BELT: BEHAVIOR-ALIGNED LOW-RANK KV CACHE COMPRESSION WITH LEARNED TOP-K RANK ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context inference makes the key-value (KV) cache a major memory bottleneck for large language models. We propose BELT, a low-rank KV cache compression framework. Its key-side objective targets attention distributions rather than raw keys or pre-softmax logits, yielding a quadratic surrogate with a closed-form low-rank solution. A learned global rank allocator uses differentiable soft selection during training and hard selection at deployment to meet a user-specified component budget. Experiments demonstrate improved model quality under aggressive compression and increased decoding throughput at comparable accuracy. Evaluated across three LLMs and multiple benchmarks, BELT achieves up to 70% low-rank KV cache compression while outperforming the strongest baseline by up to 23.3 points on LongBench, and up to $17.8\times$ overall KV cache reduction when combined with quantization. With Triton-based GPU kernels and a quantized cache, BELT delivers up to $2.7\times$ speedup for the attention module and $3.3\times$ end-to-end decoding throughput.

1 INTRODUCTION

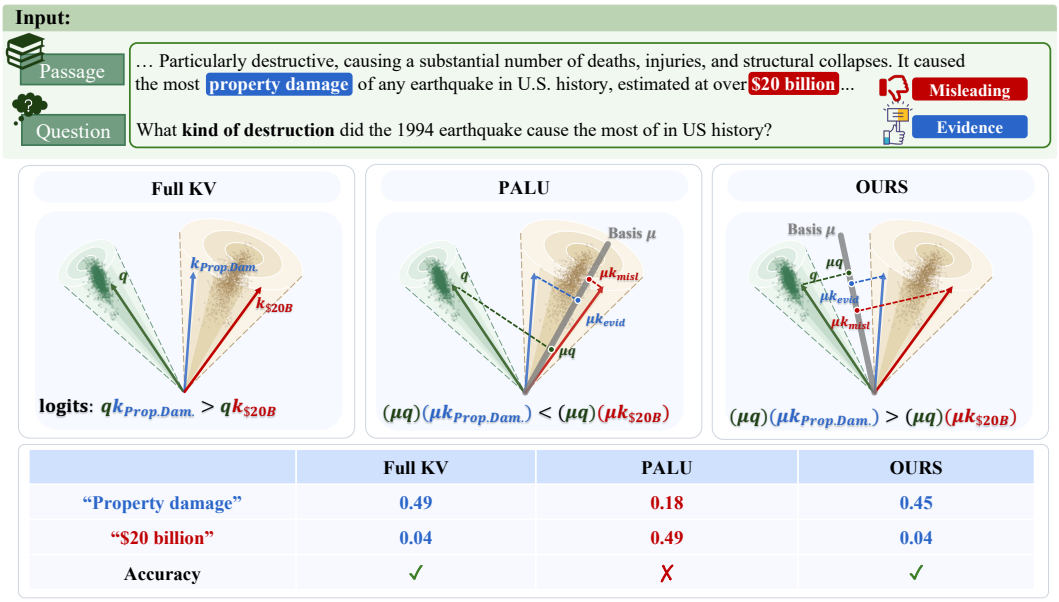
Large language models (LLMs) have become ubiquitous in modern AI applications, yet their growing scale—in both model size and context length—poses significant inference challenges. Modern LLMs support 128K-token contexts or more (Liu et al., 2024b; Stallone et al., 2024; Xu et al., 2025), making the **key-value (KV) cache** a critical inference bottleneck. During autoregressive generation, the cache stores keys and values for preceding tokens, and its memory footprint grows linearly with sequence length, potentially exceeding available GPU memory (Jiang et al., 2026). Efficient KV cache management is therefore essential, motivating approaches such as **token eviction** (Zhang et al., 2023; Liu et al., 2023), **quantization** (Hooper et al., 2024; Lin et al., 2025), and **low-rank decomposition** (Chang et al., 2025; Bhatnagar et al., 2026).

Among these, *low-rank decomposition* exploits hidden-dimension redundancy to reduce KV cache memory. Prior work follows two broad directions: adapting general-purpose decomposition methods to the KV cache (Wang et al., 2025; Yuan et al., 2023; Abbasi et al., 2026; Bai et al., 2025; 2026; Zhou et al., 2026), and developing KV-specific methods tailored to attention and caching dynamics (Zhang et al., 2024; Chang et al., 2025; Lee et al., 2025; Mu et al., 2025; Ji et al., 2026; Chang et al., 2026). STAR-KV (Bhatnagar et al., 2026) and A^3 (Wong et al., 2025) represent recent state-of-the-art advances in adaptive rank control and analytical attention approximation, respectively.

Despite this progress, two challenges still remain. First, *the compression criterion of the state-of-the-art does not reflect the effect of approximation errors on attention behavior*. Methods based on KV reconstruction do not explicitly account for how key errors interact with queries and the nonlinear softmax map. As illustrated in Figure 1, retaining dominant singular directions of the KV tensors can distort attention and change the generated answer. A^3 moves beyond raw key reconstruction by targeting *pre-softmax attention logits*, which, however, is not the same as preserving attention probabilities or the resulting attention output. *This motivates a key-compression criterion that explicitly accounts for post-softmax behavior*.

Second, *allocating ranks across heterogeneous KV projections under a prescribed global budget remains challenging*. Uniform allocation ignores differences in compression sensitivity, while

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.



	Full KV	PALU	OURS
"Property damage"	0.49	0.18	0.45
"$20 billion"	0.04	0.49	0.04
Accuracy	✓	✗	✓

Figure 1: **Left:** Full attention, where each query attends to all context tokens. **Middle:** State-of-the-art low-rank compression (e.g., vanilla SVD) retains only the dominant KV eigenvector μ, reducing the query–key product from $q^\top k$ to $(q^\top \mu)(\mu \top k)$. Since μ is derived solely from the input-context tokens and ignores subsequent queries, this distorts the attention distribution and can yield an incorrect answer. **Right:** The dominant eigen-vector derived by ours aligns the query and key decompositions, preserving the downstream attention output and yielding the correct answer.

sensitivity-based searches and learned threshold methods offer different approaches to adaptive allocation. STAR-KV (Bhatnagar et al., 2026) learns block-specific thresholds with a soft compression penalty, without directly enforcing a prescribed global component budget. Our analysis further identifies these learned thresholds concentrate under certain conditions, potentially limiting allocation effectiveness.

We address these challenges from two complementary perspectives. *First, we design compression objectives to preserve the attention output of the uncompressed model.* While our value-side construction is closely related to A^3 (Wong et al., 2025), our key-side objective targets attention-distribution preservation rather than pre-softmax logit reconstruction like A^3. By accounting for query-dependent softmax sensitivity and RoPE interactions, BELT better preserves attention behavior under aggressive compression, as demonstrated empirically. *Second, we learn ranks across KV blocks under an explicit global budget.* Our allocator learns how to distribute ranks across attention heads according to their compression sensitivity. We use a differentiable relaxation of top-B selection for end-to-end training, followed by hard selection of B components, where B is the user-specified budget.

Extensive experiments evaluate BELT across multiple models, compression budgets, and tasks. At 70% KV cache compression on Qwen3-32B, BELT improves the average score over the strongest evaluated baseline by 23.3 points on LongBench and 13.8 points on RULER. We additionally evaluate perplexity, zero-shot accuracy, and the contributions of the compression bases and rank allocation; combined with 3-bit KV quantization, BELT reaches 17.8× KV cache compression within 2 points of the uncompressed zero-shot accuracy. System-level measurements show up to 3.26× higher decoding throughput and 87% lower KV cache memory than full-cache inference on LLaMA-2-7B with KV quantization. These results demonstrate the practical benefits of attention-aware compression and globally coordinated rank allocation.

In summary, our contributions are:

- We formulate an attention-output-motivated approach to low-rank KV compression, with a key-side metric derived from attention-distribution KL rather than pre-softmax logit reconstruction.

- We derive quadratic compression surrogates with exact weighted-SVD solutions, extend the construction to RoPE and GQA, and explicitly characterize their approximation errors.
- We introduce a learned global rank allocator that enforces a relaxed budget during training and an exact component-count budget at deployment, while preserving an ordered prefix within each KV block.
- We empirically evaluate model quality, approximation accuracy, rank allocation, and decoding efficiency under aggressive KV cache compression.

2 PRELIMINARIES

2.1 ATTENTION AND KV CACHE

We consider a single attention head in a Transformer decoder layer (Vaswani et al., 2023). Let $X \in \mathbb{R}^{n\times d}$ be the stacked representations of the n preceding tokens stored in the KV cache. For any input token $x \in \mathbb{R}^{1\times d}$, each head computes its query, key, and value projections as

$$q_x = xW_Q, \quad k_x = xW_K, \quad v_x = xW_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_h}$ and d_h is the per-head dimension. During autoregressive generation, the queries, keys, and values of the n tokens are obtained from X as

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

so the KV cache stores $K, V \in \mathbb{R}^{n\times d_h}$, which is interacted with q for attention calculation as follows:

$$\begin{aligned} &o_x(W_K, W_V) = a_x(W_K)\, V\, W_O \in \mathbb{R}^{1\times d} = a_x(W_K)\, X\, W_V\, W_O \in \mathbb{R}^{1\times d}, \\ &\text{where } a_x(W_K) = \mathrm{Softmax}\left(\frac{q_x K}{\sqrt{d_h}}\right) = \mathrm{Softmax}\left(\frac{q_x XW_K}{\sqrt{d_h}}\right) \in \mathbb{R}^{1\times n}, \text{ and } W_O \in \mathbb{R}^{d_h\times d} \end{aligned} \tag{1}$$

In this formula, $a_x(W_K)$ is the attention probability vector, where we explicitly highlight its reliance on W_K. Similarly, we use $o_x(W_K, W_V)$ to denote the final attention output which relies on both W_k and W_V. In practice, Rotary Position Embedding (RoPE) (Su et al., 2024) is applied to queries and keys as a non-linear, position-dependent embedding before attention. More detailed description for attention mechanism with RoPE and how BELT is generalized to RoPE is Appendix D.

State-of-the-art large language models typically employ multiple attention heads per layer with two widely adopted strategies: Multi-head attention (MHA) runs h such heads in parallel, while grouped-query attention (GQA) shares each key–value pair among a group of query heads. Our framework aims to compress KV cache at individual heads, which thus applies to both. We start by assuming the use of MHA by default and discuss the extension of BELT to GQA in Appendix C.

2.2 LOW-RANK KV CACHE COMPRESSION

Low-rank compression reduces the KV cache size by approximating each key/value projection matrix $W_K \in \mathbb{R}^{d\times d_h}, W_V \in \mathbb{R}^{d\times d_h}$ with a rank-r factorization ($r < d_h$):

$$W_K \approx \hat{W}_K = W_K^{\mathrm{down}} W_K^{\mathrm{up}}, W_V \approx \hat{W}_V = W_V^{\mathrm{down}} W_V^{\mathrm{up}}$$

where $W_K^{\mathrm{down}}, W_V^{\mathrm{down}} \in \mathbb{R}^{d\times r}$, $W_K^{\mathrm{up}}, W_V^{\mathrm{up}} \in \mathbb{R}^{r\times d_h}$ and $\hat{W}_K, \hat{W}_V \in \mathbb{R}^{d\times d_h}$, which all have rank at most r. For an input token sequence X, we project it using $\hat{W}_K$ and $\hat{W}_V$ as follows:

$$XW_K \approx XW_K^{\mathrm{down}} W_K^{\mathrm{up}}, XW_V \approx XW_V^{\mathrm{down}} W_V^{\mathrm{up}}$$

where $XW_K^{\mathrm{down}}, XW_V^{\mathrm{down}} \in \mathbb{R}^{n\times r}$ are the compressed KV cache. Instead of storing the full $K, V \in \mathbb{R}^{n\times d_h}$, we cache only the compressed representations $K' = XW_K^{\mathrm{down}}$ and $V' = XW_V^{\mathrm{down}}$. This reduces the per-token memory footprint from d_h to r during inference. Following prior studies, e.g., Bhatnagar et al. (2026); Chang et al. (2025), W_K^{down}, W_V^{down}, W_K^{up} and W_V^{up} are learned with a small calibration dataset, $\mathcal{X}$.

2.3 Inference with Compressed Cache

The overall inference process with low-rank decomposition remains largely the same as that with no compression. The only difference is at the attention computation phase, where low-rank decomposition requires that the input sequence X is first projected down using W_K^{down} and W_V^{down}, which is then cached throughout the inference phase, and projected back to d_h dimensions using W_K^{up} and Wup_V respectively. Weight absorption (Liu et al., 2024a; Chang et al., 2025) further accelerates this process by merging W_V^{up} into the subsequent projection for value tensors. Specifically, the attention output defined in equation 1 changes from $a_x(W_K)\,V\,W_O$ to $a_x(W_K)\,V'\,W_V^{\text{up}}\,W_O$. Since both W_V^{up} and W_O are fixed, precomputing $W'_O = W_V^{\text{up}} W_O \in \mathbb{R}^{r\times d}$ yields $a_x(W_K)\,V'\,W'_O$, eliminating the need of multiplying projecting V' back to d_h dimensions at each step. In contrast, a similar absorption does not apply to the key tensor since W_K^{up} cannot be absorbed with the RoPE computation due to its non-linearity nor commute it to merge it with other matrices like the query projection W_Q.

3 The BELT Framework

In this section, we first present low-rank decomposition within one head under a given rank budget in Section 3.1. We then generalize the method to multiple heads under a global budget in Section 3.2, and finally combine rank allocation with quantization in Section 3.3.

3.1 Low-rank decomposition within one head

3.1.1 Decomposition objective

Unlike methods that directly minimize key/value reconstruction error, BELT seeks low-rank projections whose induced attention output remains close to that of the original model, which is formulated as the following loss function defined over every input x from the calibration data $\mathcal{X}$:

$$J(\widehat{W}_K,\widehat{W}_V)=\sum_x \left\|o_x(W_K,W_V)-o_x(\widehat{W}_K,\widehat{W}_V)\right\|_2^2,\ \widehat{W}_K^*,\widehat{W}_V^*=\underset{\text{rank}(\widehat{W}_K)\le r_K,\text{rank}(\widehat{W}_V)\le r_V}{\arg\min}\ \mathcal{J}(\widehat{W}_K,\widehat{W}_V).$$

However, jointly solving $\widehat{W}_K^*,\widehat{W}_V^*$ in $J(\widehat{W}_K,\widehat{W}_V)$ is difficult due to their inter-dependency. Indeed, the above objective, $\mathcal{J}(\widehat{W}_K,\widehat{W}_V)$, can be relaxed to the summation of the following two objectives, which decouples the optimization of $\widehat{W}_K^*$,$\widehat{W}_V^*$:

$$\begin{aligned}&\mathcal{J}(\widehat{W}_K,\widehat{W}_V)\le 2\mathcal{L}_V(\widehat{W}_V)+4M^2\cdot\mathcal{L}_K(\widehat{W}_K),\\&\text{where }\mathcal{L}_V(\widehat{W}_V):=\sum\nolimits_x\left\|a_xX(W_V-\widehat{W}_V)W_O\right\|_2^2,\\&\mathcal{L}_K(\widehat{W}_K):=\sum\nolimits_x D_{\text{KL}}\Big(a_x(W_K)\,\Big\|\,a_x(\widehat{W}_K)\Big),\ M_{X,t}:=\max_i\|(\widehat{Z}_{X,t})_{i,:}\|_2\le M.\end{aligned}\tag{2}$$

The detailed derivation of the above formula is delayed to Appendix A.

3.1.2 Solving $\mathcal{L}_V(\widehat{W}_V)$

Indeed, we can derive a closed-form $\widehat{W}_V$ by solving $\mathcal{L}_V(\widehat{W}_V)$. To achieve this, we first transform $\mathcal{L}_V(\widehat{W}_V)$ into the following surrogate objective.

Lemma 1. *Let $G_V^0=\sum_x X_x^\top a_x^\top a_x X_x, H_V^0=W_OW_O^\top$. For $\epsilon>0$, let $G_V=G_V^0+\epsilon I_d=L_V^\top L_V, H_V=H_V^0+\epsilon I_{d_h}=C_VC_V^\top$ be their positive-definite regularizations and Cholesky factorizations. Define*

$$\mathcal{L}_V^{\approx}(\widehat{W}_V)=\left\|L_V(W_V-\widehat{W}_V)C_V\right\|_F^2.$$

Then, with $E=W_V-\widehat{W}_V$, $\mathcal{L}_V^{\approx}(\widehat{W}_V)-\mathcal{L}_V(\widehat{W}_V)=\Delta_{\epsilon,V}(E)\ge 0$,

where $\Delta_{\epsilon,V}(E)=\epsilon\sum_x\|a_xX_xE\|_2^2+\epsilon\|EW_O\|_F^2+\epsilon^2\|E\|_F^2$. In particular, for fixed E and calibration statistics, $\Delta_{\epsilon,V}(E)=O(\epsilon)$ as $\epsilon\to 0$.

The proof of Lemma 1 is given in Appendix B.1.1. Subsequently, by performing some essential linear algebric transformations and applying Eckart–Young–Mirsky theorem Eckart & Young (1936); Mirsky (1960), The following theorem gives a closed-form global minimizer of the regularized value surrogate $\mathcal{L}_V^{\approx}$.

Theorem 1. *let $L_V W_V C_V = \bar{U}_V \bar{\Sigma}_V \bar{R}_V^\top$ be a thin singular value decomposition, where $\bar{\Sigma}_V = \mathrm{diag}(\bar{\sigma}_{V,1}, \ldots, \bar{\sigma}_{V,m}), m = \min(d, d_h)$, and $\bar{\sigma}_{V,1} \geq \cdots \geq \bar{\sigma}_{V,m} \geq 0$. Let $\bar{U}_{V,1:r}$, $\bar{\Sigma}_{V,1:r}$, and $\bar{R}_{V,1:r}$ denote the components associated with the largest r singular values. Then one global minimizer of $\min_{\mathrm{rank}(\widehat{W}_V)\leq r} \mathcal{L}_V^{\approx}(\widehat{W}_V)$ is*

$$\widehat{W}_{V,r}^* = L_V^{-1} \bar{U}_{V,r} \bar{\Sigma}_{V,r} \bar{R}_{V,r}^\top C_V^{-1}. \tag{3}$$

The corresponding projection matrix, $W_{V,1:r}^{\mathrm{down}} = L_V^{-1} \bar{U}_{V,1:r} \bar{\Sigma}_{V,1:r}^{1/2}$, and $W_{V,1:r}^{\mathrm{up}} = \bar{\Sigma}_{V,1:r}^{1/2} \bar{R}_{V,1:r}^\top C_V^{-1}$.

The proof is provided in Appendix B.1.2. The above decomposition via solving $\mathcal{L}_V(\widehat{W}_V)$ is closely related to A^3 (Wong et al., 2025); we do not claim the weighted-SVD solution itself as a distinct contribution. Our principal distinction lies in the key-side decomposition described next.

