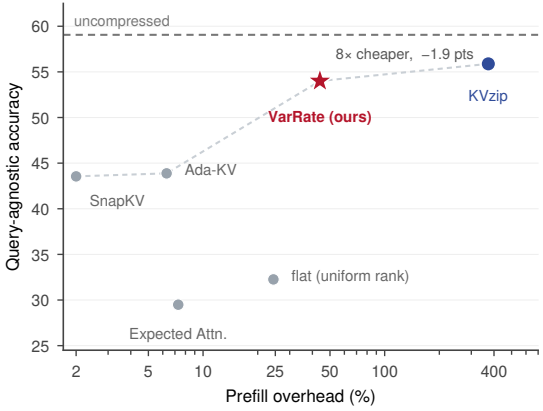
The powered KVzip comparison favors a careful claim

VarRate − KVzip	Delta	95% CI
Llama, query-aware	-0.55	$[-2.00, +0.91]$
Llama, query-agnostic	$+0.95$	$[-1.48, +3.42]$
Qwen, query-aware	-0.04	$[-1.67, +1.62]$
Qwen, query-agnostic	-5.06	$[-7.57, -2.58]$
Pooled, all four cells	-1.18	$[-2.20, -0.16]$

- About 600 paired examples/cell, 2,400 pooled.
- No significant difference in three cells; KVzip leads in the fourth and in the pooled comparison.

Paper, App. A.16, Table 18, PDF p. 17. Different aggregation from the main-text four-cell mean gap of 0.77.

About 8× lower added cost is not 8× faster



- Added prefill: VarRate +44.0%, KVzip +372.2%.
- Added-cost ratio: $372.2/44.0 \approx 8.46$.
- Including base prefill: $4.722/1.44 \approx 3.28$.

VarRate is about 1.9 points below KVzip on the plotted reuse mean.

Paper, Fig. 4 and Table 3, PDF p. 7. Cost: Llama, A100, 16K; accuracy: two-model reuse mean. Ratios: reader calculation.

The current accuracy path reconstructs full K/V

Compressed representation versus runtime memory

The reported 20.2% cache representation is not a demonstrated 5× reduction in end-to-end peak GPU memory.

- Decode: 37.4 tok/s for VarRate versus 37.5 for Full.
- A fused kernel is still needed to consume compressed states efficiently.
- End-to-end retrieval: 1.256 s → 1.781 s (+41.8%).
- End-to-end GovReport: 13.59 s → 13.95 s (+2.6%, reported within noise).

Paper, Sec. 4.4, Table 3, PDF p. 7; App. A.14, Table 16, p. 16. Wall-clock: Llama, A100, n=18.

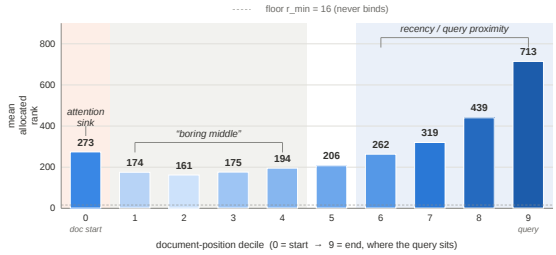
Both content scoring and graded allocation matter

Llama, 20% budget	Passage	GovReport
flat	87.5	30.1
Random score + water-filling	93.0	29.6
Salience + binary split	99.5	29.0
Salience + water-filling	99.5	30.4

- Real salience adds retrieval quality beyond random allocation.
- A binary split matches retrieval, but loses distributed summary information.
- Controls on Qwen also penalize scores assigned to the wrong tokens.

Paper, Table 4, PDF p. 7; content controls in Table 19, p. 18. Scores higher are better.

The explicit rank floor is slack in observed runs



- Removing the floor changes each tested task by less than one point.
- Logged floor occupancy is reported as 0.0% for both settings.
- The tail receives much more rank; content alignment adds value beyond position.

Preserving a representation is the design; the explicit floor is not an established active cause of the gains.

Paper, App. A.19, Table 21, PDF p. 18; Fig. 7, p. 19. Means by position do not prove a per-token minimum.

Quantization is a strong alternative and a complement

Llama, coefficient coding	Passage	MuSiQue	GovReport	Storage
VarRate fp16	99.5	32.0	30.4	$\sim 20\%$
VarRate + 3-bit	99.5	31.3	30.4	$\sim 9\%$

- The exact anchors and metadata limit how far coefficient quantization reduces storage.
- KIVI-2 is stronger on Llama GovReport: 35.2 versus 30.4.
- Near-parity to fp16 VarRate does not mean parity to Full or to KIVI.

