PROJCACHE: OUTPUT-PROJECTED KV CACHE EVICTION WITH OBS-INSPIRED COMPENSATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evicting key-value (KV) pairs reduces the memory cost of large language model inference but perturbs attention outputs. We introduce ProjCache, which unifies KV cache eviction and compensation under a shared layer-wise objective: the reconstruction error measured after the attention output projection. We parameterize both operations using token masks applied to attention weights and extend the saliency analysis of Optimal Brain Damage (OBD) to these mask coordinates. By summing all terms in the Taylor expansion of the reconstruction error, we derive an exact single-pair eviction score. This score accounts for attention renormalization and retains cross terms between query heads sharing a KV head in grouped-query attention. For compensation, we adapt Optimal Brain Surgeon (OBS) to jointly optimize the masks of retained tokens using a quadratic approximation of the same reconstruction objective. Experiments on LongBench and RULER with LLaMA-3.1-8B and Qwen3-8B show that replacing heuristic attention scores with ProjCache scores improves performance in 106 of the 108 evaluated settings, with gains of up to 3.4 points on LongBench and 25.6 points on RULER. ProjCache achieves performance comparable to state-of-the-art methods. For compensation, ProjCache lowers the reconstruction error by 17% to 44% in all 24 LongBench and RULER-4K settings, whereas prior compensation methods increase it in every setting.

1 INTRODUCTION

Large language models cache key and value states to avoid recomputing them during autoregressive decoding. The key-value (KV) cache grows linearly with sequence length and batch size (Shazeer, 2019; Kwon et al., 2023). Evicting KV pairs reduces this memory cost but perturbs attention outputs. Preserving these outputs raises two complementary problems: selecting which KV pairs to evict, and compensating for the resulting output perturbation.

For eviction, attention weights alone do not fully capture attention output perturbations, which also depend on value vectors (Guo et al., 2024) and attention renormalization. Existing methods use reconstruction error or bounds on output perturbation to score eviction candidates (Gu et al., 2026; Zhang et al., 2026; Feng et al., 2026). However, computing eviction scores separately for each query head omits cross terms between projected output perturbations of query heads sharing a KV head. For compensation, prior methods merge evicted-token information into retained KV pairs (Zhang et al., 2024; Wan et al., 2024).

We introduce ProjCache, using reconstruction error after the attention output projection as a shared objective for eviction and compensation within each layer. We parameterize both operations using token masks that scale attention weights. In grouped-query attention (GQA; Ainslie et al., 2023), evicting one KV pair perturbs the outputs of all query heads sharing its KV head. For each query, we sum the projected output perturbations of these query heads before taking the squared L_2 norm. The squared norm includes cross terms between the projected output perturbations of different query heads sharing the same KV head. For compensation, we adjust retained token masks to reduce the same reconstruction error while keeping cached keys and values fixed.

We derive eviction scores from a Taylor expansion of reconstruction error in token-mask coordinates, whereas OBCache uses an OBD-style second-order Taylor approximation in key-value coordinates (Gu et al., 2026). With queries, keys, and values fixed, we vary one KV pair's mask from one to zero

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.

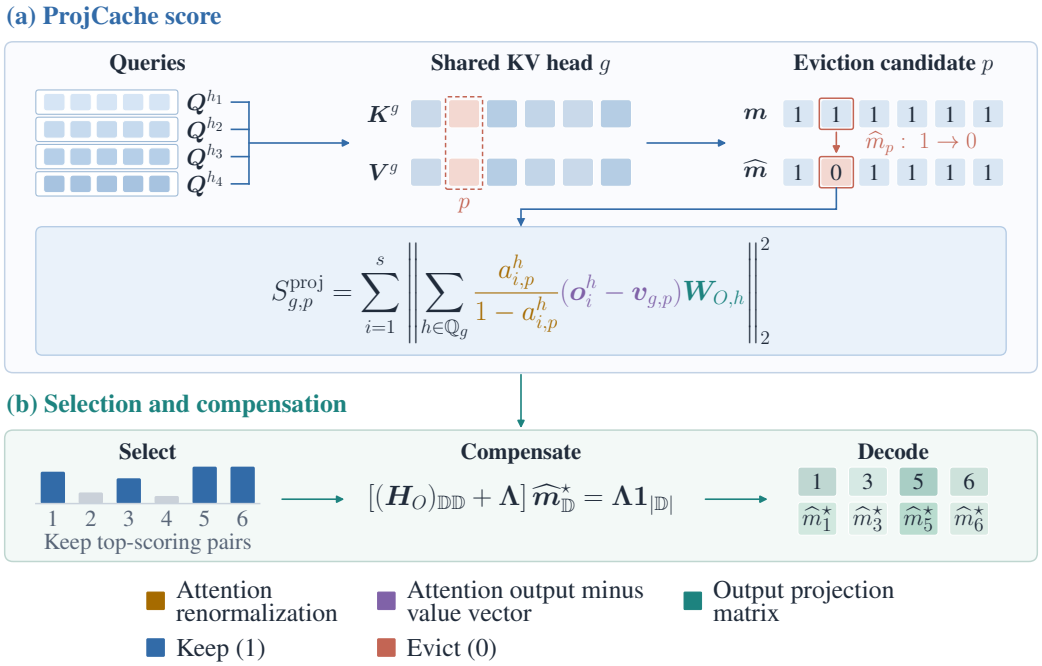

Figure 1: Overview of ProjCache. (a) Single-pair eviction scores measure reconstruction error after output projection, including cross terms between query heads sharing a KV head. (b) We retain high-scoring KV pairs and compensate by adjusting their masks, keeping keys and values fixed.

while keeping all other masks at one. For each query head, summing all Taylor terms gives the exact single-pair reconstruction error before output projection, recovering the DropKV score for the same queries. This provides a Taylor-based interpretation of DropKV. We further apply this expansion to reconstruction error after output projection to derive the exact single-pair eviction score.

For compensation, we fix the eviction set and adapt Optimal Brain Surgeon (OBS; Hassibi & Stork, 1992) to jointly adjust retained token masks. We minimize a regularized quadratic approximation of the same reconstruction objective independently for each KV head, keeping evicted token masks at zero. Figure 1 summarizes scoring, selection, and compensation.

We evaluate ProjCache on LLaMA-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) using RULER (Hsieh et al., 2024) and LongBench (Bai et al., 2024). Without compensation, replacing attention-based scores with ProjCache scores improves task accuracy in 106 of 108 settings across KV budgets and selection policies. The largest gain is 25.6 points on query-aware RULER-4K with Qwen3-8B at a 10% KV budget under AdaKV budget allocation (Feng et al., 2024), where the 13-task average increases from 37.9 to 63.5. Compensation reduces the measured reconstruction error, but gains in task accuracy are small and inconsistent across settings. Improving task accuracy through compensation remains future work. Our contributions are threefold.

- We derive exact single-pair eviction scores by summing all terms in the Taylor expansion of reconstruction error in token-mask coordinates. The derivation applies to reconstruction error measured both before and after output projection. We also develop scoring variants with lower computational cost.
- We derive OBS-inspired compensation by minimizing a regularized quadratic approximation of the reconstruction objective after output projection. The update jointly adjusts retained token masks within each KV head while keeping cached keys and values fixed.
- We conduct extensive experiments on RULER and LongBench across various experimental settings. The results show that replacing attention-based eviction scores with ProjCache scores improves task accuracy in 106 of 108 settings without compensation, and that OBS-inspired compensation reduces reconstruction error in all 24 evaluated compensation settings.

2 Related Work

KV cache compression. For eviction, existing methods select KV pairs using attention weights (Zhang et al., 2023; Liu et al., 2023; Oren et al., 2024; Li et al., 2024), retain attention sinks and recent tokens (Xiao et al., 2024), or allocate KV budgets across layers and heads (Cai et al., 2024; Wang et al., 2024; Fu et al., 2024). Ada-KV uses a bound on output perturbation to allocate head budgets (Feng et al., 2024). CriticalKV scores eviction candidates using an upper bound on the L_1 perturbation of each query head's projected output (Feng et al., 2026). OBCache uses a second-order Taylor expansion in key-value coordinates to estimate reconstruction error from zeroing a token's key and value (Gu et al., 2026). CAOTE measures the L_2 norm of the output perturbation from single-token eviction before output projection, for one query and head (Goel et al., 2026), while DropKV sums the squares of these norms across queries (Zhang et al., 2026). ProjCache measures reconstruction error after output projection, including cross terms between query heads sharing a KV head. For compensation, CaM and D2O merge information from evicted tokens into retained KV pairs (Zhang et al., 2024; Wan et al., 2024). ProjCache instead adjusts retained token masks to scale attention weights, keeping cached keys and values fixed.