3.1.3 Solving $\mathcal{L}_K(\widehat{W}_K)$

Unlike the value objective, $\mathcal{L}_K(\widehat{W}_K)$ contains the nonlinear softmax map. Hence, directly solving $\mathcal{L}_K(\widehat{W}_K)$ cannot yield a closed-form solution. To mitigate this issue, we first approximate $\mathcal{L}_K(\widehat{W}_K)$ by a quadratic objective using Taylor expansion Rudin (1976) and Kronecker decomposition Wilkinson (1979), which only incurs a small approximation error. Before formally presenting how this approximation is performed in Theorem 2, we first compute the following intermediate quantities, taking as input each query token $x \in X$ and its corresponding attention probability distribution a_x calculated using equation 1:

$$J_x = \mathrm{diag}(a_x) - a_x^\top a_x, , A_x := J_x \,\mathrm{diag}(1/a_x) J_x, G_{K,x}^{\epsilon} := \frac{1}{2d_h} X^\top A_x X + \epsilon I_d, \tag{4}$$

Theorem 2. *Let $M_K := \sum_x q_x q_x^\top \otimes G_{K,x}^{\epsilon} \in \mathbb{R}^{(d_h d)\times(d_h d)}$, and approximate it using its Kronecker rank-1 approximation $M_K \approx A_K \otimes B_K, A_K \in \mathbb{R}^{d_h \times d_h}, \quad A_K \succeq 0, B_K \in \mathbb{R}^{d\times d}, \quad B_K \succeq 0$. For $\epsilon > 0$, let $B_K + \epsilon I_d = L_K^\top L_K, A_K + \epsilon I_{d_h} = C_K C_K^\top$ be Cholesky decompositions. Then $L_K \in \mathbb{R}^{d\times d}$ and $C_K \in \mathbb{R}^{d_h \times d_h}$ are invertible. Define the surrogate objective*

$$\mathcal{L}_K^{\approx}(\widehat{W}_K) := \left\| L_K (W_K - \widehat{W}_K) C_K \right\|_F^2.$$

This is a quadratic function of $E := W_K - \widehat{W}_K$, which incurs an approximation error

$$|\mathcal{L}_K(\widehat{W}_K) - \mathcal{L}_K^{\approx}(\widehat{W}_K)| \leq \underbrace{|\,\mathrm{vec}(E)^\top (M_K - A_K \otimes B_K)\,\mathrm{vec}(E)|}_{\textit{Kronecker truncation error}} + \underbrace{\mathcal{E}_{\mathrm{Taylor}}(E)}_{\textit{Taylor approximation error}} + O(\epsilon),$$

where $\mathcal{E}_{\mathrm{Taylor}}(E)$ represents the overall Taylor approximation error.

The proof and explicit approximation error terms are provided in Appendix B.2.1, and Algorithm 2 in Appendix B.2.3 details the construction of $\mathcal{L}_K^{\approx}$. Our empirical analysis in Section 4 verifies that both Taylor and Kronecker errors are consistently small at the low-rank solutions returned by our method. These results support the accuracy of $\mathcal{L}_K^{\approx}$ as an approximation to $\mathcal{L}_K$ at those solutions under the evaluated compression settings. They do not establish a uniform approximation guarantee over all rank-constrained candidates.

Theorem 2 yields a weighted quadratic surrogate with the same algebraic form as the one for $\mathcal{L}_V(\widehat{W}_V)$. Applying the argument of Theorem 1 therefore gives a closed-form minimizer of $\mathcal{L}_K^{\approx}$ (see Corollary 1 in Appendix B.2.2).

3.2 Learned Rank Allocation across heads

An LLM contains key and value projections across many heads and layers, and their sensitivity to low-rank compression varies substantially. We therefore allocate a global singular component

budget across these heads. In this section, we concatenate all W_K's and W_V's across all heads across all layers and index them using $i \in \{1, \ldots, N\}$. Plus, we use m_i to denote the total numbers of singular components for each i, and B is the total number of retained components across all i's. We require $N \leq B \leq \sum_i m_i$ and retain at least one component for each i.

To address this problem, we propose to rank all singular components globally throughout training, as summarized in Algorithm 1.

Budget-constrained relaxation. Each block i has a learnable score s_i, and its j-th singular component receives score $t_{i,j} = s_i - j$. We reserve the first component of every block in both training and deployment. For a total budget B, the remaining budget is $b = B - N$. During training, we use

$$g_{i,1} = 1, \qquad g_{i,j} = \sigma\left((t_{i,j} - \alpha)/\tau\right), \quad j \geq 2,$$

where the shared cutoff α is solved using the Newton's method Kelley (2003) so that $\sum_i \sum_{j=2}^{m_i} g_{i,j} = b$.

We differentiate through this constraint using implicit differentiation. Thus training preserves an exact relaxed gate-mass budget, rather than an exact matrix-rank budget. More training details are delayed to Appendix F.1.

Hard allocation. After training, we retain the first component of every block and the b remaining components with the largest scores $t_{i,j}$, using deterministic tie-breaking. The resulting ranks satisfy

$$1 \leq r_i \leq m_i, \qquad \sum_i r_i = B.$$

Algorithm 1 Learned global rank allocation across KV blocks

Input: budget $N \leq B \leq \sum_i m_i$, calibration set $\mathcal{X}$, temperature $\tau > 0$.
Output: ranks $\{r_i\}_{i=1}^N$ with $\sum_i r_i = B$.
Compute and freeze all transformed SVD components using Theorem 1 and Corollary 1.
Set $b = B - N$; initialize one score s_i per block.
while not converged **do**
  Set $t_{i,j} = s_i - j$ and reserve $g_{i,1} = 1$.
  Solve the shared cutoff α so that $\sum_i \sum_{j=2}^{m_i} \sigma((t_{i,j} - \alpha)/\tau) = b$. using the Newton's method.
  Set $g_{i,j} = \sigma((t_{i,j} - \alpha)/\tau)$ for $j \geq 2$ and gate the fixed transformed singular values.
  Compute $\mathcal{L}_{\mathrm{KD}}$ on $\mathcal{X}$; update $\{s_i\}$ by implicitly differentiating through α.
end while
Recompute the scores; retain each block's first component and the top-b remaining components, breaking ties deterministically.
Set r_i to the number retained in block i.
return $\{r_i\}_{i=1}^N$.

3.3 Integration with Quantization

To further compress the KV cache, BELT can be combined with existing low-rank KV quantizers without modification. We integrate Palu (Chang et al., 2025) and STAR-KV (Bhatnagar et al., 2026). Since the leading coordinates of the compressed representation dominate the quantization range, both quantizers apply an orthogonal transform—a norm-preserving change of basis R with $R^\top R = I$ that mixes coordinates to spread outliers—before quantization. We fuse these transforms into the factors offline, adding no inference cost. Arbitrary learned ranks are handled with orthogonal matrices, and no retraining is required. Full details are in Appendix F.2.

4 Experiments

4.1 Experiment Settings

Models and Tasks. We evaluate on four LLMs: LLaMA-2-7B Touvron et al. (2023), LLaMA-3.1-8B-Instruct Grattafiori et al. (2024), Qwen3-32B Yang et al. (2025), and MPT-7B. They span both MHA (LLaMA-2-7B, MPT-7B) and GQA (LLaMA-3.1-8B-Instruct, Qwen3-32B), as well as RoPE (first three) and NoPE (MPT-7B). For accuracy, we measure perplexity on WikiText-2 Merity

et al. (2016) and C4 Dodge et al. (2021), and zero-shot accuracy on six tasks via LM-Evaluation-Harness Sutawika et al. (2024). For long-context evaluation, we additionally test on LongBench Bai et al. (2024) and RULER Hsieh et al. (2024), covering diverse document-level and structured tasks.

Baseline. We compare BELT against two types of baselines for low-rank KV cache compression: (1) general low-rank decomposition methods for LLM compression, including SVD-LLM Wang et al. (2025), ASVD Yuan et al. (2023), Zero Sum SVD Abbasi et al. (2026), and ERC-SVD Bai et al. (2025); and (2) KV-specific low-rank methods, including PALU Chang et al. (2025) and the state-of-the-art STAR-KV Bhatnagar et al. (2026) and A3 Wong et al. (2025). Other baselines such as S2VD Zhou et al. (2026) are excluded due to code unavailability.

4.2 Accuracy Analysis

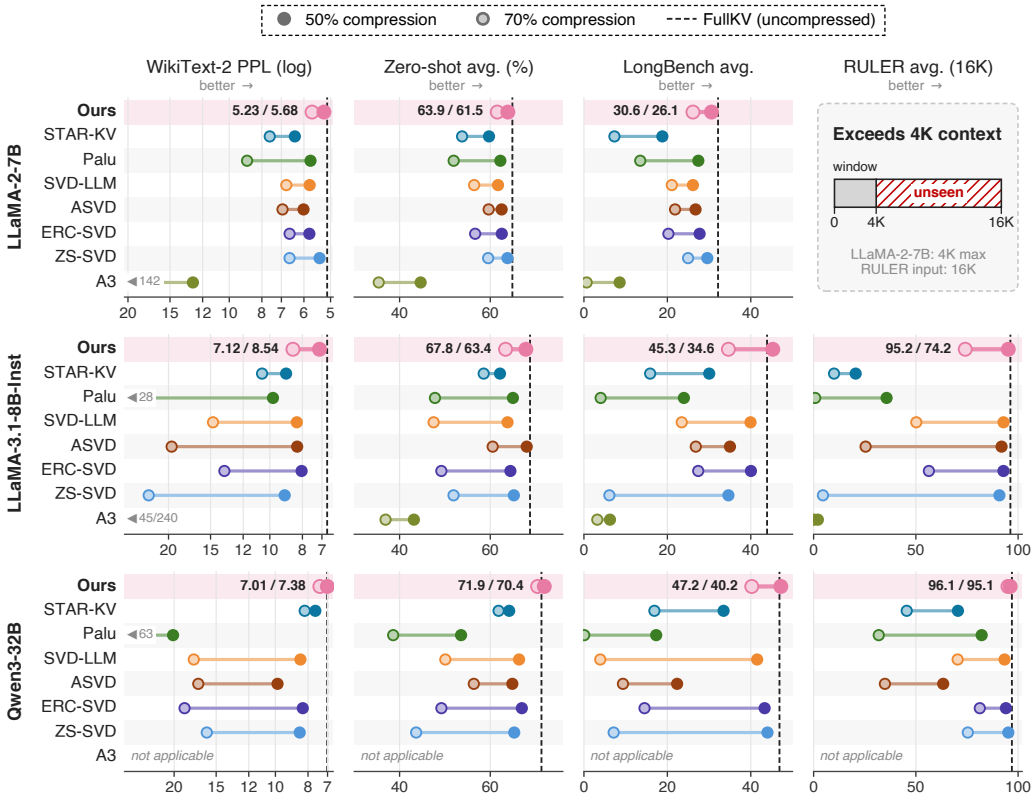


Figure 2: Main results at 50% (dark) and 70% (light) KV cache compression. Each row connects one method's two rates; the dashed line marks the uncompressed model. All axes point right toward better results; perplexity is reversed and logarithmic, with out-of-range values printed at the left end. Numbers beside BELT denote 50%/70% results. RULER uses 16K tokens, exceeding LLaMA-2-7B's 4K context. A3 is not applicable to Qwen3-32B, whose query/key normalization lies between the projections and RoPE. Per-task results are in Appendix H.4.

Perplexity and Zero-Shot Evaluation. Figure 2 shows that BELT maintains model quality under aggressive KV cache compression. On LLaMA-2-7B and LLaMA-3.1-8B-Instruct, BELT consistently achieves the best perplexity and zero-shot accuracy at both 50% and 70% compression, with only marginal degradation from the uncompressed model. **In contrast, baselines degrade sharply at 70% compression, with several collapsing entirely. These results demonstrate that BELT preserves generation quality even at high compression rates.**

Long Context Evaluation. Figure 2 shows that BELT achieves the best average score at both rates, surpassing even the uncompressed LLaMA-3.1-8B at 50% compression. **Baselines compet-**

itive on LM-Eval-Harness collapse here, confirming that short-context metrics underestimate long-context damage. Retrieval-intensive tasks (Qasper, TriviaQA) are the most compression-sensitive and where BELT gains the most, while summarization remains insensitive—consistent with our motivation that query–key matching dominates retrieval.

Comparison with Fine-tuning. BELT constructs compressed KV projections without fine-tuning model weights, whereas the methods in Table 1 additionally update weights on LLaMA-3.1-8B-Instruct. **Fine-tuning constantly improves the results of all methods, but baselines still fall behind BELT.** Direct FT trains low-rank key and value projection matrix from a random orthogonal initialization. Despite using the same objective and training data as BELT+FT, it achieves substantially lower accuracy, which can potentially be attributed to the local minimum issue.

Table 1: Fine-tuning the low-rank projection matrix on LLaMA-3.1-8B-Instruct. Bold marks the best result at each compression rate.

Method	Comp (%)	Perplexity ↓		Zero-Shot Accuracy (%) ↑						
		Wiki2	C4	OBQA	PIQA	ARC-e	ARC-c	Hella	Wino	Avg.
ZS-SVD	50	7.72	14.70	43.20	77.53	77.90	49.66	74.66	**72.61**	65.93
	70	9.34	22.51	35.60	71.98	65.36	38.99	64.90	65.51	57.06
STAR-KV	50	8.31	14.82	40.20	76.66	76.01	47.61	71.65	66.06	63.03
	70	9.73	18.64	37.20	74.16	70.24	41.89	65.71	61.64	58.48
Direct FT	50	8.66	17.56	37.20	75.95	72.43	43.34	68.56	59.67	59.53
	70	10.65	27.99	33.60	71.11	59.30	31.14	58.94	53.99	51.35
Ours+FT	50	**6.96**	**12.63**	**44.20**	**79.87**	**80.01**	**54.95**	**78.24**	72.45	**68.29**
	70	**7.75**	**15.11**	**42.80**	**77.31**	**75.34**	**49.40**	**75.00**	**69.69**	**64.92**

4.3 System-level Performance Analysis

We measure decoding speed on LLaMA-2-7B in FP16 on a single NVIDIA RTX PRO 6000 (96 GB). The baseline is PyTorch SDPA over the full cache, which dispatches to FlashAttention-2 (Dao, 2023). Because BELT at 70% compression matches or exceeds the quality of STAR-KV, Palu, and ERC-SVD at 50% (Table 9), we compare BELT at 70% against these baselines at 50%. More details about these configurations are given in Appendix G.

Attention Latency. Figure 3 reports speedups at roughly matched accuracy, where BELT's better accuracy–speed trade-off yields faster performance at the same recall. BELT consistently outperforms SDPA and all baselines, with its advantage growing at longer contexts, while STAR-Q further boosts the speedup. STAR-KV and Palu remain nearly flat below 1.5×, and BELT at 70% compression already surpasses all baselines at 50%, whereas ERC-SVD is even slower than the uncompressed baseline. **Further runtime breakdown in Appendix H.1 shows that attention computation dominates the overhead and that BELT substantially accelerates it.**

Generation Throughput. Table 3 reports full-model decoding throughput without prefill, using the largest power-of-two batch that fits in memory. By storing the BELT cache in a small fraction of the FP16 size, STAR-Q enables substantially larger batch sizes than the baselines, yielding a large throughput gain over SDPA, whereas STAR-KV improves throughput only modestly. Even without STAR-Q, BELT still achieves considerable speedups.

Table 3: End-to-end generation throughput (tokens/s). The gray text in brackets denotes batch size. Baselines use 50% KV cache compression and BELT uses 70% with STAR-KV.

Context Len.	PyTorch SDPA	STAR-KV	Palu	ERC-SVD	BELT	Gain (BELT)
1K	2091 (128)	3092 (256)	2485 (256)	855 (256)	4859 (1024)	2.32×
2K	1044 (64)	1681 (128)	1340 (128)	456 (128)	2935 (512)	2.81×
4K	530 (32)	867 (64)	688 (64)	234 (64)	1642 (256)	3.10×
8K	271 (16)	437 (32)	343 (32)	117 (32)	869 (128)	3.21×
16K	137 (8)	211 (16)	174 (16)	58 (16)	447 (64)	3.26×

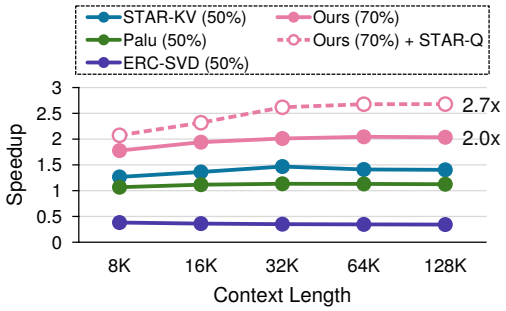


Figure 3: Speedup of the attention module over PyTorch SDPA for one decoding step of LLaMA-2-7B at batch size 16, measured over all 32 layers; layers left uncompressed run SDPA. Baselines use 50% compression and Ours uses 70%; the dashed line adds STAR-Q.

Table 2: Quantization results on Qwen3-32B: average KV bit-width, KV cache compression ratio, and average zero-shot accuracy over the six commonsense tasks.

Method	**Bit**	**Comp.**	**Avg. (%)**
Qwen3-32B	16	1×	71.29
KIVI-2-gs32-r0	2	8.0×	67.90
Palu (50%)	4	8.0×	52.02
Palu (50%)	3	10.7×	48.43
STAR-KV (50%)	3.3	9.7×	63.30
Ours (70%)	16	3.3×	70.45
Ours (70%) + Palu-Q	4	13.3×	70.12
Ours (70%) + Palu-Q	3	17.8×	69.36
Ours (70%) + STAR-Q	3.2	16.7×	70.28

4.4 Quantization.

Table 2 reports quantization results on Qwen3-32B. Applying Palu-Q and STAR-Q to BELT at 70% compression yields 13.3× and 16.7× compression at 70.12% and 70.28% accuracy, nearly matching the unquantized model, with 3-bit Palu-Q further raising compression to 17.8× at 69.36%. In contrast, Palu and STAR-KV reach only 48.43% and 63.30% at lower compression (10.7× and 9.7×), and KIVI-2 attains 67.90% at 8×. The accuracy advantage of BELT under low-rank compression thus carries over to the quantized cache.

Other experiments, e.g., ablation studies and verifying the assumption of our theoretical analysis, are included in Appendix H.

5 Related Work

Low-rank decomposition for general LLM compression. A broad line of work decomposes projection weights to reduce model memory and computation, including SVD-LLM (Wang et al., 2025), ASVD (Yuan et al., 2023), Zero Sum SVD (Abbasi et al., 2026), ERC-SVD (Bai et al., 2025), and S2VD (Zhou et al., 2026). These methods motivate data-weighted factorizations, but their layer- or weight-reconstruction objectives do not directly reflect attention behavior under KV caching. BELTinstead derives its weighting from attention outputs and attention-distribution curvature.

General KV-cache compression. KV-cache compression has also been studied along the token and precision axes: token-selection methods retain a subset of cached tokens (Zhang et al., 2023; Liu et al., 2023; Ge et al., 2024; Li et al., 2024; Cai et al., 2024; Oren et al., 2024), while quantization methods reduce cache precision (Zhao et al., 2024; Liu et al., 2024c; Lin et al., 2025; Su et al., 2025; Hooper et al., 2024; He et al., 2024). These are complementary to low-rank compression, which reduces the hidden dimension rather than the sequence dimension or bit-width. Our joint allocator further treats rank and bit-width as choices under a shared memory budget.

6 Conclusion

We revisited low-rank KV cache compression and addressed two limitations of existing methods: objectives that minimize KV reconstruction error rather than preserve attention outputs, and rank allocation that collapses to uniform thresholds. We reformulated compression to align with generation outputs, derived a closed-form solution via an approximate objective, and proposed a global rank allocation mechanism with a hard budget that preserves the truly important components across heads. Extensive experiments show consistent accuracy and efficiency gains under aggressive compression, with up to 3.26× faster decoding and 87% lower memory at comparable accuracy.