Paper, Apps. A.3–A.4, Tables 6–7, PDF pp. 10–11. Three-task evidence; runtime caveat still applies.

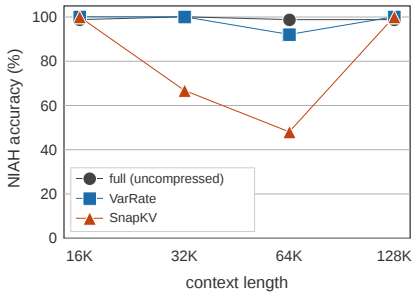
The advantage does not grow indefinitely as budgets shrink

Qwen passage retrieval, query-aware	Keep 8%	Keep 10%	Keep 12%
SnapKV	99.5	99.5	99.5
VarRate	29.6	98.8	99.5
flat	27.8	37.3	46.0

- At 8%, VarRate approaches flat's retrieval score on Qwen.
- Selection can still retain some useful tokens exactly.
- Wider anchor stride helps tight budgets, but loses on the 20% full suite.

Paper, App. A.20, selected results from Table 22, PDF p. 20; stride: Table 14, p. 15.

Long-context retrieval is promising, with a narrow scope



- Fixed needle set; 75 examples per cell; 20% budget.
- At 64K: VarRate 92%, SnapKV 48%.
- Both reach 100% at 128K; the rebound is unexplained.
- The 128K full-reconstruction run needs two GPUs.

Paper, App. A.21, Fig. 8 and Table 24, PDF pp. 19–20. Llama, query-agnostic NIAH, not a broad 128K benchmark.

The evidence supports allocation, with deployment work left

- **Strongest evidence:** adaptive versus uniform rank under a shared codec and budget.
- **Workload boundary:** reuse benefits differ from known-query, very-low-budget behavior.
- **System boundary:** offline basis, quadratic anchor search, and full K/V reconstruction remain costs.
- **Comparison boundary:** no universal win over quantization or KVzip; no direct trained-codec comparison.

Reader's next experiment

Measure a fused variable-rank implementation on real multi-query reuse: peak memory, time to first token, throughput, and quality together.

Synthesis of Secs. 3–4; Apps. A.5, A.19–A.21. Proposed experiment is reader analysis, not a reported result.

Three takeaways for the discussion

1. A cheap score can control how much information survives, not only which token survives.
2. The cleanest win is over uniform allocation; competing methods remain strong in their own regimes.
3. Storage quality, actual memory savings, and latency must be evaluated separately.

Primary source

Esmat, Shah, and Jannesari. *VarRate: Training-Free Variable-Rate KV Cache Compression for Long-Context LLMs*.
arXiv:2607.15498v1, July 16, 2026.

Read: author TeX, original figures, final 20-page PDF, Secs. 1–5 and Apps. A.1–A.21. Experiments not independently rerun.

Backup: three cross-checks support a limited generalization

Language modeling: WikiText-2

VarRate PPL 6.95 versus flat 7.26 at identical realized storage. Lower is better; detailed setup is incompletely specified.

Third family: Mistral-7B

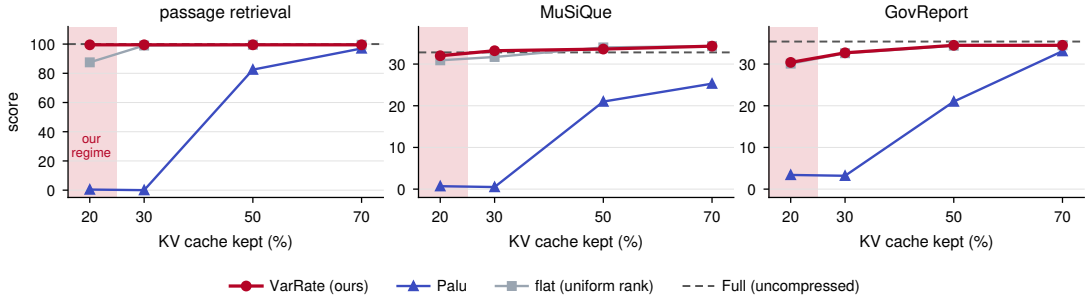
At a 15% budget, passage retrieval is 73.6 for VarRate, 29.1 for flat, and 75.6 for Full. This is a three-task cross-check.

