KVGRAD: QUERY-AGNOSTIC KV CACHE EVICTION VIA GRADIENT-BASED GLOBAL IMPORTANCE SCORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Key-Value (KV) cache eviction mitigates the memory overhead of Large Language Models (LLMs) by discarding cached KV pairs that are less relevant to future token generation. However, existing eviction methods typically rely on local heuristics, scoring cached entries by their immediate contribution to the attention output of the same layer, while overlooking their downstream impact across subsequent layers. To address this, we propose **KVgrad**, a query-agnostic eviction method that scores each cache entry by its global impact on the model's final outputs. Based on a simple analysis using a first-order Taylor approximation and the chain rule, KVgrad factorizes the impact of cached value into a linear local component and a gradient-based downstream sensitivity term. Then, we apply a magnitude-aware scoring scheme to stabilize the importance score. Evaluations on RULER and LongBench demonstrate that KVgrad achieves an average maximum compression ratio of 5.9$\times$ with negligible performance loss ($< 2\%$), significantly outperforming strong local-only state-of-the-art baselines. Furthermore, we demonstrate that our signal is highly amenable to distillation; by training a lightweight MLP to predict KVgrad scores from local hidden-state features, we enable high-fidelity, on-the-fly compression at negligible computational cost. Finally, we demonstrate the transferability of KVgrad to vision-language models; evaluations on image and long-context video benchmarks suggest a unified path for importance-aware KV cache compression.

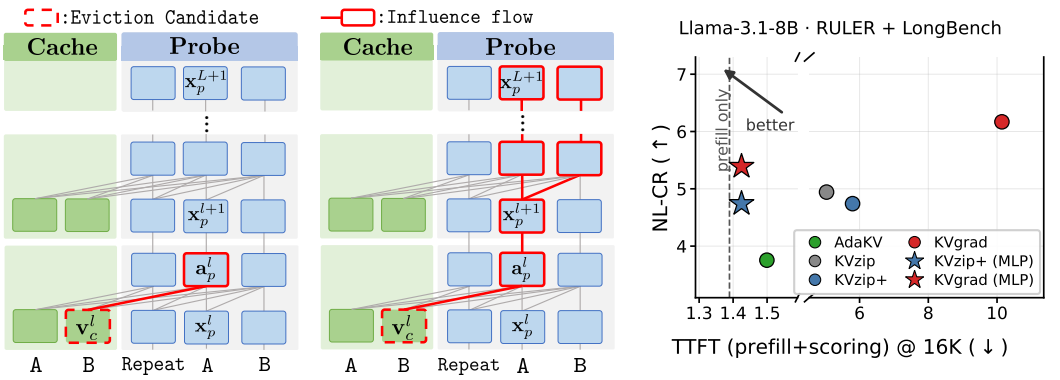
1 INTRODUCTION

Long-context autoregressive decoding in large language models is fundamentally *memory-bandwidth bound* (Shazeer, 2019; Pope et al., 2023). During inference, the entire key-value (KV) cache must be read from GPU HBM to compute attention for each new token (Shah et al., 2024). For example, in long-context scenarios like agentic workflows or multimodal processing, KV cache transfers can account for over 93% of GPU kernel time (Code Llama, 100K-token context; Jiang et al. (2025)).

To alleviate this bottleneck, post-hoc KV cache eviction methods (Zhang et al., 2023; Li et al., 2024) reduce decoding memory traffic by discarding KV pairs deemed less important for future generation, without retraining requirement and while remaining compatible with other compression techniques such as quantization (Kim et al., 2025; Ramachandran et al., 2025). In particular, query-agnostic approaches (Kim et al., 2025; Jegou & Jeblick, 2026) are highly advantageous for deployment: by computing importance scores independently of the query, they enable the compressed KV cache to be stored or offloaded with a reduced memory footprint, while allowing it to be reused across multiple queries. Furthermore, query-agnostic eviction has been reported (Kim et al., 2025) to outperform query-aware methods, particularly in multi-hop scenarios, where query-aware methods may discard intermediate evidence that is not directly relevant to the initial query.

However, existing eviction methods—whether query-aware or query-agnostic—typically evaluate the importance of cached KV pairs from a **local** view, considering each entry's direct contribution to the attention output ($\mathbf{a}_p^\ell$) at its respective layer (Figure 1 (a)). This perspective ignores the **downstream propagation** of this contribution through the subsequent feed-forward network (FFN) and higher layers to the final output. Since the model's prediction depends on the end-to-end effect of each entry, relying solely on local heuristics can lead to suboptimal eviction decisions.

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.



(a) **Prior work** (b) **KVgrad (Ours)** (c) **NL-CR vs. TTFT**

Figure 1: **Overview of KVgrad**: **(a) Prior work** (Feng et al., 2024; Jegou & Jeblick, 2026) uses a **local view**, scoring cache entry $\mathbf{v}_c^l$ solely by its immediate contribution to attention output $\mathbf{a}_p^l$. **(b) KVgrad** adopts a **global view**, quantifying the total influence flow from the entry to the final hidden states. **(c) Near-lossless compression ratio (NL-CR, $< 2\%$)[1] vs. Time-To-First-Token (TTFT).** ① Exact KVgrad attains the highest NL-CR at the longest scoring time, suiting settings where the cache is compressed once and stored. ② Distilling it into KVgrad (MLP) brings TTFT down to prefill only while retaining most of the NL-CR, above KVzip+ (MLP) under the identical recipe and AdaKV.

To bridge this gap, we propose **KVgrad**, a query-agnostic importance scoring method that estimates the **global** influence of each KV cache entry on the model's final output. We frame KV eviction as a perturbation to the cached values and use a first-order Taylor expansion to approximate its effect as an Input $\times$ Gradient (I$\times$G) score with respect to the value vector. Applying the chain rule, this I$\times$G score decomposes into the inner products between two conceptually distinct components:

(1) **Local component**: The direct, linear path from the value vector $\mathbf{v}_c^\ell$ to the attention output $\mathbf{a}_p^\ell$.
(2) **Downstream component**: The gradient-based sensitivity of the final hidden state $\mathbf{x}_p^{L+1}$ with respect to that attention output $\mathbf{a}_p^\ell$ (Figure 1 (b)).

However, direction of the downstream gradient can be noisy, and aggregating contributions across probes can dilute KV entries that are important only to specific probes. We therefore introduce two complementary mechanisms: **Magnitude-Aware Scoring (MAS)**, which removes the influence of the gradient direction on score and retains only its magnitude, and **Probe-Specific Preservation (PSP)**, which replaces summation over probes with a max operation. Together, they form a robust scoring function ($S_{\ell,c}^{\text{KVgrad}}$ in Eq. 8) that captures the end-to-end influence of each KV cache entry (Figure 1 (b)). KVgrad thereby extends the local scores of recent state-of-the-art methods like KVzip (Kim et al., 2025) and KVzip+ (Jegou & Jeblick, 2026) with a downstream sensitivity factor that they omit.

We extensively evaluate KVgrad on four LLMs using RULER (Hsieh et al., 2024) and LongBench (Bai et al., 2024). KVgrad outperforms query-agnostic baselines across most compression ratios, with the margin widening at aggressive compression (Figure 2), and reaches an average near-lossless ($< 2\%$) compression of 5.9$\times$, against 4.7$\times$ for the previous state-of-the-art KVzip+ (Table 1). While KVgrad's scoring adds a backward pass, following the distillation recipe of Jegou & Jeblick (2026) with KVgrad scores as the supervision target yields KVgrad (MLP), a lightweight surrogate that runs alongside prefill at negligible extra cost; it not only outperforms KVzip+ (MLP) trained under the identical setting but matches or surpasses exact KVzip+, suggesting that gradient-based importance is a better distillation target. The resulting eviction translates into decoding speedups: at 8:1 compression, attention-kernel latency drops by 5.6$\times$ and end-to-end generation runs 1.7$\times$ faster (2.0$\times$ with sparse prefill; Section 4.3). Finally, we provide preliminary evidence of KVgrad's generality on Vision-Language Models across image and video tasks; results on VideoMME and MLVU suggest that global importance-aware compression extends to multimodal processing.

[1] NL-CR is the largest compression ratio at which accuracy stays within τ=2% (relative) of the uncompressed model, geometric-mean over subtasks. Headline numbers average over four models and two benchmarks (Table 1); Figure 1(c) shows Llama-3.1-8B.

2 Related Work

2.1 KV Cache Compression

KV cache compression spans several orthogonal axes: token eviction Zhang et al. (2023); Li et al. (2024); Cai et al. (2024); Feng et al. (2024), low-rank projection Chang et al. (2025), quantization Liu et al. (2024c); Hooper et al. (2024), and architectural KV sharing or latent compression Ainslie et al. (2023); Liu et al. (2024a). These axes are largely composable—for instance, eviction can be applied on top of a quantized cache Kim et al. (2025); Ramachandran et al. (2025)—so progress along one axis does not preclude gains along another. Among these axes, we focus on post-hoc *token eviction*, which is complementary to quantization and low-rank methods already in deployment.

Query-aware vs. query-agnostic eviction Many eviction methods score context tokens based on the attention they receive from a set of *probe tokens*, which are used only for importance estimation and are not retained in the compressed cache. Depending on whether the user query that follows the context is used as a probe, KV cache eviction can be broadly categorized into *query-aware* and *query-agnostic* paradigms. Query-aware approaches (Li et al., 2024; Liu et al., 2026; Cai et al., 2024) estimate scores from attention weights induced by a specific query. While effective for single requests, these approaches incur deployment overhead: the uncompressed context must be retained or re-prefilled, then re-scored and re-compressed for every new query. In contrast, query-agnostic approaches (Devoto et al., 2025; Kim et al., 2025; Jegou & Jeblick, 2026; Zweiger et al., 2026; Park et al., 2026; Xiao et al., 2024) evaluate context-intrinsic importance before any query arrives, enabling the compressed KV cache to be stored, offloaded, and reused across multiple queries.

Query-agnostic approaches can be further divided based on whether they use probe tokens. Some use no probe tokens, relying on positional heuristics (Xiao et al., 2024), key geometry (Park et al., 2026), or estimated future attention patterns (Devoto et al., 2025). Others construct probe tokens from the context itself via context reconstruction (Kim et al., 2025; Jegou & Jeblick, 2026), summarization (Zweiger et al., 2026), or synthesized broad QA (Baek et al., 2026). These approaches have demonstrated stronger performance than query-aware baselines, particularly on multi-hop reasoning tasks where preserving indirect contextual dependencies is important. Furthermore, recent work (Jegou & Jeblick, 2026) addresses the latency overhead of the probe forward pass by using lightweight surrogate models to directly predict KV importance scores.

Beyond attention weights Existing KV eviction methods estimate token importance from attention weights (Zhang et al., 2023; Li et al., 2024; Liu et al., 2026; Cai et al., 2024; Kim et al., 2025; Baek et al., 2026; Zweiger et al., 2026), from attention weights combined with value vectors (Jegou & Jeblick, 2026; Devoto et al., 2025; Goel et al., 2025; Gu et al., 2025; Feng et al., 2025b), or from attention-free signals such as key geometry (Park et al., 2026) and token position (Xiao et al., 2024). However, their criteria remains local to the current layer, without accounting for the influence of subsequent FFNs and downstream layers. KVgrad addresses this limitation by using a gradient signal to capture the end-to-end sensitivity of each KV cache entry to the model's output.

2.2 Attribution Methods for KV Eviction

Input attribution in Explainable AI (XAI) quantifies how individual inputs contribute to model outputs. This offers a framework for capturing the end-to-end influence of KV cache entries, which has remained less explored in KV cache compression. Gradient-based sensitivity maps (Simonyan et al., 2013) and Input $\times$ Gradient (Shrikumar et al., 2017) are a natural starting point for input attribution, offering simple first-order estimates of input influence. However, the inherent instability of raw gradients (Balduzzi et al., 2017; Ghorbani et al., 2019) can compromise their reliability as attribution signals. Sampling- and path-integration-based gradient aggregations (Smilkov et al., 2017; Sundararajan et al., 2017) or perturbation-based methods (Lundberg & Lee, 2017; Petsiuk et al., 2018) offer high attribution fidelity, but their requirement for numerous forward or backward passes limits their applicability to KV eviction in long-context LLMs. KVgrad adapts Input $\times$ Gradient (Shrikumar et al., 2017) to LLM architectures and stabilizes its gradient signal through Magnitude-Aware Scoring (MAS), yielding an efficient yet reliable eviction criterion based on end-to-end KV influence.

3 PROPOSED METHOD

3.1 PRELIMINARY: QUERY-AGNOSTIC KV EVICTION

Repeat-reconstruction probes Query-agnostic KV cache eviction aims to evict KV pairs that are unimportant for generation, without access to the user query. KVzip Kim et al. (2025) formulates this objective as retaining the signals necessary to faithfully reconstruct the original context from the post-eviction cache. Specifically, we append a sequence of *probe tokens* consisting of a fixed instruction, such as "`Repeat the previous context:`", followed by a copy of the original context. For example, in Figure 1, the original context tokens A and B are duplicated after the instruction; together they form probe tokens used to score the cache entries corresponding to A and B. Hereafter, c and $p \in \mathcal{P}$ denote cache and probe token indices, respectively; by construction, $c < p$.

Notation Consider an L-layer Transformer with layers indexed by $\ell \in \{1, \dots, L\}$. We describe a single attention head; the extension to multi-head or grouped-query attention is straightforward, treating each head or group independently. Let $\mathbf{x}_p^\ell \in \mathbb{R}^{1\times D}$ denote the hidden state entering layer ℓ at position p, $\mathbf{v}_t^\ell \in \mathbb{R}^{1\times D}$ the value vector of token t, $\mathbf{W}_O^\ell \in \mathbb{R}^{D\times D}$ the output projection, and $A_{p,t}^\ell \in [0,1]$ the causal attention weight from position p to $t \le p$, with $\sum_{t=1}^{p} A_{p,t}^\ell = 1$.

The attention output at position p, $\mathbf{a}_p^\ell \in \mathbb{R}^{1\times D}$, and the residual update at layer ℓ are

$$\mathbf{a}_p^\ell = \sum_{t\le p} A_{p,t}^\ell \, \mathbf{v}_t^\ell \, \mathbf{W}_O^\ell, \qquad \mathbf{x}_p^{\ell+1} = \mathbf{x}_p^\ell + \mathbf{a}_p^\ell + \mathrm{FFN}^\ell\big(\mathbf{x}_p^\ell + \mathbf{a}_p^\ell\big), \tag{1}$$

where FFN^ℓ denotes the FFN block at layer ℓ (pre-normalization is omitted for brevity).

Importance scoring KVzip Kim et al. (2025) scores the importance of each cached KV pair by the maximum attention weight it receives from *any* probe token:

$$S_{\ell,c}^{\mathrm{KVzip}} \;=\; \max_{p\in\mathcal{P}} \; A_{p,c}^\ell. \tag{2}$$

Subsequently, KVzip+ (Jegou & Jeblick, 2026) measures each entry's contribution to *attention output* $\mathbf{a}_p^\ell$ relative to the residual-stream norm $\big\|\mathbf{x}_p^\ell\big\|_2$:

$$S_{\ell,c}^{\mathrm{KVzip+}} \;=\; \max_{p\in\mathcal{P}} \; A_{p,c}^\ell \, \frac{\big\|\mathbf{v}_c^\ell \, \mathbf{W}_O^\ell\big\|_2}{\big\|\mathbf{x}_p^\ell\big\|_2}. \tag{3}$$

Efficient MLP surrogate for single-query compression Because the scores above require additional forward passes over probes $\mathcal{P}$, exact calculation is practical mainly where a compressed cache is stored and reused, as in prefix caching or KV reuse (Lu et al., 2025; Yang et al., 2025b). To bypass this additional forward-pass overhead for single-query settings, KVzip+ (MLP) Jegou & Jeblick (2026)[2] distills the score into a per-layer MLP that predicts importance directly from hidden states without probe passes:

$$\hat{S}_{\ell,c}^{\mathrm{KVzip+}} \;=\; f_\theta^\ell\big(\mathbf{x}_c^\ell\big). \tag{4}$$

Trained once offline on a general pretraining corpus, this surrogate is reported to closely track the exact probe-based score $S_{\ell,c}^{\mathrm{KVzip+}}$,[3] while matching uncompressed accuracy at 2–4$\times$ compression in their setting; subsequent work adopts the same recipe Yao et al. (2026); Chen et al. (2026). Crucially, because this distillation recipe is score-agnostic, it serves as a controlled testbed for comparing underlying scoring functions. For training and inference details, please refer to Appendix A.2.

Limitations of local scoring The importance scores in Eqs. 2, 3, as well as the surrogate in Eq. 4, quantify how a cache entry locally affects the attention output $\mathbf{a}_p^\ell$ at its own layer. However, this local contribution does not necessarily reflect the impact on the final output, since the signal further propagates through the FFN layer of layer ℓ and the remaining $L-\ell$ blocks. Figure 1 illustrates this gap: prior scores terminate at layer ℓ (Figure 1(a)), whereas a faithful importance measure should account for the *influence flow* from $\mathbf{v}_c^\ell$ through the subsequent layers (Figure 1(b)).

[2] Denoted KVzap-MLP in the original paper.

[3] The surrogate is trained to predict $\log S_{\ell,c}^{\mathrm{KVzip+}}$; we omit the logarithm for notational simplicity.

3.2 KVGRAD: A GRADIENT-BASED GLOBAL IMPORTANCE SCORE

Formulating eviction via input attribution To quantify the end-to-end influence of each KV cache entry, we introduce a scalar target Φ that summarizes the model's outputs at the probe positions $p \in \mathcal{P}$ (specified in "Choice of Φ" below). Treating the prefilled cache as a fixed input to the probe forward pass, we view Φ as a function of the cached KV pairs, the probe tokens, and the model parameters. With a slight abuse of notation, we write $\Phi(\mathbf{v}_c^\ell)$ as a function of a single value vector, holding all other KVs fixed. Then, we use the counterfactual difference $\Phi(\mathbf{v}_c^\ell) - \Phi(\mathbf{0})$, as a proxy for the contribution of evicting $\mathbf{v}_c^\ell$. Approximating it with a first-order Taylor expansion, $\Phi(\mathbf{v}_c^\ell + \boldsymbol{\delta}) \approx \Phi(\mathbf{v}_c^\ell) + \langle \boldsymbol{\delta}, \nabla_{\mathbf{v}_c^\ell} \Phi \rangle$, and setting $\boldsymbol{\delta} = -\mathbf{v}_c^\ell$ yields the classic Input $\times$ Gradient (I$\times$G) (Shrikumar et al., 2017) score:

$$\Phi(\mathbf{v}_c^\ell) - \Phi(\mathbf{0}) \approx \langle \mathbf{v}_c^\ell, \nabla_{\mathbf{v}_c^\ell} \Phi \rangle. \tag{5}$$

Crucially, this formulation enables computing the end-to-end importance scores for all cached entries across all layers simultaneously with a single backward pass.

Reliability of I$\times$G score The accuracy of this first-order I$\times$G approximation depends on how close Φ is to linear in $\mathbf{v}_c^\ell$ along the segment from $\mathbf{0}$ to $\mathbf{v}_c^\ell$. Notably, the attention output $\mathbf{a}_p^\ell$ is exactly linear in the value vector $\mathbf{v}_c^\ell$. We can therefore apply the chain rule through the attention outputs to separate this exact linear contribution from the subsequent nonlinear transformations:

$$\begin{aligned} \langle \mathbf{v}_c^\ell, \nabla_{\mathbf{v}_c^\ell} \Phi \rangle &= \langle \mathbf{v}_c^\ell, \Big(\sum_{p \in \mathcal{P}} \nabla_{\mathbf{v}_c^\ell} \mathbf{a}_p^\ell (\nabla_{\mathbf{a}_p^\ell} \Phi)^T \Big) \rangle \\ &= \sum_{p \in \mathcal{P}} \underbrace{\langle A_{p,c}^\ell \mathbf{v}_c^\ell \mathbf{W}_O^\ell, \nabla_{\mathbf{a}_p^\ell} \Phi \rangle}_{T_{p,c}^\ell}, \end{aligned} \tag{6}$$

where the last equality uses the linearity of the attention output in the value vectors, *i.e.,* $\nabla_{\mathbf{v}_c^\ell} \mathbf{a}_p^\ell = A_{p,c}^\ell \mathbf{W}_O^\ell$; we place the scalar $A_{p,c}^\ell$ first for readability. In this formulation, the *local component* $A_{p,c}^\ell \mathbf{v}_c^\ell \mathbf{W}_O^\ell$ captures the exact linear contribution of the value vector to the attention output at position p. In contrast, the *downstream component* $\nabla_{\mathbf{a}_p^\ell} \Phi$ accounts for how this localized signal propagates through subsequent nonlinear layers to influence the function Φ.

From I$\times$G to KVgrad (i) Magnitude-Aware Scoring (MAS) The signed score $T_{p,c}^\ell$ depends on the downstream gradient $\nabla_{\mathbf{a}_p^\ell} \Phi$ through both its norm and its direction. The direction is known to be unstable (Smilkov et al., 2017; Ghorbani et al., 2019), which we confirm in Appendix A.4.2. We therefore drop the direction and replace $|T_{p,c}^\ell|$ by its Cauchy–Schwarz upper bound:

$$|T_{p,c}^\ell| \;\le\; A_{p,c}^\ell \left\| \mathbf{v}_c^\ell \mathbf{W}_O^\ell \right\|_2 \left\| \nabla_{\mathbf{a}_p^\ell} \Phi \right\|_2. \tag{7}$$

The bound removes the dependence on the gradient direction while keeping its magnitude, producing a more robust KV cache importance measure.

From I$\times$G to KVgrad (ii) Probe-specific preservation (PSP) While Eq. 7 removes directional cancellation *within* each probe, *i.e.,* across hidden dimensions, summing over all probes can still dilute KV entries that are important only for a subset of probes. To preserve such probe-specific dependencies, we replace $\sum_{p \in \mathcal{P}}$ with the $\max_{p \in \mathcal{P}}$ operator. Combining this PSP with MAS, we obtain the final **KVgrad** score:

$$S_{\ell,c}^{\mathrm{KVgrad}} \;=\; \max_{p \in \mathcal{P}} \; A_{p,c}^\ell \left\| \mathbf{v}_c^\ell \mathbf{W}_O^\ell \right\|_2 \left\| \nabla_{\mathbf{a}_p^\ell} \Phi \right\|_2. \tag{8}$$

Our ablations in Sec. 4.5 show that both MAS and PSP become critical to eviction performance.

Choice of Φ We define $\Phi = \sum_{p \in \mathcal{P}} \|\mathbf{x}_p^{L+1}\|_2^2$ as the summed squared norm of the last-layer hidden states before the final normalization and LM head. Unlike target-dependent output functions (e.g., next-token loss), this choice requires no target token and thus also applies to probe positions holding visual embeddings, which have no vocabulary target. Among the alternatives we tested, this choice performs best especially at high compression ratios (see Appendix A.4.4).