7 AI USE STATEMENT

In this work, we used generative AI tools for "assist in the writing of proofs" and "design or provide feedback on research methodology or experiments". We have not used generative AI tools for "Generate synthetic data sets", "help develop theoretical models or conceptual frameworks", "formulate mathematical claims", "provide critical ingredients for proving mathematical claims", "propose or refine hypotheses", "implement methods", "assist with translation", "clean and reformat dataset", "support qualitative and thematic data analysis", "interpret results". We have reviewed all AI-assisted work, and we checked LLM-generated research ideas for potential plagiarism through a manual literature survey. In addition, LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ali Abbasi, Chayne Thrash, Haoran Qin, Shansita Sharma, Sepehr Seifi, and Soheil Kolouri. Zero sum svd: Balancing loss sensitivity for low rank llm compression. *arXiv preprint arXiv:2602.02848*, 2026.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In *Advances in Neural Information Processing Systems*, 2024.

Haolei Bai, Siyong Jian, Tuo Liang, Yu Yin, and Huan Wang. Erc-svd: Error-controlled svd for large language model compression. *arXiv preprint arXiv:2505.20112*, 2025.

Haolei Bai, Siyong Jian, Tuo Liang, Yu Yin, and Huan Wang. Erc-svd: Error-controlled svd for large language model compression, 2026. URL https://arxiv.org/abs/2505.20112.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Priyansh Bhatnagar, Ashkan Moradifirouzabadi, Se-Hyun Yang, SeungJae Lee, Jungwook Choi, and Mingu Kang. Star-kv: Low-rank kv cache compression via soft thresholding for adaptive rank control, 2026. URL https://arxiv.org/abs/2606.08382.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: KV-cache compression with low-rank projection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LWMS4pk2vK.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Hung-Yueh Chiang, Yash Akhauri, Xilai Dai, Huiqiang Jiang, Yucheng Li, Luis Ceze, Kai-Chiang Wu, and Mohamed S. Abdelfattah. xkv: Cross-layer kv-cache compression via aligned singular vector extraction, 2026. URL https://arxiv.org/abs/2503.18893.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and

Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98/.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms, 2024. URL https://arxiv.org/abs/2310.01801.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu

Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, and Bohan Zhuang. Zipcache: Accurate and efficient kv cache quantization with salient token identification. *Advances in Neural Information Processing Systems*, 37:68287–68307, 2024.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL https://arxiv.org/abs/2404.06654.

Shiyu Ji, Yixuan Wang, Yijun Liu, Qingfu Zhu, and Wanxiang Che. Echokv: Efficient kv cache compression via similarity-based reconstruction, 2026. URL https://arxiv.org/abs/2603.22910.

Jiantong Jiang, Peiyu Yang, Rui Zhang, and Feng Liu. Towards efficient large language model serving: A survey on system-aware kv cache optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38450–38476, 2026.

Carl T Kelley. *Solving nonlinear equations with Newton's method*. SIAM, 2003.

Jaeseong Lee, Seung-won Hwang, Aurick Qiao, Daniel Campos, Zhewei Yao, and Yuxiong He. Tale: Token-adaptive low-rank kvcache approximation with reconstruction elimination. *Transactions of the Association for Computational Linguistics*, 13:1298–1318, 2025.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Yujun Lin, Haotian Tang, Shang Yang, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, and Song Han. Qserve: W4a8kv4 quantization and system co-design for efficient llm serving, 2025. URL https://arxiv.org/abs/2405.04532.

Aixin Liu, Hao Bei, Haowei Chen, et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Di Liu, Meng Chen, Baotong Lu, Huiqiang Jiang, Zhenhua Han, Qianxi Zhang, Qi Chen, Chengruidong Zhang, Bailu Ding, Kai Zhang, Chen Chen, Fan Yang, Yuqing Yang, and Lili Qiu. Retrievalattention: Accelerating long-context llm inference via vector retrieval, 2024b. URL https://arxiv.org/abs/2409.10516.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024c.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

Junlin Mu, Hantao Huang, Jihang Zhang, Minghui Yu, Tao Wang, and Yidong Li. Sals: Sparse attention in latent space for kv cache compression, 2025. URL https://arxiv.org/abs/2510.24273.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state rnns. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18724–18741, 2024.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL https://arxiv.org/abs/2108.12409.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. ISBN 9780070856134. URL https://books.google.com/books?id=kwqzPAAACAAJ.

Matt Stallone, Vaibhav Saxena, Leonid Karlinsky, Bridget McGinn, Tim Bula, Mayank Mishra, Adriana Meza Soria, Gaoyuan Zhang, Aditya Prasad, Yikang Shen, et al. Scaling granite code models to 128k context. *arXiv preprint arXiv:2407.13739*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zunhai Su, Zhe Chen, Wang Shen, Hanyu Wei, Linge Li, Huangqi Yu, and Kehong Yuan. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations, 2025. URL https://arxiv.org/abs/2501.16383.

Lintang Sutawika, Hailey Schoelkopf, Leo Gao, Baber Abbasi, Stella Biderman, Jonathan Tow, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Aflah, Niklas Muennighoff, Thomas Wang, sdtblck, nopperl, gakada, tttyuntian, et al. EleutherAI/lm-evaluation-harness: v0.4.3. Zenodo, version v0.4.3, July 2024. URL https://doi.org/10.5281/zenodo.12608602.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression, 2025. URL https://arxiv.org/abs/2403.07378.

James Hardy Wilkinson. Kronecker's canonical form and the qz algorithm. *Linear Algebra and its Applications*, 28:285–303, 1979.

Jeffrey TH Wong, Cheng Zhang, Xinye Cao, Pedro Gimenes, Christos-Savvas Bouganis, George A Constantinides, Wayne Luk, and Yiren Zhao. A3: an analytical low-rank approximation framework for attention. *arXiv preprint arXiv:2505.12942*, 2025.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities, 2025. URL https://arxiv.org/abs/2407.14482.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

Rongzhi Zhang, Kuang Wang, Liyuan Liu, Shuohang Wang, Hao Cheng, Chao Zhang, and Yelong Shen. Lorc: Low-rank compression for llms kv cache with a progressive compression strategy. *arXiv preprint arXiv:2410.03111*, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. Atom: Low-bit quantization for efficient and accurate llm serving, 2024. URL https://arxiv.org/abs/2310.19102.

Hao Zhou, Jiapeng Guan, Jie Zhang, Ran Wei, Xudong Zhao, Zhe Jiang, and Xiangyang Ji. S 2 vd: A subspace-aware svd method for efficient llm compression. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 20506–20510. IEEE, 2026.

A DERIVATION OF $\mathcal{L}_V(\widehat{W}_V)$ AND $\mathcal{L}_K(\widehat{W}_K)$

We derive separate value and key criteria from the single-head attention-output error using the notation of Section D.1. For each calibration query x, X denotes its corresponding context matrix; we suppress this dependence to match the main-text notation. Calibration representations and queries are held fixed, and all attention distributions are restricted to their common finite, unmasked support. All objectives below are summed over calibration queries, consistently with Equation 3.1.1.

For brevity, write

$$a_x = a_x(W_K), \qquad \widehat{a}_x = a_x(\widehat{W}_K).$$

Using Equation 1, adding and subtracting $a_x X \widehat{W}_V W_O$ gives the exact decomposition

$$\begin{aligned} &o_x(W_K, W_V) - o_x(\widehat{W}_K, \widehat{W}_V) \\ &\quad = \underbrace{a_x X (W_V - \widehat{W}_V) W_O}_{\text{value perturbation}} + \underbrace{(a_x - \widehat{a}_x) X \widehat{W}_V W_O}_{\text{key perturbation}}. \end{aligned} \tag{5}$$

Applying Young's inequality to these two terms and summing over queries yields, for every $\xi > 0$,

$$\mathcal{J}(\widehat{W}_K, \widehat{W}_V) \le (1+\xi)\mathcal{L}_V(\widehat{W}_V) + (1+\xi^{-1})\mathcal{L}_{K,\mathrm{out}}(\widehat{W}_K; \widehat{W}_V), \tag{6}$$

where

$$\mathcal{L}_V(\widehat{W}_V) = \sum_x \left\| a_x X (W_V - \widehat{W}_V) W_O \right\|_2^2, \tag{7}$$

$$\mathcal{L}_{K,\mathrm{out}}(\widehat{W}_K; \widehat{W}_V) = \sum_x \left\| (a_x - \widehat{a}_x) X \widehat{W}_V W_O \right\|_2^2. \tag{8}$$

Thus $\mathcal{L}_V$ is exactly the output error caused by compressing the value projection while keeping the original attention distributions fixed. The key-output term remains coupled to $\widehat{W}_V$.

To obtain a value-independent key criterion, define

$$\mathcal{L}_K(\widehat{W}_K) = \sum_x D_{\mathrm{KL}}\left(a_x(W_K) \,\middle\|\, a_x(\widehat{W}_K) \right). \tag{9}$$

For each query, let

$$M_x(\widehat{W}_V) = \max_{1 \le i \le n} \|(X \widehat{W}_V W_O)_{i,:}\|_2.$$

The triangle inequality gives

$$\left\| (a_x - \widehat{a}_x) X \widehat{W}_V W_O \right\|_2 \le M_x(\widehat{W}_V) \|a_x - \widehat{a}_x\|_1.$$

Pinsker's inequality, with natural logarithms, then implies

$$\left\| (a_x - \widehat{a}_x) X \widehat{W}_V W_O \right\|_2^2 \le 2 M_x(\widehat{W}_V)^2 D_{\mathrm{KL}}(a_x \,\|\, \widehat{a}_x).$$

Summing over queries yields the candidate-dependent bound

$$\mathcal{L}_{K,\mathrm{out}}(\widehat{W}_K; \widehat{W}_V) \le 2 \sum_x M_x(\widehat{W}_V)^2 D_{\mathrm{KL}}(a_x \,\|\, \widehat{a}_x). \tag{10}$$

Suppose a common finite constant M satisfies

$$M_x(\widehat{W}_V) \le M$$

for every calibration query and every value approximation in the considered candidate class. Then

$$\mathcal{L}_{K,\mathrm{out}}(\widehat{W}_K; \widehat{W}_V) \le 2M^2 \mathcal{L}_K(\widehat{W}_K). \tag{11}$$

Taking $\xi = 1$ in Equation 6 gives

$$\mathcal{J}(\widehat{W}_K, \widehat{W}_V) \le 2\mathcal{L}_V(\widehat{W}_V) + 4M^2 \mathcal{L}_K(\widehat{W}_K). \tag{12}$$

Scope of the bound. Equation 12 requires the same finite M throughout the considered candidate class; a rank constraint alone does not ensure this condition. Without a uniform norm bound, Equation 10 still holds, but its coefficients depend on $\widehat{W}_V$. In that setting, $\mathcal{L}_K$ is an upper-bound-motivated, distribution-preserving surrogate, rather than a value-independent upper bound on the unrestricted joint problem.

The value objective has an exact quadratic representation, and its positive-definite regularization admits the weighted-SVD solution in Theorem 1. For the nonlinear key objective, we derive its local quadratic term, apply pre-fitting regularization and a Kronecker approximation, and solve the resulting regularized surrogate in closed form. These solutions exactly minimize their respective surrogates, not necessarily the original joint objective. The analysis concerns attention-output errors on fixed calibration inputs, rather than a guarantee of final autoregressive generation preservation.

B THEORETICAL ANALYSIS WITHOUT ROPE

B.1 THEORETICAL ANALYSIS FOR $\mathcal{L}_V^{\approx}(\widehat{W}_V)$

B.1.1 PROOF OF LEMMA 1

Proof. Let

$$E = W_V - \widehat{W}_V, \qquad u_x = a_x X_x.$$

Then

$$G_V^0 = \sum_x u_x^\top u_x, \qquad H_V^0 = W_O W_O^\top$$

are positive semidefinite. Adding ϵI with $\epsilon > 0$ makes both matrices positive definite, so their Cholesky factors L_V and C_V exist and are invertible.

Aggregated quadratic form. Using the trace identity for a squared Euclidean norm and cyclicity of the trace,

$$\begin{aligned}
\mathcal{L}_V(\widehat{W}_V) &= \sum_x \|u_x E W_O\|_2^2 \\
&= \sum_x \operatorname{tr}\left(u_x E H_V^0 E^\top u_x^\top\right) \\
&= \sum_x \operatorname{tr}\left(u_x^\top u_x E H_V^0 E^\top\right) \\
&= \operatorname{tr}\left(G_V^0 E H_V^0 E^\top\right).
\end{aligned}$$

Regularized weighted Frobenius form. By the Cholesky factorizations,

$$\begin{aligned}
\mathcal{L}_V^{\approx}(\widehat{W}_V) &= \|L_V E C_V\|_F^2 \\
&= \operatorname{tr}\left(L_V E C_V C_V^\top E^\top L_V^\top\right) \\
&= \operatorname{tr}\left(G_V E H_V E^\top\right) \\
&= \operatorname{tr}\left((G_V^0 + \epsilon I_d) E (H_V^0 + \epsilon I_{d_h}) E^\top\right).
\end{aligned}$$

Expanding yields

$$\begin{aligned}
\mathcal{L}_V^{\approx}(\widehat{W}_V) = \operatorname{tr}(G_V^0 E H_V^0 E^\top) + \epsilon \operatorname{tr}(G_V^0 E E^\top) \\
+ \epsilon \operatorname{tr}(E H_V^0 E^\top) + \epsilon^2 \|E\|_F^2.
\end{aligned}$$

The first term is $\mathcal{L}_V(\widehat{W}_V)$. Furthermore,

$$\operatorname{tr}(G_V^0 E E^\top) = \sum_x \|u_x E\|_2^2 = \sum_x \|a_x X_x E\|_2^2,$$

and

$$\operatorname{tr}(E H_V^0 E^\top) = \|E W_O\|_F^2.$$

Therefore,

$$\mathcal{L}_V^{\approx}(\widehat{W}_V) = \mathcal{L}_V(\widehat{W}_V) + \Delta_{\epsilon,V}(E),$$

with $\Delta_{\epsilon,V}(E) = \epsilon \sum_x \|a_x X_x E\|_2^2 + \epsilon \|E W_O\|_F^2 + \epsilon^2 \|E\|_F^2$. Every term in this correction is non-negative.

Finally,

$$0 \le \Delta_{\epsilon,V}(E) \le \left[\epsilon \left(\|G_V^0\|_{\mathrm{op}} + \|H_V^0\|_{\mathrm{op}}\right) + \epsilon^2\right] \|E\|_F^2. \tag{13}$$

For fixed E and calibration statistics, this is $O(\epsilon)$ as $\epsilon \to 0$, completing the proof. □

B.1.2 Proof of Theorem 1

Proof. By Lemma 1, the rank-constrained value problem is

$$\min_{\mathrm{rank}(\widehat{W}_V) \le r} \left\| L_V W_V C_V - L_V \widehat{W}_V C_V \right\|_F^2.$$

Define

$$M_V := L_V W_V C_V, \qquad Y := L_V \widehat{W}_V C_V.$$

Since L_V and C_V are invertible, $\mathrm{rank}(Y) = \mathrm{rank}(\widehat{W}_V)$, and the transformation is bijective. The problem is therefore equivalent to

$$\min_{\mathrm{rank}(Y) \le r} \|M_V - Y\|_F^2.$$

By the Eckart–Young–Mirsky theorem Eckart & Young (1936); Mirsky (1960), one global minimizer is

$$Y_r^* = \bar{U}_{V,r} \bar{\Sigma}_{V,r} \bar{R}_{V,r}^\top,$$

and the minimum value is $\sum_{j=r+1}^m \bar{\sigma}_{V,j}^2$. Applying the inverse transformation gives

$$\widehat{W}_{V,r}^* = L_V^{-1} \bar{U}_{V,r} \bar{\Sigma}_{V,r} \bar{R}_{V,r}^\top C_V^{-1}.$$

□

B.2 Theoretical analysis for $\mathcal{L}_K^{\approx}(\widehat{W}_K)$

B.2.1 Proof of Theorem 2

Before proving Theorem 2, we prove the following lemma for the objective $\mathcal{L}_K(W_K)$ evaluated on individual input token, i.e., $\mathcal{L}_{K,x}(\widehat{W}_K) = D_{\mathrm{KL}}\Big(a_x(W_K) \,\Big\|\, a_x(\widehat{W}_K)\Big)$ such that $\mathcal{L}_K(\widehat{W}_K) = \sum_x \mathcal{L}_{K,x}(\widehat{W}_K)$.

Lemma 2 (Per-query key-side quadratic approximation)**.** *For a query x, let $J_x, A_x, G_{K,x}^\epsilon$ be defined in equation 4, and write*

$$G_{K,x} = G_{K,x}^\epsilon - \epsilon I_d, \qquad H_{K,x} = q_x q_x^\top, \qquad H_{K,x}^\epsilon = H_{K,x} + \epsilon I_{d_h}.$$

Let

$$G_{K,x}^\epsilon = L_{K,x}^\top L_{K,x}, \qquad H_{K,x}^\epsilon = C_{K,x} C_{K,x}^\top.$$

be the Choleskey decomposition of $G_{K,x}^\epsilon$ and $H_{K,x}^\epsilon$. For $E = W_K - \widehat{W}_K$, define

$$\mathcal{L}_{K,x}^{\approx}(\widehat{W}_K) = \|L_{K,x} E C_{K,x}\|_F^2.$$

Then

$$\mathcal{L}_{K,x}(\widehat{W}_K) - \mathcal{L}_{K,x}^{\approx}(\widehat{W}_K) = \mathcal{R}_x(E) - \Delta_{\epsilon,x}(E), \tag{14}$$

where

$$|\mathcal{R}_x(E)| \le \frac{1}{6 d_h^{3/2}} \omega(X_x E q_x)^3, \qquad \omega(v) := \max_n v_n - \min_n v_n,$$

and

$$\Delta_{\epsilon,x}(E) = \epsilon \,\mathrm{tr}(G_{K,x} E E^\top) + \epsilon \|E q_x\|_2^2 + \epsilon^2 \|E\|_F^2 \ge 0.$$

Consequently,

$$\left|\mathcal{L}_{K,x}(\widehat{W}_K) - \mathcal{L}^{\approx}_{K,x}(\widehat{W}_K)\right| \le \frac{1}{6d_h^{3/2}}\omega(X_xEq_x)^3 + \Delta_{\epsilon,x}(E). \tag{15}$$

For the one-sided regularization used in Kronecker fitting,

$$\mathcal{L}_{K,x}(\widehat{W}_K) = q_x^\top E^\top G^\epsilon_{K,x} E q_x + \mathcal{R}_x(E) - \epsilon\|Eq_x\|_2^2. \tag{16}$$

Proof. All attention quantities are restricted to the finite, unmasked support. Let

$$h_x = z_x(\widehat{W}_K)^\top - z_x(W_K)^\top = -\frac{1}{\sqrt{d_h}} X_x E q_x$$

be the logit-perturbation column vector, and define

$$F_x(t) = \log\left(\sum_n a_{x,n} e^{t h_{x,n}}\right), \qquad t \in [0,1].$$

The exact softmax KL identity gives

$$\mathcal{L}_{K,x}(\widehat{W}_K) = F_x(1) - F_x(0) - F_x'(0),$$

where $F_x(0) = 0$ and $F_x'(0) = a_x h_x$.