Model pruning and compensation. Model pruning removes parameters or structured components (Blalock et al., 2020; Hoefler et al., 2021; Diao et al., 2023; Le et al., 2025; Wang et al., 2026). Pruning methods assess parameter importance using global loss (Molchanov et al., 2019; Singh & Alistarh, 2020) or local reconstruction error (Frantar & Alistarh, 2023). OBD estimates the loss change from setting a parameter to zero using a diagonal second-order Taylor approximation (LeCun et al., 1989). OBS compensates for pruning by updating the remaining parameters under a quadratic approximation of the loss (Hassibi & Stork, 1992). OBCache first adapted OBD to KV cache eviction (Gu et al., 2026). Building on this approach, ProjCache expands reconstruction error in token-mask coordinates and sums all Taylor terms to obtain the exact single-pair eviction score. Furthermore, we adapt OBS to our projected reconstruction objective for compensation. We solve a regularized quadratic problem separately for each KV head to jointly adjust retained token masks.

3 ProjCache: Eviction Scoring and Compensation

We introduce notation and preliminaries in Section 3.1, then define a reconstruction objective in token-mask coordinates (Section 3.2) and sum all Taylor terms to obtain the exact single-pair reconstruction error before output projection (Section 3.3). We then derive projected eviction scores (Section 3.4) and an OBS-inspired compensation update from the same projected reconstruction objective (Section 3.5).

3.1 Notation and Preliminaries

Prefill. Let $\boldsymbol{X} \in \mathbb{R}^{s\times d}$ be the hidden states of s prompt tokens with model dimension d. The attention head (Vaswani et al., 2017) computes $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q$, $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$, and $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V$ with $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d\times d_h}$, where d_h is the head dimension. The attention weights and output are

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{s\times s}, \qquad \boldsymbol{O} = \boldsymbol{A}\boldsymbol{V} \in \mathbb{R}^{s\times d_h}, \tag{1}$$

where $\mathrm{softmax}(\cdot)$ is applied row-wise and denoted by $\sigma(\cdot)$ below.

Decoding. Prefill caches the prompt keys and values. At each decoding position $t > s$, the new token produces a query $\boldsymbol{q}_t$, key $\boldsymbol{k}_t$, and value $\boldsymbol{v}_t$, each in $\mathbb{R}^{1\times d_h}$. After appending $\boldsymbol{k}_t$ and $\boldsymbol{v}_t$ to the cache, the attention output is

$$\boldsymbol{o}_t = \sigma\left(\frac{\boldsymbol{q}_t\boldsymbol{K}^\top}{\sqrt{d_h}}\right)\boldsymbol{V},$$

where $\boldsymbol{K}$ and $\boldsymbol{V}$ denote the cache. Without eviction, the cache length increases by one per step.

Cache eviction. We limit the KV cache to N key-value pairs. If its length s exceeds this limit, we remove $s - N$ pairs and retain N pairs.

Grouped-query attention. Let g and h index KV heads and query heads, respectively. For each KV head g, let $\mathbb{Q}_g$ contain the indices h of query heads sharing its keys and values. The cardinality $|\mathbb{Q}_g|$ is the number of query heads in the group: $|\mathbb{Q}_g| > 1$ for GQA (Ainslie et al., 2023) and $|\mathbb{Q}_g| = 1$ for MHA. For $h \in \mathbb{Q}_g$, the query matrix is $\boldsymbol{Q}^h \in \mathbb{R}^{s\times d_h}$, and the shared key and value

matrices are $\boldsymbol{K}^g, \boldsymbol{V}^g \in \mathbb{R}^{s\times d_h}$. The attention weights and head output are

$$\boldsymbol{A}^h = \sigma\left(\frac{\boldsymbol{Q}^h(\boldsymbol{K}^g)^\top}{\sqrt{d_h}}\right), \qquad \boldsymbol{O}^h = \boldsymbol{A}^h\boldsymbol{V}^g. \tag{2}$$

We write $a^h_{i,j} = (\boldsymbol{A}^h)_{i,j}$. Evicting the KV pair at position p from KV head g perturbs only the outputs of query heads in $\mathbb{Q}_g$.

OBCache objective. OBCache (Gu et al., 2026) estimates token saliency using a second-order Taylor expansion of the attention-output reconstruction error in key-value coordinates. For fixed queries $\boldsymbol{Q}^h$ and a reference cache $(\boldsymbol{K}^g, \boldsymbol{V}^g)$, let $\widehat{\boldsymbol{K}}^g = \boldsymbol{K}^g + \delta\boldsymbol{K}^g$ and $\widehat{\boldsymbol{V}}^g = \boldsymbol{V}^g + \delta\boldsymbol{V}^g$ denote the perturbed key and value matrices. Zeroing row p of both matrices while keeping all other rows unchanged gives the reconstruction error

$$S^{\mathrm{OBCache}}_{g,p} = \sum_{h\in\mathbb{Q}_g}\left\|\sigma\left(\frac{\boldsymbol{Q}^h(\widehat{\boldsymbol{K}}^g)^\top}{\sqrt{d_h}}\right)\widehat{\boldsymbol{V}}^g\Big|_{\boldsymbol{e}_p^\top[\widehat{\boldsymbol{V}}^g\ \widehat{\boldsymbol{K}}^g]=0} - \sigma\left(\frac{\boldsymbol{Q}^h(\boldsymbol{K}^g)^\top}{\sqrt{d_h}}\right)\boldsymbol{V}^g\right\|_F^2, \tag{3}$$

where $\boldsymbol{e}_p \in \mathbb{R}^{s\times 1}$ is the p-th standard basis vector selecting row p.

3.2 Reconstruction in Token-Mask Coordinates

KV zeroing versus eviction. In the OBCache formulation above, zeroing a KV pair leaves its position in the softmax normalization. For a query i that can attend to position p, the zeroed key contributes $\exp(0) = 1$ to the softmax denominator, while the zeroed value contributes zero to the numerator:

$$\left[\sigma\left(\frac{\boldsymbol{Q}^h(\widehat{\boldsymbol{K}}^g)^\top}{\sqrt{d_h}}\right)\widehat{\boldsymbol{V}}^g\right]_{i,:}\Bigg|_{\boldsymbol{e}_p^\top[\widehat{\boldsymbol{V}}^g\ \widehat{\boldsymbol{K}}^g]=0} = \frac{\underbrace{\boldsymbol{0}}_{j=p} + \sum_{j\neq p}\exp\big((\boldsymbol{Q}^h(\boldsymbol{K}^g)^\top)_{i,j}/\sqrt{d_h}\big)\,\boldsymbol{V}^g_{j,:}}{\underbrace{1}_{j=p} + \sum_{j\neq p}\exp\big((\boldsymbol{Q}^h(\boldsymbol{K}^g)^\top)_{i,j}/\sqrt{d_h}\big)}. \tag{4}$$

Both sums include only positions $j \neq p$ that query i can attend to. When the KV pair at position p is evicted, the term 1 in the denominator of equation 4 is replaced by 0. To represent eviction exactly, we use token masks to set the attention weights of evicted positions to zero and renormalize the remaining weights.

Mask coordinates. Let $\boldsymbol{m} = \mathbf{1}_s \in \mathbb{R}^{s\times 1}$ be the mask before eviction, with all KV pairs retained. With queries, keys, and values fixed, we vary $\widehat{\boldsymbol{m}} = \boldsymbol{m} + \delta\boldsymbol{m}$ for KV head g, leaving other KV heads unchanged. For a binary mask $\widehat{\boldsymbol{m}}$, its j-th entry $\widehat{m}_j$ is 1 if the KV pair at position j is retained and 0 if it is evicted. This mask applies to query heads $h \in \mathbb{Q}_g$. Masking and renormalization give

$$\boldsymbol{A}^h(\widehat{\boldsymbol{m}}) := \frac{\boldsymbol{A}^h \odot \widehat{\boldsymbol{m}}^\top}{\boldsymbol{A}^h\widehat{\boldsymbol{m}}}, \qquad \boldsymbol{O}^h(\widehat{\boldsymbol{m}}) := \boldsymbol{A}^h(\widehat{\boldsymbol{m}})\boldsymbol{V}^g. \tag{5}$$

Here $\odot$ denotes element-wise multiplication. We multiply column j of $\boldsymbol{A}^h$ by $\widehat{m}_j$, then divide each resulting row by its sum. These row sums must be positive for every $h \in \mathbb{Q}_g$.