Remark 1 (Heuristic nature of KVgrad and relation to prior work): KVgrad is obtained from the counterfactual difference $\Phi(\mathbf{v}_c^\ell) - \Phi(0)$, corresponding to value-zeroing, and applies a chain of relaxations: a first-order approximation (Eq. 5), MAS (Eq. 7), and PSP. The resulting score is therefore an attribution inspired heuristic rather than an exact attribution methods. Nonetheless, KVgrad generalizes KVzip and KVzip+, both of which omit the downstream sensitivity $\|\nabla_{\mathbf{a}_p^\ell}\Phi\|_2$. This connects existing heuristic eviction scores to attribution-based analysis, providing a lens for understanding KV eviction through the end-to-end influence of each cache entry on the model output.

Remark 2 (Connection to Grad-CAM): Grad-CAM (Selvaraju et al., 2017) scores importance by multiplying a spatially aggregated gradient weight into its activation map. In our context, weighting $\mathbf{a}_p^\ell$ with its aggregated gradient and substituting with Eq. 1 yield $\mathbf{a}_p^\ell \left\|\nabla_{\mathbf{a}_p^\ell}\Phi\right\|_2 = \sum_{c\le p} A_{p,c}^\ell \, \mathbf{v}_c^\ell \mathbf{W}_O^\ell \left\|\nabla_{\mathbf{a}_p^\ell}\Phi\right\|_2$. The ℓ_2 norm of the summand corresponds to our KVgrad score $S_{\ell,c}^{\mathrm{KVgrad}}$ (before $\max$), effectively decomposing the Grad-CAM-style importance into fine-grained contributions of individual cache entries c.

Remark 3 (Max operator): KVzip is designed to preserve entries critical to at least one output token, which corresponds to $\max_{p\in\mathcal{P}} \Phi_p$ in our formulation, where $\Phi_p := \left\|\mathbf{x}_p^{L+1}\right\|_2^2$ is the per-probe target. To avoid per-probe backward pass, KVgrad evaluates importance using the gradient collapsed at the intermediate activation $\mathbf{a}_p^\ell$ during a single pass on $\Phi = \sum_p \Phi_p$, providing a position-localized proxy for per-probe importance.

4 Experimental Results

4.1 Setup

Models and datasets To assess KV cache eviction performance across model scales and families, we evaluate KVgrad on three sizes of Qwen3 (Yang et al., 2025a) (0.6B, 4B, 8B) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Our primary text evaluation uses RULER 4K/16K (Hsieh et al., 2024), comprising 13 synthetic retrieval and tracing subtasks, and LongBench (Bai et al., 2024), consisting of 16 English and code tasks spanning QA and summarization over diverse domains (e.g., Wikipedia, news, academic papers, and code repositories), with contexts typically ranging from 5K to 15K tokens. To demonstrate multimodal generalization, we further test Qwen3-VL-2B (Bai et al., 2025) on image benchmarks (MMBench (Liu et al., 2024b), TextCaps (Sidorov et al., 2020)) and long-video understanding tasks (MLVU (Zhou et al., 2025), VideoMME (Fu et al., 2025)). Detailed information regarding benchmark categories, dataset statistics, and evaluation metrics is provided in Appendix A.1.1.

Baselines and implementation We take KVzip Kim et al. (2025) and KVzip+ Jegou & Jeblick (2026), the two top-ranked query-agnostic methods on the KVpress[4] leaderboard, as primary baselines, together with KVzip+ (MLP), the efficient MLP surrogate trained to regress KVzip+ scores. Following Jegou & Jeblick (2026), we train the surrogate on 2,367 sequences of 750–1,250 tokens sub-sampled from Nemotron-Pretraining-Dataset-sample Basant et al. (2025); KVgrad (MLP) is trained with the identical recipe, which takes $\sim$50 min per 8B model on a single A100 GPU (see Appendix A.2). To cover other eviction paradigms, we also evaluate the query-aware baselines SnapKV Li et al. (2024) and AdaKV Feng et al. (2025a), despite the need for per-query repeated eviction in multi-query scenarios. For image benchmarks, we further compare against the image-centric baseline SparseMM Wang et al. (2025). We implement our method in an extended version of KVpress Devoto et al. (2025) for text and video, and in the SparseMM codebase[5] for image, so that each comparison uses the baselines' reference implementations. Framework extensions and modality-specific configurations are detailed in Appendix A.1.3.

Evaluation We follow the eviction and evaluation procedure of Kim et al. (2025) for all query-agnostic methods (KVzip, KVzip+, KVgrad). After an initial *prefill* on the shared context, we sequentially process the context in 256-token repeat chunks to update the importance score of the corresponding prefilled KV entries. Once scoring is complete, entries with the lowest importance

[4] https://github.com/NVIDIA/kvpress

[5] https://github.com/CR400AF-A/SparseMM

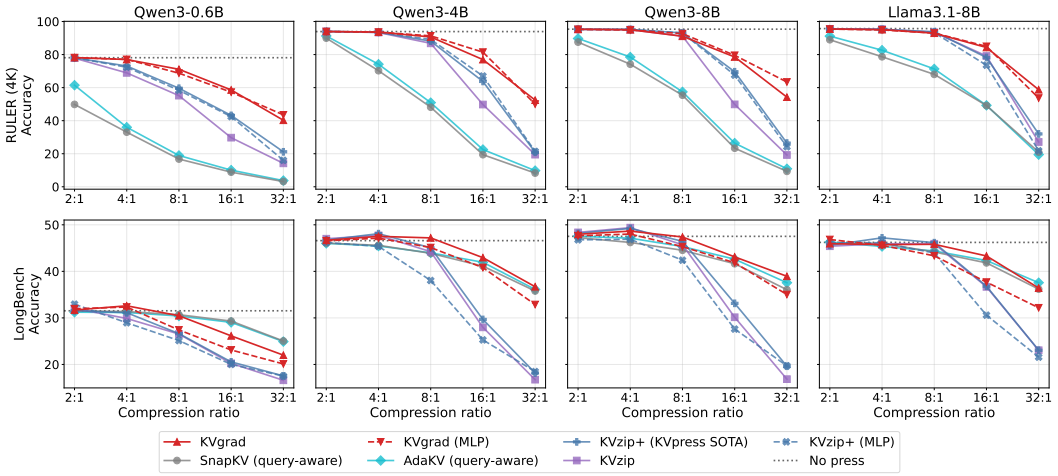


Figure 2: Average accuracy vs. compression ratio across four models (columns) and two benchmarks (rows): RULER 4K (top), LongBench (bottom).

scores are evicted to achieve the target compression ratio (CR). Note that our method is robust to the chunk size (see Appendix A.4.3). For query-aware baselines (SnapKV, AdaKV), we follow their original query-conditioned scoring pipelines. The methods differ in budget allocation granularity: query-agnostic methods allocate budget jointly across layers and heads by following Feng et al. (2025a); Kim et al. (2025), whereas SnapKV and AdaKV assign budgets per head, uniformly and adaptively, respectively.

4.2 Main Results

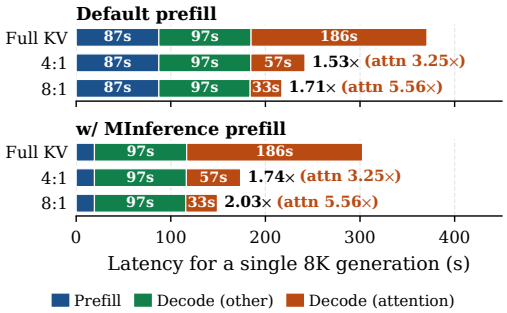
KVgrad outperforms query-agnostic baselines and surpasses query-aware methods in most settings. Figure 2 reports the average accuracy on RULER and LongBench under compression ratios ranging from 2:1–32:1. First, KVgrad establishes new state-of-the-art performance among *query-agnostic* methods in most cases. It achieves near-lossless performance at strong compression ratios (Table 1), and its performance advantage becomes particularly distinct under aggressive compression ratios (Figure 2 16:1–32:1). Second, even when compared to *query-aware* approaches (SnapKV and AdaKV) that use additional query information, KVgrad consistently exhibits higher accuracy across model scales, falling short only on LongBench for Qwen3-0.6B (16:1–32:1) and Llama-3.1-8B (32:1). Finally, this superiority consistently holds across model architectures (Qwen, Llama), scales (0.6B, 4B, 8B), data domains (*e.g.,* Wikipedia, academic papers, code), and context lengths, demonstrating its broad generalizability. In the Appendix A.6, we provide detailed per-task breakdowns for RULER (4K) and LongBench, as well as extended evaluations on RULER (16K).

KVgrad provides a superior distillation target. To isolate the efficacy of our importance score, we train lightweight MLP surrogates for both KVgrad and KVzip+ with an identical, **one-time** pre-training recipe and architecture, so the comparison directly reflects the quality of the teacher scores. As shown in Figure 2, KVgrad (MLP) tracks its exact KVgrad, matches (within 0.5 points) or outperforms KVzip+ (MLP) across all settings, and even exceeds the exact KVzip+ at 16:1 and beyond. On some models (Qwen3-4B, 8B) and needle-retrieval tasks, KVgrad (MLP) even exceeds exact KVgrad, but this is model- and task-specific; elsewhere it falls modestly below, mostly on LongBench (Table 1). Against query-aware methods, KVgrad (MLP) remains superior on RULER, particularly on multi-needle tasks (niah-mk, mq, and mv) where query-driven scores concentrate on a few positions and discard the remaining needles. On LongBench, KVgrad (MLP) performs comparably to AdaKV in most settings, with larger gaps for Qwen3-0.6B and Llama-3.1-8B concentrated on `passage-retrieval` and `code-completion`, where the query directly indicates which tokens to retain.

Appendix A.2.3 further shows that MLP surrogate training generalizes over context length, data domains (code, math, and chinese), and unseen visual modalities (video); as surrogate training is not

Phase	Time (s)	Peak mem. (GiB)
Prefill	87.32±0.13	76.11±0.04
Scoring		
KVzip	268.27±0.02	76.22±0.00
KVzip+	282.10±0.05	76.22±0.00
KVgrad	872.10±9.11	88.00±0.05
AdaKV	0.88±0.02	77.47±0.62
KVgrad (MLP)	0.19±0.00	76.60±0.00

(a) Per-phase cost (wall-clock).



(b) KVgrad (MLP) breakdown (GPU-busy).

Figure 3: **Cost of scoring and decoding on Llama-3.1-8B (256K context). (a)** Per-phase wall-clock latency and peak GPU memory; peak memory includes model weights, activations, and the full KV cache. **(b)** Breakdown of a single 8K generation with KVgrad (MLP) under default and MInference Jiang et al. (2024) prefill. (Single-batch, RTX PRO 6000 Blackwell)

our primary contribution, we defer these experiments there, which collectively establish the KVgrad score as an effective distillation target and the surrogate as a practical deployment option.

4.3 Cost Analysis

Scoring cost and use cases Figure 3(a) reports per-phase wall-clock cost on Llama-3.1-8B at 256K context. Query-agnostic scoring is inherently costlier than query-aware scoring, since the context must be re-processed with repeat chunks, and KVgrad's backward pass makes it the most expensive of all. The two variants therefore trade accuracy for scoring latency:

- **Exact KVgrad** is preferable when a context is prefilled and stored for later use, and memory rather than scoring latency is the bottleneck, e.g., precomputed KV caches for prompts and documents (Gim et al., 2024; Lu et al., 2025). Scoring is then off the request path, and exact KVgrad attains a higher NL-CR than its MLP surrogate in six of eight model–benchmark pairs (Table 1).
- **KVgrad (MLP)** is the better fit when scoring must run at request time: scoring becomes negligible relative to prefill (0.19 s vs. 87 s; below even query-aware AdaKV), with a modest accuracy loss relative to exact KVgrad, mostly on LongBench, while still matching or outperforming KVzip+ (MLP) and query-aware baselines in most settings (Figure 2 and Table 1).

Table 1: **Near-lossless compression ratio** (τ=2). Geometric-mean NL-CR over subtasks (Higher is better); the parenthetical is the change when the learned MLP surrogate replaces it. **Bold** indicates the better of KVgrad and KVzip+.

Model	Method	RULER-4K	LongBench
Qwen3-0.6B	KVzip+	3.36 (−0.19)	3.48 (+0.16)
	KVgrad	**4.49** (−0.25)	**6.06** (−2.06)
Qwen3-4B	KVzip+	5.81 (+0.32)	4.81 (−1.92)
	KVgrad	**7.19** (+0.81)	**5.28** (−0.47)
Qwen3-8B	KVzip+	**7.19** (−0.37)	4.76 (−1.39)
	KVgrad	6.82 (+3.08)	**5.66** (−0.69)
Llama-3.1-8B	KVzip+	6.13 (±0.00)	3.67 (±0.00)
	KVgrad	**8.00** (−0.42)	**4.76** (−0.93)

Decode-time gains Figure 3(b) decomposes a single 8K generation with KVgrad (MLP) in GPU-busy time. Eviction acts directly on the attention kernel, whose latency drops 5.56× (186→33 s) at 8:1; the end-to-end speedup (1.71×) is bounded by the non-attention decode components (green) and prefill (blue), which eviction leaves untouched. With MInference (Jiang et al., 2024) sparse prefill, which shrinks the prefill share, end-to-end speedup rises to 2.03× over Full KV.

4.4 Vision Dataset Results

Figure 4 explores the extension of KVgrad to multimodal settings, comparing it against KVzip, KVzip+, and SparseMM (image only) across image and video benchmarks using Qwen3-VL-2B. The results demonstrate that: (i) a naive application of *repeat instruction prompt*, though designed for text, shows surprising transferability to image and video KV cache compression; and (ii) KVgrad generally

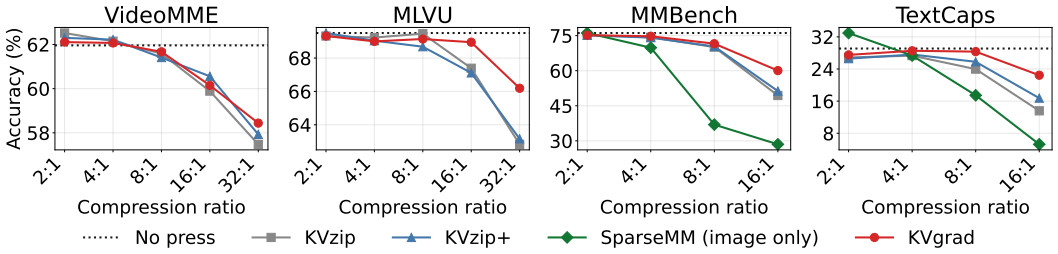


Figure 4: Multimodal dataset results across four datasets on Qwen3-VL-2B.

Table 2: Score-variant ablation study. (Qwen3-4B, RULER)

Variant	PSP	MAS	8:1	16:1	32:1
KVgrad	✓	✓	90.65	**76.96**	**52.44**
w/o MAS	✓	✗	**90.88**	74.68	34.44
w/o PSP	✗	✓	82.75	63.68	29.48
w/o Both	✗	✗	69.72	53.33	25.63

Table 3: Φ ablation study. (Qwen3-4B, RULER-subset)

Target	8:1	16:1	32:1
Last hidden state (**KVgrad**)	**90.79**	**77.25**	**54.94**
After RMSNorm	89.16	74.10	35.42
Full LM-head logits	88.39	63.07	31.27
Answer LM-head logit	90.40	77.21	51.90
Cross-entropy loss	43.97	26.45	15.29

maintains a performance edge, showing competitive results even on challenging benchmarks like VideoMME. These preliminary findings suggest that the principles of global importance are not strictly tied to text and can successfully be adapted to multimodal contexts.

4.5 Ablation Studies

We conduct ablation studies to evaluate the efficacy of Magnitude-Aware Scoring (MAS), Probe-Specific Preservation (PSP), and the impact of probe token chunk size. All experiments are performed using Qwen3-4B on the RULER benchmark.

Magnitude-aware scoring and probe-specific preservation Table 2 describes the individual contributions of MAS and PSP, which constitute the transition from the Taylor approximation form in Eq. 6 to our proposed KVgrad Eq. 8. Our results indicate that omitting the MAS or substituting the $\max$ operator with $\sum$ degrades sharply at aggressive compression. Notably, this performance gap widens significantly as the compression ratio increases.

Choice of Φ We vary the target of Φ from the last hidden state used by KVgrad to the *RMSNorm output*, *full LM-head logits*, *answer logit*, and *cross-entropy (CE) loss* on Qwen3-4B with 10% fraction of RULER-4K dataset. As shown in Table 3, the last hidden state achieves the best performance, while the answer logit remains competitive, supporting our choice of the last hidden state as the target. See Appendix A.4.4 for the formal formulations of these targets and additional ablations.

5 Conclusion

We introduce KVgrad, a query-agnostic KV cache eviction framework that evaluates the importance of cache entries through their global impact on a model's final representation. By factorizing importance into a direct local component and a downstream gradient sensitivity term, KVgrad captures the global sensitivity of the final model output to each cached entry across the entire network depth. Our evaluations on RULER and LongBench demonstrate that KVgrad achieves an average near-lossless compression ratio of $5.9\times$, outperforming existing query-agnostic baselines and even query-aware methods in most settings. Because scoring runs once per context, KVgrad suits settings where the cache is compressed and stored ahead of serving, and a lightweight MLP surrogate distilled from KVgrad brings the same score to request-time compression with negligible scoring overhead. Finally, we provide preliminary evidence that the principles of KVgrad extend to vision-language models, suggesting that global importance-aware compression is a promising direction for both text and multimodal inputs.

AI USE STATEMENT

In this work, we used generative AI tools for helping develop conceptual frameworks, providing feedback on research methodology and experiments, implementing methods, assisting in the writing of proofs, assisting with translation, and interpreting results. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses. Generating synthetic datasets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools for modifying scientific figures, creating and editing software code, identifying relevant literature, summarizing and analyzing existing literature, and editing the paper to improve readability. We have reviewed all AI-assisted work. All figures were first drafted by the authors, and AI assistance was limited to minor formatting adjustments (e.g., width, height, and font size for readability); the final figures were then reviewed by the authors for clarity and consistency with the underlying data. All AI-assisted code, including method implementations, was inspected and tested for correctness by the authors. AI-assisted proofs were independently checked step by step by the authors, and AI-assisted interpretations of results and feedback on methodology were treated as suggestions that the authors validated against the experimental evidence before adoption. All references identified or summarized with AI assistance were checked against the original publications, and every AI-edited passage was reviewed by the authors for factual accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies KV cache compression for LLM inference acceleration and does not involve human subjects, personally identifiable information, or the release of new datasets. All experiments use publicly available pretrained models and benchmarks under their respective licenses.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix A.1 specify the models, benchmarks, compression ratios, and evaluation protocol used in all experiments, including the settings for both query-agnostic and query-aware baselines. Appendix A.2 details the training and inference procedure of the MLP surrogate. Appendix A.1.2 describes the hardware and measurement protocol for the latency analysis. Full per-task results for all experiments are reported in Appendix A.6. Anonymized source code, including scripts to reproduce all tables and figures, is provided in the supplementary material.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Changwoo Baek, Seungjun Shin, and Kyeongbo Kong. Restorekv: Recovering full-cache behavior under aggressive query-agnostic kv cache eviction. *arXiv preprint arXiv:2608.01247*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International conference on machine learning*, pp. 342–350. PMLR, 2017.

Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

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.

Chuangtao Chen, Grace Li Zhang, Xunzhao Yin, Cheng Zhuo, Bing Li, and Ulf Schlichtmann. Kv packet: Recomputation-free context-independent kv caching for llms. *arXiv preprint arXiv:2604.13226*, 2026.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: Kv cache compression by estimating attention from future queries distribution. *arXiv preprint arXiv:2510.00636*, 2025. URL https://arxiv.org/abs/2510.00636.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *arXiv preprint arXiv:2407.11550*, 2024.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *Advances in Neural Information Processing Systems*, 38:113152–113188, 2025a.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Identify critical kv cache in llm inference from an output perturbation perspective. *arXiv preprint arXiv:2502.03805*, 2025b.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33013681. URL https://doi.org/10.1609/aaai.v33i01.33013681.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.

Raghavv Goel, Junyoung Park, Mukul Gagrani, Dalton Jones, Matthew Morse, Harper Langston, Mingu Lee, and Chris Lott. Caote: Kv cache selection for llms via attention output error-based token eviction. *arXiv preprint arXiv:2504.14051*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuzhe Gu, Xiyu Liang, Jiaojiao Zhao, and Enmao Diao. Obcache: Optimal brain kv cache pruning for efficient long-context llm inference. *arXiv preprint arXiv:2510.07651*, 2025.

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? *arXiv preprint arXiv:2404.06654*, 2024.

Simon Jegou and Maximilian Jeblick. Kvzap: Fast, adaptive, and faithful kv cache pruning. *arXiv preprint arXiv:2601.07891*, 2026.

Bo Jiang, Taolue Yang, Youyuan Liu, Xubin He, Sheng Di, and Sian Jin. Packkv: Reducing kv cache memory footprint through llm-aware lossy compression. *arXiv preprint arXiv:2512.24449*, 2025.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W Lee, Sangdoo Yun, and Hyun Oh Song. Kvzip: Query-agnostic kv cache compression with context reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 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.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Bo Li, Xuming Hu, and Xiaowen Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *Advances in Neural Information Processing Systems*, 38:28728–28778, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

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.

Songshuo Lu, Hua Wang, Yutian Rong, Zhi Chen, and Yaohua Tang. Turborag: Accelerating retrieval-augmented generation with precomputed kv caches for chunked text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6599–6612, 2025.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Junyoung Park, Dalton Jones, Matthew Morse, Raghavv Goel, Mingu Lee, and Christopher Lott. Keydiff: Key similarity-based kv cache eviction for long-context llm inference in resource-constrained environments. *Advances in Neural Information Processing Systems*, 38:5983–6019, 2026.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

Akshat Ramachandran, Marina Neseem, Charbel Sakr, Rangharajan Venkatesan, Brucek Khailany, and Tushar Krishna. Thinkv: Thought-adaptive kv cache compression for efficient reasoning models. *arXiv preprint arXiv:2510.01290*, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/shrikumar17a.html`.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pp. 742–758. Springer, 2020.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. Sparsemm: Head sparsity emerges from visual concept responses in mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23177–23187, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jingbo Yang, Bairu Hou, Wei Wei, Yujia Bao, and Shiyu Chang. Kvlink: Accelerating large language models via efficient kv cache reuse. *Advances in Neural Information Processing Systems*, 38: 133797–133824, 2025b.