Define the tilted probabilities and their mean by

$$a^{(t)}_{x,n} = \frac{a_{x,n} e^{t h_{x,n}}}{\sum_j a_{x,j} e^{t h_{x,j}}}, \qquad \mu_x(t) = \sum_n a^{(t)}_{x,n} h_{x,n}.$$

Differentiation yields

$$F_x''(0) = h_x^\top J_x h_x, \qquad F_x'''(t) = \sum_n a^{(t)}_{x,n}\big(h_{x,n} - \mu_x(t)\big)^3.$$

Taylor's theorem with integral remainder therefore gives

$$\mathcal{L}_{K,x}(\widehat{W}_K) = \frac{1}{2} h_x^\top J_x h_x + \mathcal{R}_x(E), \qquad \mathcal{R}_x(E) = \frac{1}{2}\int_0^1 (1-t)^2 F_x'''(t)\, dt. \tag{17}$$

Since $\mu_x(t)$ lies between the minimum and maximum entries of h_x,

$$|F_x'''(t)| \le \sum_n a^{(t)}_{x,n} |h_{x,n} - \mu_x(t)|^3 \le \omega(h_x)^3.$$

Hence

$$|\mathcal{R}_x(E)| \le \frac{1}{6}\omega(h_x)^3 = \frac{1}{6d_h^{3/2}}\omega(X_xEq_x)^3.$$

To connect this expansion to the previously defined metric, write $D_x = \mathrm{diag}(a_x)$. Because $D_x^{-1} a_x^\top = \mathbf{1}$ and $J_x \mathbf{1} = 0$,

$$A_x = J_x D_x^{-1} J_x = J_x D_x^{-1}(D_x - a_x^\top a_x) = J_x.$$

Substituting h_x into equation 17 therefore yields

$$\begin{aligned}\mathcal{L}_{K,x}(\widehat{W}_K) &= q_x^\top E^\top G_{K,x} E q_x + \mathcal{R}_x(E) \\ &= \mathrm{tr}(G_{K,x} E H_{K,x} E^\top) + \mathcal{R}_x(E).\end{aligned} \tag{18}$$

By the Cholesky factorizations and cyclicity of the trace,

$$\begin{aligned}\mathcal{L}^{\approx}_{K,x}(\widehat{W}_K) &= \mathrm{tr}\left(G^\epsilon_{K,x} E H^\epsilon_{K,x} E^\top\right) \\ &= \mathrm{tr}(G_{K,x} E H_{K,x} E^\top) + \epsilon\,\mathrm{tr}(G_{K,x} E E^\top) \\ &\quad + \epsilon\|Eq_x\|_2^2 + \epsilon^2\|E\|_F^2.\end{aligned}$$

The additional terms equal $\Delta_{\epsilon,x}(E)$ and are nonnegative. Subtracting this expression from equation 18 proves equation 14; the triangle inequality gives equation 15.

Finally,

$$q_x^\top E^\top G^\epsilon_{K,x} E q_x = q_x^\top E^\top G_{K,x} E q_x + \epsilon\|Eq_x\|_2^2,$$

which, together with equation 18, proves equation 16. □

Lemma 3 (Kronecker–vectorization identity). *Let $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$ be symmetric matrices, and let $E \in \mathbb{R}^{n\times m}$. Then*

$$\boxed{\operatorname{vec}(E)^\top (A \otimes B)\operatorname{vec}(E) = \operatorname{tr}\big(BEAE^\top\big).}$$

In particular, for any vector $q \in \mathbb{R}^m$ and any symmetric matrix $G \in \mathbb{R}^{n\times n}$,

$$\boxed{q^\top E^\top GEq = \operatorname{vec}(E)^\top (qq^\top \otimes G)\operatorname{vec}(E).}$$

Proof. We prove the general identity first.

Step 1: Vectorization identity. For matrices P, X, Q of compatible dimensions,

$$\operatorname{vec}(PXQ) = (Q^\top \otimes P)\operatorname{vec}(X).$$

Applying this with $P = B$, $X = E$, $Q = A$, and using $A = A^\top$ (since A is symmetric),

$$(A \otimes B)\operatorname{vec}(E) = \operatorname{vec}(BEA).$$

Step 2: Trace inner product. For any matrices P, Q of the same size,

$$\operatorname{vec}(P)^\top \operatorname{vec}(Q) = \operatorname{tr}(P^\top Q).$$

Applying this with $P = E$ and $Q = BEA$,

$$\operatorname{vec}(E)^\top \operatorname{vec}(BEA) = \operatorname{tr}\big(E^\top BEA\big).$$

Combining with Step 1,

$$\operatorname{vec}(E)^\top (A \otimes B)\operatorname{vec}(E) = \operatorname{tr}\big(E^\top BEA\big).$$

Step 3: Cyclicity of the trace. Using cyclicity of the trace,

$$\operatorname{tr}\big(E^\top BEA\big) = \operatorname{tr}\big(BEAE^\top\big).$$

Therefore

$$\operatorname{vec}(E)^\top (A \otimes B)\operatorname{vec}(E) = \operatorname{tr}\big(BEAE^\top\big).$$

Corollary: per-token identity. Set $A = qq^\top$ and $B = G$. Since $qq^\top$ and G are symmetric, the general identity gives

$$\operatorname{vec}(E)^\top (qq^\top \otimes G)\operatorname{vec}(E) = \operatorname{tr}\big(GEqq^\top E^\top\big).$$

On the other hand, $q^\top E^\top GEq$ is a scalar, so it equals its own trace:

$$q^\top E^\top GEq = \operatorname{tr}\big(q^\top E^\top GEq\big) = \operatorname{tr}\big(E^\top GEqq^\top\big) = \operatorname{tr}\big(GEqq^\top E^\top\big),$$

where the last equality follows from cyclicity of the trace. Therefore

$$q^\top E^\top GEq = \operatorname{vec}(E)^\top (qq^\top \otimes G)\operatorname{vec}(E).$$

□

Then we formally provide the proof of Theorem 2 below:

Proof. Let $E = W_K - \widehat{W}_K$ and $e = \operatorname{vec}(E)$. By equation 16,

$$\mathcal{L}_{K,x}(\widehat{W}_K) = q_x^\top E^\top G^\epsilon_{K,x} Eq_x + \mathcal{R}_x(E) - \epsilon\|Eq_x\|_2^2.$$

Applying Lemma 3 and summing over calibration queries gives

$$\mathcal{L}_K(\widehat{W}_K) = e^\top M_K e + \sum_x \mathcal{R}_x(E) - \Delta_{\epsilon,\mathrm{pre}}(E), \tag{19}$$

where

$$M_K = \sum_x (q_x q_x^\top) \otimes G^{\epsilon}_{K,x}, \qquad \Delta_{\epsilon,\mathrm{pre}}(E) = \epsilon \sum_x \|E q_x\|_2^2.$$

Lemma 2 also gives

$$\left| \sum_x \mathcal{R}_x(E) \right| \le \frac{1}{6 d_h^{3/2}} \sum_x \omega(X_x E q_x)^3.$$

Let A_K, B_K be the fixed positive-semidefinite Kronecker factors specified in the theorem. Their post-fitting regularizations are positive definite, so L_K and C_K exist and are invertible. Using their Cholesky factorizations and Lemma 3,

$$\begin{aligned} \mathcal{L}_K^{\approx}(\widehat{W}_K) &= \|L_K E C_K\|_F^2 \\ &= \mathrm{tr}\left((B_K + \epsilon I_d) E (A_K + \epsilon I_{d_h}) E^\top\right) \\ &= e^\top (A_K \otimes B_K) e + \Delta_{\epsilon,\mathrm{post}}(E), \end{aligned}$$

where

$$\Delta_{\epsilon,\mathrm{post}}(E) = \epsilon\, \mathrm{tr}(B_K E E^\top) + \epsilon\, \mathrm{tr}(E A_K E^\top) + \epsilon^2 \|E\|_F^2.$$

Both regularization contributions are nonnegative. Writing

$$\Delta_{\epsilon,K}(E) = \Delta_{\epsilon,\mathrm{pre}}(E) + \Delta_{\epsilon,\mathrm{post}}(E),$$

and subtracting the surrogate from equation 19 yields the exact signed decomposition

$$\begin{aligned} \mathcal{L}_K(\widehat{W}_K) - \mathcal{L}_K^{\approx}(\widehat{W}_K) = e^\top (M_K - A_K \otimes B_K) e \\ + \sum_x \mathcal{R}_x(E) - \Delta_{\epsilon,K}(E). \end{aligned} \tag{20}$$

Taking absolute values gives

$$\begin{aligned} \left| \mathcal{L}_K(\widehat{W}_K) - \mathcal{L}_K^{\approx}(\widehat{W}_K) \right| \le \left| e^\top (M_K - A_K \otimes B_K) e \right| \\ + \frac{1}{6 d_h^{3/2}} \sum_x \omega(X_x E q_x)^3 + \Delta_{\epsilon,K}(E). \end{aligned} \tag{21}$$

The three terms in equation 21 correspond to distinct approximation sources. The first, $\left|e^\top (M_K - A_K \otimes B_K) e\right|$, is the absolute Kronecker approximation error evaluated along the residual E. The second is an explicit upper bound on the total Taylor expansion error $|\sum_x \mathcal{R}_x(E)|$. The third, $\Delta_{\epsilon,K}(E)$, accounts for both the pre-fitting ridge and the post-fitting factor regularization. In particular,

$$\Delta_{\epsilon,K}(E) \le \epsilon \sum_x \|E q_x\|_2^2 + \left[\epsilon\left(\|A_K\|_{\mathrm{op}} + \|B_K\|_{\mathrm{op}}\right) + \epsilon^2\right] \|E\|_F^2.$$

For fixed E and calibration data, this term is $O(\epsilon)$ as $\epsilon \to 0$, provided the fitted factors remain bounded in operator norm. Without this boundedness condition, the explicit regularization term should be retained rather than abbreviated as $O(\epsilon)$.

□

B.2.2 COROLLARY FOR DECOMPOSING KEY TENSOR

Corollary 1. *let $L_K W_K C_K = \bar{U}_K \bar{\Sigma}_K \bar{R}_K^\top$ be a thin singular value decomposition, where $\bar{\Sigma}_K = \mathrm{diag}(\bar{\sigma}_{K,1}, \ldots, \bar{\sigma}_{K,m}), m = \min(d, d_h)$, and $\bar{\sigma}_{K,1} \ge \cdots \ge \bar{\sigma}_{K,m} \ge 0$. Let $\bar{U}_{K,1:r}$, $\bar{\Sigma}_{K,1:r}$, and $\bar{R}_{K,1:r}$ denote the components associated with the largest r singular values. Then one global minimizer of $\min_{\mathrm{rank}(\widehat{W}_K) \le r} \mathcal{L}_K^{\approx}(\widehat{W}_K)$ is*

$$\widehat{W}^*_{K,1:r} = L_K^{-1} \bar{U}_{K,1:r} \bar{\Sigma}_{K,1:r} \bar{R}_{K,1:r}^\top C_K^{-1}. \tag{22}$$

The corresponding projection matrix, $W^{\mathrm{down}}_{K,1:r} = L_V^{-1} \bar{U}_{K,1:r} \bar{\Sigma}^{1/2}_{K,1:r}$, and $W^{\mathrm{up}}_{K,1:r} = \bar{\Sigma}^{1/2}_{K,1:r} \bar{R}^\top_{K,1:r} C_K^{-1}$.

Algorithm 2 Rank-1 Kronecker-factored low-rank approximation of W_K

Input: KV cache X, query vectors $\{q_x\}$, current key matrix W_K, rank r, regularization ϵ.
for each query token x **do**
  Compute attention distribution: $p_x \leftarrow \mathrm{Softmax}\big(\frac{1}{\sqrt{d_h}} X W_K q_x\big)$.
  Compute softmax Jacobian: $J_x \leftarrow \mathrm{diag}(p_x) - p_x p_x^\top$.
  Compute Fisher weight: $A_x \leftarrow J_x \,\mathrm{diag}(1/p_x) J_x$.
  Compute per-token left weight: $G^\epsilon_{K,x} \leftarrow \frac{1}{2d_h} X^\top A_x X + \epsilon I_d$.
end for
Form exact Kronecker quadratic form: $M_K \leftarrow \sum_x q_x q_x^\top \otimes G^\epsilon_{K,x}$.
Compute rank-1 Kronecker approximation $M_K \approx A_K \otimes B_K$ via alternating least squares.
Regularize: $B^\epsilon_K \leftarrow B_K + \epsilon I_d$, $A^\epsilon_K \leftarrow A_K + \epsilon I_{d_h}$.
Cholesky: $B^\epsilon_K = L_K^\top L_K$, $A^\epsilon_K = C_K C_K^\top$.
Compute SVD $L_K W_K C_K = U \Sigma V^\top$.
Return $\widehat{W}^\star_K \leftarrow L_K^{-1} U_r \Sigma_r V_r^\top C_K^{-1}$.

B.2.3 Overall algorithm for computing $\mathcal{L}^{\approx}_K(W_K)$

Algorithm 2 summarizes the rank-1 Kronecker-factored low-rank approximation of the key projection matrix W_K. Given the KV cache X, the query vectors $\{q_x\}$, and the current key matrix W_K, the algorithm aims to find a rank-r approximation $\widehat{W}_K$ that minimizes a quadratic surrogate of the token-summed KL divergence between the original and approximated attention distributions. The algorithm proceeds in three main stages. First, for each query token x, it computes the attention distribution p_x, the softmax Jacobian J_x, the Fisher weight A_x, and the regularized per-token left weight $G^\epsilon_{K,x}$. Second, it assembles the exact Kronecker quadratic form $M_K = \sum_x q_x q_x^\top \otimes G^\epsilon_{K,x}$ and approximates it by a rank-1 Kronecker product $A_K \otimes B_K$ using alternating least squares. Third, it regularizes and Cholesky-factorizes A_K and B_K to obtain L_K and C_K, and then solves the resulting weighted low-rank approximation problem in closed form via truncated SVD of $L_K W_K C_K$. The output is the optimal rank-r key matrix $\widehat{W}^\star_K$. The alternating least squares step avoids forming the full $(d_h d) \times (d_h d)$ matrix M_K, making the algorithm scalable.

C Extension to Grouped-Query Attention

Grouped-query attention (GQA) shares one key projection and one value projection among multiple query heads. Consequently, compression must assign a single pair of low-rank projections to each KV group, while accounting for the behavior of all query heads in that group.

We first consider one KV group with query-head index set $\mathcal{H}$ and size $H = |\mathcal{H}|$. The group shares

$$W_K, W_V \in \mathbb{R}^{d \times d_h}.$$

Each query head $h \in \mathcal{H}$ has its own query $q_{x,h} \in \mathbb{R}^{d_h}$ and output-projection block $O_h \in \mathbb{R}^{d_h \times d}$. For calibration query x, let $X_x \in \mathbb{R}^{n_x \times d}$ contain its causal context representations, and let $p_{x,h}(W_K) \in \mathbb{R}^{n_x}$ denote the corresponding attention probability column vector. All attention distributions are restricted to the finite, unmasked support.

The contribution of this group to the layer output is

$$o_x(W_K, W_V) = \sum_{h \in \mathcal{H}} p_{x,h}(W_K)^\top X_x W_V O_h. \tag{23}$$

All derivations below condition on the calibration inputs and query vectors.

C.1 Output-Error Decomposition

Let

$$E_V = W_V - \widehat{W}_V, \qquad p_{x,h} = p_{x,h}(W_K), \qquad \widehat{p}_{x,h} = p_{x,h}(\widehat{W}_K).$$

Adding and subtracting the output obtained with original attention and compressed values yields

$$o_x(W_K, W_V) - o_x(\widehat{W}_K, \widehat{W}_V) = \underbrace{\sum_{h\in\mathcal{H}} p_{x,h}^\top X_x E_V O_h}_{v_x} + \underbrace{\sum_{h\in\mathcal{H}} (p_{x,h} - \widehat{p}_{x,h})^\top X_x \widehat{W}_V O_h}_{k_x}. \tag{24}$$

Define

$$\mathcal{J}_{\mathrm{GQA}} = \sum_x \|o_x(W_K, W_V) - o_x(\widehat{W}_K, \widehat{W}_V)\|_2^2, \tag{25}$$

$$\mathcal{L}_{V,\mathrm{GQA}}(\widehat{W}_V) = \sum_x \|v_x\|_2^2, \tag{26}$$

$$\mathcal{L}_{K,\mathrm{GQA}}(\widehat{W}_K) = \sum_x \sum_{h\in\mathcal{H}} D_{\mathrm{KL}}(p_{x,h} \,\|\, \widehat{p}_{x,h}). \tag{27}$$

Suppose that, on the considered candidate class,

$$\max_{x,h,n} \|(X_x \widehat{W}_V O_h)_{n,:}\|_2 \le M$$

for a common finite constant M. Then

$$\mathcal{J}_{\mathrm{GQA}} \le 2\mathcal{L}_{V,\mathrm{GQA}} + 4HM^2 \mathcal{L}_{K,\mathrm{GQA}}. \tag{28}$$

To verify this, write

$$k_{x,h} = (p_{x,h} - \widehat{p}_{x,h})^\top X_x \widehat{W}_V O_h.$$

The row-norm bound and Pinsker's inequality imply

$$\|k_{x,h}\|_2^2 \le M^2 \|p_{x,h} - \widehat{p}_{x,h}\|_1^2 \le 2M^2 D_{\mathrm{KL}}(p_{x,h} \,\|\, \widehat{p}_{x,h}).$$

Using $\|\sum_h k_{x,h}\|_2^2 \le H \sum_h \|k_{x,h}\|_2^2$ and $\|v_x + k_x\|_2^2 \le 2\|v_x\|_2^2 + 2\|k_x\|_2^2$ gives equation 28.

As in the single-head case, this is a conditional upper bound: a rank constraint alone does not provide the required uniform row-norm bound.

C.2 Shared-Key Compression

The shared key projection affects every query head in the group. Accordingly, the key objective equation 27 aggregates their attention KL divergences before constructing the low-rank approximation.

Let

$$E_K = W_K - \widehat{W}_K, \qquad e_K = \mathrm{vec}(E_K).$$

For each query-head pair (x, h), define

$$J_{x,h} = \mathrm{diag}(p_{x,h}) - p_{x,h} p_{x,h}^\top.$$

With RoPE, define the logit-perturbation map

$$(\mathcal{B}_{x,h})_{n,:} = \mathrm{vec}\left(X_{x,n}^\top q_{x,h}^\top R_{n-m_x}\right)^\top.$$

The corresponding logit perturbation is

$$\Delta z_{x,h} = -\frac{1}{\sqrt{d_h}} \mathcal{B}_{x,h} e_K.$$

Applying the logit-space KL expansion separately to each query head gives

$$\mathcal{L}_{K,\mathrm{GQA}}(\widehat{W}_K) = e_K^\top M_{K,\mathrm{GQA}} e_K + \sum_{x,h} \mathcal{R}_{x,h}, \tag{29}$$

where

$$M_{K,\mathrm{GQA}} = \frac{1}{2d_h} \sum_x \sum_{h \in \mathcal{H}} \mathcal{B}_{x,h}^\top J_{x,h} \mathcal{B}_{x,h} \succeq 0. \tag{30}$$

Using $\omega(v) = \max_n v_n - \min_n v_n$, the remainder obeys

$$|\mathcal{R}_{x,h}| \leq \frac{1}{6d_h^{3/2}} \omega(\mathcal{B}_{x,h} e_K)^3. \tag{31}$$

Without RoPE, $\mathcal{B}_{x,h} = q_{x,h}^\top \otimes X_x$, and therefore

$$M_{K,\mathrm{GQA}} = \frac{1}{2d_h} \sum_{x,h} (q_{x,h} q_{x,h}^\top) \otimes (X_x^\top J_{x,h} X_x).$$

Thus the shared-key extension consists of aggregating the curvature of all associated query heads. The Kronecker fitting and low-rank optimization are performed once for the shared key projection, rather than independently for each query head.