Mask-based reconstruction objective. With all KV pairs retained, $\boldsymbol{A}^h(\boldsymbol{m}) = \boldsymbol{A}^h$ and $\boldsymbol{O}^h(\boldsymbol{m}) = \boldsymbol{O}^h$. To evict only the KV pair at position p from KV head g, we set $\boldsymbol{e}_p^\top\widehat{\boldsymbol{m}} = \widehat{m}_p = 0$ and keep $\widehat{m}_j = 1$ for $j \neq p$. The exact unprojected reconstruction error for this single-pair eviction is

$$S_{g,p} := \sum_{h\in\mathbb{Q}_g}\left\|\boldsymbol{O}^h(\widehat{\boldsymbol{m}})\big|_{\boldsymbol{e}_p^\top\widehat{\boldsymbol{m}}=0} - \boldsymbol{O}^h\right\|_F^2 = \sum_{h\in\mathbb{Q}_g}\left\|\frac{\boldsymbol{A}^h\odot\widehat{\boldsymbol{m}}^\top}{\boldsymbol{A}^h\widehat{\boldsymbol{m}}}\boldsymbol{V}^g\Big|_{\boldsymbol{e}_p^\top\widehat{\boldsymbol{m}}=0} - \boldsymbol{A}^h\boldsymbol{V}^g\right\|_F^2. \tag{6}$$

Before setting $\widehat{m}_p = 0$, we expand the reconstruction error as a function of $\widehat{m}_p$ about $\widehat{m}_p = 1$, keeping all other mask entries at one. Evaluating the infinite Taylor series at $\widehat{m}_p = 0$ gives $S_{g,p}$.

3.3 Exact Unprojected Eviction Score

Proposition 3.1 (Infinite-order recovery of unprojected reconstruction error)**.** *Fix the queries, keys, and values, and set $\widehat{m}_j = 1$ for $j \neq p$. Consider single-pair evictions with $a^h_{i,p} < 1$ for every evaluated query row and every $h \in \mathbb{Q}_g$. The reconstruction error and its first derivative are zero at*

$\widehat{m}_p = 1$, *so the Taylor series starts at second order. For* $K \geq 2$, *the degree-*K *Taylor polynomial of* $\sum_{h\in\mathbb{Q}_g} \|\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h\|_F^2$ *about* $\widehat{m}_p = 1$, *evaluated at* $\widehat{m}_p = 0$, *yields*

$$S_{g,p}^{(K)} = \sum_{h\in\mathbb{Q}_g} \sum_{i=1}^{s} \left[\sum_{r=2}^{K} (r-1)\big(a_{i,p}^h\big)^r \right] \|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2. \tag{7}$$

Here $\boldsymbol{o}_i^h$ *is the output of query head* h *for query* i, *and* $\boldsymbol{v}_{g,p}$ *is the value vector at position* p *in KV head* g. *As* $K \to \infty$, $S_{g,p}^{(K)}$ *converges to the exact unprojected reconstruction error for single-pair eviction:*

$$\lim_{K\to\infty} S_{g,p}^{(K)} = S_{g,p} = \sum_{h\in\mathbb{Q}_g} \sum_{i=1}^{s} \left(\frac{a_{i,p}^h}{1 - a_{i,p}^h} \right)^2 \|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2. \tag{8}$$

The proof is given in Section A.2.

Relation to OBCache. OBCache derives eviction scores from a second-order Taylor approximation in key-value coordinates (Gu et al., 2026). The omitted higher-order terms can affect eviction decisions. We instead sum all Taylor terms in token-mask coordinates to recover the exact single-pair reconstruction error before output projection without Taylor truncation error.

Relation to DropKV. DropKV derives the same exact eviction score as equation 8 (Zhang et al., 2026). Proposition 3.1 expresses the score as the sum of all Taylor terms in token-mask coordinates. The Taylor-based interpretation extends to other reconstruction objectives. In Section 3.4, we incorporate the output projection $\boldsymbol{W}_O$ into the reconstruction objective and derive the corresponding exact projected eviction score.

3.4 Exact Projected Eviction Score

Reconstruction objective. We measure reconstruction error after output projection to preserve the attention output passed to subsequent layers. The objective accounts for $\boldsymbol{W}_O$ and cross terms between the projected output perturbations of query heads sharing a KV head. Since $\boldsymbol{W}_O$ is fixed and linear, summing all Taylor terms still yields the exact projected reconstruction error for single-pair eviction. For query head h, let $\boldsymbol{W}_{O,h} \in \mathbb{R}^{d_h\times d}$ denote the corresponding row block of $\boldsymbol{W}_O$, so the projected attention output is $\sum_h \boldsymbol{O}^h \boldsymbol{W}_{O,h}$. Evicting the KV pair at position p from KV head g perturbs only the outputs of query heads in $\mathbb{Q}_g$. We define the reconstruction error as the squared Frobenius norm of the sum of their projected output perturbations:

$$S_{g,p}^{\mathrm{proj}} := \left\| \sum_{h\in\mathbb{Q}_g} \left(\boldsymbol{O}^h(\widehat{\boldsymbol{m}})\big|_{\boldsymbol{e}_p^\top \widehat{\boldsymbol{m}}=0} - \boldsymbol{O}^h \right) \boldsymbol{W}_{O,h} \right\|_F^2. \tag{9}$$

Here $\widehat{m}_p = 0$, with all other mask entries fixed at one.

Proposition 3.2 (Infinite-order recovery of projected reconstruction error)**.** *Under the assumptions of Proposition 3.1, with the output projection also fixed, define* $S_{g,p}^{\mathrm{proj},(K)}$ *analogously for the projected reconstruction error* $\|\sum_{h\in\mathbb{Q}_g}(\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h)\boldsymbol{W}_{O,h}\|_F^2$. *As* $K \to \infty$, *this approximation converges to the exact projected reconstruction error:*

$$\lim_{K\to\infty} S_{g,p}^{\mathrm{proj},(K)} = S_{g,p}^{\mathrm{proj}} = \sum_{i=1}^{s} \left\| \sum_{h\in\mathbb{Q}_g} \frac{a_{i,p}^h}{1 - a_{i,p}^h} (\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}) \boldsymbol{W}_{O,h} \right\|_2^2. \tag{10}$$

The proof is given in Section A.2.

Output projection and cross-head terms. Expanding the squared norm in equation 10 gives a sum over pairs of query heads $h, h' \in \mathbb{Q}_g$:

$$S_{g,p}^{\mathrm{proj}} = \sum_{i=1}^{s} \sum_{h,h'\in\mathbb{Q}_g} \frac{a_{i,p}^h}{1 - a_{i,p}^h} \frac{a_{i,p}^{h'}}{1 - a_{i,p}^{h'}} (\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}) \boldsymbol{M}_{h,h'} (\boldsymbol{o}_i^{h'} - \boldsymbol{v}_{g,p})^\top, \tag{11}$$

where $\boldsymbol{M}_{h,h'} = \boldsymbol{W}_{O,h}\boldsymbol{W}_{O,h'}^\top \in \mathbb{R}^{d_h\times d_h}$ are the blocks of the Gram matrix $\boldsymbol{M} = \boldsymbol{W}_O\boldsymbol{W}_O^\top \in \mathbb{R}^{(Hd_h)\times(Hd_h)}$. Here H denotes the total number of query heads in the layer. For KV head g, we use only the blocks $\boldsymbol{M}_{h,h'}$ with $h, h' \in \mathbb{Q}_g$. Terms with $\boldsymbol{M}_{h,h}$ give the squared norm of each query head's projected output perturbation. Terms with $\boldsymbol{M}_{h,h'}$, $h \neq h'$, give inner products between the projected output perturbations of different query heads in $\mathbb{Q}_g$.

Scoring variants. Approximating the Gram matrix $\boldsymbol{M}$ reduces the computational cost of scoring. ALL-FULL uses all entries of the blocks $\boldsymbol{M}_{h,h'}$ with $h, h' \in \mathbb{Q}_g$ and computes the exact projected eviction score in equation 10. SELF-FULL approximates this score using only the blocks $\boldsymbol{M}_{h,h}$, omitting cross terms between query heads. SELF-DIAG further approximates each block $\boldsymbol{M}_{h,h}$ by its diagonal. For each variant, the required blocks or diagonal entries can be precomputed from $\boldsymbol{W}_O$. Appendix B.1 describes additional variants and reports their computational costs.

3.5 OBS-Inspired Compensation

Relation to classical OBS. Classical OBS uses a Lagrange multiplier to enforce the pruning constraint when minimizing a second-order Taylor approximation (Hassibi & Stork, 1992). Setting the Lagrangian derivatives with respect to the parameter updates and the multiplier to zero gives a linear system for compensation updates and pruning scores. However, retaining all Taylor orders gives a nonlinear system for the updates and the multiplier, so the standard OBS closed-form solution cannot be applied directly. We therefore select an eviction set using the exact projected eviction score, then jointly adjust retained token masks within each KV head using a regularized quadratic approximation of the projected reconstruction objective.