Jiayi Yao, Samuel Shen, Kuntai Du, Shaoting Feng, Dongjoo Seo, Rui Zhang, Yuyang Huang, Yuhan Liu, Shan Lu, and Junchen Jiang. Vericache: Turning lossy kv cache into lossless llm inference. *arXiv preprint arXiv:2605.17613*, 2026.

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.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Adam Zweiger, Xinghong Fu, Han Guo, and Yoon Kim. Fast kv compaction via attention matching. *arXiv preprint arXiv:2602.16284*, 2026.

A Appendix

A.1 Detailed Experiment Settings

A.1.1 Benchmark specification

Table 4: Summary of benchmarks used for text and multimodal evaluation

Benchmark	Modality	Samples	Key Metrics	Context Range (Avg)
RULER-4K/16K (Hsieh et al., 2024)	Text	6,500	Recall, EM	4,096/16,384 tokens
LongBench (Bai et al., 2024)	Text	3,750	F1, ROUGE-L, EM	5K–15K tokens
MMBench (Liu et al., 2024b)	Image	4,329	MCQ Accuracy	∼192 tokens
TextCaps (Sidorov et al., 2020)	Image	3,166	CIDEr, BLEU-4	∼998 tokens
MLVU (Zhou et al., 2025)	Video	2,174	MCQ Accuracy	3m∼2h (∼25k†)
VideoMME (Fu et al., 2025)	Video	2,700	MCQ Accuracy	11s∼1h (∼26k†)

† Mean vision tokens per instance under our 256-frame sampling configuration.

RULER Comprises 13 synthetic subtasks across four categories: Retrieval (Single/Multi-query/Multi-key Needle-in-a-Haystack), Multi-hop Tracing (variable tracking), Aggregation (unique Words, frequent words), and Question Answering (Single/Multi-hop Reasoning). Following previous work Kim et al. (2025); Jegou & Jeblick (2026) for establishing baseline long-context robustness, we use the 4096-token context length configuration for all tasks. Each task has 500 samples each, which in total has 6500 samples. Evaluation is based on Recall (for retrieval tasks) and Exact Match (for tracking and QA).

LongBench While the full LongBench suite includes bilingual tasks, we evaluate on the 16 English and code tasks following the selection in KVpress (Devoto et al., 2025). The subset comprises 3,750 samples (150∼500 per task) spanning Single-doc QA, Multi-doc QA, Summarization, Few-shot Learning, Synthetic Tasks, and Code Completion. Each task has $150 \sim 500$ samples per task, which in total has 3,750 samples. Context lengths range from 1K to 18K words, with most tasks centered between 5K and 15K tokens. Evaluation metrics include F1-score (for QA), ROUGE-L (for summarization), and Exact Match or Edit Similarity (for code and synthetic tasks).

MMBench A comprehensive evaluation suite designed to assess fine-grained multimodal abilities. We use the MMBench-v1.1 English split, which contains 4,329 evaluation instances (both test and dev split) across 20 ability dimensions. The images in this dataset have a mean resolution of approximately 388×388 pixels ($150,368$ total pixels). Given the 2×2 patch-merging strategy of Qwen3-VL-2B, each image accounts for an average of 192 vision tokens. Evaluation follows the Circular Evaluation strategy, and results are reported as MCQ Accuracy.

TextCaps Focuses on image captioning with reading comprehension, requiring models to interpret written text within its visual context. We evaluate on the validation set, which consists of 3,166 images and 15,830 associated ground-truth captions. Compared to MMBench, this subset has a higher mean resolution of 782,202 pixels (approximately 884×884). Following the 2×2 patch-merging architecture of Qwen3-VL-2B, each image contributes an average of 998 vision tokens to the KV cache. Performance is assessed using standard captioning metrics, specifically CIDEr and BLEU-4.

MLVU A multi-discipline benchmark for long video understanding, featuring videos ranging from 3 minutes to 2 hours. It comprises 9 tasks, including plot question answering, anomaly recognition, and video summarization. We evaluate on the Multiple Choice subset across 2,174 instances (covering 1,122 unique videos). For our evaluation, we uniformly sample up to 256 frames. Given the 2×2 patch-merging strategy of Qwen3-VL-2B, this results in an average of $24,647$ vision tokens per video. Evaluation results are reported as MCQ Accuracy.

VideoMME The first full-spectrum multi-modal evaluation benchmark for video analysis, covering 6 primary domains. It contains 900 videos categorized by duration: Short (<2 min), Medium (4–15 min), and Long (30–60 min). Our evaluation includes $2,700$ questions total. We uniformly sample up to 256 frames with a total pixel constraint of $28,000$ per frame, resulting in an average of $25,767$ vision tokens for the Qwen3-VL-2B model. Evaluation is performed in a multiple-choice format, reporting MCQ Accuracy.

A.1.2 Hardware and Software Environment

We evaluate KVgrad across two primary compute environments. The larger models—including Qwen3-0.6B, Qwen3-4B, Qwen3-8B, Llama-3.1-8B-Instruct—were evaluated on NVIDIA A100 GPU. The multimodal Qwen3-VL-2B model was evaluated on NVIDIA RTX A5000 GPU. Both environments were standardized on PyTorch 2.6.0 and CUDA 12.4 to ensure consistent performance and KV cache handling.

Table 5: Hardware and software specifications

Component	**NVIDIA A100 (80GB)**	**NVIDIA RTX A5000 (24GB)**
CPU	AMD EPYC 74F3	AMD Threadripper PRO 3975WX
System RAM	2.0 TiB	256 GiB
Operating System	Ubuntu 24.04 LTS	Ubuntu 22.04 LTS
PyTorch / CUDA	2.6.0 / 12.4	2.6.0 / 12.4

A.1.3 Implementation details

In this section, we provide additional technical details regarding the implementation of KVgrad and the configuration of our evaluation frameworks across text, image, and video modalities.

Text Tasks For all text-centric evaluations (RULER and LongBench), we use KVpress (Devoto et al., 2025), an open-source library designed for the modular evaluation of KV-cache eviction policies. We implemented KVgrad as a new eviction policy within this framework, enabling a direct and fair comparison against the built-in implementations of KVzip (Kim et al., 2025) and KVzip+ (Jegou & Jeblick, 2026).

Image Tasks For image-based benchmarks (MMBench and TextCaps), we chose to implement our method directly within the codebase of SparseMM Wang et al. (2025), which serves as our primary image-centric baseline. By integrating KVgrad into this existing repository, we ensured that intricate yet critical vision pre-processing configurations—such as visual patch-merging, resolution-dependent attention masking, and the standard 2×2 token reduction logic—remained identical between the baseline and our method. This allows for a rigorous head-to-head comparison of cache-pruning efficiency without confounding factors from different vision-language processing pipelines.

Video Tasks To evaluate long-video understanding on MLVU and VideoMME, we extended the KVpress framework to handle sequential multi-frame inputs. We integrated the data-loading and preprocessing pipeline from VLMEvalKit Duan et al. (2024), following their official protocol for frame extraction and resolution management. Specifically, we uniformly sample up to 256 frames at a rate of 2 FPS, with a total pixel constraint of 28,000 per frame. To ensure the reproducibility of our setup, we verified that our framework—when running without compression ('No press' baseline in 4)—matches the results reported in the Qwen3-VL technical report Bai et al. (2025) within a small margin. All video models were prompted using the standard VLMEvalKit multiple-choice template:

Prompt for MLVU and VideoMME

```
<video>
Select the best answer to the following multiple-choice question
    based on the video.
Respond with only the letter (A, B, C, or D) of the correct option.
```

```
Question: {question}
Possible answer choices: {options}
The best answer is:
```

A.1.4 CHUNKING-BASE COMPRESSION

As described in Section 4.1, our importance scoring use a chunk-based reconstruction strategy. This approach ensures that the KV cache is scored in a query-agnostic manner by forcing the model to attend to and "reconstruct" the context. For the Qwen3 model family, we use the specific chat-template format shown below:

Prompt for chunk-based reconstruction (Qwen3 families)

```
{context}

Repeat the previous context exactly, starting with {previous
    chunk[-8:]} <|im_end|>
<|im_start|>assistant
<think>

</think>

{chunk}
```

A.2 DISTILLATION DETAILS FOR KVGRAD (MLP)

This section provides training and inference details for the KVgrad (MLP). In short, we adopt the KVzap training and evaluation protocol of Jegou & Jeblick (2026) verbatim[6], with the only change being the regression target: the data, features, architecture, optimizer, and hyperparameters are identical, and only the supervision signal is swapped from KVzip+ to KVgrad. Results are reported for both KVgrad (MLP) and KVzip+ (MLP); because the two MLP variants differ only in supervision signal, the gap between them isolates the contribution of the regression target from that of the distillation recipe itself.

A.2.1 METHOD

Computing either score — $S_{\ell,c}^{\text{kvzip+}}$ in Eq. 3 or $S_{\ell,c}^{\text{kvgrad}}$ in Eq. 8 — requires multi-pass forward passes over the probe context (and, for KVgrad, an additional backward pass), which dominates scoring latency (Figure 3 (a)). Following Jegou & Jeblick (2026), we replace this multi-pass scorer with a per-layer model f_θ^ℓ that predicts the score from the cache-token hidden state in a single forward pass:

$$\hat{S}_{\ell,c} \;=\; f_\theta^\ell\big(\mathbf{x}_c^\ell\big), \qquad f_\theta^\ell : \mathbb{R}^{1\times D} \to \mathbb{R} \text{ (single KV head).} \tag{9}$$

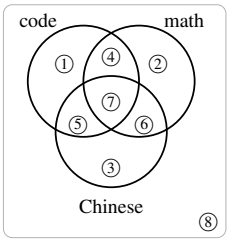
where, for multi-head or grouped-query attention, the output is instead $\mathbb{R}^{H_{\text{kv}}}$ so that one forward pass yields the scores of all H_{kv} KV heads at layer ℓ.

At inference, we keep the top-scoring entries matching the target ratio and evict the rest, pooled across layers and KV heads, so every method is compared at the same cache size. As with KVzip and KVzip+, and KVgrad itself, the first 4 tokens are protected as attention sinks and no recent window is reserved. As shown in Table 7, this inherited rule also outperforms the KVzap default, which protects the recent 128-token window.

A.2.2 IMPLEMENTATION DETAILS

Training data We use the `Nemotron-Pretraining-Dataset-sample` (Basant et al., 2025), the same pretraining mixture as Jegou & Jeblick (2026), drawing from all nine sub-corpora

[6]Reference implementation: `https://github.com/NVIDIA/kvpress/tree/main/kvzap`.



(a) Prompt regions.

Mixture	Removed	①	②	③	④	⑤–⑦	⑧	Total
DEFAULT	—	121	942	58	273	15	913	2,322
CONTROL-1K	—	55	414	21	119	9	382	1,000
ABLATION-1K (CONTROL-1K minus one domain)								
NO-CODE	①④⑤⑦	0	529	24	0	8	439	1,000
NO-MATH	②④⑥⑦	113	0	45	0	1	841	1,000
NO-ZH	③⑤⑥⑦	55	419	0	120	0	406	1,000

(b) Mixture composition (prompt counts per region).

Table 6: Distillation mixtures. (a) Each prompt is flagged as code, math and/or Chinese by regex and LLM judge, giving eight regions; ⑧ is none of the three. (b) DEFAULT is the training split of Section A.2.2 (test prompts excluded) and trains the main-text surrogates. CONTROL-1K is a 1,000-prompt subset of DEFAULT with the same per-sub-corpus proportions, sized to match the ablations. Each ABLATION-1K mixture drops every region touching its domain and is re-sampled to 1,000 with CONTROL-1K's per-sub-corpus proportions, so only the held-out domain differs. In subsequent evaluations, MLP surrogates are denoted by their training mixture using the "MLP-" prefix (*i.e.,* MLP-CON1K is distilled on CONTROL-1K, and MLP-ABL1K is distilled on the respective ABLATION-1K mixture).

(math, code, factual QA, etc.). Prompts whose tokenized length lies outside $[750, 1250]$ are filtered, leaving $\approx 2{,}367$ prompts balanced as 500 train + 5 test per sub-corpus; the $2{,}322$-prompt training split is the DEFAULT mixture of Table 6, on which both KVzip+ (MLP) and KVgrad (MLP) reported throughout the paper are trained. Every prompt is wrapped with the model's chat template so captured hidden states match the evaluation distribution; we sub-sample 500 token positions per prompt with a fixed seed.

Training pairs $(\mathbf{x}_c^\ell,\, S_{\ell,c})$ For each prompt and each layer ℓ, the input feature is the hidden state $\mathbf{x}_c^\ell$ at the sampled cache position c, paired with the score $S_{\ell,c}$ at the same (ℓ, c). Both KVzip+ and KVgrad targets are computed under the repeat-prompt protocol; KVgrad additionally requires one backward pass through the probe context to obtain the gradient term $\nabla_{\mathbf{a}_p^\ell}\Phi$ in Eq. 8. Targets are log-transformed (clamped for numerical stability) to compress their dynamic range.

Table 7: Protected tokens (Qwen3-4B, RULER-4K). Underlined: KVzap default; bold: best within each surrogate.

	Protected		Compression ratio			
Surrogate	rec.	sink	4:1	8:1	16:1	32:1
KVzip+ (MLP)	<u>128</u>	<u>0</u>	93.7	85.2	24.7	12.1
	<u>128</u>	<u>4</u>	**94.6**	89.3	36.5	18.3
	0	0	93.9	**90.8**	63.7	27.1
	0	4	94.4	90.6	**69.1**	**30.6**
KVgrad (MLP)	<u>128</u>	<u>0</u>	94.5	91.6	59.8	16.5
	<u>128</u>	<u>4</u>	94.6	91.6	59.9	18.3
	0	0	**94.7**	**92.3**	83.2	63.2
	0	4	**94.7**	**92.3**	**83.5**	**63.4**

Model and optimization For each layer ℓ, f_θ^ℓ is a two-layer MLP, $\mathbb{R}^{1\times D} \xrightarrow{\text{Linear}} \mathbb{R}^{1\times 512} \xrightarrow{\text{GELU}} \mathbb{R}^{1\times 512} \xrightarrow{\text{Linear}} \mathbb{R}$, producing the scalar score $\hat{S}_{\ell,c}$. For multi-head or grouped-query attention, we extend the final linear to match the number of KV heads, so that the MLP predicts all per-KV-head scores at layer ℓ jointly in a single forward pass. Following Jegou & Jeblick (2026), we optimize with AdamW (learning rate 5×10^{-3}, cosine annealing over 15 epochs, batch size 512, gradient-norm clipping at 1.0, MSE loss).

A.2.3 Additional result: Robustness to train–test shift

Figure 5 tests whether the surrogate still tracks KVgrad, the score it was distilled to predict, when the evaluation distribution differs from the distillation mixture along five axes: context length, modality, math, code and language. Two comparisons guide this interpretation. First, the gap between MLP-CON1K and MLP-ABL1K, which differ only in whether the tested domain is in the mixture, isolates the effect of that domain's absence. Second, the gap from either surrogate to KVgrad reflects the fidelity of the MLP surrogate for KVgrad score regression.

Two observations hold throughout. First, on the three axes with an ablated arm, MLP-CON1K and MLP-ABL1K coincide at every ratio, so removing the evaluated domain from the mixture does not change the surrogate. Second, where the surrogates fall below KVgrad, they do so by the same

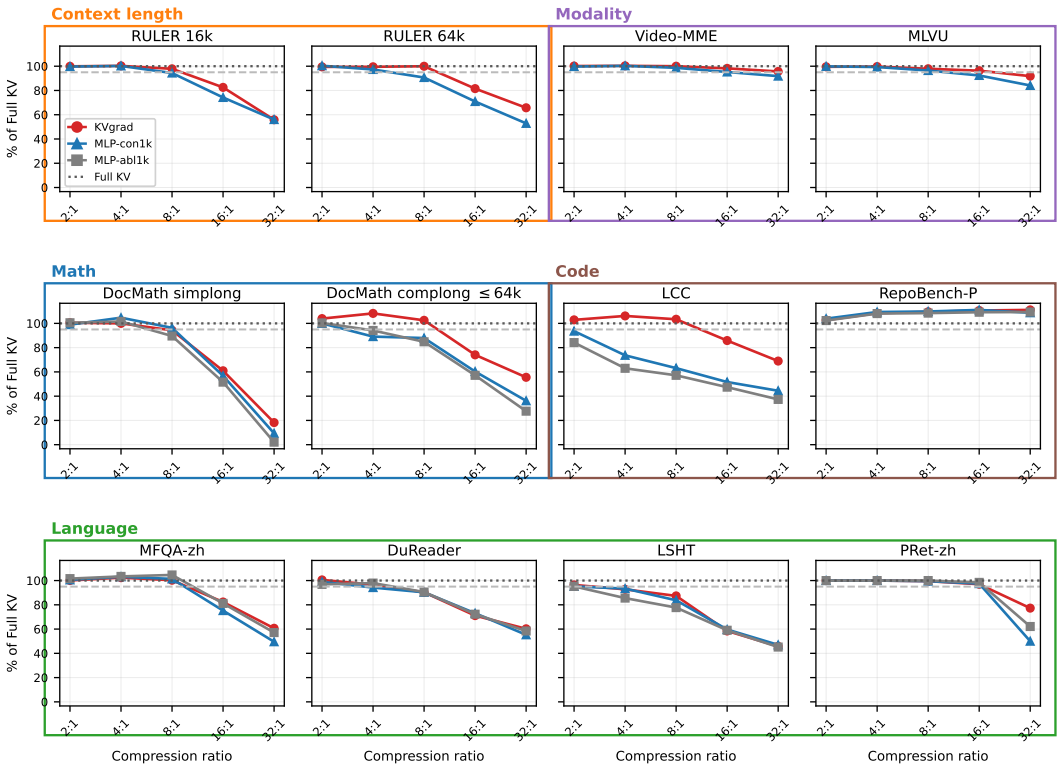


Figure 5: Train–test shift across context length, modality, math, code and language, as accuracy relative to full KV (dotted: full KV; dashed: 95%) against compression ratio. **KVgrad**: score from Eq. 8, computed on the evaluation context. **MLP-CON1K**: surrogate distilled on CONTROL-1K (Table 6). **MLP-ABL1K**: surrogate distilled on the ABLATION-1K mixture with the tested domain removed (NO-MATH, NO-CODE, NO-ZH). Context length and modality have no ablated arm; because all distillation prompts are strictly 750–1,250 tokens of text, evaluating MLP-CON1K on these axes inherently tests out-of-distribution generalization.

margin with and without that domain in the mixture, so the shortfall is a property of the single-pass regression rather than of the mixture. In detail:

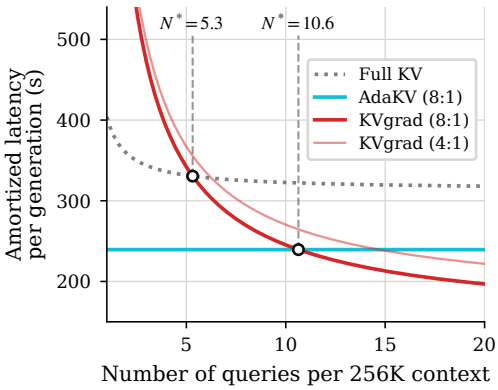
- *Context length* (RULER-16K/64K). MLP-CON1K, distilled on $\leq$ 1.25k-token prompts, stays within a few points of KVgrad at every ratio at both lengths; all three degrade together at 16:1 and 32:1.
- *Modality* (VideoMME, MLVU). Distilled on text only, MLP-CON1K stays within a few points of KVgrad on visual tokens at every ratio.
- *Language* (MultiFieldQA-zh, DuReader, LSHT, PassageRetrieval-zh). The two surrogates coincide at every ratio and stay within a few points of KVgrad, except at 32:1 on PassageRetrieval-zh where both fall below it.
- *Math* (DocMath). The two surrogates coincide; both sit a few points below KVgrad at 16:1 and 32:1.
- *Code* (LCC, RepoBench-P). The two surrogates coincide. On LCC both fall below KVgrad from 8:1 on, the largest margin in the grid; on RepoBench-P all three stay above 95% of full KV through 32:1.

A.3 KVGRAD AMORTIZATION ANALYSIS

We characterize the cost of KVgrad through a controlled profiling study. Our goal is to illustrate how costs scale with the number of queries sharing a context. Because KVgrad is query-agnostic, prefill and scoring run once per context, and only the compressed cache is retained across queries, shrinking

Phase	Time (s)	Peak mem. (GiB)
Prefill	87.32±0.13	76.11±0.04
Scoring		
KVgrad	872.10±9.11	88.00±0.05
AdaKV	0.88±0.02	77.47±0.62
KVgrad (MLP)	0.19±0.00	76.60±0.00
Decode	**Time (ms/tok)**	**KV cache (GiB)**
Full KV	38.28±0.03	32.0
4:1	21.18±0.29	8.0
8:1	18.12±0.28	4.0

(a) Per-phase latency and memory.



(b) Amortized latency vs. num. of queries.

Figure 6: **Cost profile and crossover scaling for Llama-3.1-8B (256K context, 8K generation). (a)** Wall-clock latency and GPU memory per phase; peak mem. includes model weights, activations, and the full KV cache, while KV cache is the persistent cache retained for decoding. **(b)** Amortized latency per generation of exact KVgrad versus the number of queries sharing a context

the persistent KV cache at 256K context from 32 GiB to 8 GiB (4:1) and 4 GiB (8:1) (Table 6a). Query-aware caches (*e.g.,* AdaKV), in contrast, are query-specific: under the same cache memory budget, they must re-run prefill and scoring for every query. The amortized latency per generation over N queries is thus

$$T(N) = \begin{cases} \text{Prefill}/N + \text{Decode}_{\text{full}} & \text{(Full KV)}, \\ (\text{Prefill} + \text{Scoring})/N + \text{Decode}_{\text{comp}} & \text{(KVgrad)}, \\ \text{Prefill} + \text{Scoring} + \text{Decode}_{\text{comp}} & \text{(query-aware)}. \end{cases}$$

KVgrad's one-time scoring cost is absorbed after $N^* = 5.3$ queries against Full KV at 8:1 (6.3 at 4:1) and $N^* = 10.6$ against AdaKV (Figure 6b). Compression does not lower the peak memory during prefill (76 GiB), and exact scoring adds 12 GiB to it.