C.3 Shared-Value Compression

Sharing values requires additional care because the group output sums the contributions of multiple query heads before its error is measured. Define

$$u_{x,h} = p_{x,h}^\top X_x \in \mathbb{R}^{1 \times d}, \qquad e_V = \mathrm{vec}(E_V).$$

Then

$$v_x = \sum_{h \in \mathcal{H}} u_{x,h} E_V O_h.$$

By column vectorization,

$$v_x^\top = T_x e_V, \qquad T_x = \sum_{h \in \mathcal{H}} (O_h^\top \otimes u_{x,h}) \in \mathbb{R}^{d \times d d_h}.$$

Consequently, the exact shared-value objective is

$$\mathcal{L}_{V,\mathrm{GQA}}(\widehat{W}_V) = e_V^\top M_{V,\mathrm{GQA}} e_V, \qquad M_{V,\mathrm{GQA}} = \sum_x T_x^\top T_x \succeq 0. \tag{32}$$

Expanding the metric gives

$$M_{V,\mathrm{GQA}} = \sum_x \sum_{h,k \in \mathcal{H}} (O_h O_k^\top) \otimes (u_{x,h}^\top u_{x,k}). \tag{33}$$

The terms with $h \neq k$ capture cross-head interactions in the summed group output.

In general, equation 32 is not a single Kronecker product. Therefore, the single-head weighted-SVD theorem does not apply unchanged to the exact shared-value objective. A tractable extension approximates $M_{V,\mathrm{GQA}}$ by a positive-semidefinite Kronecker product before applying weighted SVD.

For comparison, a headwise surrogate that discards output cross terms would be

$$\mathcal{L}_{V,\mathrm{head}} = \sum_{x,h} \|u_{x,h} E_V O_h\|_2^2,$$

with metric

$$M_{V,\mathrm{head}} = \sum_{x,h} (O_h O_h^\top) \otimes (u_{x,h}^\top u_{x,h}).$$

It satisfies

$$\mathcal{L}_{V,\mathrm{GQA}} \leq H \mathcal{L}_{V,\mathrm{head}},$$

but remains a sum of Kronecker products in general. Thus even this alternative does not automatically reduce to one exact two-sided weighted Frobenius objective.

When $H = 1$, the exact value metric reduces to

$$M_{V,\mathrm{GQA}} = (O_1 O_1^\top) \otimes \left(\sum_x u_{x,1}^\top u_{x,1} \right),$$

recovering the single-head formulation.

C.4 Common Kronecker Surrogate and Closed-Form Solution

For $S \in \{K, V\}$, let $M_{S,\mathrm{GQA}}$ denote the corresponding metric above. Fit fixed positive-semidefinite factors

$$M_{S,\mathrm{GQA}} \approx A_S \otimes B_S, \qquad A_S \in \mathbb{R}^{d_h \times d_h}, \quad B_S \in \mathbb{R}^{d \times d}.$$

For $\epsilon > 0$, factor

$$B_S + \epsilon I_d = L_S^\top L_S, \qquad A_S + \epsilon I_{d_h} = C_S C_S^\top.$$

The surrogate is

$$\widehat{\mathcal{L}}_{S,\mathrm{GQA}}(\widehat{W}_S) = \|L_S(W_S - \widehat{W}_S)C_S\|_F^2.$$

If

$$L_S W_S C_S = U_S \Sigma_S V_S^\top,$$

then a global rank-r_S minimizer is

$$\widehat{W}_{S,r_S} = L_S^{-1} U_{S,r_S} \Sigma_{S,r_S} V_{S,r_S}^\top C_S^{-1}. \tag{34}$$

This follows because the invertible change of variables $Y = L_S \widehat{W}_S C_S$ preserves rank, reducing the problem to ordinary truncated SVD.

For $E_S = W_S - \widehat{W}_S$ and $e_S = \mathrm{vec}(E_S)$, define

$$\Delta_{\epsilon,S}(E_S) = \epsilon \,\mathrm{tr}(B_S E_S E_S^\top) + \epsilon \,\mathrm{tr}(E_S A_S E_S^\top) + \epsilon^2 \|E_S\|_F^2.$$

The value approximation satisfies

$$\begin{aligned}\left|\mathcal{L}_{V,\mathrm{GQA}} - \widehat{\mathcal{L}}_{V,\mathrm{GQA}}\right| \le \left|e_V^\top (M_{V,\mathrm{GQA}} - A_V \otimes B_V) e_V\right| \\ + \Delta_{\epsilon,V}(E_V).\end{aligned} \tag{35}$$

The key approximation satisfies

$$\begin{aligned}\left|\mathcal{L}_{K,\mathrm{GQA}} - \widehat{\mathcal{L}}_{K,\mathrm{GQA}}\right| \le \left|e_K^\top (M_{K,\mathrm{GQA}} - A_K \otimes B_K) e_K\right| \\ + \frac{1}{6 d_h^{3/2}} \sum_{x,h} \omega(\mathcal{B}_{x,h} e_K)^3 + \Delta_{\epsilon,K}(E_K).\end{aligned} \tag{36}$$

Thus the GQA value construction incurs Kronecker and regularization errors, while the key construction additionally incurs the KL Taylor remainder. Equation equation 34 exactly minimizes the constructed surrogate, not necessarily the original grouped objective.

C.5 Multiple Groups and Cache Accounting

The construction is applied separately to each KV group. If a layer has G groups and their output errors are $\Delta o_{x,1}, \ldots, \Delta o_{x,G}$, then

$$\left\| \sum_{g=1}^{G} \Delta o_{x,g} \right\|_2^2 \le G \sum_{g=1}^{G} \|\Delta o_{x,g}\|_2^2.$$

Independent groupwise optimization therefore need not exactly minimize the total layer-output error, whose expansion also contains cross-group terms.

A shared KV group stores only one compressed key representation and one compressed value representation per token. If their ranks are $r_{K,g}$ and $r_{V,g}$, its cache stores $r_{K,g} + r_{V,g}$ latent coordinates per token, regardless of the number of query heads in the group. Accordingly, at a common precision and context length, the global rank budget is imposed over KV groups:

$$\sum_{\ell,g} (r_{K,\ell,g} + r_{V,\ell,g}) \le B_{\mathrm{cache}},$$

rather than counting the shared projections once per query head.

D EXTENSION TO ROPE

D.1 ROPE-AWARE ATTENTION AND KV CACHE

Rotary Position Embedding (RoPE) (Su et al., 2024) injects position information by rotating the query and key vectors before the attention dot product. Let m_x denote the position of the query token x, and let n index the positions of the keys stored in the KV cache. For each position m, let $R_m \in \mathbb{R}^{d_h \times d_h}$ be the orthogonal rotation matrix associated with position m. For a query row vector $q_x = xW_Q \in \mathbb{R}^{1\times d_h}$ and a key row vector $k_n = X_n W_K \in \mathbb{R}^{1\times d_h}$, where $X_n \in \mathbb{R}^{1\times d}$ is the n-th row of X, RoPE applies

$$\tilde{q}_x = q_x R_{m_x}^\top, \qquad \tilde{k}_n = k_n R_n^\top.$$

The attention logit between query x and key n is therefore

$$z_{x,n}(W_K) = \frac{\tilde{q}_x \tilde{k}_n^\top}{\sqrt{d_h}} = \frac{q_x R_{m_x}^\top R_n W_K^\top X_n^\top}{\sqrt{d_h}} = \frac{q_x R_{n-m_x} W_K^\top X_n^\top}{\sqrt{d_h}},$$

where the relative rotation is

$$R_{n-m_x} := R_{m_x}^\top R_n.$$

The RoPE attention probability vector is

$$p_x(W_K) = \mathrm{Softmax}\big(z_{x,1}(W_K), \ldots, z_{x,n}(W_K)\big) \in \mathbb{R}^{1\times n}.$$

The attention output is

$$o_x(W_K, W_V) = p_x(W_K) X W_V W_O \in \mathbb{R}^{1\times d},$$

with $p_x(W_K)$ computed with RoPE. When RoPE is disabled, $R_{n-m_x} = I$, and $p_x(W_K)$ reduces to the NoPE attention vector $a_x(W_K)$.

For the key-side analysis, under $W_K \to \widehat{W}_K = W_K - E$, the logit perturbation is

$$\Delta z_{x,n} = -\frac{1}{\sqrt{d_h}} q_x R_{n-m_x} E^\top X_n^\top.$$

Collecting over n,

$$\Delta z_x = -\frac{1}{\sqrt{d_h}} B_x \,\mathrm{vec}(E),$$

where $B_x \in \mathbb{R}^{n\times d d_h}$ has rows

$$(B_x)_{n,:} = \mathrm{vec}\big(X_n^\top q_x R_{n-m_x}\big)^\top.$$

When RoPE is disabled, $R_{n-m_x} = I$, and the above notation reduces to the NoPE case with

$$z_{x,n}(W_K) = \frac{1}{\sqrt{d_h}} q_x W_K^\top X_n^\top, \qquad B_x \text{ replaced by } (q_x \otimes X_n^\top)^\top.$$

D.2 EXTENSION OF LEMMA 1 AND THEOREM 1

Let $a_x^{\mathrm{RoPE}} = a_x^{\mathrm{RoPE}}(W_K)$ denote the original attention probability row vector computed with RoPE. As in Appendix A, X denotes the context associated with query x, with this dependence suppressed. The attention output is

$$o_x(W_K, W_V) = a_x^{\mathrm{RoPE}} X W_V W_O.$$

RoPE acts on queries and keys, but not on the value or output projections. Holding the original attention distributions fixed, the value objective is therefore

$$\mathcal{L}_V(\widehat{W}_V) = \sum_x \Big\| a_x^{\mathrm{RoPE}} X (W_V - \widehat{W}_V) W_O \Big\|_2^2.$$

Writing $E = W_V - \widehat{W}_V$, define the aggregated statistics

$$G_V^0 = \sum_x X^\top (a_x^{\mathrm{RoPE}})^\top a_x^{\mathrm{RoPE}} X, \qquad H_V^0 = W_O W_O^\top.$$

The same trace calculation as in Lemma 1 gives

$$\mathcal{L}_V(\widehat{W}_V) = \mathrm{tr}(G_V^0 E H_V^0 E^\top).$$

For $\epsilon > 0$, factor

$$G_V = G_V^0 + \epsilon I_d = L_V^\top L_V, \qquad H_V = H_V^0 + \epsilon I_{d_h} = C_V C_V^\top.$$

The regularized surrogate satisfies

$$\mathcal{L}_V^{\approx}(\widehat{W}_V) = \|L_V E C_V\|_F^2 = \mathcal{L}_V(\widehat{W}_V) + \Delta_{\epsilon,V}(E),$$

where

$$\Delta_{\epsilon,V}(E) = \epsilon \sum_x \|a_x^{\mathrm{RoPE}} X E\|_2^2 + \epsilon \|E W_O\|_F^2 + \epsilon^2 \|E\|_F^2 \geq 0.$$

Thus Lemma 1 holds with the RoPE-enabled attention distributions used to compute G_V^0. For fixed W_O and ϵ, H_V and its factor C_V are unchanged.

Since L_V and C_V are invertible, Theorem 1 applies to the regularized surrogate without further modification. If

$$L_V W_V C_V = \bar{U}_V \bar{\Sigma}_V \bar{R}_V^\top,$$

then a global minimizer over $\mathrm{rank}(\widehat{W}_V) \leq r$ is

$$\widehat{W}_{V,r}^* = L_V^{-1} \bar{U}_{V,1:r} \bar{\Sigma}_{V,1:r} \bar{R}_{V,1:r}^\top C_V^{-1},$$

with factors

$$W_{V,1:r}^{\mathrm{down}} = L_V^{-1} \bar{U}_{V,1:r} \bar{\Sigma}_{V,1:r}^{1/2}, \qquad W_{V,1:r}^{\mathrm{up}} = \bar{\Sigma}_{V,1:r}^{1/2} \bar{R}_{V,1:r}^\top C_V^{-1}.$$

Even when keys are also compressed, a_x^{RoPE} in this value objective remains the original attention distribution. The additional output error caused by changing attention is accounted for separately by the key-output term in Appendix A. Accordingly, the solution above exactly minimizes the regularized fixed-attention value surrogate, not the joint key–value objective.

D.3 Extension of Theorem 2

Lemma 4 (Rank-1 Kronecker weighted Frobenius form with RoPE)**.** *For each calibration query x at position m_x, let $q_x \in \mathbb{R}^{d_h}$ be its unrotated query column vector, and let $X_x \in \mathbb{R}^{n_x \times d}$ contain the context representations on its finite, unmasked attention support. Write $X_{x,n}$ for the n-th row of X_x, and define*

$$z_{x,n}(W_K) = \frac{1}{\sqrt{d_h}} q_x^\top R_{n-m_x} W_K^\top X_{x,n}^\top, \qquad p_x = \mathrm{softmax}(z_x(W_K)).$$

Here z_x and p_x are column vectors and $p_{x,n} > 0$. Define

$$J_x = \mathrm{diag}(p_x) - p_x p_x^\top, \qquad (B_x)_{n,:} = \mathrm{vec}\left(X_{x,n}^\top q_x^\top R_{n-m_x}\right)^\top,$$

where vec *stacks matrix columns.*

The exact, unregularized second-order quadratic metric is

$$M_K^0 = \frac{1}{2d_h} \sum_x B_x^\top J_x B_x = \frac{1}{2d_h} \sum_x B_x^\top J_x \,\mathrm{diag}(1/p_x) J_x B_x. \tag{37}$$

For $\epsilon > 0$, define the pre-fitting regularized metric

$$P_K = \sum_x (q_x q_x^\top) \otimes I_d, \qquad M_K^\epsilon = M_K^0 + \epsilon P_K. \tag{38}$$

This regularizer is defined in the unrotated query coordinates.

Let $A_K \succeq 0$ and $B_K \succeq 0$ be fixed factors of a rank-one Kronecker approximation

$$M_K^\epsilon \approx A_K \otimes B_K,$$

with dimensions $d_h \times d_h$ and $d \times d$, respectively. The factors may depend on ϵ through the fitting target. Define the post-fitting regularizations

$$A_K^\epsilon = A_K + \epsilon I_{d_h}, \qquad B_K^\epsilon = B_K + \epsilon I_d,$$

and factor

$$B_K^\epsilon = L_K^\top L_K, \qquad A_K^\epsilon = C_K C_K^\top.$$

Then the surrogate

$$\widehat{\mathcal{L}}_K(\widehat{W}_K) = \left\| L_K(W_K - \widehat{W}_K) C_K \right\|_F^2$$

has a global minimizer over $\operatorname{rank}(\widehat{W}_K) \le r$ *given by*

$$\widehat{W}_K^\star = L_K^{-1} U_r \Sigma_r V_r^\top C_K^{-1}, \tag{39}$$

where $L_K W_K C_K = U \Sigma V^\top$ and U_r, Σ_r, V_r are its top-r components.

For $E = W_K - \widehat{W}_K$, $e = \operatorname{vec}(E)$, and

$$\mathcal{L}_{K,x}(\widehat{W}_K) = D_{\mathrm{KL}}\left(p_x \,\|\, \operatorname{softmax}(z_x(\widehat{W}_K)) \right),$$

the approximation error satisfies

$$\left| \sum_x \mathcal{L}_{K,x}(\widehat{W}_K) - \widehat{\mathcal{L}}_K(\widehat{W}_K) \right| \le \underbrace{\left| e^\top (M_K^\epsilon - A_K \otimes B_K) e \right|}_{\textit{Kronecker approximation}} + \underbrace{\frac{1}{6 d_h^{3/2}} \sum_x \omega(B_x e)^3}_{\textit{Taylor remainder bound}} + \underbrace{\Delta_\epsilon(E)}_{\textit{total regularization}}, \tag{40}$$

where

$$\omega(v) = \max_n v_n - \min_n v_n$$

and

$$\Delta_\epsilon(E) = \Delta_{\epsilon,\mathrm{pre}}(E) + \Delta_{\epsilon,\mathrm{post}}(E), \tag{41}$$

$$\Delta_{\epsilon,\mathrm{pre}}(E) = \epsilon e^\top P_K e = \epsilon \sum_x \|E q_x\|_2^2, \tag{42}$$

$$\Delta_{\epsilon,\mathrm{post}}(E) = \epsilon \operatorname{tr}(B_K E E^\top) + \epsilon \operatorname{tr}(E A_K E^\top) + \epsilon^2 \|E\|_F^2. \tag{43}$$

Both regularization contributions are nonnegative.

When RoPE is disabled, $R_{n-m_x} = I$ and $B_x = q_x^\top \otimes X_x$. Consequently,

$$M_K^\epsilon = \sum_x (q_x q_x^\top) \otimes \left(\frac{1}{2 d_h} X_x^\top J_x X_x + \epsilon I_d \right), \tag{44}$$

recovering the non-RoPE construction with a pre-fitting ridge in the per-query left metric.

Proof. Below is the proof:

Step 1: RoPE-aware logit perturbation. For $E = W_K - \widehat{W}_K$, define

$$h_{x,n} = z_{x,n}(\widehat{W}_K) - z_{x,n}(W_K) = -\frac{1}{\sqrt{d_h}} q_x^\top R_{n-m_x} E^\top X_{x,n}^\top.$$

Set $u = X_{x,n}^\top \in \mathbb{R}^d$ and $v = R_{n-m_x}^\top q_x \in \mathbb{R}^{d_h}$. The column-vectorization identity gives

$$u^\top E v = \operatorname{vec}(E)^\top (v \otimes u) = \operatorname{vec}(E)^\top \operatorname{vec}(u v^\top).$$

Therefore,

$$q_x^\top R_{n-m_x} E^\top X_{x,n}^\top = e^\top \operatorname{vec}\left(X_{x,n}^\top q_x^\top R_{n-m_x} \right).$$

Stacking over the attention support yields

$$h_x = -\frac{1}{\sqrt{d_h}} B_x e. \tag{45}$$

Step 2: Direct KL expansion in logit space. For a fixed query x, define

$$F_x(t) = \log\left(\sum_{n=1}^{n_x} p_{x,n} e^{t h_{x,n}}\right), \qquad t \in [0,1].$$

The exact softmax KL identity gives

$$\mathcal{L}_{K,x}(\widehat{W}_K) = F_x(1) - F_x(0) - F_x'(0). \tag{46}$$

Here $F_x(0) = 0$ and $F_x'(0) = p_x^\top h_x$.