Reconstruction objective. For KV head g, let $\mathbb{D} \subseteq \{1, \dots, s\}$ be the nonempty set of retained token indices. During compensation, we fix $\widehat{m}_j = 0$ for $j \notin \mathbb{D}$. We use the same projected reconstruction objective as in scoring:

$$\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}}) := \left\| \sum_{h \in \mathbb{Q}_g} \left(\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h \right) \boldsymbol{W}_{O,h} \right\|_F^2 . \tag{12}$$

The reference outputs $\boldsymbol{O}^h$ are those before eviction.

Quadratic compensation. We expand equation 12 to second order about the mask before eviction, $\boldsymbol{m} = \mathbf{1}_s$. Let $\delta\boldsymbol{m} = \widehat{\boldsymbol{m}} - \boldsymbol{m}$ be the mask change and let $\boldsymbol{H}_O \in \mathbb{R}^{s \times s}$ be the positive semidefinite Hessian of $\mathcal{L}_g^{\mathrm{proj}}$ with respect to $\widehat{\boldsymbol{m}}$, evaluated at $\widehat{\boldsymbol{m}} = \boldsymbol{m}$. At $\widehat{\boldsymbol{m}} = \mathbf{1}_s$, the reconstruction error and its gradient are zero, so the second-order Taylor approximation contains only the quadratic term. We jointly adjust the retained token masks by minimizing the regularized quadratic objective:

$$\min_{\delta\boldsymbol{m}} \ \frac{1}{2} \delta\boldsymbol{m}^\top \boldsymbol{H}_O \delta\boldsymbol{m} + \frac{1}{2} \delta\boldsymbol{m}_{\mathbb{D}}^\top \boldsymbol{\Lambda} \delta\boldsymbol{m}_{\mathbb{D}}, \quad \text{s.t.} \quad \delta m_j = -1 \ (j \notin \mathbb{D}). \tag{13}$$

Here $\delta\boldsymbol{m}_{\mathbb{D}}$ contains the changes to the retained token masks. The positive diagonal matrix $\boldsymbol{\Lambda}$ penalizes deviations of the retained masks from one. Solving equation 13 and using $\widehat{\boldsymbol{m}}_{\mathbb{D}}^\star = \mathbf{1}_{|\mathbb{D}|} + \delta\boldsymbol{m}_{\mathbb{D}}^\star$ gives the compensated masks:

$$\widehat{\boldsymbol{m}}_{\mathbb{D}}^\star = \left[(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda} \right]^{-1} \boldsymbol{\Lambda} \, \mathbf{1}_{|\mathbb{D}|}, \tag{14}$$

where $(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} \in \mathbb{R}^{|\mathbb{D}| \times |\mathbb{D}|}$ is the Hessian block for the retained token masks, computed directly at $\widehat{\boldsymbol{m}} = \mathbf{1}_s$. If the solution contains any negative mask value, we skip compensation for that KV head. Appendix A.3 provides the derivation.

4 Experiments

Experimental setup. We evaluate LLaMA-3.1-8B (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) on 16 LongBench tasks (Bai et al., 2024) and 13 RULER tasks in the 4K and 32K settings (Hsieh et al., 2024). For prefill compression, we retain approximately 10%, 20%, or 30% of the prompt KV pairs in both query-aware and query-agnostic settings. We choose three attention-based scoring methods: H2O (Zhang et al., 2023), SnapKV (Li et al., 2024), and AdaKV (Feng et al., 2024). For each method, we compare its attention-based scores with OBCache-VK, DropKV, CriticalKV, and ProjCache scores. For compensation, we compare our OBS-inspired update with CaM (Zhang et al., 2024) and D2O (Wan et al., 2024), keeping the ProjCache eviction sets fixed. Following the experimental setup of OBCache (Gu et al., 2026), we use the last 64 queries as the observation window. ProjCache uses ALL-FULL scoring by default. We assess ProjCache under static KV cache eviction at the prefill stage and dynamic eviction during decoding (see Streaming decoding). Appendix B lists the experimental settings and results for each task.

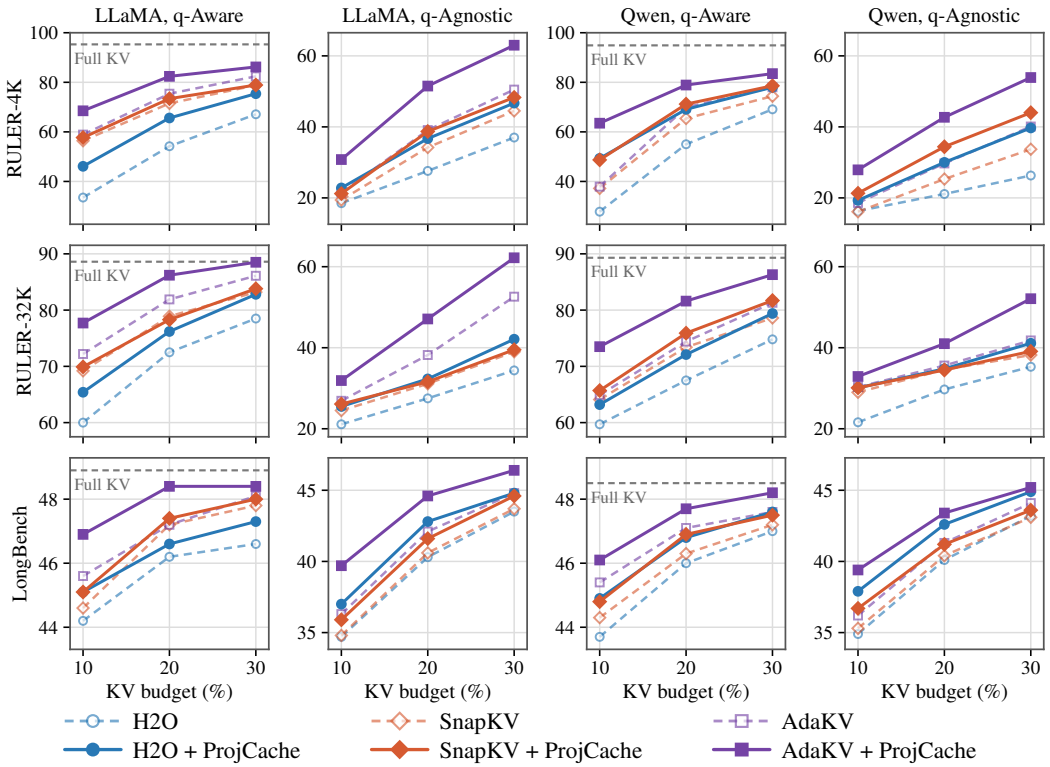

Figure 2: Prefill-stage KV cache eviction results for LLaMA-3.1-8B and Qwen3-8B on LongBench and RULER. ProjCache consistently improves task-average scores over attention-based scoring across KV budgets. LLaMA-3.1-8B attention and Full KV results are from Gu et al. (2026).

Table 1: Prefill-stage KV cache eviction results on LLaMA-3.1-8B. Scores are averaged over 16 LongBench tasks or 13 RULER tasks. Columns show KV budgets (%). Bold marks the best score within each method and column. Wins counts columns in which each row achieves the best score.

	LongBench				RULER-4K				RULER-32K				
	Q-Aware		Q-Agnostic		Q-Aware		Q-Agnostic		Q-Aware		Q-Agnostic		
Method	10	20	10	20	10	20	10	20	10	20	10	20	Wins
Full KV	48.9		48.9		95.3		95.3		88.6		88.6		—
H2O	44.2	46.2	34.7	40.3	33.5	54.2	18.5	27.6	60.0	72.5	21.1	27.5	0
+ DropKV	**45.1**	**46.7**	36.3	42.1	44.4	65.3	22.4	34.9	64.9	75.8	24.7	32.1	2
+ OBCache-VK	44.9	46.4	36.0	41.7	40.7	62.4	21.2	32.5	**65.8**	**76.7**	**25.9**	**33.5**	4
+ CriticalKV	44.6	46.3	34.9	41.0	39.8	60.3	20.2	30.4	63.1	73.8	23.0	30.1	0
+ ProjCache	**45.1**	46.6	**37.0**	**42.8**	**46.1**	**65.6**	**22.8**	**36.7**	65.4	76.2	25.5	32.3	**7**
SnapKV	44.6	47.2	34.8	40.6	56.4	71.5	19.4	34.2	69.2	78.9	24.5	31.1	0
+ DropKV	45.1	47.4	35.6	41.4	58.1	73.4	22.1	37.9	69.8	78.5	25.9	31.7	0
+ OBCache-VK	45.0	47.4	35.7	42.0	54.8	72.0	20.4	36.2	**70.0**	**79.6**	26.2	33.9	2
+ CriticalKV	**45.4**	**47.5**	**36.8**	**43.2**	**61.3**	**75.3**	**25.0**	**43.1**	69.4	78.7	**29.6**	**48.8**	**10**
+ ProjCache	45.1	47.4	35.9	41.6	57.7	73.4	21.2	38.7	69.9	78.3	26.1	31.6	0
AdaKV	45.6	47.2	36.3	42.1	58.9	75.4	21.5	39.1	72.2	81.9	26.9	38.2	0
+ DropKV	46.7	48.2	39.6	**44.6**	68.3	82.3	28.9	51.1	77.5	85.6	31.7	47.1	1
+ OBCache-VK	46.8	48.0	38.9	43.7	63.4	81.1	26.8	49.7	**78.2**	**86.4**	32.1	49.4	2
+ CriticalKV	46.2	47.6	38.1	44.0	62.4	78.4	27.0	48.0	72.4	82.6	**34.1**	**56.9**	2
+ ProjCache	**46.9**	**48.4**	**39.7**	**44.6**	**68.5**	**82.4**	**30.8**	**51.5**	77.7	86.2	31.9	47.1	**8**

Prefill-stage KV cache eviction. ProjCache improves task-average scores over attention-based scoring in 106 of 108 settings, including all 36 H2O settings, as shown in Figure 2. For H2O, SnapKV, and AdaKV, we replace only the eviction scores at each KV budget, without compensation.