A.4 Analysis and Ablation Studies

A.4.1 Score Visualization

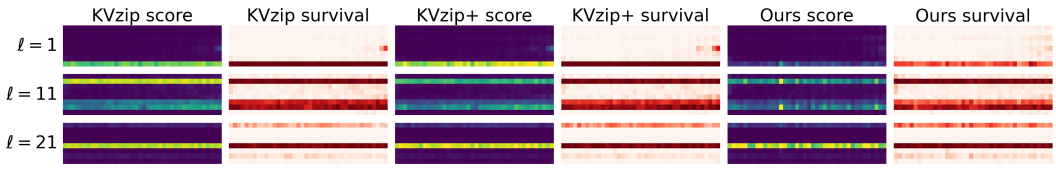


Figure 7: Importance score and survived KV entries of KVzip, KVzip+, and KVgrad

Figure 7 show the importance score of KVzip, KVzip+, and KVgrad. While all three method show similar eviction pattern overall, KVgrad assigns lower importance to the first layer compared to the others. We attribute this difference to KVzip+, which appears to overestimate the importance of lower layers due to the application of $1/\|\mathbf{x}_p^\ell\|_2$.

A.4.2 Sensitivity of Gradient-Based Scores to Input Noise

Section 3.2 notes that the gradient signal can be noisy. To measure how this noise propagates to eviction decisions, we add isotropic Gaussian noise to the replay input embeddings, with the noise level ρ set as a fraction of the mean embedding norm, and report the overlap between the kept sets

Table 8: Ranking stability on Qwen3-8B, RULER-4K

ρ	Score	Kept-set overlap 8:1	16:1	32:1
1%	A only (ref.)	0.987 ± 0.004	0.984 ± 0.006	0.981 ± 0.007
	Signed $\lvert T_{p,c}\rvert$	0.954 ± 0.011	0.938 ± 0.015	0.921 ± 0.021
	KVgrad	0.969 ± 0.008	0.959 ± 0.012	0.948 ± 0.016
5%	A only (ref.)	0.985 ± 0.005	0.983 ± 0.008	0.979 ± 0.008
	Signed $\lvert T_{p,c}\rvert$	0.949 ± 0.011	0.934 ± 0.016	0.916 ± 0.023
	KVgrad	0.966 ± 0.008	0.954 ± 0.012	0.943 ± 0.017
20%	A only (ref.)	0.979 ± 0.010	0.976 ± 0.013	0.972 ± 0.014
	Signed $\lvert T_{p,c}\rvert$	0.917 ± 0.015	0.893 ± 0.022	0.867 ± 0.031
	KVgrad	0.940 ± 0.013	0.918 ± 0.021	0.897 ± 0.028

selected before and after perturbation at three budgets. Attention-only scoring, which has no gradient term, serves as a reference.

Table 8 shows that the attention-only reference is the most stable at every setting, so the gradient term is the source of most of the sensitivity. Among the two scores that retain it, MAS is consistently more stable than the signed score, by 2.3–3.0 points, with the largest margin at 32:1. Dropping the alignment factor thus recovers part of the stability lost to the gradient term, while the rmaining gap to the attention-only reference is due to the gradient norm, which MAS retains.

A.4.3 Ablation on Chunk Size

Table 9: Ablation study on chunk size (Qwen3-8B, RULER-4K)

Chunk size	2:1	4:1	8:1	16:1	32:1
256 (KVgrad)	**95.31**	**95.05**	**91.08**	**78.55**	54.24
512	95.28	94.84	90.36	78.46	**54.87**
1024	95.25	94.83	90.08	77.24	53.72

Table 9 shows that KVgrad is insensitive to the chunk size: across chunk sizes from 256 to 1024, accuracy varies by at most 1.3 points across all compression ratios, demonstrating high robustness to this hyperparameter. We therefore use a chunk size of 256 for all main experiments without extensive tuning.

A.4.4 Φ Ablation

We conduct an ablation study on the Φ objective by varying both the target representation and its scalarization. For each probe position p, we define a per-position scalar objective Φ_p and compute gradients from $\Phi = \sum_{p \in \mathcal{P}} \Phi_p$. We consider five targets: (i) the last hidden state $\mathbf{x}_p^{L+1}$ before the final RMSNorm, (ii) the normalized hidden state after RMSNorm $\bar{\mathbf{x}}_p^{L+1}$, (iii) the output logits $\mathbf{l}_p$, (iv) the logit of the ground-truth next token $l_{p,\pi(p)}$, where $\pi(p)$ denotes the next token at position p, and (v) the next-token cross-entropy loss. For the vector-valued targets (i)–(iii), we compare three scalarizations: the squared ℓ_2 norm, the ℓ_1 norm, and the element-wise sum; for the answer logit (iv), we compare its signed, absolute, and squared values.

As shown in Table 10, all objectives perform similarly at mild compression, but they diverge at high compression ratios. The last hidden state with the squared ℓ_2 norm achieves the best accuracy at 8:1 and 32:1. The answer-logit variants remain competitive at high compression ratios, but they require the LM head and ground-truth tokens, which are unavailable at visual probe positions. Moreover, logit-based targets require computing and backpropagating through the LM head over the full vocabulary, which incurs additional computation and memory compared with the last hidden state. We therefore adopt the squared ℓ_2 norm of the last hidden state as Φ.

Table 10: Ablation study of the Φ objective on the 10% fraction of RULER (4K) using Qwen3-4B.

ϕ_p objective	2:1	4:1	8:1	16:1	32:1
$\|\mathbf{x}_p^{L+1}\|_2^2$	93.50	92.75	**90.79**	77.25	**54.94**
$\|\mathbf{x}_p^{L+1}\|_1$	93.64	**93.25**	90.04	**77.90**	40.15
$\mathbf{1}^\top \mathbf{x}_p^{L+1}$	**93.66**	92.63	90.61	73.28	44.86
$\|\bar{\mathbf{x}}_p^{L+1}\|_2^2$	93.45	92.43	89.16	74.10	35.42
$\|\bar{\mathbf{x}}_p^{L+1}\|_1$	93.35	92.42	89.75	74.83	52.77
$\mathbf{1}^\top \bar{\mathbf{x}}_p^{L+1}$	93.24	92.91	90.70	73.57	38.69
$\|\mathbf{l}_p\|_2^2$	93.43	92.78	88.39	63.07	31.27
$\|\mathbf{l}_p\|_1$	93.38	92.76	89.32	74.94	39.49
$\mathbf{1}^\top \mathbf{l}_p$	93.45	92.84	89.35	75.93	37.81
$l_{p,\pi(p)}$	93.35	92.77	89.85	77.42	49.54
$\|l_{p,\pi(p)}\|$	93.35	92.83	90.24	77.52	49.39
$l_{p,\pi(p)}^2$	93.40	92.72	90.40	77.21	51.90
$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(l_{p,\pi(p)})}{\sum_{v\in\mathcal{V}} \exp(l_{p,v})}$	80.11	59.98	43.97	26.45	15.29

A.5 Limitations and Future Work

Despite the novelty of the proposed Taylor-based compression algorithm and its strong empirical performance, several limitations remain to be addressed in future research.

Contextual redundancy The Taylor-based approach focuses on individual token importance but does not explicitly account for information redundancy within the context. This can lead to suboptimal compression where important but redundant information is retained. Addressing this, particularly for video datasets containing temporally redundant frames, could significantly enhance the compression ratio.

Implementation efficiency Our current implementation involves redundant computations of the attention matrix. Transitioning to a specialized kernel-level implementation would drastically improve throughput. Future iterations should explore deep-level optimizations, such as intervening in the backpropagation process and integration with high-performance inference frameworks like vLLM.

Evaluation scope in multimodality While we adhered to standard Large Language Model (LLM) benchmarks, our evaluation on Vision-Language Models (VLMs) remains preliminary. Although we have demonstrated the feasibility of our method across different modalities, more extensive benchmarking on diverse image and video datasets, including the comparison with the SOTA baseline methods, is required to fully validate its generalizability.

Future works Several promising directions for future research emerge from this study. By integrating our method with KV-reuse settings, we can investigate hybrid strategies for KV compression, offloading, and re-entry. This includes handling KV pairs from independent sources that may contain causally independent compressed data. Furthermore, applying Explainable AI (XAI) techniques to compute end-to-end importance scores represents a compelling avenue for achieving more interpretable and efficient context management.

A.6 Full Experimental Results

A.6.1 RULER Results - Qwen3 0.6B

Table 11: Per-task RULER accuracy and NL-CR for Qwen3-0.6B (1/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

			Accuracy (per CR)					**NL-CR (per τ)**	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
	SnapKV		**1.2**/0.2	2.8/0.2	3.9/0.1	1.3/0.0	1.2/0.1	**32:1**/4:1	**32:1**/4:1
	AdaKV		0.5/**0.3**	**4.1/0.3**	**6.8**/0.1	**3.1**/0.1	1.3/0.0	**32:1**/4:1	**32:1**/4:1
	KVzip		0.4/0.2	0.3/0.2	0.3/**0.2**	0.6/0.2	0.4/0.1	**32:1**/16:1	**32:1**/16:1
cwe	KVzip+	0.4/0.2	0.5/0.2	0.4/0.2	0.6/**0.2**	0.5/**0.3**	0.5/**0.4**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		0.4/0.2	0.3/0.2	0.3/**0.2**	0.3/0.2	0.3/0.1	2:1/16:1	2:1/16:1
	KVgrad		0.4/0.2	0.4/0.2	1.1/**0.2**	1.9/0.2	0.4/0.1	16:1/16:1	16:1/16:1
	KVgrad (MLP)		0.4/0.2	0.4/0.2	1.5/**0.2**	0.7/**0.3**	**1.5**/0.2	**32:1/32:1**	**32:1/32:1**
	SnapKV		71.1/85.7	56.2/86.7	35.5/77.7	17.9/56.3	6.1/41.0	1:1/4:1	1:1/4:1
	AdaKV		75.1/86.7	67.2/87.7	52.1/83.3	28.3/72.3	7.3/**57.0**	1:1/4:1	**2:1/8:1**
	KVzip		77.1/88.3	**73.8**/88.3	60.6/81.3	40.4/58.0	23.1/23.0	1:1/4:1	**2:1**/4:1
fwe	KVzip+	79.1/87.7	**77.2**/88.0	73.1/**90.0**	**63.7**/85.3	**52.1**/64.7	34.1/41.7	1:1/4:1	**2:1/8:1**
	KVzip+ (MLP)		77.1/88.0	72.5/83.7	63.4/79.0	47.3/59.3	24.3/8.0	1:1/2:1	**2:1**/4:1
	KVgrad		**77.2/88.7**	72.6/89.3	62.6/**86.3**	50.1/**77.3**	36.5/54.7	1:1/**8:1**	**2:1/8:1**
	KVgrad (MLP)		76.8/88.3	72.3/85.3	61.7/80.0	47.7/63.0	**41.6**/55.0	1:1/2:1	**2:1**/4:1
	SnapKV		77.2/74.0	40.2/55.0	14.0/25.0	9.6/11.0	5.0/10.0	1:1/1:1	1:1/1:1
	AdaKV		80.6/84.0	29.2/56.0	13.4/22.0	8.8/13.0	5.4/10.0	1:1/1:1	1:1/2:1
	KVzip		96.4/**87.0**	79.2/41.0	60.6/44.0	27.6/14.0	2.0/0.0	2:1/2:1	2:1/2:1
niah-mk1	KVzip+	97.4/86.0	97.2/86.0	84.4/45.0	65.6/43.0	43.6/32.0	8.0/12.0	2:1/2:1	2:1/2:1
	KVzip+ (MLP)		98.0/86.0	85.6/47.0	72.4/40.0	32.2/24.0	4.4/9.0	2:1/2:1	2:1/2:1
	KVgrad		97.4/**87.0**	96.0/**88.0**	85.6/43.0	68.6/**40.0**	**44.4/31.0**	**4:1/4:1**	**4:1/4:1**
	KVgrad (MLP)		**98.2**/86.0	**97.6**/83.0	**89.4/55.0**	**71.4**/34.0	43.4/23.0	**4:1**/2:1	**4:1/4:1**
	SnapKV		49.0/25.0	21.6/15.0	7.6/10.0	1.8/4.0	0.2/2.0	1:1/1:1	1:1/1:1
	AdaKV		78.8/38.0	38.2/22.0	15.2/13.0	3.2/7.0	0.4/**3.0**	1:1/1:1	1:1/1:1
	KVzip		**92.4/57.0**	70.0/**58.0**	35.4/13.0	7.6/0.0	0.6/0.0	2:1/**4:1**	2:1/**4:1**
niah-mk2	KVzip+	91.6/55.0	92.2/56.0	90.0/55.0	43.8/17.0	8.0/2.0	0.6/0.0	**4:1/4:1**	**4:1/4:1**
	KVzip+ (MLP)		92.0/53.0	86.4/21.0	31.2/9.0	2.6/0.0	1.4/0.0	2:1/1:1	2:1/2:1
	KVgrad		92.2/**57.0**	**92.8**/56.0	**84.2/50.0**	**45.4/17.0**	**6.0**/1.0	**4:1/4:1**	**4:1/4:1**
	KVgrad (MLP)		**92.4**/53.0	91.0/48.0	62.0/16.0	32.0/1.0	4.0/0.0	**4:1**/1:1	**4:1**/2:1
	SnapKV		26.6/6.0	1.8/1.0	0.4/1.0	0.0/0.0	0.0/**0.0**	1:1/1:1	1:1/1:1
	AdaKV		60.0/10.0	4.6/5.0	0.6/3.0	0.0/1.0	0.0/**0.0**	1:1/1:1	1:1/1:1
	KVzip		90.6/**13.0**	45.4/6.0	21.2/4.0	4.4/1.0	0.0/**0.0**	**2:1/2:1**	2:1/**2:1**
niah-mk3	KVzip+	90.0/12.0	90.2/**13.0**	65.4/6.0	20.4/5.0	5.2/0.0	0.4/**0.0**	**2:1/2:1**	2:1/**2:1**
	KVzip+ (MLP)		90.6/10.0	59.4/5.0	21.6/1.0	6.6/0.0	0.0/**0.0**	**2:1**/1:1	2:1/1:1
	KVgrad		**90.8**/12.0	**85.6/8.0**	**63.0/6.0**	**20.4/4.0**	2.6/**0.0**	**2:1/2:1**	**4:1/2:1**
	KVgrad (MLP)		90.4/9.0	83.8/5.0	43.4/1.0	14.0/0.0	**3.8/0.0**	**2:1**/1:1	2:1/1:1
	SnapKV		93.8/89.8	75.0/69.5	23.6/41.8	10.2/19.0	3.5/9.2	1:1/1:1	2:1/1:1
	AdaKV		94.5/93.2	57.0/76.5	14.7/43.8	9.2/16.2	5.2/9.5	1:1/2:1	2:1/2:1
	KVzip		97.8/94.5	95.2/81.8	83.7/71.0	37.0/26.0	2.9/1.2	2:1/2:1	4:1/2:1
niah-mq	KVzip+	97.9/94.5	98.0/94.0	96.5/85.5	86.1/77.5	68.6/46.5	11.5/15.0	4:1/2:1	4:1/2:1
	KVzip+ (MLP)		98.1/94.0	95.5/79.0	91.5/83.0	63.1/51.5	3.5/13.0	2:1/2:1	4:1/2:1
	KVgrad		**98.2/94.8**	98.0/**94.0**	96.3/82.8	90.8/**75.2**	72.0/**58.5**	**8:1/4:1**	**8:1/4:1**
	KVgrad (MLP)		98.2/94.2	**98.2**/93.0	**97.0/85.0**	**91.5**/73.8	**76.5**/50.5	**8:1/4:1**	**8:1/4:1**
	SnapKV		60.8/61.0	25.5/26.8	10.1/16.0	7.5/9.8	2.5/5.8	1:1/1:1	1:1/1:1
	AdaKV		49.5/80.5	16.1/32.2	9.5/15.8	7.3/9.0	2.9/6.0	1:1/1:1	1:1/1:1
	KVzip		88.8/86.8	75.7/63.2	57.2/55.0	22.1/25.5	2.1/2.2	1:1/2:1	2:1/2:1
niah-mv	KVzip+	90.8/86.5	89.5/87.2	79.8/69.2	62.1/54.5	35.7/38.2	9.6/15.0	**2:1**/2:1	2:1/2:1
	KVzip+ (MLP)		89.2/**89.5**	76.0/63.0	62.5/56.5	40.3/43.5	2.2/12.0	**2:1**/2:1	2:1/2:1
	KVgrad		90.1/85.2	**87.2**/85.0	81.1/67.0	**72.3/65.2**	56.7/**55.0**	**2:1/4:1**	**4:1/4:1**
	KVgrad (MLP)		**91.0**/86.0	86.3/**85.5**	**81.8/68.2**	69.8/58.5	**66.9**/51.8	**2:1/4:1**	**4:1/4:1**

Table 12: Per-task RULER accuracy and NL-CR for Qwen3-0.6B (2/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

		Accuracy (per CR)						NL-CR (per τ)	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
niah-s1	SnapKV	100.0/100.0	40.0/89.0	26.2/70.0	16.4/42.0	7.0/27.0	0.8/18.0	1:1/1:1	1:1/1:1
	AdaKV		97.4/**100.0**	70.0/97.0	32.0/84.0	8.2/60.0	0.6/27.0	1:1/2:1	2:1/4:1
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	99.8/**100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0**/96.0	**32:1**/16:1	**32:1/32:1**
niah-s2	SnapKV	100.0/100.0	88.8/93.0	60.2/75.0	22.0/44.0	3.8/24.0	1.2/9.0	1:1/1:1	1:1/1:1
	AdaKV		94.6/99.0	59.8/82.0	12.4/53.0	2.4/17.0	1.2/4.0	1:1/2:1	1:1/2:1
	KVzip		**100.0/100.0**	99.0/99.0	98.6/93.0	20.8/11.0	4.8/0.0	**8:1**/4:1	8:1/4:1
	KVzip+		**100.0/100.0**	99.6/**100.0**	**100.0**/96.0	72.8/59.0	16.8/9.0	**8:1**/4:1	8:1/**8:1**
	KVzip+ (MLP)		99.2/**100.0**	**99.8/100.0**	**100.0**/97.0	**97.4**/80.0	11.0/6.0	**8:1**/4:1	**16:1/8:1**
	KVgrad		**100.0/100.0**	99.6/**100.0**	99.8/**99.0**	93.0/77.0	60.2/52.0	**8:1/8:1**	8:1/**8:1**
	KVgrad (MLP)		**100.0/100.0**	99.4/**100.0**	99.2/**99.0**	97.0/**81.0**	**82.6/61.0**	**8:1/8:1**	**16:1/8:1**
niah-s3	SnapKV	99.4/99.0	9.0/9.0	3.0/5.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	AdaKV		14.2/36.0	2.6/6.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	KVzip		**99.4/99.0**	99.0/**99.0**	94.6/98.0	58.2/38.0	2.0/3.0	4:1/**8:1**	8:1/8:1
	KVzip+		**99.4/99.0**	**99.4/99.0**	95.2/97.0	85.0/76.0	42.6/31.0	4:1/4:1	8:1/8:1
	KVzip+ (MLP)		**99.4/99.0**	**99.4/99.0**	99.2/**99.0**	84.0/81.0	18.2/26.0	8:1/**8:1**	8:1/8:1
	KVgrad		**99.4/99.0**	**99.4/99.0**	98.8/98.0	96.6/**96.0**	56.8/**82.0**	8:1/**8:1**	**16:1/16:1**
	KVgrad (MLP)		**99.4/99.0**	**99.4/99.0**	**99.4/99.0**	**98.8/96.0**	**62.4**/45.0	**16:1/8:1**	**16:1/16:1**
qa1	SnapKV	58.2/43.0	56.4/43.0	54.0/**46.0**	47.8/**42.0**	30.8/**36.0**	7.0/**32.0**	1:1/**4:1**	2:1/**8:1**
	AdaKV		56.4/42.0	52.4/43.0	**48.0/42.0**	**33.4**/34.0	8.6/29.0	1:1/**4:1**	2:1/**8:1**
	KVzip		58.4/44.0	56.0/30.0	37.6/14.0	18.2/10.0	7.0/5.0	**2:1**/2:1	**4:1**/2:1
	KVzip+		**59.4**/44.0	**56.8**/29.0	43.4/16.0	23.6/12.0	8.8/7.0	**2:1**/2:1	**4:1**/2:1
	KVzip+ (MLP)		59.0/44.0	56.0/27.0	35.4/14.0	19.8/6.0	6.2/6.0	**2:1**/2:1	**4:1**/2:1
	KVgrad		58.6/43.0	**56.8**/42.0	47.8/24.0	28.0/14.0	14.4/10.0	**2:1**/2:1	**4:1**/4:1
	KVgrad (MLP)		57.6/**45.0**	**56.8**/42.0	46.6/27.0	28.6/11.0	**15.2**/6.0	**2:1**/2:1	**4:1**/4:1
qa2	SnapKV	33.4/31.0	33.0/32.0	30.0/31.0	27.8/**30.0**	20.0/**29.0**	11.0/23.0	2:1/**4:1**	2:1/**8:1**
	AdaKV		32.4/32.0	28.6/**33.0**	26.8/**30.0**	**21.4**/27.0	12.6/**26.0**	1:1/**4:1**	2:1/**8:1**
	KVzip		33.4/30.0	28.2/19.0	19.8/19.0	14.0/15.0	9.6/8.0	2:1/1:1	2:1/2:1
	KVzip+		33.0/30.0	28.4/27.0	21.4/19.0	15.8/20.0	9.4/9.0	2:1/1:1	2:1/2:1
	KVzip+ (MLP)		33.6/**33.0**	28.4/25.0	21.4/18.0	14.4/14.0	9.8/11.0	2:1/2:1	2:1/2:1
	KVgrad		**33.8**/31.0	**33.6**/29.0	26.8/25.0	20.4/19.0	**14.2**/12.0	**4:1**/2:1	**4:1**/2:1
	KVgrad (MLP)		32.4/30.0	32.2/32.0	**28.2**/23.0	17.8/17.0	11.2/13.0	1:1/**4:1**	**4:1**/4:1
vt	SnapKV	77.3/79.8	41.0/59.8	33.3/47.4	6.8/27.0	2.8/11.6	0.3/0.6	1:1/1:1	1:1/1:1
	AdaKV		65.0/76.2	38.6/63.8	12.2/52.2	3.1/10.8	0.7/0.2	1:1/1:1	1:1/2:1
	KVzip		77.1/80.4	74.2/79.6	47.4/74.0	35.3/60.6	29.8/46.0	2:1/4:1	4:1/4:1
	KVzip+		78.1/79.0	76.1/79.4	72.9/77.4	50.9/75.6	32.8/56.8	4:1/4:1	4:1/8:1
	KVzip+ (MLP)		79.1/**80.6**	81.0/77.0	60.2/64.8	43.9/63.6	26.9/34.6	4:1/2:1	4:1/4:1
	KVgrad		78.5/80.4	79.0/81.4	77.1/**79.2**	74.6/**78.0**	**59.2/69.4**	**8:1/8:1**	**16:1/16:1**
	KVgrad (MLP)		**79.3**/80.4	**84.6/81.6**	**83.4**/69.2	**74.9**/59.6	56.6/21.0	**8:1**/4:1	**16:1**/4:1