Define

$$p_{x,n}^{(t)} = \frac{p_{x,n} e^{t h_{x,n}}}{\sum_j p_{x,j} e^{t h_{x,j}}}, \qquad \mu_x(t) = \sum_n p_{x,n}^{(t)} h_{x,n}.$$

Differentiation yields

$$F_x''(0) = h_x^\top J_x h_x, \qquad F_x'''(t) = \sum_n p_{x,n}^{(t)} \big(h_{x,n} - \mu_x(t)\big)^3.$$

Taylor's theorem with integral remainder therefore gives

$$\mathcal{L}_{K,x}(\widehat{W}_K) = \frac{1}{2} h_x^\top J_x h_x + \mathcal{R}_x, \qquad \mathcal{R}_x = \frac{1}{2}\int_0^1 (1-t)^2 F_x'''(t)\, dt. \tag{47}$$

Since $\mu_x(t)$ lies between the minimum and maximum entries of h_x,

$$|F_x'''(t)| \le \sum_n p_{x,n}^{(t)} |h_{x,n} - \mu_x(t)|^3 \le \omega(h_x)^3.$$

It follows that

$$|\mathcal{R}_x| \le \frac{1}{6}\omega(h_x)^3 = \frac{1}{6 d_h^{3/2}} \omega(B_x e)^3. \tag{48}$$

Step 3: Exact metric and pre-fitting ridge. Substituting equation 45 into equation 47 gives

$$\mathcal{L}_{K,x}(\widehat{W}_K) = \frac{1}{2 d_h} e^\top B_x^\top J_x B_x e + \mathcal{R}_x.$$

To verify the alternative expression for M_K^0, let $D_x = \mathrm{diag}(p_x)$. Since $D_x^{-1} p_x = \mathbf{1}$ and $J_x \mathbf{1} = 0$,

$$J_x D_x^{-1} J_x = J_x D_x^{-1}(D_x - p_x p_x^\top) = J_x - J_x \mathbf{1} p_x^\top = J_x.$$

Thus the two expressions in equation 37 coincide, and

$$\sum_x \mathcal{L}_{K,x}(\widehat{W}_K) = e^\top M_K^0 e + \sum_x \mathcal{R}_x. \tag{49}$$

The pre-fitting regularization contributes

$$\epsilon e^\top P_K e = \epsilon \sum_x e^\top \big((q_x q_x^\top) \otimes I_d\big) e = \epsilon \sum_x \|E q_x\|_2^2.$$

Hence

$$\sum_x \mathcal{L}_{K,x}(\widehat{W}_K) = e^\top M_K^\epsilon e + \sum_x \mathcal{R}_x - \Delta_{\epsilon,\mathrm{pre}}(E). \tag{50}$$

The subtraction is essential: the pre-fitting ridge is added to the fitting target, not to the original KL loss.

Step 4: Kronecker approximation and post-fitting ridge. The vectorization identity gives

$$e^\top (A_K \otimes B_K)e = \mathrm{tr}(B_K E A_K E^\top).$$

Using the Cholesky factorizations,

$$\begin{aligned}\widehat{\mathcal{L}}_K(\widehat{W}_K) &= \mathrm{tr}\left((B_K + \epsilon I_d)E(A_K + \epsilon I_{d_h})E^\top\right) \\ &= e^\top (A_K \otimes B_K)e + \Delta_{\epsilon,\mathrm{post}}(E).\end{aligned}$$

The pre-fitting correction is nonnegative because $P_K \succeq 0$, and the post-fitting correction is nonnegative because $A_K, B_K \succeq 0$.

Combining this equality with equation 50 yields the exact signed error decomposition

$$\begin{aligned}\sum_x \mathcal{L}_{K,x}(\widehat{W}_K) - \widehat{\mathcal{L}}_K(\widehat{W}_K) = e^\top (M_K^\epsilon - A_K \otimes B_K)e \\ + \sum_x \mathcal{R}_x - \Delta_\epsilon(E).\end{aligned} \tag{51}$$

Taking absolute values and applying equation 48 proves equation 40.

When RoPE is disabled,

$$B_x = q_x^\top \otimes X_x, \qquad B_x^\top J_x B_x = (q_x q_x^\top) \otimes (X_x^\top J_x X_x).$$

Adding ϵP_K therefore gives equation 44.

Step 5: Closed-form surrogate minimizer. Since $A_K + \epsilon I_{d_h}$ and $B_K + \epsilon I_d$ are positive definite, L_K and C_K are invertible. Define

$$T_K = L_K W_K C_K, \qquad Y = L_K \widehat{W}_K C_K.$$

The map $\widehat{W}_K \mapsto Y$ is bijective and preserves rank. Consequently,

$$\min_{\mathrm{rank}(\widehat{W}_K)\le r} \widehat{\mathcal{L}}_K(\widehat{W}_K) = \min_{\mathrm{rank}(Y)\le r} \|T_K - Y\|_F^2.$$

By the Eckart–Young–Mirsky theorem, a global minimizer is $Y^\star = U_r \Sigma_r V_r^\top$. Applying the inverse transformation proves equation 39. □

E Mathematical Comparison with A^3

A^3 (Wong et al., 2025) and BELT both use analytical low-rank approximation, but differ in their treatment of key compression. A^3 approximates the attention logits before softmax, whereas BELT derives its key metric from preservation of the attention distributions after softmax. The distinction lies in the approximation objective and its induced metric, rather than in the truncated-SVD operation itself. Our value-side construction is closely related to the attention-aware weighted approximation in A^3; we therefore focus this comparison on keys.

E.1 Notation and the Pre-Softmax Attention Operator

We use the projection convention of Section D.1: $W_Q, W_K \in \mathbb{R}^{d\times d_h}$ and token representations are row vectors. Without RoPE, the logits for query x and context X are

$$z_x(W_K) = \frac{1}{\sqrt{d_h}} x W_Q W_K^\top X^\top.$$

Thus the bilinear attention operator in representation space is

$$P = W_Q W_K^\top \in \mathbb{R}^{d\times d}.$$

Under the alternative convention in which projection matrices act on column-vector representations, the same operator is written as $W_Q^\top W_K$. The apparent transpose difference is therefore purely notational.

To describe covariance-weighted logit reconstruction, let Y stack calibration query representations and let X stack context representations. The corresponding logit matrix is

$$Z = \frac{1}{\sqrt{d_h}} Y P X^\top.$$

E.2 A^3: Covariance-Weighted Logit Approximation

The pre-softmax reconstruction criterion underlying the A^3 key construction can be written as

$$\mathcal{L}_{\text{logit}}(\widehat{P}) = \frac{1}{d_h}\|Y(P-\widehat{P})X^\top\|_F^2. \tag{52}$$

Let

$$G_Q = Y^\top Y, \qquad G_X = X^\top X,$$

and write covariance factorizations

$$G_Q = U_Q U_Q^\top, \qquad G_X = U_X U_X^\top.$$

Then trace cyclicity gives

$$\begin{aligned}\mathcal{L}_{\text{logit}}(\widehat{P}) &= \frac{1}{d_h}\operatorname{tr}\left(G_Q(P-\widehat{P})G_X(P-\widehat{P})^\top\right)\\ &= \frac{1}{d_h}\|U_Q^\top(P-\widehat{P})U_X\|_F^2.\end{aligned} \tag{53}$$

When the covariance factors are invertible, minimizing equation 53 over $\operatorname{rank}(\widehat{P}) \le r$ reduces to truncated SVD of

$$U_Q^\top P U_X = U_Q^\top W_Q W_K^\top U_X.$$

In particular,

$$\widehat{P}_r = U_Q^{-\top}\left(U_Q^\top P U_X\right)_r U_X^{-1}.$$

When query and context statistics coincide, $U_Q = U_X$, giving the same-covariance transformed operator. With column-oriented projection matrices, this becomes the expression $U_X^\top W_Q^\top W_K U_X$.

The covariance weighting accounts for the calibration distribution of token representations. However, the reconstruction criterion remains Euclidean in logit space: for a fixed query, its error contribution is the sum of squared perturbations of the logits. It does not explicitly distinguish logit perturbations according to their effect on the resulting softmax distribution.

The formula above describes the unregularized covariance-weighted objective. If positive-definite covariance regularization is used, the SVD instead exactly solves the corresponding regularized weighted objective.

E.3 Expressing Logit Reconstruction in Key-Weight Space

For a direct comparison with BELT, hold W_Q fixed and write

$$E = W_K - \widehat{W}_K, \qquad e = \operatorname{vec}(E).$$

The logit-perturbation column vector for query x is

$$h_x = z_x(\widehat{W}_K)^\top - z_x(W_K)^\top = -\frac{1}{\sqrt{d_h}} X E q_x^\top.$$

Here $q_x = xW_Q$ retains the row-vector convention of the preliminaries, and X denotes the context of the current query.

The resulting logit-reconstruction objective is

$$\mathcal{L}_{\text{logit}}(\widehat{W}_K) = \sum_x \|h_x\|_2^2 = e^\top M_{\text{logit}} e, \tag{54}$$

where

$$M_{\text{logit}} = \frac{1}{d_h}\sum_x (q_x^\top q_x) \otimes (X^\top X). \tag{55}$$

If all queries share the same context matrix, this metric factorizes exactly:

$$M_{\text{logit}} = \frac{1}{d_h}\left(\sum_x q_x^\top q_x\right) \otimes (X^\top X).$$

This separability explains why covariance whitening naturally leads to a weighted-SVD solution for logit reconstruction.

For query-dependent causal supports, X varies with x, and the metric is generally a sum of Kronecker products rather than one product. Thus a single covariance-factor construction corresponds to the particular calibration weighting and support treatment used in that construction.

The bilinear-operator formulation equation 52 and the fixed-query key-weight formulation equation 54 should not be identified without specifying the feasible parameterization. The former optimizes a rank-constrained operator $\widehat{P}$, whereas the latter restricts that operator to $W_Q\widehat{W}_K^\top$ with fixed W_Q. Our comparison of error metrics does not require the two parameterizations to be identical.

E.4 BELT: Attention-Distribution Curvature

BELT instead starts from

$$\mathcal{L}_K(\widehat{W}_K) = \sum_x D_{\mathrm{KL}}\left(a_x(W_K) \,\middle\|\, a_x(\widehat{W}_K)\right). \tag{56}$$

For the original attention row vector a_x, let

$$J_x = \mathrm{diag}(a_x) - a_x^\top a_x.$$

The direct logit-space expansion gives

$$\mathcal{L}_K(\widehat{W}_K) = \frac{1}{2}\sum_x h_x^\top J_x h_x + \sum_x \mathcal{R}_x(E), \tag{57}$$

with the explicit remainder bound derived in Lemma 2.

The unregularized quadratic metric is therefore

$$M_K^0 = \frac{1}{2d_h}\sum_x (q_x^\top q_x) \otimes (X^\top J_x X). \tag{58}$$

Comparing equation 55 and equation 58, the essential difference is

$$\boxed{X^\top X \quad \longrightarrow \quad X^\top J_x X.}$$

Apart from the overall factor $1/2$, logit reconstruction uses the identity metric on logit perturbations, whereas BELT uses the query-dependent softmax curvature J_x.

The weighting can also be expressed as

$$X^\top J_x X = \sum_n a_{x,n} X_n^\top X_n - (a_x X)^\top (a_x X). \tag{59}$$

Thus BELT uses an attention-weighted, centered covariance of context representations. This differs from the uncentered context second moment $X^\top X$ used by Euclidean logit reconstruction.

E.5 What the Different Metrics Preserve

Uniform logit shifts. Softmax is invariant to adding the same scalar to every logit:

$$\mathrm{softmax}(z + c\mathbf{1}) = \mathrm{softmax}(z).$$

For $h_x = c\mathbf{1}$, logit reconstruction incurs

$$\|h_x\|_2^2 = n_x c^2,$$

whereas

$$h_x^\top J_x h_x = 0.$$

In this case the original attention KL is also exactly zero, not merely zero to second order. BELT's unregularized curvature therefore respects a symmetry of the attention distribution that raw logit reconstruction does not.

Attention-weighted variation. For an arbitrary perturbation,

$$h_x^\top J_x h_x = \sum_n a_{x,n} h_{x,n}^2 - \left(\sum_n a_{x,n} h_{x,n} \right)^2 . \tag{60}$$

Equivalently,

$$h_x^\top J_x h_x = \frac{1}{2} \sum_{n,m} a_{x,n} a_{x,m} (h_{x,n} - h_{x,m})^2 . \tag{61}$$

The metric therefore measures changes in relative logits, weighted by the original attention probabilities, rather than absolute logit discrepancies.

For example, with two attended positions and $a_x = (p, 1-p)$,

$$h_x^\top J_x h_x = p(1-p)(h_{x,1} - h_{x,2})^2,$$

while Euclidean logit reconstruction penalizes $h_{x,1}^2 + h_{x,2}^2$. The two criteria thus need not select the same low-rank subspace.

Uniform-attention special case. If $a_x = \mathbf{1}^\top / n_x$, then

$$J_x = \frac{1}{n_x} \left(I - \frac{1}{n_x} \mathbf{1}\mathbf{1}^\top \right).$$

Hence BELT's local objective becomes

$$\frac{1}{2} h_x^\top J_x h_x = \frac{1}{2n_x} \left\| h_x - \frac{\mathbf{1}^\top h_x}{n_x} \mathbf{1} \right\|_2^2 .$$

In this special case, it is proportional to centered logit reconstruction. For nonuniform attention, however, the weighting is generally anisotropic and query dependent.

Local sensitivity and its limitation. The attention-dependent metric can assign little quadratic cost to directions that are locally insensitive under the original attention distribution. This does not imply that arbitrarily large perturbations in those directions are harmless. For example, a sufficiently large change to a low-probability logit can substantially alter attention. The Taylor remainder is therefore essential when assessing the accuracy of the quadratic approximation.

Pre-fitting regularization and subsequent Kronecker approximation can also modify the null directions and relative weights of the exact curvature. The interpretations above describe M_K^0; the implemented surrogate approximates this metric with the explicit error contributions analyzed in our theorems.

E.6 RoPE and Position-Dependent Key Sensitivity

Without RoPE, a single bilinear operator $P = W_Q W_K^\top$ maps query and context representations to logits. With RoPE, the corresponding operator depends on the relative position:

$$z_{x,n}(W_K) = \frac{1}{\sqrt{d_h}} x W_Q R_{n-m_x} W_K^\top X_n^\top .$$

In general, there is no single position-independent matrix P representing all these interactions.

Let B_x be the RoPE-aware perturbation map defined in Lemma 4, so that

$$h_x = -\frac{1}{\sqrt{d_h}} B_x e.$$

A RoPE-aware Euclidean logit objective would have metric

$$M_{\text{logit}}^{\text{RoPE}} = \frac{1}{d_h} \sum_x B_x^\top B_x, \tag{62}$$

whereas BELT's exact local KL metric is

$$M_K^{0,\text{RoPE}} = \frac{1}{2d_h} \sum_x B_x^\top J_x B_x. \tag{63}$$

Thus the distinction between identity weighting and softmax-curvature weighting persists even when both criteria account for positional rotations.

Equation equation 62 describes a RoPE-aware logit-reconstruction criterion for comparison; it is not an assertion that A^3 implements this extension. BELT explicitly incorporates the relative rotations into the perturbation map used to construct its key metric.

E.7 Tractability and the Scope of Optimality

A^3 obtains an analytical approximation by applying SVD to a covariance-weighted pre-softmax operator. BELT's query-dependent attention curvature is generally not separable into a single left and right weighting matrix. To obtain a tractable weighted-SVD problem, BELT first adds the specified pre-fitting ridge,

$$M_K^\epsilon = M_K^0 + \epsilon \sum_x (q_x^\top q_x) \otimes I_d,$$

and fits

$$M_K^\epsilon \approx A_K \otimes B_K.$$

After regularizing the fitted factors, the resulting surrogate is

$$\mathcal{L}_K^{\approx}(\widehat{W}_K) = \|L_K(W_K - \widehat{W}_K)C_K\|_F^2.$$

Its global rank-r minimizer is

$$\widehat{W}_{K,r} = L_K^{-1}(L_K W_K C_K)_r C_K^{-1}.$$

Both methods therefore use standard weighted low-rank approximation once their respective metrics are fixed. The distinction is that A^3 targets pre-softmax reconstruction, while BELT constructs its metric from post-softmax distribution preservation and then approximates that metric to recover a tractable SVD solution. The Kronecker fitting stage is separate from the closed-form SVD stage.

BELT does not establish that its solution globally minimizes the original attention KL, nor that it universally dominates logit reconstruction. Its approximation involves Taylor, Kronecker, and regularization errors, which must be assessed at the returned solutions. Moreover, attention KL does not directly incorporate the value vectors or downstream network sensitivity; its connection to attention-output preservation is through the conditional bound in Appendix A.

F Additional method details

F.1 Budget-Constrained Gates and Their Gradients

Let

$$\mathcal{F} = \{(i,j) : 1 \le i \le N,\ 2 \le j \le m_i\}$$

be the set of components eligible for allocation after reserving the first component of every block. Write

$$M_{\text{free}} = |\mathcal{F}|, \qquad b = B - N, \qquad t_{i,j} = s_i - j.$$

For an interior budget $0 < b < M_{\text{free}}$, define $\alpha(\boldsymbol{s})$ through

$$F(\alpha, \boldsymbol{s}) = \sum_{(i,j)\in\mathcal{F}} \sigma\left(\frac{s_i - j - \alpha}{\tau}\right) - b = 0. \tag{64}$$

Existence and uniqueness. For finite scores and $\tau > 0$, F is continuous and strictly decreasing in α. Furthermore,

$$\lim_{\alpha\to-\infty} F(\alpha, \boldsymbol{s}) = M_{\text{free}} - b > 0, \qquad \lim_{\alpha\to+\infty} F(\alpha, \boldsymbol{s}) = -b < 0.$$

Therefore, equation 64 has a unique finite solution.

Implicit differentiation. For $(i,j) \in \mathcal{F}$, let

$$g_{i,j} = \sigma\left(\frac{s_i - j - \alpha}{\tau}\right), \qquad w_{i,j} = \frac{g_{i,j}(1-g_{i,j})}{\tau},$$

and define

$$W_i = \sum_{j=2}^{m_i} w_{i,j}, \qquad W = \sum_i W_i > 0.$$

Holding α fixed,

$$\frac{\partial F}{\partial s_k} = W_k, \qquad \frac{\partial F}{\partial \alpha} = -W.$$

The implicit function theorem gives

$$\frac{\partial \alpha}{\partial s_k} = -\frac{\partial F/\partial s_k}{\partial F/\partial \alpha} = \frac{W_k}{W}. \tag{65}$$

Consequently,

$$\frac{\partial g_{i,j}}{\partial s_k} = w_{i,j}\left(\mathbf{1}_{\{i=k\}} - \frac{W_k}{W}\right). \tag{66}$$

Summing over the free components yields

$$\sum_{(i,j)\in\mathcal{F}} \frac{\partial g_{i,j}}{\partial s_k} = W_k - \frac{W_k}{W} W = 0,$$

consistent with the constant relaxed budget.

If $\ell_{i,j} = \partial \mathcal{L}_{\mathrm{KD}}/\partial g_{i,j}$ and the scores affect the loss only through the gates, the score gradient is

$$\frac{\partial \mathcal{L}_{\mathrm{KD}}}{\partial s_k} = \sum_{j=2}^{m_k} \ell_{k,j} w_{k,j} - \frac{W_k}{W} \sum_{(i,j)\in\mathcal{F}} \ell_{i,j} w_{i,j}. \tag{67}$$

The reserved gates $g_{i,1} = 1$ have zero derivatives.