Table 2: LongBench scores of ProjCache with ALL-FULL and SELF-DIAG scoring. Columns indicate KV budgets (%). Bold marks the higher score of the two variants.

	LLaMA-3.1-8B						Qwen3-8B					
	Q-Aware			Q-Agnostic			Q-Aware			Q-Agnostic		
Method	10	20	30	10	20	30	10	20	30	10	20	30
H2O												
+ ALL-FULL	45.1	**46.6**	**47.3**	**37.0**	**42.8**	**44.8**	**44.9**	**46.8**	**47.6**	**37.9**	**42.6**	**44.9**
+ SELF-DIAG	**45.2**	**46.6**	47.1	36.5	41.8	44.5	44.8	46.7	**47.6**	37.2	41.6	43.9
SnapKV												
+ ALL-FULL	**45.1**	**47.4**	48.0	**35.9**	**41.6**	**44.6**	**44.8**	**46.9**	**47.5**	**36.7**	**41.2**	**43.6**
+ SELF-DIAG	**45.1**	**47.4**	**48.2**	35.8	41.4	**44.6**	44.6	46.8	**47.5**	**36.7**	**41.2**	43.5
AdaKV												
+ ALL-FULL	**46.9**	**48.4**	**48.4**	39.7	**44.6**	**46.4**	**46.1**	**47.7**	**48.2**	**39.4**	**43.4**	**45.2**
+ SELF-DIAG	**46.9**	48.2	48.2	**40.1**	44.4	46.2	45.9	47.6	**48.2**	38.9	43.1	44.9

The largest gain is 25.6 points, from 37.9 to 63.5, on RULER-4K with Qwen3-8B and AdaKV in the query-aware setting at a 10% KV budget. On LongBench, ProjCache improves scores in all 36 settings. The largest gain is 3.4 points, from 36.3 to 39.7, with LLaMA-3.1-8B and AdaKV in the query-agnostic setting at a 10% KV budget.

Prefill-stage comparison with state-of-the-art eviction scores. ProjCache achieves the most wins on LLaMA-3.1-8B in Table 1, with the best score in 15 of 36 settings, compared with 12 for CriticalKV, 8 for OBCache-VK, and 3 for DropKV. ProjCache wins 7 of 12 H2O settings and 8 of 12 AdaKV settings, including all eight RULER-4K settings across these two methods. CriticalKV's advantage in Table 1 is concentrated in SnapKV, where it wins 10 of 12 settings. ProjCache outperforms CriticalKV in all 12 H2O settings and 10 of 12 AdaKV settings. Appendix B.2 reports the full comparisons across these methods and KV budgets.

Scoring variants. ProjCache with ALL-FULL or SELF-DIAG achieves similar LongBench scores without compensation, as shown in Table 2. Across the 36 evaluated settings, replacing ALL-FULL with SELF-DIAG changes the task-average score by between -1.0 and $+0.4$ points. The absolute difference is at most 0.5 points in 32 settings. The largest decreases occur with H2O in the query-agnostic setting. For SnapKV, all differences are within 0.2 points. SELF-DIAG reduces the computational cost of scoring by retaining only the diagonal entries of the query-head blocks $M_{h,h}$. Appendix B.1 describes additional variants and reports their computational costs.

Streaming decoding. ProjCache with ALL-FULL or SELF-DIAG achieves slightly lower final cumulative perplexity than DropKV on PG19 test set, as shown in Figure 3. SELF-DIAG reaches 10.566, compared with 10.585 for DropKV. We evaluate LLaMA-3.1-8B on 32,768 next-token predictions using teacher forcing and a 1,024-token cache per KV head. Following Xiao et al. (2024), we retain the first four tokens as attention sinks for all methods. H2O, DropKV, and ProjCache also retain the 256 most recent tokens and select the remaining 764 using accumulated per-query eviction scores. Sink fills the remaining 1,020 cache slots with the most recent tokens. DropKV and ProjCache differ only in their eviction scores.

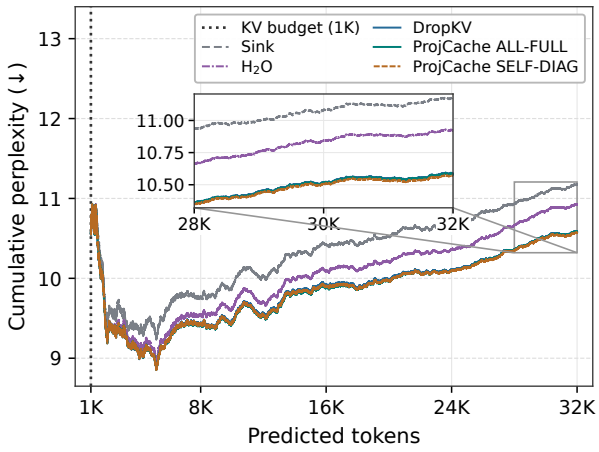

Figure 3: Cumulative perplexity during streaming decoding on PG19 test set with LLaMA-3.1-8B and a 1,024-token cache per KV head.

Table 3: Compensation at a 20% KV budget with fixed ProjCache eviction sets on LongBench and RULER-4K. Avg. is the average score over 16 LongBench tasks or 13 RULER-4K tasks. ΔMSE is the average percentage change in reconstruction MSE relative to uncompensated ProjCache; – indicates no compensation. Bold marks the best value in each column within each model and method.

	RULER-4K				**LongBench**			
	Q-Aware		Q-Agnostic		Q-Aware		Q-Agnostic	
Method	Avg.↑	ΔMSE↓ (%)	Avg.↑	ΔMSE↓ (%)	Avg.↑	ΔMSE↓ (%)	Avg.↑	ΔMSE↓ (%)
LLaMA-3.1-8B								
H2O w/ ProjCache	65.6	–	**36.7**	–	46.6	–	**42.8**	–
+ OBS	66.0	**−28.32**	**36.7**	**−30.63**	**46.8**	**−42.15**	42.5	**−43.75**
+ CaM	**66.2**	+0.51	36.5	+0.90	46.4	+1.78	42.6	+2.37
+ D2O	59.9	>+100	28.5	>+100	45.5	>+100	37.6	>+100
SnapKV w/ ProjCache	73.4	–	**38.7**	–	47.4	–	41.6	–
+ OBS	**73.9**	**−23.63**	38.6	**−21.55**	**47.5**	**−29.97**	**41.7**	**−31.37**
+ CaM	73.6	+0.43	38.4	+0.67	**47.5**	+1.14	41.6	+1.50
+ D2O	64.0	>+100	30.4	>+100	46.0	>+100	37.9	>+100
AdaKV w/ ProjCache	82.4	–	51.5	–	**48.4**	–	**44.6**	–
+ OBS	**82.5**	**−29.16**	**51.7**	**−27.34**	**48.4**	**−37.63**	44.5	**−39.25**
+ CaM	82.3	+0.42	**51.7**	+0.75	48.2	+1.43	**44.6**	+1.92
+ D2O	80.1	>+100	47.0	>+100	46.9	>+100	41.9	>+100
Qwen3-8B								
H2O w/ ProjCache	**69.1**	–	**30.0**	–	46.8	–	**42.6**	–
+ OBS	68.9	**−23.06**	29.9	**−24.87**	46.8	**−32.37**	42.4	**−33.00**
+ CaM	68.8	+16.47	**30.0**	+20.54	**46.9**	+31.34	42.5	+36.66
+ D2O	56.1	>+100	24.4	>+100	44.1	>+100	35.3	>+100
SnapKV w/ ProjCache	71.2	–	34.4	–	**46.9**	–	41.2	–
+ OBS	**71.8**	**−19.20**	**34.7**	**−17.17**	**46.9**	**−24.37**	**41.7**	**−24.46**
+ CaM	71.0	+13.71	34.4	+16.98	46.5	+26.20	41.1	+29.76
+ D2O	55.0	>+100	26.1	>+100	43.1	>+100	34.4	>+100
AdaKV w/ ProjCache	78.9	–	42.7	–	**47.7**	–	43.4	–
+ OBS	**79.1**	**−23.29**	**43.4**	**−21.39**	**47.7**	**−30.82**	**43.7**	**−30.98**
+ CaM	78.4	+19.12	43.1	+23.64	47.4	+35.28	42.8	+40.51
+ D2O	67.9	>+100	37.8	>+100	45.1	>+100	38.5	>+100