A.6.2 RULER RESULTS - QWEN3 4B

Table 13: Per-task RULER accuracy and NL-CR for Qwen3-4B (1/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

			Accuracy (per CR)					**NL-CR (per τ)**	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
cwe	SnapKV	94.1/69.6	92.5/68.5	81.2/62.5	53.6/45.0	31.1/21.8	17.3/7.3	2:1/2:1	2:1/2:1
	AdaKV		93.3/68.3	86.4/62.2	62.6/43.5	35.7/22.4	20.4/7.5	2:1/2:1	2:1/2:1
	KVzip		94.4/70.7	93.5/72.4	79.2/56.3	41.9/15.2	0.5/0.2	**4:1/4:1**	**4:1/4:1**
	KVzip+		94.5/**71.6**	**94.3**/72.8	85.1/64.2	64.5/36.6	2.5/0.4	**4:1/4:1**	**4:1/4:1**
	KVzip+ (MLP)		94.2/70.5	92.8/72.5	78.5/53.1	35.1/8.1	1.1/0.3	**4:1/4:1**	**4:1/4:1**
	KVgrad		**94.6/71.6**	93.9/**75.2**	**88.2/66.0**	**79.2/42.3**	**55.3/9.2**	**4:1/4:1**	**4:1/4:1**
	KVgrad (MLP)		93.9/70.8	91.8/71.2	84.1/57.6	54.6/14.5	15.4/0.6	2:1/**4:1**	**4:1/4:1**
fwe	SnapKV	87.4/95.7	83.8/**97.7**	79.5/**98.0**	68.4/**99.0**	53.8/96.3	19.1/89.0	1:1/**16:1**	2:1/16:1
	AdaKV		85.5/96.7	82.5/97.3	73.7/98.3	60.3/**98.7**	22.7/**93.0**	1:1/**16:1**	2:1/**32:1**
	KVzip		87.7/96.3	86.0/96.0	78.3/92.7	45.4/54.3	0.0/0.0	**4:1**/4:1	**4:1**/8:1
	KVzip+		87.8/95.7	86.4/95.7	78.0/94.0	66.1/78.3	0.0/0.7	**4:1**/8:1	**4:1**/8:1
	KVzip+ (MLP)		**87.9**/96.0	**87.8**/92.0	78.2/90.0	69.2/34.0	0.2/0.0	**4:1**/2:1	**4:1**/4:1
	KVgrad		87.8/95.7	85.1/95.7	78.6/97.0	**73.6**/93.7	**61.1**/75.0	2:1/8:1	**4:1**/16:1
	KVgrad (MLP)		87.7/96.3	86.5/95.0	**80.8**/91.0	71.7/70.7	10.0/1.0	**4:1**/4:1	**4:1**/8:1
niah-mk1	SnapKV	100.0/99.0	**100.0**/99.0	97.0/**100.0**	54.0/99.0	10.2/**93.0**	5.2/54.0	2:1/**8:1**	4:1/**8:1**
	AdaKV		**100.0**/99.0	92.0/99.0	37.4/**100.0**	10.4/89.0	6.8/42.0	2:1/**8:1**	2:1/**8:1**
	KVzip		**100.0**/99.0	99.6/99.0	93.4/85.0	51.4/49.0	0.8/5.0	4:1/4:1	4:1/4:1
	KVzip+		**100.0**/99.0	**100.0**/99.0	97.4/95.0	60.4/67.0	3.6/13.0	4:1/4:1	8:1/**8:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0**/97.0	99.2/93.0	67.0/47.0	3.8/13.0	**8:1**/2:1	8:1/4:1
	KVgrad		**100.0**/99.0	**100.0**/99.0	**99.6**/99.0	86.6/89.0	44.0/53.0	**8:1/8:1**	8:1/**8:1**
	KVgrad (MLP)		**100.0**/99.0	**100.0**/99.0	99.2/98.0	**95.0**/89.0	**67.2/63.0**	**8:1/8:1**	**16:1/8:1**
niah-mk2	SnapKV	100.0/100.0	77.0/89.0	43.6/59.0	19.6/32.0	6.0/13.0	1.4/6.0	1:1/1:1	1:1/1:1
	AdaKV		95.0/99.0	67.8/85.0	35.2/55.0	10.8/21.0	2.0/10.0	1:1/2:1	2:1/2:1
	KVzip		**100.0**/99.0	**100.0**/99.0	96.8/**96.0**	77.6/38.0	9.0/0.0	4:1/**4:1**	**8:1/8:1**
	KVzip+		**100.0**/99.0	**100.0/100.0**	94.0/93.0	**89.8**/55.0	4.6/1.0	4:1/**4:1**	4:1/4:1
	KVzip+ (MLP)		**100.0/100.0**	99.6/99.0	96.0/92.0	81.0/48.0	17.2/0.0	4:1/**4:1**	**8:1**/4:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**99.0/96.0**	73.8/**63.0**	**32.8/17.0**	**8:1/4:1**	**8:1/8:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0**/98.0	97.2/91.0	79.0/43.0	8.0/1.0	4:1/**4:1**	**8:1**/4:1
niah-mk3	SnapKV	100.0/100.0	96.0/96.0	37.0/74.0	2.2/31.0	0.0/7.0	0.0/0.0	1:1/1:1	2:1/2:1
	AdaKV		99.2/98.0	56.0/78.0	6.4/45.0	0.0/14.0	0.0/**3.0**	2:1/2:1	2:1/2:1
	KVzip		**100.0/100.0**	**100.0**/97.0	97.6/71.0	22.6/3.0	0.0/0.0	4:1/2:1	**8:1/4:1**
	KVzip+		**100.0/100.0**	**100.0**/99.0	98.2/76.0	24.4/11.0	0.0/0.0	**8:1/4:1**	**8:1/4:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0**/99.0	**99.8**/86.0	40.8/3.0	0.2/0.0	**8:1/4:1**	**8:1/4:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	98.6/**92.0**	49.8/**26.0**	**2.6**/0.0	**8:1/4:1**	**8:1/4:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**99.8**/86.0	**82.2**/5.0	0.0/0.0	**8:1/4:1**	**8:1/4:1**
niah-mq	SnapKV	100.0/99.8	**100.0/99.8**	**99.7**/99.5	83.0/99.5	10.6/**99.5**	3.7/**90.2**	**4:1/16:1**	4:1/**16:1**
	AdaKV		**100.0/99.8**	99.4/99.5	68.6/**99.8**	10.4/**99.5**	7.3/84.0	**4:1/16:1**	4:1/**16:1**
	KVzip		**100.0/99.8**	97.0/**99.8**	92.3/87.2	39.0/36.2	0.1/1.0	2:1/4:1	4:1/4:1
	KVzip+		**100.0/99.8**	97.2/**99.8**	93.1/87.8	56.2/57.2	5.7/13.2	2:1/4:1	4:1/4:1
	KVzip+ (MLP)		99.2/**99.8**	96.9/97.5	87.4/86.2	69.0/53.8	2.2/10.8	2:1/2:1	4:1/4:1
	KVgrad		**100.0/99.8**	98.8/**99.8**	**97.7**/96.8	80.9/89.5	37.1/49.0	**4:1**/4:1	**8:1**/8:1
	KVgrad (MLP)		**100.0/99.8**	97.4/**99.8**	96.7/94.5	**93.3**/86.5	**57.3**/72.8	2:1/4:1	**8:1**/4:1
niah-mv	SnapKV	100.0/100.0	91.2/99.8	49.5/91.0	23.8/58.5	9.3/31.2	1.1/15.2	1:1/2:1	1:1/2:1
	AdaKV		89.2/**100.0**	43.3/89.8	17.9/49.5	9.2/25.8	3.2/13.2	1:1/2:1	1:1/2:1
	KVzip		100.0/**100.0**	98.5/95.5	72.4/70.2	24.7/29.8	0.6/1.8	4:1/2:1	4:1/**4:1**
	KVzip+		100.0/**100.0**	98.3/94.8	78.0/79.2	38.0/43.8	7.9/20.0	4:1/2:1	4:1/2:1
	KVzip+ (MLP)		**100.0/100.0**	97.8/90.2	93.8/89.5	54.2/52.2	3.1/6.8	2:1/2:1	4:1/2:1
	KVgrad		**100.0/100.0**	**99.8/100.0**	96.0/**93.5**	81.3/77.8	51.5/53.5	4:1/**4:1**	**8:1/4:1**
	KVgrad (MLP)		**100.0/100.0**	99.4/93.8	**98.3**/91.2	**92.3/85.8**	**78.8/70.0**	**8:1**/2:1	**8:1**/2:1

Table 14: Per-task RULER accuracy and NL-CR for Qwen3-4B (2/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

		Accuracy (per CR)						NL-CR (per τ)	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
niah-s1	SnapKV	100.0/100.0	**100.0/100.0**	98.4/**100.0**	59.6/**100.0**	7.4/**100.0**	0.6/88.0	4:1/16:1	4:1/16:1
	AdaKV		**100.0/100.0**	**100.0/100.0**	94.8/**100.0**	23.4/**100.0**	0.8/**100.0**	4:1/**32:1**	4:1/**32:1**
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
niah-s2	SnapKV	100.0/100.0	**100.0/100.0**	**100.0/100.0**	94.2/**100.0**	5.6/99.0	1.2/**80.0**	4:1/**16:1**	4:1/**16:1**
	AdaKV		**100.0/100.0**	99.8/**100.0**	79.8/**100.0**	2.8/94.0	1.2/48.0	4:1/8:1	4:1/8:1
	KVzip		**100.0/100.0**	**100.0/100.0**	96.0/**100.0**	50.8/53.0	0.4/1.0	4:1/8:1	8:1/8:1
	KVzip+		**100.0/100.0**	**100.0/100.0**	99.4/**100.0**	78.8/81.0	2.6/16.0	8:1/8:1	8:1/8:1
	KVzip+ (MLP)		**100.0/100.0**	99.8/**100.0**	99.4/99.0	88.4/90.0	1.2/5.0	8:1/8:1	8:1/8:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	96.4/96.0	58.2/55.0	8:1/8:1	**16:1/16:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**99.8/100.0**	**62.6**/69.0	**16:1/16:1**	**16:1/16:1**
niah-s3	SnapKV	99.8/100.0	92.4/72.0	13.4/11.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	AdaKV		85.4/87.0	10.4/11.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	KVzip		**99.8/100.0**	**99.8/100.0**	96.6/97.0	22.4/28.0	0.0/0.0	4:1/4:1	8:1/8:1
	KVzip+		**99.8/100.0**	**99.8/100.0**	97.0/98.0	62.8/73.0	0.0/1.0	4:1/8:1	8:1/8:1
	KVzip+ (MLP)		**99.8/100.0**	**99.8/100.0**	**99.8/100.0**	80.0/84.0	0.0/0.0	8:1/8:1	8:1/8:1
	KVgrad		**99.8/100.0**	**99.8/100.0**	**99.8/100.0**	**99.8/99.0**	81.8/83.0	16:1/16:1	16:1/16:1
	KVgrad (MLP)		**99.8/100.0**	**99.8/100.0**	**99.8/100.0**	**99.8/99.0**	**98.0/99.0**	**32:1/32:1**	**32:1/32:1**
qa1	SnapKV	79.6/76.0	80.2/74.0	79.2/73.0	74.2/72.0	60.2/71.0	25.2/62.0	**4:1**/1:1	4:1/4:1
	AdaKV		80.0/**76.0**	79.2/74.0	**76.0/74.0**	**64.0/73.0**	25.6/**63.0**	**4:1**/2:1	**8:1/16:1**
	KVzip		80.2/75.0	81.0/75.0	72.4/63.0	37.6/41.0	22.8/24.0	**4:1/4:1**	4:1/4:1
	KVzip+		79.4/75.0	80.2/**77.0**	72.8/70.0	47.0/44.0	22.8/22.0	**4:1/4:1**	4:1/4:1
	KVzip+ (MLP)		**81.2**/75.0	**82.8**/75.0	72.6/65.0	46.4/39.0	27.8/28.0	**4:1/4:1**	4:1/4:1
	KVgrad		79.4/74.0	79.8/74.0	65.6/68.0	36.6/49.0	26.2/33.0	**4:1**/1:1	4:1/4:1
	KVgrad (MLP)		79.8/74.0	81.8/73.0	72.6/68.0	45.4/44.0	**28.2**/34.0	**4:1**/1:1	4:1/4:1
qa2	SnapKV	60.0/55.0	60.0/54.0	58.8/53.0	55.2/52.0	43.8/**53.0**	29.8/48.0	4:1/2:1	4:1/**16:1**
	AdaKV		59.8/55.0	58.8/53.0	54.6/52.0	**46.8**/52.0	**33.2/50.0**	4:1/2:1	4:1/4:1
	KVzip		**60.2**/54.0	58.2/**58.0**	53.4/48.0	33.6/31.0	19.4/18.0	2:1/4:1	4:1/4:1
	KVzip+		**60.2**/55.0	58.8/57.0	52.8/48.0	40.6/32.0	22.2/19.0	4:1/4:1	4:1/4:1
	KVzip+ (MLP)		59.8/55.0	59.0/54.0	54.2/49.0	40.2/36.0	22.4/19.0	4:1/4:1	4:1/4:1
	KVgrad		60.0/**56.0**	**59.8**/56.0	55.4/**55.0**	42.6/44.0	31.8/38.0	4:1/**8:1**	4:1/8:1
	KVgrad (MLP)		59.8/54.0	**59.8**/55.0	**58.8**/48.0	46.0/43.0	24.8/28.0	**8:1**/4:1	**8:1**/4:1
vt	SnapKV	100.0/100.0	96.1/98.2	76.0/98.0	36.0/96.6	12.0/85.2	0.6/33.8	1:1/4:1	2:1/8:1
	AdaKV		99.2/99.6	88.3/96.4	52.3/97.2	16.8/92.2	1.4/59.6	2:1/2:1	2:1/8:1
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	100.0/**100.0**	99.2/**99.8**	**32:1/32:1**	**32:1/32:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0**/99.8	**100.0**/99.8	99.8/**99.8**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0**/99.8	99.1/96.0	**32:1**/16:1	**32:1/32:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	100.0/99.8	99.4/**99.8**	**32:1/32:1**	**32:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0**/99.8	**100.0**/97.0	**32:1**/16:1	**32:1/32:1**

A.6.3 RULER RESULTS - QWEN3 8B

Table 15: Per-task RULER accuracy and NL-CR for Qwen3-8B (1/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

		Accuracy (per CR)						NL-CR (per τ)	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
cwe	SnapKV	99.0/84.0	98.3/77.9	92.6/63.5	61.7/33.2	29.9/12.5	16.5/4.1	2:1/1:1	2:1/1:1
	AdaKV		98.8/81.0	96.9/66.1	76.5/36.5	36.6/14.2	17.5/4.6	2:1/1:1	4:1/2:1
	KVzip		99.1/83.9	**99.1**/82.7	94.3/76.3	43.9/19.4	0.6/0.4	4:1/**4:1**	**8:1**/4:1
	KVzip+		**99.2**/84.5	**99.1**/**84.7**	**97.4**/**80.1**	81.3/54.7	4.1/0.8	**8:1**/**4:1**	**8:1**/**8:1**
	KVzip+ (MLP)		99.0/**85.1**	**99.1**/83.0	96.0/68.7	45.7/10.3	4.3/0.1	4:1/**4:1**	**8:1**/4:1
	KVgrad		99.1/84.4	98.5/82.0	96.0/77.5	**89.0**/**65.9**	**66.4**/**35.7**	4:1/2:1	**8:1**/4:1
	KVgrad (MLP)		99.0/84.2	98.2/80.7	93.7/73.1	75.1/48.5	36.3/21.4	4:1/2:1	4:1/4:1
fwe	SnapKV	95.3/92.3	93.1/**94.0**	85.3/**94.7**	75.9/94.0	66.6/**95.0**	22.9/87.3	1:1/16:1	2:1/16:1
	AdaKV		94.3/92.3	89.5/94.3	80.7/**94.3**	69.7/**95.0**	27.5/**91.3**	2:1/**32:1**	2:1/**32:1**
	KVzip		95.5/92.0	**95.8**/91.0	89.3/90.7	61.9/73.3	0.0/2.3	**4:1**/8:1	4:1/8:1
	KVzip+		95.5/92.0	95.7/90.3	91.5/92.0	77.9/83.0	2.1/7.0	**4:1**/8:1	**8:1**/8:1
	KVzip+ (MLP)		**95.6**/91.7	**95.8**/89.7	90.7/90.7	63.9/57.7	2.0/1.3	**4:1**/8:1	**8:1**/8:1
	KVgrad		**95.6**/91.7	95.5/91.7	88.9/91.0	76.8/89.3	56.2/70.3	**4:1**/8:1	4:1/16:1
	KVgrad (MLP)		95.5/92.7	95.3/91.0	**91.9**/89.0	**82.5**/84.7	**61.6**/67.0	**4:1**/4:1	**8:1**/8:1
niah-mk1	SnapKV	100.0/100.0	**100.0**/**100.0**	98.6/**100.0**	78.8/**100.0**	10.4/**99.0**	5.4/72.0	4:1/**16:1**	4:1/**16:1**
	AdaKV		99.8/**100.0**	98.4/**100.0**	59.4/**100.0**	10.8/92.0	8.0/51.0	4:1/8:1	4:1/8:1
	KVzip		**100.0**/**100.0**	99.8/99.0	90.8/80.0	47.0/40.0	0.2/2.0	4:1/4:1	4:1/4:1
	KVzip+		**100.0**/**100.0**	**100.0**/**100.0**	97.4/94.0	47.4/49.0	21.4/27.0	4:1/4:1	8:1/4:1
	KVzip+ (MLP)		**100.0**/**100.0**	99.8/**100.0**	96.4/90.0	51.6/48.0	13.2/27.0	4:1/4:1	8:1/4:1
	KVgrad		**100.0**/**100.0**	**100.0**/**100.0**	98.8/**100.0**	89.6/96.0	60.8/**73.0**	8:1/8:1	8:1/**16:1**
	KVgrad (MLP)		**100.0**/**100.0**	99.8/**100.0**	**100.0**/**100.0**	**98.8**/90.0	**85.0**/68.0	**16:1**/8:1	**16:1**/8:1
niah-mk2	SnapKV	100.0/100.0	80.8/94.0	48.6/70.0	21.0/43.0	6.2/18.0	1.0/10.0	1:1/1:1	1:1/1:1
	AdaKV		97.6/**100.0**	69.2/88.0	37.8/69.0	9.8/36.0	1.2/14.0	1:1/2:1	2:1/2:1
	KVzip		**100.0**/**100.0**	**100.0**/**99.0**	99.6/97.0	74.8/32.0	2.0/1.0	**8:1**/4:1	**8:1**/**8:1**
	KVzip+		**100.0**/**100.0**	**100.0**/**99.0**	**100.0**/98.0	**87.4**/51.0	1.2/0.0	**8:1**/**8:1**	**8:1**/**8:1**
	KVzip+ (MLP)		**100.0**/99.0	**100.0**/**99.0**	98.6/**99.0**	78.8/50.0	8.4/1.0	**8:1**/**8:1**	**8:1**/**8:1**
	KVgrad		**100.0**/**100.0**	**100.0**/**99.0**	96.8/98.0	62.6/**67.0**	**24.4**/**18.0**	4:1/**8:1**	**8:1**/**8:1**
	KVgrad (MLP)		**100.0**/**100.0**	**100.0**/85.0	89.8/31.0	24.8/9.0	3.0/2.0	4:1/2:1	4:1/2:1
niah-mk3	SnapKV	100.0/100.0	98.2/**100.0**	59.8/94.0	8.0/81.0	0.0/55.0	0.0/9.0	2:1/2:1	2:1/2:1
	AdaKV		99.8/99.0	85.4/**99.0**	17.2/87.0	0.2/**72.0**	0.0/**21.0**	2:1/**4:1**	2:1/4:1
	KVzip		**100.0**/99.0	99.8/**99.0**	98.4/73.0	24.2/7.0	0.0/0.0	**8:1**/**4:1**	**8:1**/4:1
	KVzip+		**100.0**/99.0	99.8/**99.0**	99.2/81.0	55.2/21.0	0.0/0.0	**8:1**/**4:1**	**8:1**/4:1
	KVzip+ (MLP)		**100.0**/99.0	99.6/96.0	98.2/92.0	48.0/5.0	0.0/0.0	**8:1**/2:1	**8:1**/4:1
	KVgrad		**100.0**/**100.0**	**100.0**/**99.0**	85.0/**95.0**	39.2/41.0	**2.2**/3.0	4:1/**4:1**	4:1/**8:1**
	KVgrad (MLP)		99.8/99.0	99.6/92.0	**99.4**/40.0	**61.6**/10.0	0.8/0.0	**8:1**/2:1	**8:1**/2:1
niah-mq	SnapKV	99.9/100.0	**100.0**/**100.0**	**100.0**/99.8	92.3/99.8	10.9/**100.0**	4.1/**94.5**	4:1/**16:1**	4:1/**16:1**
	AdaKV		**100.0**/**100.0**	99.9/99.8	81.0/99.8	11.0/99.8	7.8/87.0	4:1/**16:1**	4:1/**16:1**
	KVzip		100.0/**100.0**	100.0/99.2	99.2/98.5	41.2/46.2	0.1/0.2	8:1/8:1	8:1/8:1
	KVzip+		100.0/**100.0**	99.9/**100.0**	99.7/99.5	68.0/72.5	13.9/22.0	8:1/8:1	8:1/8:1
	KVzip+ (MLP)		100.0/**100.0**	**100.0**/99.5	99.5/96.5	73.6/73.5	8.3/21.5	8:1/4:1	8:1/8:1
	KVgrad		99.8/**100.0**	**100.0**/**100.0**	98.8/**100.0**	89.0/96.5	55.2/79.8	8:1/8:1	8:1/**16:1**
	KVgrad (MLP)		100.0/**100.0**	100.0/**100.0**	**100.0**/99.8	**99.0**/96.5	**85.2**/86.8	**16:1**/8:1	**16:1**/**16:1**
niah-mv	SnapKV	100.0/99.0	87.0/**99.8**	49.5/94.5	26.4/71.5	9.3/37.0	2.0/17.5	1:1/2:1	1:1/4:1
	AdaKV		89.2/99.5	41.9/94.5	18.7/57.0	9.4/25.5	3.1/10.2	1:1/2:1	1:1/4:1
	KVzip		100.0/99.5	99.2/99.5	84.8/83.2	28.9/41.8	0.1/1.5	4:1/4:1	4:1/4:1
	KVzip+		**100.0**/99.5	99.7/99.5	89.1/88.2	50.4/57.5	22.1/39.0	4:1/4:1	4:1/4:1
	KVzip+ (MLP)		100.0/**99.8**	99.8/99.2	97.7/95.0	60.6/72.2	8.8/27.8	4:1/4:1	8:1/8:1
	KVgrad		100.0/99.5	99.8/**99.8**	98.9/**99.0**	93.4/93.2	69.8/77.5	8:1/**8:1**	8:1/8:1
	KVgrad (MLP)		**100.0**/99.2	**100.0**/99.2	**99.9**/98.8	**99.5**/**95.8**	**95.3**/**85.0**	**16:1**/**8:1**	**32:1**/**16:1**