Numerical solution. The forward cutoff can be computed using bisection or safeguarded Newton iteration. The backward pass uses equation 65–equation 67, rather than treating the cutoff as constant. In finite precision, the relaxed budget is satisfied up to the stated root-solving tolerance. Very small temperatures can cause sigmoid saturation and an ill-conditioned cutoff derivative; the implementation should monitor the residual and numerical stability.

Endpoint budgets. When $b = 0$, all free gates are set to zero. When $b = M_{\mathrm{free}}$, all free gates are set to one. These cases require no cutoff solve or allocation training.

Hard allocation and prefix preservation. For deployment, retain the first component of every block and select exactly b elements of $\mathcal{F}$ with the largest scores, breaking ties deterministically. For two free components in the same block with $j < k$,

$$t_{i,j} = s_i - j > s_i - k = t_{i,k}.$$

Thus selecting component k necessarily selects all preceding free components. Together with the reserved first component, each block retains a prefix, and the final allocation has exactly B components.

The relaxed gate-mass constraint does not imply a hard rank constraint during training. Nor does exact budget satisfaction establish optimality of the learned hard allocation. Its quality and the soft-to-hard discrepancy are evaluated empirically.

F.2 Integration with Quantization

To further enhance compression beyond low-rank decomposition, we quantize the compressed KV cache. BELT determines only the factors and the rank of each block, and block i still caches the compressed representation $Z_i = XW_i^{\mathrm{down}} \in \mathbb{R}^{n\times r_i}$ as in prior low-rank methods. Hence existing quantizers for low-rank KV caches apply to BELT without modification. We combine BELT with

the quantizers of Palu (Chang et al., 2025) and STAR-KV (Bhatnagar et al., 2026), denoted Palu-Q and STAR-Q.

Since the j-th column of W_i^{down} is scaled by the square root of the j-th largest singular value (Theorem 1 and Corollary 1), the leading coordinates of Z_i are much larger than the others and dominate the per-token quantization range. Both quantizers therefore apply an orthogonal rotation that spreads these outliers across coordinates (Ashkboos et al., 2024), followed by per-token asymmetric quantization. With $\mathcal{Q}_\beta$ denoting β-bit quantization, Palu-Q and STAR-Q cache

$$\hat{Z}_i = \mathcal{Q}_\beta(Z_i R_i) \quad \text{and} \quad \hat{Z}_i = \left[\, \mathcal{Q}_{\beta_{\text{out}}}(Z_i^{\text{out}} R_i^{\text{out}}) \;\; \mathcal{Q}_{\beta_{\text{in}}}(Z_i^{\text{in}} R_i^{\text{in}}) \,\right], \tag{68}$$

respectively, where R_i, R_i^{out}, and R_i^{in} are orthogonal matrices. STAR-Q quantizes the leading 20% coordinates Z_i^{out} with $\beta_{\text{out}} = 4$ bits and the rest Z_i^{in} with $\beta_{\text{in}} = 3$ bits, which requires the leading coordinates to be the dominant ones. BELT satisfies this requirement because every hard allocation retains a prefix of the transformed SVD of each block (Section 3.2). The rotations are fused into the factors offline, i.e., $W_i^{\text{down}} \leftarrow W_i^{\text{down}} R_i$ and $W_i^{\text{up}} \leftarrow R_i^\top W_i^{\text{up}}$ with a block-diagonal R_i for STAR-Q, and thus add no computation at inference. As the learned ranks are arbitrary integers, we use the normalized Walsh–Hadamard matrix when the rotated width is a power of two and a fixed random orthogonal matrix otherwise. As a rank can be as small as one, some blocks are single-signed, so we do not clip the zero point of $\mathcal{Q}_\beta$ to $[0, 2^\beta - 1]$.

G Additional Experimental Settings

Configurations for System-level Performance Analysis All low-rank methods share STAR-KV's Triton kernels (Bhatnagar et al., 2026), which we generalize to per-head, grouped, and joint layouts: keys are rebuilt with RoPE on chip, attention weights are applied directly to the compressed values, and packed codes are dequantized on chip with STAR-Q. Since latency depends only on tensor shapes, each cache holds random values in the layout and ranks of the corresponding checkpoint.

H Additional Experimental Results

H.1 Additional System-level Analysis Results

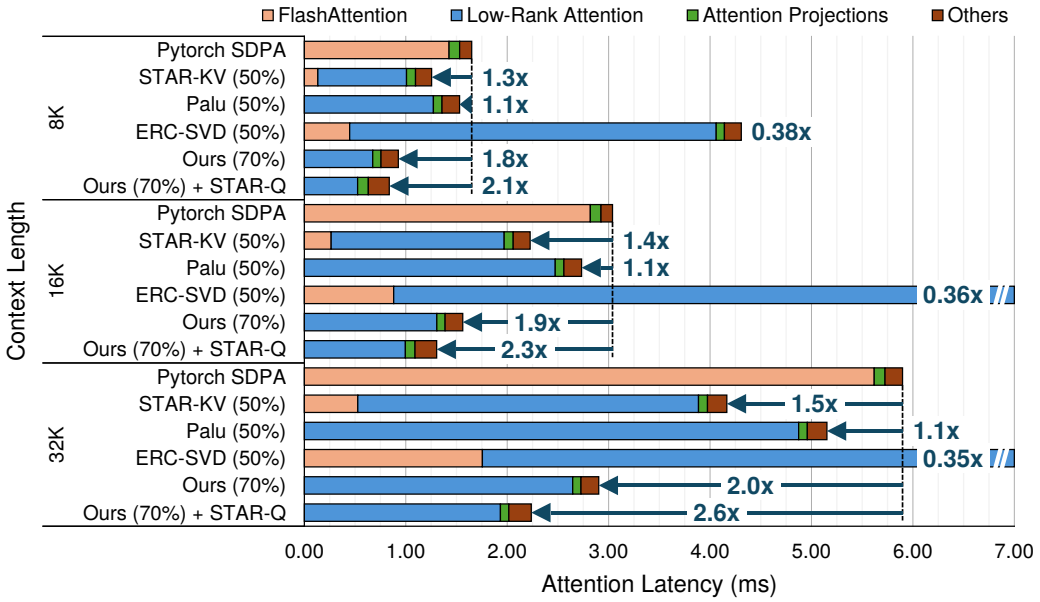


Figure 4: Latency breakdown of the attention module in Figure 3, averaged over all 32 layers. Arrows give the speedup over SDPA; the ERC-SVD bars at 16K and 32K are truncated.

H.2 Additional ablation studies

Attention without RoPE To test BELT without rotary embeddings, we follow A3 (Wong et al., 2025) and use MPT-7B, which encodes position via ALiBi (Press et al., 2022). We keep the main pipeline and data, setting all sequence lengths to 2048. Table 4 shows that BELT incurs accuracy losses close to those on LLaMA-2-7B, the other multi-head attention model, at both compression rates. **These results confirm that BELT generalizes to attention without RoPE.**

Table 4: BELT on MPT-7B, which encodes position with ALiBi instead of RoPE. Perplexity is measured on 2048-token windows.

Model	Comp (%)	Perplexity ↓		Zero-Shot Accuracy (%) ↑						
		Wiki2	C4	OBQA	PIQA	ARC-e	ARC-c	Hella	Wino	Avg.
MPT-7B	0	7.70	9.99	42.60	79.22	74.66	41.81	76.15	69.14	63.93
Ours	50	7.83	10.76	41.20	78.78	73.23	40.36	74.83	67.72	62.69
	70	8.34	12.06	38.80	76.44	68.98	39.08	71.13	66.06	60.08

Effect of Rank Allocation. To isolate rank allocation, we fix the key and value bases of BELT on LLaMA-3.1-8B-Instruct and vary only the rank distribution under the same KV cache budget, which is reported in Table 6. Baselines include Static (uniform fractions), Layer-wise search (ASVD/A3 sensitivity-based), and STAR-KV (learned soft thresholds). **BELT achieves the lowest perplexity and highest average accuracy at both compression rates.** Layer-wise search improves over Static but still trails BELT, while the state-of-the-art STAR-KV falls below Static. We give an analysis of why START-KV fails in Appendix I. A key pattern is that **learned allocation shifts ranks from keys to values**, retaining only 39–41% in keys versus about half for layer-wise search, and **its benefit grows with compression**: at 70%, it raises average accuracy by over 10 points over Static with identical bases.

Table 5: Compressing only the keys or only the values on LLaMA-3.1-8B-Instruct. Each variant takes the BELT checkpoint at 50% or 70% compression, keeps its low-rank projections for the compressed cache and the original projections for the other cache, and is not retrained. Comp is the resulting KV cache compression.

Method	Compressed cache	Comp (%)	Perplexity ↓		Zero-Shot Accuracy (%) ↑						
			Wiki2	C4	OBQA	PIQA	ARC-e	ARC-c	Hella	Wino	Avg.
LLaMA-3.1-8B-Inst	–	0	6.75	11.10	43.00	79.92	81.82	55.12	79.20	73.56	68.77
Ours (50%)	Keys and values	50.0	7.10	12.99	44.20	79.60	79.63	53.92	77.67	71.82	67.81
	Keys only	30.3	6.90	11.99	44.20	80.30	81.40	54.78	78.37	73.24	68.72
	Values only	19.7	6.97	12.17	44.80	79.22	80.89	54.52	78.10	73.56	68.52
Ours (70%)	Keys and values	70.0	8.54	17.51	41.80	76.66	73.32	47.01	72.70	68.75	63.37
	Keys only	37.6	7.46	13.39	42.20	80.14	79.38	52.05	76.76	70.48	66.83
	Values only	32.4	8.12	16.44	43.20	76.71	77.31	50.60	73.91	71.51	65.54

Effect of Key and Value Compression. To isolate each cache, we compress only the keys or only the values from our checkpoints at 50% and 70% compression, without retraining. The results of this experiment are included in Table 5. First, we observe that **Values are substantially more sensitive to compression than keys.** At 70%, compressing only values degrades accuracy and perplexity more than compressing only keys, even though the latter removes more of the cache. This matches our learned allocation, which assigns fewer ranks to keys (39–41%). At high compression, the combined loss of both caches roughly equals the sum of individual losses, confirming that most degradation comes from the value cache and that keys tolerate much lower ranks.

Effect of Calibration Data. We evaluate the sensitivity of BELT on the calibration data, which is performed on WikiText-2 and reported in Table 7. For this experiment, we rebuild BELT for Qwen3-32B at 50% compression using FineWeb-Edu (Penedo et al., 2024) and the C4 (Dodge et al., 2021) training split instead, keeping all else fixed. Table 7 suggests that **accuracy is nearly invariant, varying by only 0.26 points across corpora and consistently exceeding the uncompressed model.** Perplexity follows the calibration domain: WikiText-2 yields the lowest WikiText-2 perplexity, while web corpora lower C4 perplexity. The learned rank allocation also preserves its structure. **These results show that BELT is robust to the choice of calibration data.**

Table 6: Rank allocation on LLaMA-3.1-8B-Instruct with the key and value bases of BELT fixed. All methods share the same KV cache budget; STAR-KV stops once its budgets are met and uses 1.5% and 0.3% fewer ranks at 50% and 70% compression. Bold marks the best result at each compression rate.

Method	**Comp (%)**	**Perplexity ↓**		**Zero-Shot Accuracy (%) ↑**						
		Wiki2	C4	OBQA	PIQA	ARC-e	ARC-c	Hella	Wino	Avg.
LLaMA-3.1-8B-Inst	0	6.75	11.10	43.00	79.92	81.82	55.12	79.20	73.56	68.77
Static	50	7.66	14.92	42.20	78.35	76.73	50.77	74.44	72.53	65.84
	70	14.33	38.64	34.80	71.06	62.12	34.56	55.05	61.64	53.20
Layer-wise search	50	7.52	14.65	42.40	78.51	78.58	53.33	75.62	**73.01**	66.91
	70	11.57	32.02	37.20	72.80	68.77	41.04	61.84	65.19	57.81
STAR-KV	50	8.18	16.21	40.60	77.31	75.51	46.59	71.07	68.90	63.33
	70	15.42	40.21	35.00	71.33	60.61	31.40	54.45	58.64	51.90
Ours	50	**7.10**	**12.99**	**44.20**	**79.60**	**79.63**	**53.92**	**77.67**	71.82	**67.81**
	70	**8.54**	**17.51**	**41.80**	**76.66**	**73.32**	**47.01**	**72.70**	**68.75**	**63.37**

Table 7: Sensitivity of BELT to the calibration corpus on Qwen3-32B at 50% KV cache compression. Each corpus supplies both the calibration set of the key and value bases and the rank-allocation training data. WikiText-2 is the corpus of the main experiments (Table 9).

Model	**Calibration data**	**Perplexity ↓**		**Zero-Shot Accuracy (%) ↑**						
		Wiki2	C4	OBQA	PIQA	ARC-e	ARC-c	Hella	Wino	Avg.
Qwen3-32B	–	7.02	11.98	45.80	80.90	84.43	60.84	82.61	73.16	71.29
Ours (50%)	WikiText-2	7.01	12.35	45.40	80.25	84.18	64.08	82.50	75.06	71.91
	FineWeb-Edu	7.11	12.10	46.00	80.58	85.10	63.57	82.23	74.19	71.94
	C4	7.10	12.07	47.20	80.09	84.72	62.29	82.06	73.72	71.68

H.3 Empirical verification for theoretical claims

Evaluation of the bounds for approximating key-side objectives We empirically evaluate the error terms in the bound used for bounding the approximation error in Theorem 2 when $\widehat{W_K} = \widehat{W_K}^*$. The evaluation results are included in Table 8, which suggests that these error terms are all quite small, and the sum of Kronecker and Taylor approximation error is also very close to the true gap between $L_K^{\approx}$ and L_K, thus empirically justifying the theoretical analysis in Theorem 2.

H.4 Per-task main results

I Why STAR-KV fails?

StarKV (Bhatnagar et al., 2026) proposes a differentiable low-rank compression strategy. It performs SVD-based decomposition and introduces a learnable threshold α_i for each i, which aims to retain components whose singular values exceed α_i. Considering that this is non-differentiable, StarKV replaces it with a differentiable tanh-based soft-thresholding operator parameterized by α_i. This operator smoothly approximates the hard threshold while admitting gradients with respect to α_i, making α_i fully learnable via backpropagation. The thresholds are jointly optimized with a hybrid objective combining a KL divergence between the original and compressed output distributions,

Kronecker truncation error	0.015933
Taylor approximation error	-0.004157
ϵ	6.6e-5
$\lvert\mathcal{L}_K(\widehat{W}_K - \mathcal{L}_K^{\approx}(\widehat{W}_K)\rvert$	0.011775

Table 8: Evaluation of the bounds used in Theorem 2 at $\widehat{W_K}^*$

$\mathcal{L}_{KD}(\boldsymbol{\alpha})$, and a compression regularizer $\mathcal{L}_{\text{cmp}} = \sum_i e^{-\alpha_i}$ that drives α_i upward:

$$\mathcal{L}_{\text{StarKV}} = \mathcal{L}_{KD}(\boldsymbol{\alpha}) + \lambda \mathcal{L}_{\text{cmp}}, \tag{69}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ and D_{KL} denotes the sequence-level KL divergence.

However, the following analysis suggests that this strategy produces nearly identical thresholds across all heads, regardless of their eigenvalue distributions:

Theorem 3. *At any finite stationary point $\boldsymbol{\alpha}^*$ of equation 69, define $g_i^* = \partial_{\alpha_i} \mathcal{L}_{\text{KD}}(\boldsymbol{\alpha}^*), \bar{g}^* = \frac{1}{N}\sum_i g_i^*, \rho^* = \frac{\max_i |g_i^* - \bar{g}^*|}{\bar{g}^*}$. Then $g_i^* > 0$ and*

$$\max_i \alpha_i^* - \min_i \alpha_i^* = \log \frac{\max_i g_i^*}{\min_i g_i^*} \le \log \frac{1+\rho^*}{1-\rho^*}, \tag{70}$$

where the inequality holds whenever $\rho^* < 1$.

Appendix I provides the proof and an extension to nonstationary checkpoints. Our empirical studies suggest that the value of ρ^* is small at convergence.

We analyze the unconstrained objective

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}_{\text{KD}}(\boldsymbol{\alpha}) + \lambda \sum_{i=1}^{N} e^{-\alpha_i}, \qquad \lambda > 0.$$

The positive sign of the exponential regularizer is essential:

$$\partial_{\alpha_i} \mathcal{L} = \partial_{\alpha_i} \mathcal{L}_{\text{KD}} - \lambda e^{-\alpha_i}.$$

Constrained thresholds or alternative parameterizations require their corresponding stationarity conditions.

I.1 Proof of Theorem 3

Proof. At a finite stationary point $\boldsymbol{\alpha}^*$,

$$0 = \partial_{\alpha_i} \mathcal{L}(\boldsymbol{\alpha}^*) = g_i^* - \lambda e^{-\alpha_i^*}.$$

Therefore,

$$g_i^* = \lambda e^{-\alpha_i^*} > 0, \qquad \alpha_i^* = \log \lambda - \log g_i^*.$$

In particular, $\bar{g}^* > 0$, so ρ^* is well defined. Taking the maximum and minimum over i gives

$$\max_i \alpha_i^* = \log \lambda - \log \min_i g_i^*, \qquad \min_i \alpha_i^* = \log \lambda - \log \max_i g_i^*.$$

Subtracting yields the exact identity

$$\max_i \alpha_i^* - \min_i \alpha_i^* = \log \frac{\max_i g_i^*}{\min_i g_i^*}.$$

By the definition of ρ^*,

$$(1-\rho^*)\bar{g}^* \le g_i^* \le (1+\rho^*)\bar{g}^*.$$

If $\rho^* < 1$, the lower bound is positive, and hence

$$\max_i \alpha_i^* - \min_i \alpha_i^* \le \log \frac{1+\rho^*}{1-\rho^*}.$$

□

The exact identity holds at every finite stationary point. The condition $\rho^* < 1$ is sufficient for the stated upper bound but is not necessary for positivity of all gradients. For small ρ^*,

$$\log \frac{1+\rho^*}{1-\rho^*} = 2\rho^* + O((\rho^*)^3).$$

Our empirical studies suggest that ρ^* lies between 0.5 and 0.72. Hence, the bound $\log(\frac{1+\rho^*}{1-\rho^*}) \approx 0.1$, which is a small value.

I.2 A Measurable Bound Away from Stationarity

At a finite checkpoint $\boldsymbol{\alpha}$, define the signed distillation gradients and their mean by

$$g_i = \partial_{\alpha_i} \mathcal{L}_{\mathrm{KD}}(\boldsymbol{\alpha}), \qquad \bar{g} = \frac{1}{N} \sum_{i=1}^{N} g_i.$$

Whenever $\bar{g} > 0$, define

$$\rho = \frac{\max_i |g_i - \bar{g}|}{\bar{g}}, \qquad \varepsilon = \frac{\|\nabla \mathcal{L}(\boldsymbol{\alpha})\|_\infty}{\bar{g}}. \tag{71}$$

Thus ρ measures relative gradient dispersion, while ε measures the stationarity residual relative to the mean distillation gradient.