Compensation. ProjCache with OBS-inspired compensation reduces reconstruction MSE, while our prefill adaptations of CaM and D2O increase it, as shown in Table 3. All three methods use the same ProjCache eviction set in each setting, at a 20% KV budget. Across 24 settings, OBS reduces MSE by 17.17%–43.75%, CaM increases it by 0.42%–40.51%, and D2O increases it by more than 100%. Relative to uncompensated ProjCache, task-average scores change by -0.3 to $+0.7$ points with OBS and -0.6 to $+0.6$ points with CaM; D2O lowers scores by 1.1–16.2 points. OBS consistently reduces reconstruction MSE, while its effect on task-average scores is small and varies across settings. We suspect that this reflects a gap between the local reconstruction objective and final task performance. Future work aims to narrow this gap. See Appendix B for details.

5 Conclusion

We introduced ProjCache, using reconstruction error after output projection as a shared layer-wise objective for KV cache eviction and compensation. We derive exact single-pair eviction scores before and after output projection by summing all Taylor terms in token-mask coordinates. The projected score accounts for attention renormalization and retains cross terms between the projected output perturbations of different query heads sharing a KV head. We also provide scoring variants that reduce computational cost. For a fixed eviction set, our OBS-inspired compensation jointly adjusts retained token masks within each KV head by minimizing a regularized quadratic approximation of the same projected reconstruction objective, while keeping cached keys and values fixed. ProjCache achieves strong results on LongBench and RULER, improving task performance through eviction scoring and reducing reconstruction error through OBS-inspired compensation. Future work will explore joint optimization of KV cache eviction and compensation and design compensation objectives that produce more consistent improvements in task performance.

AI USE STATEMENT

We used generative AI tools to help revise the manuscript, check the mathematical derivations, and organize the experimental results. We have reviewed all AI-assisted content and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 presents ProjCache, and Appendix A derives its eviction scores and compensation update. Section 4 and Appendix B describe the experimental setup and report the results. Appendix B provides the compensation settings, reconstruction MSE calculation, and per-task results, with links to the CaM and D2O implementations. We will release our code.

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. *arXiv preprint arXiv:2305.13245*, 2023. URL https://arxiv.org/abs/2305.13245.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? In *Proceedings of Machine Learning and Systems*, volume 2, pp. 129–146, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/hash/6c44dc73014d66ba49b28d483a8f8b0d-Abstract.html.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Enmao Diao, Ganghua Wang, Jiawei Zhang, Yuhong Yang, Jie Ding, and Vahid Tarokh. Pruning deep neural networks from a sparsity perspective. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=i-DleYh34BM.

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. URL https://arxiv.org/abs/2407.11550.

Yuan Feng, Junlin Lv, Haoyu Guo, Yukun Cao, S. Kevin Zhou, and Xike Xie. CriticalKV: Optimizing KV cache eviction from an output perturbation perspective. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2502.03805.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning (ICML)*, 2023.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not All Heads Matter: A Head-Level KV Cache Compression Method with Integrated Retrieval and Reasoning. *arXiv preprint arXiv:2410.19258*, 2024. URL https://arxiv.org/abs/2410.19258.

Raghavv Goel, Junyoung Park, Mukul Gagrani, Dalton Jones, Matthew Morse, Matthew Harper Langston, Christopher Lott, and Mingu Lee. CAOTE: Optimizing KV cache memory through attention output error-based token eviction. In *Workshop on Memory for LLM-Based Agentic Systems (MemAgents) at ICLR*, 2026. URL https://openreview.net/forum?id=yolbP0NoZv.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yuzhe Gu, Xiyu Liang, Jiaojiao Zhao, and Enmao Diao. OBCache: Optimal brain KV cache pruning for efficient long-context LLM inference. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2510.07651.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 21158–21166, 2024.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021. URL https://www.jmlr.org/papers/v22/21-0366.html.

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. URL https://arxiv.org/abs/2404.06654.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. *arXiv preprint arXiv:2309.06180*, 2023. URL https://arxiv.org/abs/2309.06180.

Qi Le, Enmao Diao, Ziyan Wang, Xinran Wang, Jie Ding, Li Yang, and Ali Anwar. Probe Pruning: Accelerating LLMs through dynamic pruning via model-probing. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=WOt1owGfuN.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1989.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time. *arXiv preprint arXiv:2305.17118*, 2023. URL https://arxiv.org/abs/2305.17118.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance Estimation for Neural Network Pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1906.10771.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Noam Shazeer. Fast Transformer Decoding: One Write-Head is All You Need. *arXiv preprint arXiv:1911.02150*, 2019. URL https://arxiv.org/abs/1911.02150.

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient Second-Order Approximation for Neural Network Compression. *arXiv preprint arXiv:2004.14340*, 2020. URL https://arxiv.org/abs/2004.14340.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv preprint arXiv:1706.03762*, 2017. URL https://arxiv.org/abs/1706.03762.

Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, Longyue Wang, and Mi Zhang. D2O: Dynamic Discriminative Operations for Efficient Long-Context Inference of Large Language Models. *arXiv preprint arXiv:2406.13035*, 2024. URL https://arxiv.org/abs/2406.13035.

Zihao Wang, Bin Cui, and Shaoduo Gan. SqueezeAttention: 2D Management of KV-Cache in LLM Inference via Layer-wise Optimal Budget. *arXiv preprint arXiv:2404.04793*, 2024. URL https://arxiv.org/abs/2404.04793.

Ziyan Wang, Enmao Diao, Qi Le, Pu Wang, Minwoo Lee, Shu-ping Yeh, Evgeny Stupachenko, Hao Feng, and Li Yang. From local to global: Revisiting structured pruning paradigms for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35720–35739. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1653. URL https://aclanthology.org/2026.acl-long.1653/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Aozhong Zhang, Selcuk Gurses, Yanxia Deng, Naigang Wang, Chi-Chun Liu, Davis Wertheimer, Derrick Liu, Xin Li, Zi Yang, Felix X.-F. Ye, and Penghang Yin. DropKV: Decoupling residual-output perturbation for near-optimal KV-cache eviction. In *Workshop on Resource-Adaptive Foundation Model Inference (AdaptFM) at ICML*, 2026. URL https://openreview.net/forum?id=MqfNzH3TVH.

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. CaM: Cache merging for memory-efficient LLMs inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58840–58850, 2024. URL https://proceedings.mlr.press/v235/zhang24n.html.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Appendix

A Derivations of Eviction Scores and Compensation

We derive the eviction scores and compensation in Section 3, keeping queries, cached keys and values, and $\boldsymbol{W}_O$ fixed. The exact scores describe single-pair eviction; compensation jointly adjusts retained token masks within each KV head.

A.1 Reconstruction in Token-Mask Coordinates

For KV head g, the mask before eviction is $\boldsymbol{m} = \mathbf{1}_s$. We write the changed mask as $\widehat{\boldsymbol{m}} = \boldsymbol{m} + \delta\boldsymbol{m}$, shared by all query heads in $\mathbb{Q}_g$. We sum over query rows $i = 1, \dots, s$. The attention weights $a^h_{i,j}$ include the causal attention mask. Masks used in attention must have nonnegative entries, and the attention denominator must satisfy $\sum_j a^h_{i,j}\widehat{m}_j > 0$ for every query row i and query head $h \in \mathbb{Q}_g$.