Table 16: Per-task RULER accuracy and NL-CR for Qwen3-8B (2/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

		Accuracy (per CR)						NL-CR (per τ)	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
niah-s1	SnapKV	100.0/100.0	**100.0/100.0**	97.8/**100.0**	73.4/**100.0**	23.0/**100.0**	0.6/98.0	2:1/**32:1**	4:1/**32:1**
	AdaKV		**100.0/100.0**	**100.0/100.0**	91.0/**100.0**	41.2/**100.0**	0.6/**100.0**	4:1/**32:1**	4:1/**32:1**
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	99.8/**100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
niah-s2	SnapKV	100.0/100.0	**100.0/100.0**	**100.0/100.0**	98.4/**100.0**	7.0/**100.0**	1.2/94.0	8:1/**16:1**	8:1/16:1
	AdaKV		**100.0/100.0**	**100.0/100.0**	85.2/**100.0**	6.6/**100.0**	1.2/73.0	4:1/**16:1**	4:1/16:1
	KVzip		**100.0/100.0**	**100.0/100.0**	97.2/**100.0**	24.2/25.0	0.0/0.0	4:1/8:1	8:1/8:1
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	75.0/81.0	22.2/44.0	8:1/8:1	8:1/8:1
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	81.4/88.0	11.8/21.0	8:1/8:1	8:1/8:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	94.6/96.0	54.4/80.0	8:1/8:1	8:1/16:1
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**98.8/97.0**	**32:1/16:1**	**32:1/32:1**
niah-s3	SnapKV	100.0/100.0	40.2/60.0	5.2/10.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	AdaKV		44.4/86.0	3.6/10.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	KVzip		**100.0/100.0**	**100.0/100.0**	98.6/**100.0**	26.8/38.0	0.0/0.0	8:1/8:1	8:1/8:1
	KVzip+		**100.0/100.0**	**100.0/100.0**	99.0/**100.0**	70.6/85.0	3.6/11.0	8:1/8:1	8:1/8:1
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	79.6/95.0	2.0/11.0	8:1/8:1	8:1/16:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	99.4/**100.0**	58.2/71.0	16:1/16:1	16:1/16:1
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**99.8/100.0**	**32:1/32:1**	**32:1/32:1**
qa1	SnapKV	81.8/75.0	80.2/75.0	79.4/75.0	77.6/**74.0**	65.2/**69.0**	28.0/**61.0**	2:1/**8:1**	4:1/**8:1**
	AdaKV		80.6/75.0	80.4/75.0	77.0/**74.0**	**67.6/69.0**	30.2/**61.0**	**4:1/8:1**	4:1/**8:1**
	KVzip		80.2/75.0	**81.4/78.0**	76.6/65.0	43.0/31.0	25.6/22.0	**4:1**/4:1	4:1/4:1
	KVzip+		80.8/75.0	81.2/76.0	77.6/64.0	51.6/39.0	**30.4**/24.0	**4:1**/4:1	4:1/4:1
	KVzip+ (MLP)		80.8/**76.0**	**81.4**/72.0	**78.2**/63.0	53.0/38.0	30.0/29.0	**4:1**/2:1	**8:1**/4:1
	KVgrad		81.0/75.0	79.0/75.0	63.8/58.0	41.8/47.0	28.8/30.0	2:1/4:1	4:1/4:1
	KVgrad (MLP)		**81.4**/74.0	79.0/70.0	71.4/58.0	46.2/43.0	28.0/32.0	2:1/2:1	4:1/2:1
qa2	SnapKV	63.8/61.0	63.4/60.0	62.0/**59.0**	56.8/56.0	48.4/**58.0**	35.0/57.0	2:1/**2:1**	**4:1**/16:1
	AdaKV		62.0/58.0	61.4/**59.0**	58.6/**58.0**	**50.8**/55.0	**38.2/58.0**	1:1/1:1	**4:1/32:1**
	KVzip		63.4/60.0	61.0/56.0	56.2/52.0	32.4/34.0	22.6/22.0	2:1/**2:1**	**4:1**/2:1
	KVzip+		**63.6/61.0**	62.2/57.0	**59.0**/52.0	40.0/38.0	22.8/19.0	2:1/**2:1**	**4:1**/2:1
	KVzip+ (MLP)		63.0/60.0	61.6/54.0	55.6/46.0	42.0/46.0	25.8/24.0	2:1/**2:1**	**4:1**/2:1
	KVgrad		**63.6/61.0**	**62.8**/57.0	57.0/57.0	45.8/46.0	31.2/33.0	**4:1/2:1**	**4:1**/2:1
	KVgrad (MLP)		62.8/60.0	61.8/56.0	58.6/56.0	46.8/41.0	30.6/27.0	2:1/**2:1**	**4:1**/2:1
vt	SnapKV	100.0/100.0	96.8/**100.0**	85.4/99.8	47.7/98.6	24.2/91.6	2.3/53.4	1:1/8:1	2:1/8:1
	AdaKV		99.8/99.8	94.0/**100.0**	61.1/99.6	27.2/96.8	3.7/61.4	2:1/8:1	2:1/16:1
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	100.0/**100.0**	98.8/**100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	100.0/**100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0**/99.6	**32:1/32:1**	**32:1/32:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	97.5/**100.0**	16:1/**32:1**	**32:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	99.9/**100.0**	**32:1/32:1**	**32:1/32:1**

A.6.4 RULER RESULTS - LLAMA-3.1-8B

Table 17: Per-task RULER accuracy and NL-CR for Llama-3.1-8B-Instruct (1/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

			Accuracy (per CR)					**NL-CR (per τ)**	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
	SnapKV		97.6/83.9	81.5/65.7	47.4/30.5	13.2/6.7	0.1/1.8	1:1/1:1	2:1/2:1
	AdaKV		99.2/85.2	95.0/73.6	68.7/46.6	26.0/**17.0**	0.3/**4.0**	2:1/2:1	4:1/2:1
	KVzip		99.6/**86.9**	99.0/86.8	91.8/61.3	45.8/11.0	0.4/0.2	**4:1/4:1**	4:1/**4:1**
cwe	KVzip+	99.6/86.7	99.6/86.7	**99.2/87.3**	**95.6/73.1**	**50.4**/14.5	0.0/0.2	**4:1/4:1**	**8:1/4:1**
	KVzip+ (MLP)		**99.7**/82.2	98.6/86.3	93.0/37.2	32.2/0.2	0.1/0.5	**4:1/4:1**	4:1/**4:1**
	KVgrad		99.5/86.7	97.3/81.9	83.7/51.1	46.8/11.5	**6.0**/1.3	2:1/2:1	4:1/2:1
	KVgrad (MLP)		99.4/85.1	97.1/76.6	85.7/42.4	47.8/2.5	0.4/0.0	2:1/2:1	4:1/2:1
	SnapKV		89.7/89.3	77.6/89.3	65.8/84.0	44.9/75.3	15.5/65.0	1:1/4:1	1:1/**8:1**
	AdaKV		93.6/90.0	87.5/**91.3**	74.5/85.0	57.7/**79.0**	21.4/**67.3**	2:1/4:1	2:1/**8:1**
	KVzip		94.3/88.7	94.5/87.7	**93.1**/83.3	82.8/74.7	1.6/10.0	**8:1**/4:1	**8:1**/4:1
fwe	KVzip+	94.8/88.3	94.5/88.7	94.6/87.7	92.9/85.3	83.5/72.3	25.8/45.3	4:1/4:1	**8:1/8:1**
	KVzip+ (MLP)		**95.0/90.7**	**95.1**/88.3	91.7/**86.7**	81.7/67.3	5.9/6.7	4:1/**8:1**	**8:1/8:1**
	KVgrad		94.2/88.7	93.9/88.7	91.9/85.0	**88.7**/77.3	**50.6**/41.0	4:1/4:1	**8:1/8:1**
	KVgrad (MLP)		94.2/88.0	92.1/84.0	88.7/78.0	79.1/60.7	15.9/19.3	2:1/2:1	4:1/4:1
	SnapKV		99.8/**100.0**	99.8/**100.0**	99.2/**100.0**	97.8/99.0	29.6/**100.0**	8:1/**32:1**	**16:1/32:1**
	AdaKV		99.8/**100.0**	**100.0/100.0**	99.0/**100.0**	93.4/99.0	13.0/99.0	8:1/**32:1**	8:1/**32:1**
	KVzip		**100.0/100.0**	**100.0/100.0**	99.8/97.0	91.4/86.0	16.8/20.0	8:1/4:1	8:1/8:1
niah-mk1	KVzip+	99.8/100.0	**100.0/100.0**	**100.0/100.0**	99.8/96.0	93.4/92.0	24.6/38.0	8:1/4:1	8:1/8:1
	KVzip+ (MLP)		99.8/**100.0**	**100.0**/99.0	99.4/97.0	90.0/93.0	11.2/41.0	8:1/4:1	8:1/8:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	98.4/**100.0**	78.6/94.0	**16:1**/16:1	**16:1**/16:1
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**99.2/100.0**	**93.2**/91.0	**16:1**/16:1	**16:1**/16:1
	SnapKV		99.4/**100.0**	95.0/**100.0**	90.0/97.0	67.0/85.0	9.2/59.0	2:1/4:1	4:1/8:1
	AdaKV		**100.0/100.0**	98.4/**100.0**	91.2/95.0	61.4/77.0	11.6/64.0	4:1/4:1	4:1/8:1
	KVzip		**100.0/100.0**	**100.0/100.0**	**99.8**/99.0	89.8/85.0	48.8/46.0	8:1/8:1	8:1/8:1
niah-mk2	KVzip+	100.0/100.0	**100.0/100.0**	99.8/**100.0**	99.2/99.0	95.4/85.0	18.2/26.0	8:1/8:1	**16:1**/8:1
	KVzip+ (MLP)		**100.0/100.0**	99.8/**100.0**	99.2/**100.0**	90.4/90.0	11.2/16.0	8:1/8:1	8:1/8:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**99.8/100.0**	98.0/**100.0**	63.4/**72.0**	**16:1/16:1**	**16:1/16:1**
	KVgrad (MLP)		**100.0/100.0**	99.8/**100.0**	99.2/**100.0**	**99.0**/92.0	**68.0**/33.0	**16:1**/8:1	**16:1**/8:1
	SnapKV		85.4/98.0	26.6/72.0	2.4/36.0	0.0/6.0	0.0/0.0	1:1/2:1	1:1/2:1
	AdaKV		96.4/99.0	46.8/91.0	4.0/48.0	0.0/8.0	0.0/0.0	1:1/2:1	2:1/2:1
	KVzip		**99.8**/99.0	**100.0**/99.0	98.6/85.0	79.0/41.0	0.0/0.0	**8:1**/4:1	**8:1**/4:1
niah-mk3	KVzip+	99.8/100.0	**99.8**/99.0	**100.0/100.0**	98.2/84.0	61.6/19.0	0.0/0.0	**8:1**/4:1	**8:1**/4:1
	KVzip+ (MLP)		**99.8/100.0**	**100.0**/99.0	99.0/93.0	40.2/17.0	0.0/0.0	**8:1**/4:1	**8:1**/4:1
	KVgrad		**99.8/100.0**	99.8/99.0	99.8/**99.0**	80.6/**86.0**	**7.4/10.0**	**8:1/8:1**	**8:1/8:1**
	KVgrad (MLP)		**99.8**/99.0	99.6/99.0	**100.0**/98.0	**92.0**/68.0	0.6/0.0	**8:1/8:1**	**8:1/8:1**
	SnapKV		99.9/98.8	**100.0**/98.8	97.2/99.2	49.4/96.8	9.7/92.8	4:1/8:1	8:1/16:1
	AdaKV		99.8/99.2	99.7/98.8	96.9/99.5	39.5/98.5	10.1/**95.2**	4:1/**16:1**	8:1/**32:1**
	KVzip		**100.0/100.0**	**100.0/100.0**	99.7/97.2	90.7/81.0	7.2/10.8	8:1/4:1	8:1/8:1
niah-mq	KVzip+	99.9/99.2	**100.0/100.0**	**100.0**/99.0	99.5/98.0	90.3/88.8	37.0/49.8	8:1/8:1	8:1/8:1
	KVzip+ (MLP)		99.8/99.2	99.8/99.2	98.5/95.8	84.3/82.5	6.6/26.0	8:1/4:1	8:1/8:1
	KVgrad		100.0/**100.0**	**100.0/100.0**	99.9/**99.8**	98.3/98.8	79.4/92.2	**16:1/16:1**	16:1/16:1
	KVgrad (MLP)		100.0/**100.0**	**100.0/100.0**	**100.0**/99.2	**99.8/99.2**	**95.7**/94.2	**16:1/16:1**	**32:1**/16:1
	SnapKV		99.4/99.5	97.5/99.5	72.2/96.8	17.1/85.0	4.2/48.5	2:1/4:1	4:1/8:1
	AdaKV		99.3/**99.8**	97.0/**100.0**	74.7/98.5	11.2/82.0	7.5/39.8	2:1/8:1	4:1/8:1
	KVzip		**99.9**/99.5	**100.0**/99.2	97.8/94.8	61.9/52.8	7.8/10.8	4:1/4:1	8:1/8:1
niah-mv	KVzip+	99.9/99.5	**99.9**/99.2	99.8/99.5	97.8/95.5	67.5/50.5	20.9/20.8	4:1/4:1	8:1/8:1
	KVzip+ (MLP)		**99.9/99.8**	99.9/99.8	95.9/94.0	62.7/67.2	7.5/16.8	4:1/4:1	8:1/4:1
	KVgrad		**99.9**/99.5	99.9/99.5	99.6/99.0	95.1/**97.8**	70.2/**83.5**	8:1/**16:1**	**16:1/16:1**
	KVgrad (MLP)		**99.9**/99.5	99.9/99.2	**99.9/99.5**	**99.5**/97.2	**86.0**/82.8	**16:1**/8:1	**16:1/16:1**

Table 18: Per-task RULER accuracy and NL-CR for Llama-3.1-8B-Instruct (2/2), reported as *4K/16K*: the left value is RULER-4K, the right RULER-16K (0.2-fraction index), and '-' marks a cell with no run. NL-CR at tolerance τ is the largest compression ratio whose accuracy stays within $\tau\%$ (relative) of No press.

		Accuracy (per CR)						NL-CR (per τ)	
Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
niah-s1	SnapKV	100.0/100.0	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	99.8/**100.0**	82.4/**100.0**	16:1/**32:1**	16:1/**32:1**
	AdaKV		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	89.4/**100.0**	16:1/**32:1**	16:1/**32:1**
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0**/99.0	**32:1/32:1**	**32:1/32:1**
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**32:1/32:1**	**32:1/32:1**
niah-s2	SnapKV	100.0/100.0	**100.0/100.0**	**100.0/100.0**	99.0/99.0	92.6/97.0	42.0/96.0	8:1/8:1	8:1/**32:1**
	AdaKV		**100.0/100.0**	**100.0/100.0**	99.4/99.0	88.4/98.0	10.8/97.0	8:1/16:1	8:1/**32:1**
	KVzip		99.8/**100.0**	**100.0/100.0**	**100.0/100.0**	98.0/97.0	26.4/13.0	**16:1**/8:1	**16:1**/16:1
	KVzip+		99.8/**100.0**	**100.0/100.0**	99.8/**100.0**	95.2/98.0	39.8/56.0	8:1/16:1	**16:1**/16:1
	KVzip+ (MLP)		**100.0/100.0**	**100.0/100.0**	99.4/**100.0**	92.8/94.0	5.8/21.0	8:1/8:1	8:1/8:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	99.8/**100.0**	98.8/**100.0**	89.6/**99.0**	**16:1/32:1**	**16:1/32:1**
	KVgrad (MLP)		**100.0/100.0**	99.8/**100.0**	99.8/**100.0**	**99.4/100.0**	**93.2**/95.0	**16:1**/16:1	**16:1/32:1**
niah-s3	SnapKV	100.0/100.0	37.4/70.0	5.4/19.0	2.4/6.0	2.4/5.0	2.4/5.0	1:1/1:1	1:1/1:1
	AdaKV		48.0/98.0	5.8/32.0	2.4/5.0	2.4/5.0	2.4/5.0	1:1/2:1	1:1/2:1
	KVzip		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	93.2/95.0	1.6/1.0	8:1/8:1	8:1/**16:1**
	KVzip+		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	88.4/89.0	10.0/21.0	8:1/8:1	8:1/8:1
	KVzip+ (MLP)		**100.0/100.0**	99.8/**100.0**	**100.0/100.0**	91.4/91.0	1.2/14.0	8:1/8:1	8:1/8:1
	KVgrad		**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**98.8/100.0**	**65.8/83.0**	**16:1/16:1**	**16:1/16:1**
	KVgrad (MLP)		**100.0/100.0**	**100.0/100.0**	99.8/**100.0**	93.2/96.0	14.6/19.0	8:1/8:1	8:1/**16:1**
qa1	SnapKV	87.8/85.0	**88.0/85.0**	**87.8/84.0**	82.2/**81.0**	73.6/**80.0**	31.6/**77.0**	**4:1/4:1**	4:1/**8:1**
	AdaKV		87.6/84.0	87.4/83.0	**84.2**/80.0	**75.6**/78.0	**36.8/77.0**	**4:1**/2:1	**8:1**/4:1
	KVzip		87.0/81.0	85.0/83.0	77.8/73.0	53.6/47.0	22.0/24.0	2:1/1:1	4:1/4:1
	KVzip+		87.0/82.0	84.6/80.0	80.6/73.0	47.2/48.0	20.4/22.0	2:1/1:1	4:1/2:1
	KVzip+ (MLP)		86.2/**85.0**	84.2/82.0	79.0/77.0	48.6/51.0	21.6/21.0	2:1/2:1	4:1/4:1
	KVgrad		86.6/84.0	84.4/82.0	76.4/72.0	48.0/59.0	24.8/35.0	2:1/2:1	4:1/4:1
	KVgrad (MLP)		86.8/**85.0**	85.0/79.0	78.0/74.0	54.4/44.0	18.2/22.0	2:1/2:1	4:1/2:1
qa2	SnapKV	62.8/50.0	**62.4/50.0**	**62.8**/49.0	**60.2**/50.0	55.0/**49.0**	40.2/**49.0**	**4:1/32:1**	**8:1/32:1**
	AdaKV		61.8/**50.0**	61.6/49.0	59.8/50.0	**55.6/49.0**	**46.0**/48.0	**4:1**/16:1	**8:1/32:1**
	KVzip		61.2/48.0	61.8/**51.0**	57.2/**55.0**	43.8/47.0	25.8/25.0	**4:1**/8:1	4:1/8:1
	KVzip+		61.2/48.0	60.0/50.0	56.8/51.0	45.4/39.0	23.4/28.0	1:1/8:1	4:1/8:1
	KVzip+ (MLP)		61.6/49.0	60.4/**51.0**	58.2/52.0	45.0/42.0	23.0/26.0	2:1/8:1	4:1/8:1
	KVgrad		61.2/47.0	61.0/47.0	57.4/48.0	44.8/44.0	30.8/32.0	1:1/1:1	4:1/8:1
	KVgrad (MLP)		60.2/47.0	61.0/50.0	58.4/45.0	43.0/39.0	29.0/29.0	1:1/4:1	4:1/4:1
vt	SnapKV	99.9/100.0	97.8/99.6	88.8/97.2	66.3/94.2	26.2/90.4	5.7/65.6	1:1/2:1	2:1/4:1
	AdaKV		99.5/**100.0**	95.0/99.8	72.3/97.2	29.3/90.0	3.5/68.0	2:1/4:1	4:1/8:1
	KVzip		**99.9/100.0**	**99.9/100.0**	99.7/99.8	98.6/99.0	94.1/95.6	**16:1/16:1**	16:1/**32:1**
	KVzip+		**99.9/100.0**	**99.9**/99.6	**99.9**/99.4	98.6/99.0	95.6/**97.2**	**16:1/16:1**	**32:1/32:1**
	KVzip+ (MLP)		99.8/**100.0**	99.7/99.8	98.9/99.0	97.0/94.8	89.2/92.0	8:1/8:1	16:1/8:1
	KVgrad		**99.9/100.0**	**99.9/100.0**	**99.9/100.0**	**99.6/99.4**	**97.0**/96.0	**16:1/16:1**	**32:1/32:1**
	KVgrad (MLP)		**99.9/100.0**	**99.9/100.0**	99.8/99.4	98.5/95.8	83.0/86.0	**16:1**/8:1	16:1/16:1