Calibration choice

Three independently built bases vary by at most 1.8 points on the tested Llama tasks. This is not an arbitrary-domain guarantee.

Paper, App. A.1, PDF p. 10; Apps. A.7–A.8, Tables 9–10, p. 13.

Backup: the Palu gap also includes representation design



- Flat already beats Palu: residual/anchor coding contributes to the gap.
- Palu footprint 20.41% versus VarRate 20.22% at the tightest point.
- The reported Palu kernel does not support these GQA shapes.

Paper, Fig. 3, PDF p. 6; App. A.6, Table 8 and implementation discussion, pp. 11–12.

Backup: summary gains survive a small budget sweep

Llama GovReport	Keep 15%	Keep 20%	Keep 25%
Best selection score in Table 5	27.8	28.8	29.5
VarRate	28.6	30.4	32.2
Difference	+0.8	+1.6	+2.7

- Selection set: SnapKV, PyramidKV, Ada-KV.
- Reported significance at 20% and 25%; retrieval and multi-hop QA are not significantly separated.
- This is a specific summarization task, not a win over quantization.

Paper, App. A.2, Table 5, PDF p. 10. Best-selection rows and differences computed from the table.

Backup: anchor stride changes the exactness–rank tradeoff

Llama protocol	Stride 16	Stride 64
10%, query-agnostic, 3 tasks	29.92	48.88
20%, query-agnostic, 3 tasks	56.21	57.34
20%, AVG-16	45.69	44.98

- Fewer exact anchors free more coefficients for coded tokens.
- Full-suite losses concentrate in few-shot and code tasks.
- Default remains 16; the wider setting is an ablation.

Paper, App. A.12, Table 14, PDF p. 15. App. A.12's 256-example rerun wording differs from the main n=200 setup.

Backup: implementation dominates much of the measured tax

Share of compression time

- Anchor search: 32.5%.
- PCA projection: 21.8%.
- Reconstruction: 19.9%.
- Water-filling: 6.4%.

Separate optimized run

- Reference overhead: +41.7%.
- GPU-cached basis: +40.3%.
- Additional bf16 compute: +25.1%; max score change 0.07.

Reimplementation detail

Keep anchor distances numerically stable; the tested bf16 distance shortcut failed. Preserve salience precision when comparing allocations.

Paper, App. A.11, Tables 12–13, PDF p. 15. FLOPs share is not a measured latency lower bound.

Backup: content alignment survives a stronger control

Qwen, 20% budget	Passage	GovReport	MuSiQue
flat	75.0	29.6	8.5
Mass + water-filling	99.5	31.3	28.7
Shuffled mass	64.7	27.4	7.5
Position-only	89.0	26.5	7.7
Fisher + water-filling	100.0	30.8	27.9

- Shuffling preserves the score multiset but destroys token alignment.
- Position alone does not recover multi-hop QA; it does exceed flat on retrieval.
- The second-order Fisher heuristic is reported statistically tied with mass.

Paper, Apps. A.17–A.18, Tables 19–20, PDF p. 18. Later controls use a separately calibrated basis.

Backup: RULER broadens tasks, not the tested length range

Llama, RULER at 16K	Score
Full	90.0
SnapKV, 20%	81.7
VarRate, 20%	82.5

- Only 40 examples; VarRate–SnapKV difference is within noise.
- Both compressors lose about eight points to Full.
- No RULER results beyond 16K are reported here.
- The paper's 128K evidence comes from controlled NIAH retrieval.

Paper, App. A.15, Table 17, PDF p. 16. Uncertainty widths omitted here because their precise type is not clearly defined.

Backup: details that matter when reusing this v1

- KVzip: task-averaged and example-pooled gaps differ; pooled analysis favors KVzip by 1.18 points.
- Fig. 6 labels Qwen flat retrieval as 87.5; Table 19 gives 75.0. This deck uses the table.
- Table 19's broad caption overstates the negative controls: position-only beats flat on Qwen retrieval.
- The paper discloses a fixed cross-model table-aggregation bug. Raw result files were not independently audited for this reading.

Read quantitative tables, statistical units, and implementation limits together.

Paper, App. A.9, PDF p. 12; Fig. 6 and Tables 18–19, pp. 17–18. Source: arXiv:2607.15498v1 only.