Output perturbation. Substituting $\widehat{m}_j = 1 + \delta m_j$ into equation 5 gives

$$\begin{aligned}\boldsymbol{o}^h_i(\widehat{\boldsymbol{m}}) &= \frac{\sum_j a^h_{i,j}(1+\delta m_j)\boldsymbol{v}_{g,j}}{\sum_j a^h_{i,j}(1+\delta m_j)} \\ &= \frac{\sum_j a^h_{i,j}\boldsymbol{v}_{g,j} + \sum_j a^h_{i,j}\delta m_j \boldsymbol{v}_{g,j}}{\sum_j a^h_{i,j} + \sum_j a^h_{i,j}\delta m_j} \\ &= \frac{\boldsymbol{o}^h_i + \sum_j a^h_{i,j}\delta m_j \boldsymbol{v}_{g,j}}{1 + \sum_j a^h_{i,j}\delta m_j}.\end{aligned} \tag{15}$$

The last equality uses $\sum_j a^h_{i,j} = 1$ and $\sum_j a^h_{i,j}\boldsymbol{v}_{g,j} = \boldsymbol{o}^h_i$. Subtracting the output before eviction, $\boldsymbol{o}^h_i$, gives

$$\begin{aligned}\boldsymbol{o}^h_i(\widehat{\boldsymbol{m}}) - \boldsymbol{o}^h_i &= \frac{\boldsymbol{o}^h_i + \sum_j a^h_{i,j}\delta m_j \boldsymbol{v}_{g,j} - \boldsymbol{o}^h_i\left(1 + \sum_j a^h_{i,j}\delta m_j\right)}{1 + \sum_j a^h_{i,j}\delta m_j} \\ &= \frac{\sum_j a^h_{i,j}\delta m_j(\boldsymbol{v}_{g,j} - \boldsymbol{o}^h_i)}{1 + \sum_j a^h_{i,j}\delta m_j}.\end{aligned} \tag{16}$$

The denominator accounts for attention renormalization, including when several mask entries change.

Reconstruction objectives. The unprojected reconstruction objective used in Section 3.2 is

$$\begin{aligned}\mathcal{L}_g(\widehat{\boldsymbol{m}}) &:= \sum_{h\in\mathbb{Q}_g} \|\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h\|_F^2 \\ &= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s} \|\boldsymbol{o}^h_i(\widehat{\boldsymbol{m}}) - \boldsymbol{o}^h_i\|_2^2.\end{aligned} \tag{17}$$

The projected reconstruction objective in equation 12 (Section 3.5) is

$$\begin{aligned}\mathcal{L}^{\mathrm{proj}}_g(\widehat{\boldsymbol{m}}) &:= \left\| \sum_{h\in\mathbb{Q}_g} \left(\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h\right)\boldsymbol{W}_{O,h} \right\|_F^2 \\ &= \sum_{i=1}^{s} \left\| \sum_{h\in\mathbb{Q}_g} \left(\boldsymbol{o}^h_i(\widehat{\boldsymbol{m}}) - \boldsymbol{o}^h_i\right)\boldsymbol{W}_{O,h} \right\|_2^2.\end{aligned} \tag{18}$$

A.2 Exact Eviction Scores

We vary only $\widehat{m}_p$ and keep $\widehat{m}_j = 1$ for $j \neq p$. We assume $a^h_{i,p} < 1$ for every query row i and query head $h \in \mathbb{Q}_g$, so the attention denominator remains positive after eviction.

Before eviction, all token masks equal one, so the reconstruction error is zero. Evicting the KV pair at position p sets only $\widehat{m}_p$ to zero. The unprojected and projected eviction scores therefore equal their corresponding reconstruction errors after eviction:

$$S_{g,p} = \mathcal{L}_g(\widehat{\boldsymbol{m}})|_{\widehat{m}_p=0} - \underbrace{\mathcal{L}_g(\widehat{\boldsymbol{m}})|_{\widehat{m}_p=1}}_{=0} = \mathcal{L}_g(\widehat{\boldsymbol{m}})|_{\widehat{m}_p=0}\,, \tag{19}$$

$$S^{\mathrm{proj}}_{g,p} = \mathcal{L}^{\mathrm{proj}}_g(\widehat{\boldsymbol{m}})\big|_{\widehat{m}_p=0} - \underbrace{\mathcal{L}^{\mathrm{proj}}_g(\widehat{\boldsymbol{m}})\big|_{\widehat{m}_p=1}}_{=0} = \mathcal{L}^{\mathrm{proj}}_g(\widehat{\boldsymbol{m}})\big|_{\widehat{m}_p=0}\,. \tag{20}$$

Proposition 3.1 (restated). *Fix the queries, keys, and values, and set $\widehat{m}_j = 1$ for $j \neq p$. Consider single-pair evictions with $a^h_{i,p} < 1$ for every evaluated query row and every $h \in \mathbb{Q}_g$. The reconstruction error and its first derivative are zero at $\widehat{m}_p = 1$, so the Taylor series starts at second order. For $K \geq 2$, the degree-K Taylor polynomial of $\sum_{h\in\mathbb{Q}_g} \|\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h\|_F^2$ about $\widehat{m}_p = 1$, evaluated at $\widehat{m}_p = 0$, yields*

$$S^{(K)}_{g,p} = \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left[\sum_{r=2}^{K}(r-1)\big(a^h_{i,p}\big)^r\right]\|\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p}\|_2^2. \tag{7}$$

Here $\boldsymbol{o}^h_i$ is the output of query head h for query i, and $\boldsymbol{v}_{g,p}$ is the value vector at position p in KV head g. As $K \to \infty$, $S^{(K)}_{g,p}$ converges to the exact unprojected reconstruction error for single-pair eviction:

$$\lim_{K\to\infty} S^{(K)}_{g,p} = S_{g,p} = \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left(\frac{a^h_{i,p}}{1-a^h_{i,p}}\right)^2\|\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p}\|_2^2. \tag{8}$$

Proof of Proposition 3.1.

Reconstruction error. Substituting $\delta m_p = \widehat{m}_p - 1$ and $\delta m_j = 0$ for $j \neq p$ into equation 16 reduces each sum to its $j = p$ term:

$$\boldsymbol{o}^h_i(\widehat{\boldsymbol{m}}) - \boldsymbol{o}^h_i = \frac{a^h_{i,p}(\widehat{m}_p - 1)(\boldsymbol{v}_{g,p} - \boldsymbol{o}^h_i)}{1 + a^h_{i,p}(\widehat{m}_p - 1)} = \frac{a^h_{i,p}(1-\widehat{m}_p)}{1 - a^h_{i,p}(1-\widehat{m}_p)}(\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p}). \tag{21}$$

Substituting equation 21 into equation 17 gives the reconstruction error below. In the second equality, the scalar coefficient is squared when moved outside the norm:

$$\begin{aligned}
\mathcal{L}_g(\widehat{\boldsymbol{m}}) &= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left\|\frac{a^h_{i,p}(1-\widehat{m}_p)}{1-a^h_{i,p}(1-\widehat{m}_p)}(\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p})\right\|_2^2 \\
&= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left[\frac{a^h_{i,p}(1-\widehat{m}_p)}{1-a^h_{i,p}(1-\widehat{m}_p)}\right]^2\|\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p}\|_2^2 \qquad (22)\\
&= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}(a^h_{i,p})^2\underbrace{(\widehat{m}_p-1)^2}_{\text{first factor}}\times\underbrace{\big[1+a^h_{i,p}(\widehat{m}_p-1)\big]^{-2}}_{\text{second factor}}\|\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p}\|_2^2.
\end{aligned}$$

Taylor expansion. We expand $\mathcal{L}_g$ about $\widehat{m}_p = 1$, keeping $a^h_{i,p}$ and $\|\boldsymbol{o}^h_i - \boldsymbol{v}_{g,p}\|_2^2$ fixed. For $r \geq 2$, applying the general Leibniz rule to the two labeled factors in each summand gives

$$\begin{aligned}
&\frac{\partial^r}{\partial\widehat{m}_p^r}\Big\{(\widehat{m}_p-1)^2\big[1+a^h_{i,p}(\widehat{m}_p-1)\big]^{-2}\Big\} \\
&\qquad = \sum_{\ell=0}^{r}\binom{r}{\ell}\frac{\partial^\ell(\widehat{m}_p-1)^2}{\partial\widehat{m}_p^\ell}\,\frac{\partial^{r-\ell}\big[1+a^h_{i,p}(\widehat{m}_p-1)\big]^{-2}}{\partial\widehat{m}_p^{r-\ell}}.
\end{aligned} \tag{23}$$

At $\widehat{m}_p = 1$, the sum reduces to the $\ell = 2$ term: the first factor and its first derivative are zero, its second derivative is 2, and all higher derivatives are zero. Repeatedly differentiating the second factor gives

$$\frac{\partial^{r-2}}{\partial \widehat{m}_p^{r-2}} \left[1 + a_{i,p}^h(\widehat{m}_p - 1)\right]^{-2}\Big|_{\widehat{m}_p=1} = (-1)^{r-2}(r-1)!(a_{i,p}^h)^{r-2}. \tag{24}$$

Substituting equation 24 into the Leibniz formula in equation 23 gives the rth derivative of the product at $\widehat{m}_p = 1$. Multiplying by the fixed coefficient $(a_{i,p}^h)^2\|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2$ in equation 22 and summing over query rows and query heads gives

$$\begin{aligned}
&\frac{\partial^r \mathcal{L}_g(\widehat{\boldsymbol{m}})}{\partial \widehat{m}_p^r}\Big|_{\widehat{m}_p=1} \\
&= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\binom{r}{2} 2(a_{i,p}^h)^2(-1)^{r-2}(r-1)!(a_{i,p}^h)^{r-2}\|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2 \\
&= (-1)^r r!(r-1)\sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}(a_{i,p}^h)^r\|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2.
\end{aligned} \tag{25}$$