A.6.5 LONGBENCH RESULTS - QWEN3 0.6B

Table 19: Per-task LongBench score and NL-CR for Qwen3-0.6B (1/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						**NL-CR (per τ)**	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
S-Doc QA	narrativeqa (F1)	SnapKV	16.0	15.1	15.8	**17.2**	**16.5**	**14.8**	**16:1**	**16:1**
		AdaKV		15.8	15.8	16.9	15.9	14.5	**16:1**	**16:1**
		KVzip		15.9	14.3	12.6	10.4	8.9	2:1	2:1
		KVzip+		15.9	15.3	13.4	10.9	10.6	2:1	4:1
		KVzip+ (MLP)		15.6	**16.9**	14.1	10.7	8.6	4:1	4:1
		KVgrad		**16.4**	16.0	15.1	14.0	12.1	4:1	4:1
		KVgrad (MLP)		15.1	15.5	12.8	10.3	6.9	1:1	4:1
	qasper (F1)	SnapKV	18.6	18.4	18.0	16.2	14.6	**13.1**	2:1	4:1
		AdaKV		17.8	17.8	16.3	**15.0**	12.5	1:1	4:1
		KVzip		19.3	14.7	15.6	11.8	8.9	2:1	2:1
		KVzip+		19.1	17.8	13.7	12.2	9.7	2:1	4:1
		KVzip+ (MLP)		**19.6**	**20.6**	13.8	14.9	10.4	**4:1**	4:1
		KVgrad		19.1	18.8	**17.9**	11.0	11.3	**4:1**	**8:1**
		KVgrad (MLP)		19.1	16.3	14.3	12.9	11.9	2:1	2:1
	multifieldqa_en (F1)	SnapKV	47.1	46.2	45.0	41.4	37.5	26.9	2:1	4:1
		AdaKV		46.7	44.6	41.7	36.9	27.6	2:1	2:1
		KVzip		48.5	46.7	39.8	28.1	19.6	4:1	4:1
		KVzip+		48.2	48.5	44.1	30.7	20.8	4:1	4:1
		KVzip+ (MLP)		**49.6**	48.0	42.2	29.7	21.2	4:1	4:1
		KVgrad		47.5	**49.5**	45.0	**41.7**	**31.6**	4:1	**8:1**
		KVgrad (MLP)		47.5	48.2	**46.7**	37.3	27.8	**8:1**	**8:1**
M-Doc QA	hotpotqa (F1)	SnapKV	36.6	**37.9**	37.1	**37.9**	35.6	29.2	8:1	**16:1**
		AdaKV		36.8	**37.4**	37.3	**36.0**	**31.8**	**16:1**	**16:1**
		KVzip		36.2	32.1	23.8	12.2	10.6	2:1	2:1
		KVzip+		36.0	34.3	23.7	13.0	12.8	2:1	2:1
		KVzip+ (MLP)		36.6	31.9	22.4	12.1	9.6	2:1	2:1
		KVgrad		35.7	35.5	27.7	24.5	18.1	1:1	4:1
		KVgrad (MLP)		34.4	33.6	27.8	20.8	13.8	1:1	1:1
	2wikimqa (F1)	SnapKV	28.9	30.9	29.6	30.5	26.5	25.0	8:1	8:1
		AdaKV		30.3	29.4	27.7	27.0	25.0	4:1	8:1
		KVzip		**31.4**	31.3	30.2	16.3	10.8	8:1	8:1
		KVzip+		30.1	31.3	27.5	16.6	11.8	4:1	8:1
		KVzip+ (MLP)		29.5	29.2	26.9	17.5	11.9	4:1	4:1
		KVgrad		31.2	**32.1**	**32.7**	**29.1**	26.4	**16:1**	16:1
		KVgrad (MLP)		30.8	28.1	28.4	24.9	**27.4**	8:1	**32:1**
	musique (F1)	SnapKV	14.8	14.9	13.2	13.4	**11.7**	9.3	2:1	2:1
		AdaKV		14.6	**14.5**	**14.0**	11.6	**10.0**	**4:1**	**4:1**
		KVzip		14.4	11.2	7.1	5.5	5.1	1:1	2:1
		KVzip+		14.7	11.2	6.8	6.5	6.0	2:1	2:1
		KVzip+ (MLP)		**15.3**	11.2	7.2	5.9	5.1	2:1	2:1
		KVgrad		15.0	14.5	6.6	6.6	6.2	2:1	**4:1**
		KVgrad (MLP)		14.2	13.2	7.3	6.9	5.8	1:1	2:1

Table 20: Per-task LongBench score and NL-CR for Qwen3-0.6B (2/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Summ	gov_report (R)	SnapKV	27.3	25.2	22.0	19.0	15.4	12.6	1:1	1:1
		AdaKV		24.9	21.7	19.1	15.9	12.8	1:1	1:1
		KVzip		27.3	26.6	25.1	15.9	10.2	2:1	4:1
		KVzip+		27.5	26.6	25.6	15.6	9.7	2:1	4:1
		KVzip+ (MLP)		27.2	26.2	24.0	13.2	9.1	2:1	4:1
		KVgrad		**27.6**	**27.1**	**26.1**	**23.0**	**19.4**	**4:1**	**8:1**
		KVgrad (MLP)		27.4	26.8	25.0	22.1	17.8	2:1	4:1
	qmsum (R)	SnapKV	20.4	20.4	20.2	19.7	**19.4**	18.6	4:1	**16:1**
		AdaKV		**20.8**	20.5	19.8	19.2	**18.8**	4:1	8:1
		KVzip		20.4	20.5	19.9	18.5	17.2	4:1	8:1
		KVzip+		20.2	**20.9**	**20.4**	18.5	17.2	**8:1**	8:1
		KVzip+ (MLP)		20.4	20.5	19.8	18.6	18.1	4:1	8:1
		KVgrad		20.3	20.6	20.2	19.2	18.8	**8:1**	8:1
		KVgrad (MLP)		20.3	20.6	20.1	18.6	17.5	**8:1**	8:1
	multi_news (R)	SnapKV	24.0	21.8	19.0	16.1	11.0	5.2	1:1	1:1
		AdaKV		21.6	19.5	15.9	11.8	5.7	1:1	1:1
		KVzip		**24.5**	24.0	22.5	11.1	9.1	4:1	4:1
		KVzip+		24.3	24.2	23.2	11.9	8.6	4:1	**8:1**
		KVzip+ (MLP)		24.2	**24.7**	**24.9**	9.9	9.1	**8:1**	**8:1**
		KVgrad		24.1	24.2	23.7	20.6	**16.2**	**8:1**	**8:1**
		KVgrad (MLP)		24.3	24.5	24.1	**21.2**	15.9	**8:1**	**8:1**
Few-shot	trec (cls)	SnapKV	20.5	28.0	30.5	25.0	25.5	17.5	16:1	16:1
		AdaKV		21.5	26.0	22.5	20.5	17.0	16:1	16:1
		KVzip		26.0	**52.0**	**53.0**	35.5	9.0	16:1	16:1
		KVzip+		22.0	48.5	52.0	35.5	17.5	16:1	16:1
		KVzip+ (MLP)		**49.5**	47.0	48.5	36.0	24.0	**32:1**	**32:1**
		KVgrad		23.0	34.5	50.0	**45.0**	20.5	**32:1**	**32:1**
		KVgrad (MLP)		33.0	49.0	41.0	34.0	**26.5**	**32:1**	**32:1**
	triviaqa (F1)	SnapKV	69.7	70.1	70.5	**71.2**	71.1	68.1	16:1	**32:1**
		AdaKV		69.9	70.6	70.8	**72.5**	66.5	16:1	**32:1**
		KVzip		**71.4**	69.7	67.2	63.6	61.0	4:1	8:1
		KVzip+		70.5	68.7	67.9	64.3	64.2	4:1	8:1
		KVzip+ (MLP)		71.4	69.5	65.7	60.4	60.2	4:1	4:1
		KVgrad		71.0	**71.5**	69.4	69.2	**69.5**	**32:1**	**32:1**
		KVgrad (MLP)		70.3	70.4	69.5	66.4	67.7	8:1	**32:1**
	samsum (R)	SnapKV	35.2	33.3	32.9	32.9	32.2	31.1	1:1	1:1
		AdaKV		35.0	33.5	33.2	32.3	29.4	2:1	4:1
		KVzip		34.4	34.5	33.2	29.7	29.8	1:1	4:1
		KVzip+		35.1	**35.0**	31.6	29.2	28.6	4:1	4:1
		KVzip+ (MLP)		34.3	34.3	30.0	28.8	29.9	1:1	4:1
		KVgrad		**35.1**	35.0	**34.6**	**32.9**	**31.2**	**8:1**	**8:1**
		KVgrad (MLP)		34.6	34.3	32.1	31.2	30.8	2:1	4:1

Table 21: Per-task LongBench score and NL-CR for Qwen3-0.6B (3/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Synth	passage_count (cnt)	SnapKV	0.5	**0.5**	0.5	0.5	1.0	1.0	**32:1**	**32:1**
		AdaKV		**0.5**	0.5	**1.0**	**1.5**	**2.0**	**32:1**	**32:1**
		KVzip		**0.5**	0.5	0.0	0.0	0.0	4:1	4:1
		KVzip+		**0.5**	**1.0**	0.0	0.0	0.0	4:1	4:1
		KVzip+ (MLP)		**0.5**	0.0	0.0	0.0	0.0	2:1	2:1
		KVgrad		**0.5**	0.5	0.0	0.0	0.0	4:1	4:1
		KVgrad (MLP)		**0.5**	0.5	0.5	0.2	0.1	8:1	8:1
	passage_retrieval_en (ret)	SnapKV	76.0	**74.0**	75.5	**75.0**	**73.5**	**68.5**	**8:1**	**16:1**
		AdaKV		73.5	73.5	72.0	70.0	63.0	1:1	4:1
		KVzip		**74.0**	30.5	13.5	9.0	6.4	1:1	2:1
		KVzip+		73.5	47.0	14.5	9.1	5.0	1:1	2:1
		KVzip+ (MLP)		72.5	28.0	8.5	6.0	3.7	1:1	2:1
		KVgrad		**74.0**	73.0	50.5	17.0	11.5	1:1	4:1
		KVgrad (MLP)		**74.0**	**76.0**	34.5	12.5	5.0	4:1	4:1
Code	lcc (code)	SnapKV	35.2	34.1	**36.9**	37.9	38.7	21.2	**16:1**	**16:1**
		AdaKV		**35.7**	36.4	**39.4**	**39.1**	**21.9**	**16:1**	**16:1**
		KVzip		34.3	35.1	27.2	21.0	21.2	4:1	4:1
		KVzip+		33.8	33.6	26.9	19.4	21.1	1:1	4:1
		KVzip+ (MLP)		28.5	22.0	19.5	20.5	21.2	1:1	1:1
		KVgrad		33.9	35.0	33.0	27.6	21.4	4:1	4:1
		KVgrad (MLP)		33.0	26.8	19.2	13.4	10.3	1:1	1:1
	repobench-p (code)	SnapKV	33.5	33.6	35.2	37.2	38.9	38.5	**32:1**	**32:1**
		AdaKV		**34.3**	**37.0**	**38.9**	**39.5**	**40.1**	**32:1**	**32:1**
		KVzip		33.9	34.3	33.9	35.4	37.9	**32:1**	**32:1**
		KVzip+		33.5	33.9	34.5	35.4	37.0	**32:1**	**32:1**
		KVzip+ (MLP)		32.4	33.3	34.7	36.0	38.2	**32:1**	**32:1**
		KVgrad		32.8	34.0	34.9	36.6	37.5	**32:1**	**32:1**
		KVgrad (MLP)		33.1	33.6	35.2	37.0	36.6	**32:1**	**32:1**

A.6.6 LONGBENCH RESULTS - QWEN3 4B

Table 22: Per-task LongBench score and NL-CR for Qwen3-4B (1/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
S-Doc QA	narrativeqa (F1)	SnapKV	27.0	26.3	27.7	**27.0**	**26.0**	**23.5**	**8:1**	**16:1**
		AdaKV		24.0	23.6	24.0	23.7	22.0	1:1	1:1
		KVzip		**27.1**	27.0	24.7	15.2	6.8	4:1	4:1
		KVzip+		26.5	26.7	25.0	15.6	8.6	4:1	4:1
		KVzip+ (MLP)		25.5	25.3	19.8	14.0	8.7	1:1	1:1
		KVgrad		27.0	26.7	26.7	24.0	18.6	**8:1**	8:1
		KVgrad (MLP)		26.0	**28.1**	24.7	20.8	14.3	4:1	4:1
	qasper (F1)	SnapKV	40.6	38.4	35.8	31.7	29.4	22.5	1:1	1:1
		AdaKV		38.4	35.6	31.4	29.5	23.1	1:1	1:1
		KVzip		41.6	40.4	37.5	9.0	6.0	**4:1**	4:1
		KVzip+		41.0	**42.1**	**39.5**	12.1	6.5	**4:1**	**8:1**
		KVzip+ (MLP)		**42.4**	39.9	35.0	16.9	8.7	**4:1**	4:1
		KVgrad		41.4	40.9	38.6	**31.2**	**24.1**	**4:1**	**8:1**
		KVgrad (MLP)		41.4	41.2	37.4	30.9	18.4	**4:1**	4:1
	multifieldqa_en (F1)	SnapKV	56.6	55.1	55.0	49.8	45.3	34.0	1:1	**4:1**
		AdaKV		55.5	54.0	50.7	**47.1**	**35.8**	2:1	**4:1**
		KVzip		**56.3**	56.7	50.7	26.3	15.4	**4:1**	**4:1**
		KVzip+		56.2	**56.8**	**52.4**	30.8	19.6	**4:1**	**4:1**
		KVzip+ (MLP)		56.1	56.4	50.6	29.7	19.9	**4:1**	**4:1**
		KVgrad		56.1	55.4	52.3	40.2	30.8	2:1	**4:1**
		KVgrad (MLP)		55.5	55.0	52.2	39.8	27.3	2:1	**4:1**
M-Doc QA	hotpotqa (F1)	SnapKV	60.8	60.8	61.8	60.9	57.9	55.3	8:1	**16:1**
		AdaKV		60.8	61.1	**61.8**	**60.0**	**56.4**	**16:1**	**16:1**
		KVzip		61.0	**63.1**	57.4	15.7	5.7	4:1	4:1
		KVzip+		60.8	62.8	60.1	20.6	6.0	8:1	8:1
		KVzip+ (MLP)		**61.3**	59.9	48.9	21.0	9.8	4:1	4:1
		KVgrad		60.5	61.4	61.3	56.4	44.4	8:1	8:1
		KVgrad (MLP)		60.3	60.6	60.2	51.2	38.5	8:1	8:1
	2wikimqa (F1)	SnapKV	46.3	45.0	42.6	42.1	34.2	28.2	1:1	2:1
		AdaKV		45.3	44.7	42.7	37.2	29.7	1:1	4:1
		KVzip		46.3	45.5	37.0	11.2	8.8	**4:1**	4:1
		KVzip+		**47.1**	45.1	41.0	11.1	8.9	2:1	4:1
		KVzip+ (MLP)		45.5	43.4	31.4	14.2	9.4	2:1	2:1
		KVgrad		46.1	44.6	**45.0**	38.2	30.4	2:1	**8:1**
		KVgrad (MLP)		46.1	**46.1**	43.8	**38.6**	**35.0**	**4:1**	4:1
	musique (F1)	SnapKV	35.6	35.3	34.5	**36.3**	31.2	26.9	**8:1**	**8:1**
		AdaKV		35.1	35.0	34.0	**32.0**	**27.3**	4:1	**8:1**
		KVzip		35.2	32.4	26.7	6.9	4.0	2:1	2:1
		KVzip+		35.4	**35.8**	26.7	8.0	4.1	4:1	4:1
		KVzip+ (MLP)		**36.2**	33.5	21.2	6.7	4.6	2:1	2:1
		KVgrad		34.5	33.8	33.2	25.4	18.1	1:1	2:1
		KVgrad (MLP)		34.4	35.6	33.9	27.2	16.5	4:1	**8:1**

Table 23: Per-task LongBench score and NL-CR for Qwen3-4B (2/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						**NL-CR (per τ)**	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Summ	gov_report (R)	SnapKV	31.6	30.0	28.0	25.2	20.9	17.1	1:1	1:1
		AdaKV		30.1	27.9	24.6	21.3	17.4	1:1	2:1
		KVzip		31.2	30.2	27.5	21.1	11.5	**2:1**	**4:1**
		KVzip+		31.1	30.2	27.5	21.2	12.4	**2:1**	**4:1**
		KVzip+ (MLP)		30.9	30.1	26.5	18.8	11.9	1:1	**4:1**
		KVgrad		**31.2**	**30.8**	**28.6**	**25.6**	**21.6**	**2:1**	**4:1**
		KVgrad (MLP)		31.0	30.6	28.1	24.7	19.7	**2:1**	**4:1**
	qmsum (R)	SnapKV	25.0	24.4	24.5	24.0	22.9	22.0	**4:1**	**8:1**
		AdaKV		24.6	24.8	**24.1**	**23.4**	**22.3**	**4:1**	**8:1**
		KVzip		**25.0**	24.4	23.4	21.1	17.8	2:1	4:1
		KVzip+		24.8	**24.8**	23.0	20.8	19.1	**4:1**	4:1
		KVzip+ (MLP)		24.5	24.0	23.2	21.7	20.2	2:1	4:1
		KVgrad		24.8	24.6	23.9	21.9	19.7	**4:1**	**8:1**
		KVgrad (MLP)		24.6	24.7	23.4	22.0	19.7	**4:1**	4:1
	multi_news (R)	SnapKV	25.6	23.7	21.1	18.4	14.2	7.9	1:1	1:1
		AdaKV		23.8	21.4	18.6	14.4	7.4	1:1	1:1
		KVzip		**25.6**	24.9	23.5	14.3	9.8	2:1	**4:1**
		KVzip+		25.3	24.9	23.3	15.9	10.0	2:1	**4:1**
		KVzip+ (MLP)		25.4	25.0	23.5	13.6	9.9	2:1	**4:1**
		KVgrad		25.1	25.1	**23.8**	**20.8**	**17.6**	**4:1**	**4:1**
		KVgrad (MLP)		25.3	**25.2**	23.2	20.0	15.2	**4:1**	**4:1**
Few-shot	trec (cls)	SnapKV	48.0	44.5	43.5	31.5	28.2	18.5	1:1	1:1
		AdaKV		45.0	43.0	35.0	30.8	21.0	1:1	1:1
		KVzip		49.5	65.5	62.5	47.0	0.0	8:1	**16:1**
		KVzip+		46.0	65.0	64.0	52.2	1.0	**16:1**	**16:1**
		KVzip+ (MLP)		**54.5**	66.0	59.0	39.5	3.0	8:1	8:1
		KVgrad		48.5	60.5	**66.5**	**62.2**	**39.4**	**16:1**	**16:1**
		KVgrad (MLP)		54.0	**70.0**	**66.5**	58.0	21.8	**16:1**	**16:1**
	triviaqa (F1)	SnapKV	88.3	87.3	87.7	**88.1**	86.9	83.4	**16:1**	16:1
		AdaKV		87.3	87.3	**88.1**	**87.7**	81.7	**16:1**	16:1
		KVzip		86.8	87.3	86.0	81.2	63.0	4:1	8:1
		KVzip+		**87.5**	87.4	86.3	79.3	66.3	4:1	8:1
		KVzip+ (MLP)		87.3	86.7	84.8	78.8	70.4	4:1	8:1
		KVgrad		**87.5**	**87.9**	87.6	86.4	**86.5**	8:1	**32:1**
		KVgrad (MLP)		86.9	86.8	87.0	86.1	85.5	8:1	**32:1**
	samsum (R)	SnapKV	40.0	39.3	39.4	39.0	38.7	36.5	4:1	**16:1**
		AdaKV		39.8	39.9	**39.9**	**39.2**	36.0	**16:1**	**16:1**
		KVzip		39.6	39.6	38.0	35.9	33.2	4:1	8:1
		KVzip+		39.5	39.6	39.2	36.5	34.2	4:1	8:1
		KVzip+ (MLP)		39.5	38.9	36.7	34.8	33.7	2:1	4:1
		KVgrad		**39.8**	**40.3**	38.8	37.1	**37.3**	4:1	8:1
		KVgrad (MLP)		39.5	39.1	37.9	37.3	36.7	2:1	4:1

Table 24: Per-task LongBench score and NL-CR for Qwen3-4B (3/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Synth	passage_count (cnt)	SnapKV	5.5	6.0	5.0	6.0	4.0	**7.5**	**32:1**	**32:1**
		AdaKV		5.0	5.0	6.0	5.0	7.0	**32:1**	**32:1**
		KVzip		**7.0**	5.5	5.0	2.0	0.0	4:1	4:1
		KVzip+		**7.0**	**7.0**	4.5	2.0	0.5	4:1	4:1
		KVzip+ (MLP)		6.5	3.0	4.0	5.5	2.5	16:1	16:1
		KVgrad		6.5	6.5	**6.5**	**6.5**	4.5	16:1	16:1
		KVgrad (MLP)		6.0	6.5	4.0	2.5	3.5	4:1	4:1
	passage_retrieval_en (ret)	SnapKV	99.0	**99.0**	99.0	**99.0**	**99.3**	96.7	16:1	**32:1**
		AdaKV		**99.0**	99.0	**99.0**	98.8	**98.3**	**32:1**	**32:1**
		KVzip		**99.0**	99.0	94.5	46.5	5.0	4:1	8:1
		KVzip+		**99.0**	98.2	97.1	60.5	11.0	8:1	8:1
		KVzip+ (MLP)		98.5	**99.5**	57.0	6.5	4.5	4:1	4:1
		KVgrad		**99.0**	98.5	**99.0**	98.0	91.5	16:1	16:1
		KVgrad (MLP)		**99.0**	98.0	**99.0**	98.5	85.0	16:1	16:1
Code	lcc (code)	SnapKV	60.8	**64.0**	**64.8**	62.3	58.0	34.9	8:1	**16:1**
		AdaKV		64.0	64.4	**63.0**	**60.0**	33.5	**16:1**	**16:1**
		KVzip		63.9	62.9	52.4	32.3	21.7	4:1	4:1
		KVzip+		63.8	62.5	50.3	26.1	21.9	4:1	4:1
		KVzip+ (MLP)		46.7	34.9	26.6	23.6	21.6	1:1	1:1
		KVgrad		62.6	64.4	63.0	52.0	**41.8**	8:1	8:1
		KVgrad (MLP)		58.0	47.6	40.6	34.8	28.5	1:1	2:1
	repobench-p (code)	SnapKV	55.0	58.1	59.3	60.6	60.7	57.8	**32:1**	**32:1**
		AdaKV		**58.5**	60.2	**60.9**	61.1	59.0	**32:1**	**32:1**
		KVzip		57.1	59.3	**60.9**	61.6	58.8	**32:1**	**32:1**
		KVzip+		56.7	**60.8**	60.8	**62.0**	60.6	**32:1**	**32:1**
		KVzip+ (MLP)		58.1	59.7	60.3	59.1	56.7	**32:1**	**32:1**
		KVgrad		56.2	58.8	60.4	61.0	**60.7**	**32:1**	**32:1**
		KVgrad (MLP)		56.9	59.0	59.7	60.5	60.4	**32:1**	**32:1**