Proposition 1 (Checkpoint-level concentration bound)**.** *If $\bar{g} > 0$ and $\rho + \varepsilon < 1$, then*

$$D(\boldsymbol{\alpha}) := \max_i \alpha_i - \min_i \alpha_i \le \log \frac{1 + \rho + \varepsilon}{1 - \rho - \varepsilon}. \tag{72}$$

Proof. Write

$$r_i = \partial_{\alpha_i} \mathcal{L}(\boldsymbol{\alpha}) = g_i - \lambda e^{-\alpha_i}.$$

By the definitions in equation 71,

$$|g_i - \bar{g}| \le \rho \bar{g}, \qquad |r_i| \le \varepsilon \bar{g}.$$

Consequently,

$$(1 - \rho - \varepsilon)\bar{g} \le g_i - r_i = \lambda e^{-\alpha_i} \le (1 + \rho + \varepsilon)\bar{g}.$$

The lower endpoint is positive because $\rho + \varepsilon < 1$. Taking logarithms gives

$$\log \lambda - \log\big((1 + \rho + \varepsilon)\bar{g}\big) \le \alpha_i \le \log \lambda - \log\big((1 - \rho - \varepsilon)\bar{g}\big).$$

Subtracting the endpoints proves equation 72. □

At exact stationarity, $\varepsilon = 0$, recovering the bound in Theorem 3. If $\rho + \varepsilon \ge 1$, the proposition provides no finite certificate; this does not imply that the thresholds are dispersed.

I.3 Empirical Evaluation

All quantities in the bound can be evaluated at a saved checkpoint. For a calibration objective

$$\mathcal{L}_{\mathrm{KD}}(\boldsymbol{\alpha}) = \frac{1}{|\mathcal{C}|} \sum_{s \in \mathcal{C}} \ell_s(\boldsymbol{\alpha}),$$

compute the signed gradients

$$g_i = \frac{1}{|\mathcal{C}|} \sum_{s \in \mathcal{C}} \partial_{\alpha_i} \ell_s(\boldsymbol{\alpha}).$$

The model and threshold parameters are held fixed while these gradients are accumulated. The averaging and weighting must match those used to define the training objective.

Next, compute

$$\bar{g} = \frac{1}{N} \sum_i g_i, \qquad \eta = \max_i |g_i - \lambda e^{-\alpha_i}|, \qquad D = \max_i \alpha_i - \min_i \alpha_i.$$

When $\bar{g} > 0$, report

$$\rho = \frac{\max_i |g_i - \bar{g}|}{\bar{g}}, \qquad \varepsilon = \frac{\eta}{\bar{g}}.$$

When $\rho + \varepsilon < 1$, also report the certified upper bound

$$B = \log \frac{1 + \rho + \varepsilon}{1 - \rho - \varepsilon}$$

and compare it with the observed diameter D.

These statistics should be evaluated over training and across regularization strengths and random seeds. Gradients should be averaged before computing their dispersion; averaging per-minibatch values of ρ does not produce the same quantity. If only a subset of calibration data is used, the resulting statistics estimate those of the full objective and are not a rigorous certificate for it without controlling estimation error. If $\bar{g} \leq 0$, the relative statistics above are not applicable. If $\bar{g}$ is close to zero, numerical and sampling errors require particular care.

I.4 Scope and Interpretation

Theorem 3 is a stationary-point characterization, not a convergence theorem. At stationarity, concentrated positive distillation gradients and concentrated thresholds are linked by an exact identity. Observing both at the same stationary checkpoint therefore does not independently establish that gradient balancing caused threshold concentration.

Although λ cancels explicitly from the stationary spread identity, it can affect the stationary point and its gradients. The result consequently does not imply concentration independent of the regularization strength.

Finally, concentrated thresholds need not imply uniform ranks: the same threshold applied to different singular spectra can retain different numbers of components. Claims about rank-allocation quality must therefore be supported separately.

J Other related works

KV-specific low-rank compression. KV-specific methods include LoRC (Zhang et al., 2024) and PALU (Chang et al., 2025) (factorized key/value projections), TALE (Lee et al., 2025) (token-adaptive ranks), SALS (Mu et al., 2025) (latent sparse attention), EchoKV (Ji et al., 2026) (on-demand reconstruction), and xKV (Chang et al., 2026) (shared subspaces across layers). A^3 (Wong et al., 2025) develops analytical low-rank approximations for attention components, while STAR-KV (Bhatnagar et al., 2026) learns soft thresholds for adaptive rank control. Unlike reconstruction-centered methods, BELT weights approximation error by attention behavior; unlike independent threshold learning, its rank allocator satisfies a hard model-wide budget.

Table 9: Zero-shot evaluation on commonsense reasoning benchmarks using LM-Eval-Harness. We report accuracy (%) under different KV cache compression rates, together with perplexity on WikiText-2 and C4. A3 does not apply to Qwen3-32B. Bold marks the best result at each compression rate.

Model	**Comp (%)**	**Perplexity ↓**		**Zero-Shot Accuracy (%) ↑**						
		Wiki2	C4	OBQA	PIQA	ARC-e	ARC-c	Hella	Wino	Avg.
LLaMA-2-7B	0	5.12	7.04	44.00	77.64	76.26	46.16	75.98	69.22	64.88
ASVD	50	6.02	8.22	**43.80**	76.22	72.77	42.66	72.18	67.40	62.51
	70	6.95	9.48	40.80	74.54	70.08	38.65	69.15	64.64	59.64
ERC-SVD	50	5.78	8.27	43.60	75.95	71.76	42.24	73.17	68.51	62.54
	70	6.62	9.46	38.20	72.14	62.50	36.86	67.78	62.51	56.66
SVD-LLM	50	5.77	7.99	41.80	75.79	72.52	41.81	72.49	65.82	61.70
	70	6.77	9.90	37.60	72.47	63.93	36.52	67.35	60.77	56.44
ZS-SVD	50	5.39	7.71	42.60	76.88	75.04	**44.54**	74.39	**69.30**	63.80
	70	6.62	10.36	41.40	73.34	65.24	39.16	70.50	67.56	59.53
A3	50	12.83	32.71	29.60	63.22	43.94	25.60	46.00	59.51	44.64
	70	142.15	1581.78	26.80	54.35	29.76	20.99	28.80	51.70	35.40
Palu	50	5.74	8.36	42.20	76.44	73.27	41.98	73.31	66.14	62.22
	70	8.85	14.89	33.00	69.64	57.83	31.48	59.32	60.30	51.93
STAR-KV	50	6.38	9.15	42.00	75.46	66.79	38.65	70.79	64.64	59.72
	70	7.58	11.26	35.60	72.36	58.50	33.87	62.07	60.30	53.78
Ours	50	**5.23**	**7.37**	43.60	**77.64**	**75.21**	43.86	**74.77**	68.19	**63.88**
	70	**5.68**	**8.46**	**41.80**	**76.55**	**71.09**	**40.10**	**70.97**	**68.75**	**61.54**
LLaMA-3.1-8B-Inst	0	6.75	11.10	43.00	79.92	81.82	55.12	79.20	73.56	68.77
ASVD	50	8.29	14.03	**43.60**	79.49	**80.22**	**54.52**	76.93	**73.48**	**68.04**
	70	19.57	26.68	38.00	74.81	70.96	42.92	67.84	68.67	60.53
ERC-SVD	50	8.05	15.93	40.00	76.12	75.42	49.23	73.50	72.38	64.44
	70	13.65	37.88	32.40	65.34	54.34	29.35	49.74	64.09	49.21
SVD-LLM	50	8.31	17.87	42.00	76.28	73.91	47.35	72.06	71.27	63.81
	70	14.75	50.91	32.60	62.51	53.96	29.44	43.94	62.59	47.51
ZS-SVD	50	9.03	17.02	43.20	77.20	76.98	49.40	72.37	72.06	65.20
	70	22.90	185.78	33.20	69.70	61.78	32.59	53.18	60.93	51.90
A3	50	44.68	135.68	32.00	61.53	47.90	26.11	37.18	54.22	43.16
	70	240.31	460.40	28.20	58.27	33.59	21.50	29.11	50.75	36.90
Palu	50	9.79	37.56	42.40	77.86	77.74	49.66	70.89	71.43	65.00
	70	27.78	119.11	33.40	66.43	49.41	30.80	50.10	56.67	47.80
STAR-KV	50	8.95	15.87	39.40	76.06	74.87	45.65	71.05	65.82	62.14
	70	10.54	18.72	38.00	75.19	69.19	40.44	66.03	62.43	58.55
Ours	50	**7.12**	**13.02**	**43.60**	**79.71**	79.67	54.44	**77.77**	71.90	67.85
	70	**8.54**	**17.51**	**41.80**	**76.66**	**73.32**	**47.01**	**72.70**	**68.75**	**63.37**
Qwen3-32B	0	7.02	11.98	45.80	80.90	84.43	60.84	82.61	73.16	71.29
ASVD	50	9.86	15.96	42.40	77.91	78.91	54.18	72.02	63.77	64.87
	70	16.95	25.64	38.60	72.58	69.36	42.32	60.04	55.33	56.37
ERC-SVD	50	8.29	15.27	43.80	76.93	77.53	57.25	77.13	69.14	66.96
	70	18.62	45.32	34.00	65.51	55.93	34.39	49.70	55.80	49.22
SVD-LLM	50	8.43	16.10	43.80	77.15	77.40	55.72	76.08	67.96	66.35
	70	17.48	43.04	32.20	66.65	58.04	35.49	50.44	57.70	50.09
ZS-SVD	50	8.47	16.64	42.00	75.03	78.32	52.30	74.85	69.22	65.29
	70	15.99	38.33	28.40	59.68	46.59	27.56	43.47	56.20	43.65
Palu	50	20.15	43.55	32.80	66.87	64.94	41.04	55.58	60.14	53.56
	70	63.34	142.59	25.80	57.18	36.28	24.57	36.13	51.54	38.58
STAR-KV	50	7.61	14.42	40.80	76.01	77.40	51.19	72.54	67.09	64.17
	70	8.19	16.45	38.40	74.70	74.12	48.29	71.53	63.93	61.83
Ours	50	**7.01**	**12.35**	**45.40**	**80.25**	**84.18**	**64.08**	**82.50**	**75.06**	**71.91**
	70	**7.38**	**14.16**	**46.80**	**78.67**	**81.86**	**62.71**	**79.78**	**72.85**	**70.45**

Table 10: LongBench evaluation under different KV cache compression rates. A3 does not apply to Qwen3-32B. Bold marks the best result at each compression rate.

Model	**Comp (%)**	Qasper	QMSum	TriviaQA	MultiQA	TREC	MultiNews	VCSum	Avg.
LLaMA-2-7B	0	7.37	20.28	89.02	20.68	68.00	11.55	8.14	32.15
ASVD	50	1.84	16.51	79.32	22.62	58.00	4.51	4.50	26.76
	70	2.54	12.62	67.37	15.69	44.50	5.24	4.96	21.85
ERC-SVD	50	4.86	17.49	84.46	20.73	57.00	**6.75**	2.81	27.73
	70	1.00	14.47	65.76	11.44	40.50	6.95	1.60	20.25
SVD-LLM	50	2.86	18.34	83.75	17.97	56.00	3.97	0.02	26.13
	70	3.84	14.56	63.37	12.45	44.50	**7.48**	1.29	21.07
ZS-SVD	50	7.24	**20.02**	**88.18**	20.70	**65.50**	5.24	0.15	29.58
	70	4.67	15.87	**77.83**	17.28	49.00	3.60	**6.49**	24.96
A3	50	6.46	5.80	6.61	13.82	25.50	0.08	1.67	8.56
	70	0.16	1.66	0.32	0.61	0.50	1.29	0.02	0.65
Palu	50	6.32	19.69	81.92	**26.23**	52.50	1.39	3.91	27.42
	70	1.17	13.76	21.92	7.15	44.50	5.75	0.25	13.50
STAR-KV	50	4.03	7.22	65.71	16.25	38.00	0.05	0.34	18.80
	70	1.59	8.60	19.37	9.33	10.00	1.88	0.39	7.31
Ours	50	**7.66**	19.63	84.58	25.68	64.50	4.51	**7.47**	**30.57**
	70	**6.35**	**18.89**	70.30	**20.40**	**59.50**	2.34	5.13	**26.13**
LLaMA-3.1-8B-Inst	0	34.46	22.96	92.75	39.48	75.00	26.74	15.93	43.90
ASVD	50	11.88	23.09	90.58	25.65	54.00	24.67	15.20	35.01
	70	6.32	16.63	78.06	13.70	50.00	17.51	**5.48**	26.81
ERC-SVD	50	24.93	23.42	87.64	38.06	72.50	24.99	8.69	40.03
	70	6.51	14.84	72.13	23.79	63.50	10.92	0.08	27.40
SVD-LLM	50	23.50	**24.61**	88.23	37.61	70.00	24.68	10.92	39.93
	70	2.73	13.62	60.88	18.66	62.00	6.18	0.03	23.44
ZS-SVD	50	10.55	23.91	88.82	20.46	63.50	22.18	13.07	34.64
	70	0.84	14.09	2.01	1.35	18.50	5.61	0.32	6.10
A3	50	3.14	9.63	5.36	6.74	7.50	9.35	1.96	6.24
	70	1.17	8.90	3.16	1.16	0.00	4.92	3.08	3.20
Palu	50	9.39	22.80	28.43	19.86	65.00	19.15	3.10	23.96
	70	1.95	12.77	2.85	2.90	0.50	6.97	0.14	4.01
STAR-KV	50	10.50	18.82	77.25	20.13	55.50	18.59	9.46	30.03
	70	5.02	16.22	38.57	11.14	23.00	16.17	0.99	15.87
Ours	50	**34.08**	24.09	**91.70**	**47.54**	**76.50**	**26.56**	**16.46**	**45.28**
	70	**14.00**	**24.02**	**78.32**	**31.72**	**67.00**	**24.20**	3.16	**34.63**
Qwen3-32B	0	46.69	24.02	90.91	52.26	75.00	24.91	14.36	46.88
ASVD	50	17.52	19.80	19.12	29.59	35.00	21.31	14.01	22.34
	70	7.47	13.05	3.50	13.65	5.25	14.85	7.71	9.35
ERC-SVD	50	42.64	**24.21**	87.77	47.64	67.00	25.76	8.19	43.32
	70	18.86	17.77	21.70	26.82	1.50	14.93	0.06	14.52
SVD-LLM	50	41.41	24.20	87.39	47.72	59.00	**26.08**	4.90	41.53
	70	6.34	6.32	0.33	7.62	0.00	6.77	0.00	3.91
ZS-SVD	50	43.74	21.71	86.85	47.17	69.00	24.30	**15.24**	44.00
	70	2.27	12.22	2.13	23.08	0.00	9.99	0.20	7.13
Palu	50	22.29	17.66	18.08	28.46	7.00	25.77	2.00	17.32
	70	0.18	0.28	0.00	0.00	0.00	0.00	0.00	0.07
STAR-KV	50	35.28	23.17	62.58	37.71	36.00	24.74	14.87	33.48
	70	11.83	20.06	29.37	24.63	0.00	22.86	**9.51**	16.89
Ours	50	**47.28**	23.71	**90.67**	**52.28**	**77.50**	25.80	13.21	**47.21**
	70	**40.09**	**23.37**	**79.87**	**48.17**	**64.25**	**25.51**	0.37	**40.23**

Table 11: RULER evaluation at a context length of 16K under different KV cache compression rates. LLaMA-2-7B, whose context window is 4K tokens, is not evaluated, and A3 does not apply to Qwen3-32B. MK1/MK2: NIAH multi-key 1/2; MQ: NIAH multi-query; MV: NIAH multi-value; S1–S3: NIAH single 1–3; FWE: frequent words extraction; SQ: SQuAD QA. Bold marks the best result at each compression rate.

Model	**Comp (%)**	MK1	MK2	MQ	MV	S1	S2	S3	FWE	SQ	Avg.
LLaMA-3.1-8B-Inst	0	100.00	100.00	99.95	99.25	100.00	100.00	100.00	95.00	71.90	96.23
ASVD	50	99.40	97.80	99.60	98.15	**100.00**	99.80	98.60	73.33	60.57	91.92
	70	44.60	2.60	4.30	7.10	40.60	60.40	21.00	22.33	25.85	25.42
ERC-SVD	50	97.80	94.80	97.85	95.95	99.80	**100.00**	99.40	87.40	62.52	92.84
	70	79.80	36.80	70.70	73.85	97.40	88.00	18.80	6.93	34.98	56.36
SVD-LLM	50	98.80	96.40	96.10	95.95	99.80	99.60	99.60	84.20	65.45	92.88
	70	66.00	**38.80**	49.15	61.25	97.80	78.20	3.60	31.47	25.68	50.22
ZS-SVD	50	**99.80**	96.20	93.50	86.15	**100.00**	**100.00**	99.20	82.40	60.45	90.86
	70	0.00	0.00	0.00	0.00	22.60	0.00	0.00	18.27	0.13	4.56
A3	50	1.60	0.00	0.00	1.10	0.00	0.00	0.00	13.67	2.42	2.09
	70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.40	0.13
Palu	50	15.80	92.80	7.05	10.15	99.60	22.80	2.60	62.33	8.00	35.68
	70	0.00	0.00	0.00	0.00	6.20	0.00	0.00	0.00	0.53	0.75
STAR-KV	50	24.00	0.00	7.45	4.65	67.60	33.20	19.20	10.00	18.90	20.56
	70	2.20	0.00	0.10	0.45	54.40	15.00	1.20	4.27	12.33	9.99
Ours	50	99.00	**98.40**	**99.65**	**99.90**	**100.00**	**100.00**	**99.80**	**94.33**	**65.70**	**95.20**
	70	**88.80**	8.60	**97.00**	**96.55**	**100.00**	**99.80**	**73.80**	**58.13**	**44.72**	**74.16**
Qwen3-32B	0	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.93	77.77	96.97
ASVD	50	70.40	9.60	67.80	66.65	**100.00**	98.00	98.20	57.13	2.50	63.36
	70	31.20	0.00	14.05	34.65	94.20	92.20	21.60	18.13	7.77	34.87
ERC-SVD	50	98.80	94.40	**100.00**	99.90	95.60	97.60	**100.00**	94.33	65.70	94.04
	70	97.40	71.60	91.95	85.55	**100.00**	**100.00**	97.80	49.53	37.57	81.27
SVD-LLM	50	92.00	95.20	99.80	**100.00**	98.60	95.80	99.40	94.20	65.28	93.36
	70	87.80	49.40	87.85	75.20	**100.00**	**100.00**	52.20	45.93	35.55	70.44
ZS-SVD	50	99.40	99.60	**100.00**	99.65	**100.00**	99.80	**100.00**	**96.60**	61.82	95.21
	70	85.60	77.60	87.60	68.55	97.20	99.60	84.00	41.60	36.85	75.40
Palu	50	99.00	79.60	80.15	72.10	**100.00**	**100.00**	99.40	65.07	44.28	82.18
	70	56.00	0.80	16.30	25.70	99.20	55.00	23.20	0.20	10.25	31.85
STAR-KV	50	89.00	25.00	89.20	92.05	**100.00**	**100.00**	84.20	5.73	49.98	70.57
	70	36.40	0.00	42.00	41.10	99.40	94.80	65.80	1.80	28.47	45.53
Ours	50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.80	93.47	**71.95**	**96.14**
	70	**100.00**	**99.40**	**99.90**	**99.65**	**100.00**	99.80	**99.60**	**94.13**	**62.98**	**95.05**