A.6.7 LONGBENCH RESULTS - QWEN3 8B

Table 25: Per-task LongBench score and NL-CR for Qwen3-8B (1/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						**NL-CR (per τ)**	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
S-Doc QA	narrativeqa (F1)	SnapKV	28.6	28.9	28.1	27.7	27.9	**25.9**	2:1	**16:1**
		AdaKV		**29.0**	28.6	**29.3**	**28.1**	25.9	**16:1**	**16:1**
		KVzip		28.7	**29.4**	26.6	15.1	8.6	4:1	4:1
		KVzip+		28.7	28.1	25.9	19.2	11.8	4:1	4:1
		KVzip+ (MLP)		28.1	25.9	24.1	17.8	12.1	2:1	2:1
		KVgrad		28.5	28.8	27.3	24.6	21.2	4:1	8:1
		KVgrad (MLP)		28.8	27.8	24.4	21.9	17.7	2:1	4:1
	qasper (F1)	SnapKV	43.5	43.6	41.5	38.5	32.1	22.5	2:1	**4:1**
		AdaKV		43.9	41.9	39.2	**32.2**	24.0	2:1	**4:1**
		KVzip		43.1	**44.2**	39.6	12.3	6.0	**4:1**	**4:1**
		KVzip+		43.4	42.6	**41.1**	17.9	7.2	2:1	**4:1**
		KVzip+ (MLP)		**45.7**	43.3	38.9	23.9	9.9	**4:1**	**4:1**
		KVgrad		43.7	42.2	38.7	29.6	**24.2**	2:1	**4:1**
		KVgrad (MLP)		43.6	43.2	39.0	30.0	18.4	**4:1**	**4:1**
	multifieldqa_en (F1)	SnapKV	55.6	55.3	52.5	**52.9**	45.8	34.1	2:1	**8:1**
		AdaKV		55.6	53.0	51.1	**46.0**	**36.2**	2:1	4:1
		KVzip		55.3	54.8	52.4	28.7	16.6	**4:1**	4:1
		KVzip+		55.2	**56.1**	51.3	33.2	22.1	**4:1**	4:1
		KVzip+ (MLP)		55.4	53.3	51.8	36.3	21.4	2:1	4:1
		KVgrad		55.0	53.8	48.9	40.7	29.8	2:1	4:1
		KVgrad (MLP)		**56.2**	53.8	50.3	39.9	29.5	2:1	4:1
M-Doc QA	hotpotqa (F1)	SnapKV	62.8	61.6	62.2	61.4	57.9	54.0	4:1	**8:1**
		AdaKV		61.7	61.7	60.0	**59.0**	**56.5**	4:1	**8:1**
		KVzip		62.8	63.4	55.1	12.7	6.6	4:1	4:1
		KVzip+		62.7	**64.7**	61.5	25.6	9.0	4:1	**8:1**
		KVzip+ (MLP)		**63.4**	63.4	58.6	24.8	7.4	4:1	4:1
		KVgrad		62.8	62.5	**63.6**	57.0	48.7	**8:1**	**8:1**
		KVgrad (MLP)		62.4	62.8	61.0	54.6	45.4	4:1	**8:1**
	2wikimqa (F1)	SnapKV	48.9	48.8	47.2	42.8	36.4	30.6	2:1	**4:1**
		AdaKV		48.2	47.8	45.0	38.2	33.5	2:1	**4:1**
		KVzip		**49.7**	49.8	42.8	10.8	9.0	**4:1**	**4:1**
		KVzip+		48.9	48.0	43.1	13.3	9.0	**4:1**	**4:1**
		KVzip+ (MLP)		48.7	49.8	41.8	14.8	9.2	**4:1**	**4:1**
		KVgrad		49.0	**50.9**	**45.7**	39.5	33.5	**4:1**	**4:1**
		KVgrad (MLP)		48.6	48.6	43.9	**45.9**	**35.1**	**4:1**	**4:1**
	musique (F1)	SnapKV	35.9	36.1	35.3	**34.0**	32.7	27.5	**4:1**	**4:1**
		AdaKV		**36.2**	35.3	32.9	**34.0**	**30.5**	**4:1**	**4:1**
		KVzip		35.3	34.7	29.3	7.7	4.9	2:1	**4:1**
		KVzip+		35.7	34.1	28.9	10.2	6.5	2:1	**4:1**
		KVzip+ (MLP)		35.3	32.6	30.0	8.6	4.9	2:1	2:1
		KVgrad		35.2	34.7	33.0	28.3	24.7	2:1	**4:1**
		KVgrad (MLP)		35.9	**36.4**	32.1	30.3	22.7	**4:1**	**4:1**

Table 26: Per-task LongBench score and NL-CR for Qwen3-8B (2/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Summ	gov_report (R)	SnapKV	33.3	32.1	29.9	26.7	23.4	19.4	1:1	2:1
		AdaKV		32.4	29.6	26.5	23.5	20.3	1:1	2:1
		KVzip		**33.7**	33.2	31.5	23.9	12.6	**4:1**	4:1
		KVzip+		33.6	33.3	31.2	24.8	16.4	**4:1**	4:1
		KVzip+ (MLP)		33.6	**33.5**	30.5	23.8	16.4	**4:1**	4:1
		KVgrad		33.7	33.3	31.2	27.8	**23.7**	**4:1**	4:1
		KVgrad (MLP)		33.7	33.3	**31.8**	**28.1**	23.0	**4:1**	**8:1**
	qmsum (R)	SnapKV	24.4	24.2	23.6	23.5	22.2	21.4	2:1	**8:1**
		AdaKV		**24.9**	24.3	23.5	**22.7**	**22.0**	**4:1**	**8:1**
		KVzip		24.7	23.9	**23.8**	21.5	17.4	2:1	**8:1**
		KVzip+		24.4	24.2	23.6	21.2	19.8	**4:1**	**8:1**
		KVzip+ (MLP)		24.3	24.1	23.2	21.6	19.8	**4:1**	**8:1**
		KVgrad		24.7	24.1	23.2	22.2	21.0	**4:1**	4:1
		KVgrad (MLP)		24.7	**24.6**	23.3	21.6	19.4	**4:1**	**8:1**
	multi_news (R)	SnapKV	24.7	23.1	20.7	18.3	14.3	8.2	1:1	1:1
		AdaKV		23.2	20.9	18.8	14.4	7.8	1:1	1:1
		KVzip		24.7	24.3	**24.0**	13.8	8.6	**4:1**	**8:1**
		KVzip+		**24.8**	24.6	23.9	15.0	9.2	**4:1**	**8:1**
		KVzip+ (MLP)		24.8	**24.9**	23.8	14.3	9.2	**4:1**	**8:1**
		KVgrad		24.7	24.4	22.9	19.2	**15.9**	**4:1**	4:1
		KVgrad (MLP)		24.8	24.5	23.1	**19.6**	15.1	**4:1**	4:1
Few-shot	trec (cls)	SnapKV	41.0	39.5	35.0	27.5	18.0	13.5	1:1	2:1
		AdaKV		43.5	44.8	37.2	26.5	16.8	4:1	4:1
		KVzip		**51.5**	**73.0**	61.0	43.0	1.5	16:1	16:1
		KVzip+		**51.5**	70.0	65.2	46.0	3.5	16:1	16:1
		KVzip+ (MLP)		50.0	71.5	64.8	38.0	6.5	8:1	8:1
		KVgrad		48.0	62.0	**72.0**	**61.0**	**46.0**	**32:1**	**32:1**
		KVgrad (MLP)		49.0	72.5	69.0	60.0	28.2	16:1	16:1
	triviaqa (F1)	SnapKV	90.6	89.5	90.2	90.0	90.0	83.8	16:1	16:1
		AdaKV		89.7	90.2	**90.2**	**90.2**	85.8	16:1	16:1
		KVzip		**90.7**	89.6	90.0	81.2	60.1	8:1	8:1
		KVzip+		90.5	89.9	88.1	83.5	65.4	4:1	8:1
		KVzip+ (MLP)		90.1	87.5	87.1	83.6	78.2	2:1	8:1
		KVgrad		90.5	**90.7**	89.6	88.2	**88.9**	**32:1**	**32:1**
		KVgrad (MLP)		**90.7**	90.4	90.2	88.4	88.8	**32:1**	**32:1**
	samsum (R)	SnapKV	40.0	40.5	40.4	39.7	39.2	36.1	8:1	16:1
		AdaKV		40.6	41.0	40.2	**39.5**	36.2	**16:1**	16:1
		KVzip		40.7	41.0	38.9	34.9	32.6	4:1	8:1
		KVzip+		40.3	**41.5**	39.1	35.5	32.6	4:1	8:1
		KVzip+ (MLP)		**41.0**	40.8	38.4	35.2	33.1	4:1	8:1
		KVgrad		40.2	40.9	**40.6**	39.2	**38.2**	**16:1**	**32:1**
		KVgrad (MLP)		40.0	41.4	39.8	38.3	37.7	8:1	16:1

Table 27: Per-task LongBench score and NL-CR for Qwen3-8B (3/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Synth	passage_count (cnt)	SnapKV	10.5	8.5	7.5	5.6	**7.1**	5.6	1:1	1:1
		AdaKV		9.5	9.1	6.0	6.6	7.5	1:1	1:1
		KVzip		**11.0**	**9.5**	8.0	4.5	0.0	2:1	2:1
		KVzip+		9.5	**9.5**	**11.5**	5.0	2.0	**8:1**	**8:1**
		KVzip+ (MLP)		9.6	5.0	6.0	5.0	1.0	1:1	1:1
		KVgrad		10.0	9.0	4.5	5.5	**8.5**	1:1	2:1
		KVgrad (MLP)		9.0	7.5	5.5	5.0	6.0	1:1	1:1
	passage_retrieval_en (ret)	SnapKV	92.0	92.5	93.0	92.3	92.5	**93.3**	**32:1**	**32:1**
		AdaKV		91.9	94.1	93.9	**93.0**	93.2	**32:1**	**32:1**
		KVzip		**93.9**	92.1	92.6	76.7	4.5	8:1	8:1
		KVzip+		93.3	95.5	**96.1**	86.2	18.7	8:1	8:1
		KVzip+ (MLP)		92.5	**97.0**	68.6	7.8	5.5	4:1	4:1
		KVgrad		93.6	93.1	91.6	89.8	90.0	8:1	**32:1**
		KVgrad (MLP)		92.9	92.4	91.4	91.5	80.4	16:1	16:1
Code	lcc (code)	SnapKV	66.8	**68.2**	67.2	66.0	62.3	39.3	**8:1**	8:1
		AdaKV		67.6	67.1	**66.4**	**63.8**	42.0	**8:1**	**16:1**
		KVzip		66.6	64.3	52.0	32.1	16.4	2:1	4:1
		KVzip+		66.8	63.2	49.7	29.8	17.6	2:1	2:1
		KVzip+ (MLP)		44.8	34.4	28.3	23.3	17.1	1:1	1:1
		KVgrad		66.3	**67.3**	62.9	54.0	**46.0**	4:1	4:1
		KVgrad (MLP)		59.7	46.6	37.5	32.0	29.9	1:1	1:1
	repobench-p (code)	SnapKV	61.7	**63.5**	**65.1**	**65.5**	**64.4**	61.9	**32:1**	**32:1**
		AdaKV		62.4	64.7	65.1	64.1	62.6	**32:1**	**32:1**
		KVzip		61.5	62.8	63.9	63.5	**64.1**	**32:1**	**32:1**
		KVzip+		61.1	62.2	62.8	63.7	63.1	**32:1**	**32:1**
		KVzip+ (MLP)		61.1	63.3	62.5	63.0	63.5	**32:1**	**32:1**
		KVgrad		61.6	60.8	62.4	63.1	62.8	**32:1**	**32:1**
		KVgrad (MLP)		62.5	62.3	62.7	62.1	63.0	**32:1**	**32:1**

A.6.8 LONGBENCH RESULTS - LLAMA-3.1-8B

Table 28: Per-task LongBench score and NL-CR for Llama-3.1-8B-Instruct (1/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						**NL-CR (per τ)**	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
S-Doc QA	narrativeqa (F1)	SnapKV	31.0	30.4	29.6	31.1	**29.8**	28.4	**8:1**	**16:1**
		AdaKV		30.4	29.9	30.8	29.5	**29.4**	**8:1**	**16:1**
		KVzip		30.9	**31.7**	31.0	28.6	15.3	**8:1**	8:1
		KVzip+		30.8	31.4	**31.1**	28.0	18.4	**8:1**	8:1
		KVzip+ (MLP)		30.1	29.2	30.8	25.7	16.6	**8:1**	8:1
		KVgrad		**31.0**	30.3	30.8	29.6	22.5	**8:1**	**16:1**
		KVgrad (MLP)		30.8	30.1	30.8	25.3	17.5	**8:1**	8:1
	qasper (F1)	SnapKV	47.4	46.7	44.8	39.5	29.9	21.9	2:1	2:1
		AdaKV		46.0	43.8	40.0	31.9	24.9	1:1	2:1
		KVzip		47.3	45.1	45.3	26.1	11.3	2:1	**8:1**
		KVzip+		47.4	46.0	44.4	24.6	13.2	2:1	4:1
		KVzip+ (MLP)		47.0	45.2	41.9	22.1	11.5	2:1	4:1
		KVgrad		**49.3**	47.2	**46.1**	**39.4**	**27.0**	**4:1**	**8:1**
		KVgrad (MLP)		48.7	**47.7**	44.2	33.1	24.8	**4:1**	4:1
	multifieldqa_en (F1)	SnapKV	56.3	54.6	54.9	52.1	49.6	35.0	1:1	4:1
		AdaKV		55.7	55.3	53.4	**49.8**	**37.5**	4:1	4:1
		KVzip		57.0	54.9	54.7	48.3	20.5	2:1	**8:1**
		KVzip+		56.6	55.6	**56.0**	42.3	18.8	**8:1**	**8:1**
		KVzip+ (MLP)		55.8	55.6	52.9	34.2	18.4	4:1	4:1
		KVgrad		57.2	**57.4**	55.9	46.2	29.8	**8:1**	**8:1**
		KVgrad (MLP)		**57.4**	56.2	54.2	41.9	26.6	4:1	**8:1**
M-Doc QA	hotpotqa (F1)	SnapKV	59.8	**59.1**	**58.8**	**58.6**	57.2	54.0	**8:1**	**16:1**
		AdaKV		58.8	58.6	58.1	**57.8**	**55.6**	2:1	**16:1**
		KVzip		57.7	55.5	55.9	48.0	29.6	1:1	2:1
		KVzip+		57.8	57.6	56.5	45.8	26.3	1:1	4:1
		KVzip+ (MLP)		57.3	57.4	56.9	40.8	23.2	1:1	8:1
		KVgrad		58.8	57.3	56.8	53.1	45.0	2:1	8:1
		KVgrad (MLP)		58.3	55.2	54.5	53.1	38.2	1:1	2:1
	2wikimqa (F1)	SnapKV	51.3	50.8	49.4	46.0	43.0	31.1	2:1	**4:1**
		AdaKV		51.2	50.4	**47.4**	**43.1**	**35.3**	**4:1**	**4:1**
		KVzip		48.1	47.2	44.6	24.0	16.9	1:1	1:1
		KVzip+		48.8	47.0	46.1	27.5	21.9	1:1	2:1
		KVzip+ (MLP)		50.4	49.9	44.1	28.4	21.9	2:1	**4:1**
		KVgrad		50.8	49.4	45.2	41.3	30.5	2:1	**4:1**
		KVgrad (MLP)		**52.1**	**50.5**	44.1	30.5	27.1	**4:1**	**4:1**
	musique (F1)	SnapKV	33.5	32.1	**32.9**	29.9	**28.6**	24.7	**4:1**	**4:1**
		AdaKV		**32.4**	32.9	**31.6**	28.2	**28.4**	**4:1**	**4:1**
		KVzip		31.3	27.7	29.3	20.3	8.8	1:1	1:1
		KVzip+		31.4	29.0	27.4	23.6	10.2	1:1	1:1
		KVzip+ (MLP)		29.2	30.0	27.0	19.3	7.7	1:1	1:1
		KVgrad		32.2	30.9	27.2	27.8	20.0	1:1	2:1
		KVgrad (MLP)		31.8	29.8	27.9	19.1	11.8	1:1	1:1

Table 29: Per-task LongBench score and NL-CR for Llama-3.1-8B-Instruct (2/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						**NL-CR (per τ)**	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Summ	gov_report (R)	SnapKV	35.2	32.7	29.7	26.7	23.9	20.6	1:1	1:1
		AdaKV		32.4	29.5	26.8	24.2	21.2	1:1	1:1
		KVzip		34.8	34.5	32.8	27.2	14.8	2:1	**4:1**
		KVzip+		35.0	34.2	**33.1**	24.8	14.0	2:1	**4:1**
		KVzip+ (MLP)		**35.2**	**34.7**	32.7	22.9	13.6	**4:1**	**4:1**
		KVgrad		34.7	34.1	33.0	**28.9**	**22.9**	2:1	**4:1**
		KVgrad (MLP)		34.3	33.8	31.2	25.2	20.1	1:1	**4:1**
	qmsum (R)	SnapKV	25.3	24.5	24.6	23.8	23.2	22.4	1:1	4:1
		AdaKV		25.2	24.7	24.1	23.7	**23.0**	2:1	4:1
		KVzip		**25.2**	24.8	24.0	22.9	18.5	4:1	4:1
		KVzip+		25.2	24.9	24.4	22.6	18.0	4:1	**8:1**
		KVzip+ (MLP)		25.1	25.1	24.3	22.3	18.4	4:1	**8:1**
		KVgrad		25.1	24.8	25.0	**23.9**	22.4	**8:1**	**8:1**
		KVgrad (MLP)		25.2	**25.1**	**25.1**	23.3	20.5	**8:1**	**8:1**
	multi_news (R)	SnapKV	27.0	25.5	23.5	20.8	16.3	8.7	1:1	1:1
		AdaKV		26.0	23.7	21.5	17.0	8.7	1:1	2:1
		KVzip		27.1	27.0	26.2	**23.5**	8.4	**4:1**	**8:1**
		KVzip+		27.2	26.8	26.3	23.0	8.0	**4:1**	**8:1**
		KVzip+ (MLP)		26.8	27.0	26.3	17.8	7.6	**4:1**	**8:1**
		KVgrad		27.2	**27.1**	**26.4**	23.4	**19.2**	**4:1**	**8:1**
		KVgrad (MLP)		**27.3**	26.8	26.1	22.1	16.3	**4:1**	**8:1**
Few-shot	trec (cls)	SnapKV	29.0	36.0	43.0	35.0	27.0	18.5	8:1	8:1
		AdaKV		40.5	37.0	36.5	35.0	22.5	**16:1**	**16:1**
		KVzip		23.0	53.0	**59.0**	28.0	6.5	8:1	**16:1**
		KVzip+		25.5	**62.5**	56.0	22.5	7.5	8:1	8:1
		KVzip+ (MLP)		**45.5**	58.5	58.5	39.0	5.5	**16:1**	**16:1**
		KVgrad		21.0	27.5	45.0	**44.0**	**24.0**	**16:1**	**16:1**
		KVgrad (MLP)		**45.5**	50.0	44.5	20.0	11.0	8:1	8:1
	triviaqa (F1)	SnapKV	91.7	91.6	91.8	92.0	89.6	87.9	8:1	**32:1**
		AdaKV		91.9	91.8	91.1	90.8	87.0	**16:1**	16:1
		KVzip		90.9	88.9	91.7	**92.1**	88.0	**16:1**	**32:1**
		KVzip+		89.8	88.6	92.0	91.9	87.3	**16:1**	**32:1**
		KVzip+ (MLP)		91.7	**92.9**	**92.2**	88.8	87.4	8:1	**32:1**
		KVgrad		**92.4**	89.8	89.0	88.8	**88.7**	2:1	**32:1**
		KVgrad (MLP)		91.9	91.3	88.4	88.5	86.7	4:1	16:1
	samsum (R)	SnapKV	40.8	40.5	40.6	**40.2**	39.3	35.8	**8:1**	**16:1**
		AdaKV		41.6	**41.3**	39.6	39.0	35.9	4:1	**16:1**
		KVzip		39.4	35.3	37.8	37.7	35.0	1:1	2:1
		KVzip+		39.3	37.9	39.1	36.5	32.8	1:1	8:1
		KVzip+ (MLP)		**41.8**	37.7	36.7	31.0	30.4	2:1	2:1
		KVgrad		38.2	39.8	39.4	**39.4**	**38.2**	1:1	**16:1**
		KVgrad (MLP)		39.3	39.9	38.4	37.6	36.9	1:1	4:1

Table 30: Per-task LongBench score and NL-CR for Llama-3.1-8B-Instruct (3/3). Each task uses its own metric (given after the task name: F1, R = ROUGE, cls, cnt, ret, code). NL-CR at tolerance τ is the largest compression ratio whose score stays within $\tau\%$ (relative) of No press.

			Score (per CR)						NL-CR (per τ)	
Category	Task	Press	No press	2:1	4:1	8:1	16:1	32:1	τ=2	τ=5
Synth	passage_count (cnt)	SnapKV	10.7	10.7	11.1	10.1	**10.5**	6.0	**16:1**	**16:1**
		AdaKV		10.2	10.1	**11.1**	9.6	7.0	8:1	8:1
		KVzip		11.2	11.8	9.3	5.0	2.5	4:1	4:1
		KVzip+		11.3	**12.2**	9.5	5.5	2.5	4:1	4:1
		KVzip+ (MLP)		11.8	11.8	8.0	5.5	1.5	4:1	4:1
		KVgrad		**12.4**	11.2	10.8	9.0	**8.0**	8:1	8:1
		KVgrad (MLP)		11.5	11.7	9.2	9.3	6.9	4:1	4:1
	passage_retrieval_en (ret)	SnapKV	100.0	**100.0**	99.5	99.5	**99.5**	**99.0**	**32:1**	**32:1**
		AdaKV		**100.0**	99.5	99.5	**99.5**	**99.0**	**32:1**	**32:1**
		KVzip		**100.0**	**100.0**	93.5	66.0	14.5	4:1	4:1
		KVzip+		**100.0**	**100.0**	98.0	84.0	12.5	8:1	8:1
		KVzip+ (MLP)		**100.0**	99.5	93.5	10.5	4.5	4:1	4:1
		KVgrad		**100.0**	99.5	99.5	**99.5**	96.5	16:1	**32:1**
		KVgrad (MLP)		**100.0**	**100.0**	**100.0**	**99.5**	98.5	**32:1**	**32:1**
Code	lcc (code)	SnapKV	53.3	53.2	53.9	52.0	**50.0**	32.1	4:1	**8:1**
		AdaKV		51.6	50.8	50.4	48.3	34.4	1:1	4:1
		KVzip		54.1	49.3	47.4	34.6	23.8	2:1	2:1
		KVzip+		55.0	52.0	45.5	31.1	22.9	2:1	4:1
		KVzip+ (MLP)		42.3	32.6	24.7	26.2	22.9	1:1	1:1
		KVgrad		**56.4**	**56.6**	**53.4**	48.0	**37.2**	**8:1**	**8:1**
		KVgrad (MLP)		47.1	32.3	25.3	23.2	21.0	1:1	1:1
	repobench-p (code)	SnapKV	47.2	47.1	47.8	49.2	51.6	52.8	**32:1**	**32:1**
		AdaKV		46.9	46.3	48.3	49.9	51.2	**32:1**	**32:1**
		KVzip		48.4	48.3	52.1	54.4	54.4	**32:1**	**32:1**
		KVzip+		48.3	49.5	**53.2**	**55.1**	**54.6**	**32:1**	**32:1**
		KVzip+ (MLP)		**48.9**	**51.3**	52.6	54.8	54.0	**32:1**	**32:1**
		KVgrad		48.5	49.2	49.6	50.6	50.7	**32:1**	**32:1**
		KVgrad (MLP)		48.3	48.8	49.7	51.1	50.9	**32:1**	**32:1**