The reconstruction error and its first derivative are zero at $\widehat{m}_p = 1$, so the Taylor expansion starts at $r = 2$. For $K \geq 2$, the degree-K Taylor approximation about $\widehat{m}_p = 1$ is

$$\mathcal{L}_g(\widehat{\boldsymbol{m}}) \approx \sum_{r=2}^{K} \frac{(\widehat{m}_p - 1)^r}{r!} \frac{\partial^r \mathcal{L}_g(\widehat{\boldsymbol{m}})}{\partial \widehat{m}_p^r}\Big|_{\widehat{m}_p=1}. \tag{26}$$

Using equation 19, we obtain the finite-order eviction score by evaluating this Taylor approximation at $\widehat{m}_p = 0$. The factor $(\widehat{m}_p - 1)^r/r!$ then becomes $(0-1)^r/r!$. Substituting equation 25 gives

$$\begin{aligned}
S_{g,p}^{(K)} &= \sum_{r=2}^{K} \frac{(0-1)^r}{r!} \frac{\partial^r \mathcal{L}_g(\widehat{\boldsymbol{m}})}{\partial \widehat{m}_p^r}\Big|_{\widehat{m}_p=1} \\
&= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left[\sum_{r=2}^{K}\frac{(0-1)^r}{r!}(-1)^r r!(r-1)(a_{i,p}^h)^r\right]\|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2 \\
&= \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left[\sum_{r=2}^{K}(r-1)(a_{i,p}^h)^r\right]\|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2.
\end{aligned} \tag{27}$$

This is the finite-order eviction score in equation 7.

Summing all terms. Since $0 \leq a_{i,p}^h < 1$, the Taylor series converges absolutely at $\widehat{m}_p = 0$. As $K \to \infty$, the inner sum in equation 27 converges to

$$\lim_{K\to\infty}\sum_{r=2}^{K}(r-1)(a_{i,p}^h)^r = \sum_{r=2}^{\infty}(r-1)(a_{i,p}^h)^r = \left(\frac{a_{i,p}^h}{1-a_{i,p}^h}\right)^2. \tag{28}$$

Taking $K \to \infty$ in equation 27 and substituting equation 28 gives

$$\lim_{K\to\infty} S_{g,p}^{(K)} = \sum_{h\in\mathbb{Q}_g}\sum_{i=1}^{s}\left(\frac{a_{i,p}^h}{1-a_{i,p}^h}\right)^2\|\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}\|_2^2 = S_{g,p}. \tag{29}$$

This is the exact unprojected eviction score in equation 8. □

Proposition 3.2 (restated)**.** *Under the assumptions of Proposition 3.1, with the output projection also fixed, define $S_{g,p}^{\mathrm{proj},(K)}$ analogously for the projected reconstruction error $\|\sum_{h\in\mathbb{Q}_g}(\boldsymbol{O}^h(\widehat{\boldsymbol{m}}) - \boldsymbol{O}^h)\boldsymbol{W}_{O,h}\|_F^2$. As $K \to \infty$, this approximation converges to the exact projected reconstruction error:*

$$\lim_{K\to\infty} S_{g,p}^{\mathrm{proj},(K)} = S_{g,p}^{\mathrm{proj}} = \sum_{i=1}^{s}\left\|\sum_{h\in\mathbb{Q}_g}\frac{a_{i,p}^h}{1-a_{i,p}^h}(\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p})\boldsymbol{W}_{O,h}\right\|_2^2. \tag{10}$$

Proof of Proposition 3.2.

Reconstruction error. Substituting the output perturbation in equation 21 into the projected objective in equation 18 and expanding the squared norm gives

$$\begin{aligned}\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}}) &= \sum_{i=1}^{s}\left\|\sum_{h\in\mathbb{Q}_g}\frac{a_{i,p}^h(1-\widehat{m}_p)}{1-a_{i,p}^h(1-\widehat{m}_p)}(\boldsymbol{o}_i^h-\boldsymbol{v}_{g,p})\boldsymbol{W}_{O,h}\right\|_2^2 \\ &= \sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g} a_{i,p}^h a_{i,p}^{h'}\underbrace{(\widehat{m}_p-1)^2}_{\text{first factor}} \\ &\quad\times\underbrace{\left[1+a_{i,p}^h(\widehat{m}_p-1)\right]^{-1}\left[1+a_{i,p}^{h'}(\widehat{m}_p-1)\right]^{-1}}_{\text{second factor}} \\ &\quad\times(\boldsymbol{o}_i^h-\boldsymbol{v}_{g,p})\boldsymbol{M}_{h,h'}(\boldsymbol{o}_i^{h'}-\boldsymbol{v}_{g,p})^\top.\end{aligned} \tag{30}$$

Here $\boldsymbol{M}_{h,h'}=\boldsymbol{W}_{O,h}\boldsymbol{W}_{O,h'}^\top$, and the terms with $h\neq h'$ are the cross terms between query heads.

Taylor expansion. The first factor is again $(\widehat{m}_p-1)^2$, as in equation 22, so the reconstruction error and its first derivative are zero at $\widehat{m}_p=1$. For $r\geq 2$, applying the Leibniz rule as in equation 23 at $\widehat{m}_p=1$ leaves only the term that differentiates the first factor twice. This gives $2\binom{r}{2}$ times the $(r-2)$th derivative of the second factor, which is now a product of two reciprocals.

For $\ell\geq 0$, repeated differentiation of one reciprocal factor gives

$$\frac{\partial^\ell}{\partial\widehat{m}_p^\ell}\left[1+a_{i,p}^h(\widehat{m}_p-1)\right]^{-1}\Big|_{\widehat{m}_p=1}=(-1)^\ell\ell!(a_{i,p}^h)^\ell. \tag{31}$$

Applying the Leibniz rule to the product assigns ℓ derivatives to the factor for head h and $r-2-\ell$ to the factor for head h':

$$\begin{aligned}&\frac{\partial^{r-2}}{\partial\widehat{m}_p^{r-2}}\left\{\left[1+a_{i,p}^h(\widehat{m}_p-1)\right]^{-1}\left[1+a_{i,p}^{h'}(\widehat{m}_p-1)\right]^{-1}\right\}\Big|_{\widehat{m}_p=1} \\ &=\sum_{\ell=0}^{r-2}\binom{r-2}{\ell}\left[(-1)^\ell\ell!(a_{i,p}^h)^\ell\right]\left[(-1)^{r-2-\ell}(r-2-\ell)!(a_{i,p}^{h'})^{r-2-\ell}\right] \\ &=(-1)^{r-2}(r-2)!\sum_{\ell=0}^{r-2}(a_{i,p}^h)^\ell(a_{i,p}^{h'})^{r-2-\ell}.\end{aligned} \tag{32}$$

Multiplying equation 32 by $2\binom{r}{2}a_{i,p}^h a_{i,p}^{h'}$ and the fixed inner product in equation 30, then summing over query rows and pairs of query heads, gives

$$\begin{aligned}&\frac{\partial^r\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_p^r}\Big|_{\widehat{m}_p=1} \\ &=(-1)^r r!\sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}\left[a_{i,p}^h a_{i,p}^{h'}\sum_{\ell=0}^{r-2}(a_{i,p}^h)^\ell(a_{i,p}^{h'})^{r-2-\ell}\right] \\ &\quad\times(\boldsymbol{o}_i^h-\boldsymbol{v}_{g,p})\boldsymbol{M}_{h,h'}(\boldsymbol{o}_i^{h'}-\boldsymbol{v}_{g,p})^\top \\ &=(-1)^r r!\sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}\left[\sum_{\ell=1}^{r-1}(a_{i,p}^h)^\ell(a_{i,p}^{h'})^{r-\ell}\right] \\ &\quad\times(\boldsymbol{o}_i^h-\boldsymbol{v}_{g,p})\boldsymbol{M}_{h,h'}(\boldsymbol{o}_i^{h'}-\boldsymbol{v}_{g,p})^\top.\end{aligned} \tag{33}$$

In the last equality, multiplying by $a_{i,p}^h a_{i,p}^{h'}$ increases both exponents by one. We then shift the summation index from $0,\ldots,r-2$ to $1,\ldots,r-1$. For $h=h'$, the inner sum becomes $(r-1)(a_{i,p}^h)^r$, matching the coefficient in equation 25.

For $K \geq 2$, we use the Taylor approximation in equation 26, replacing $\mathcal{L}_g$ by $\mathcal{L}_g^{\mathrm{proj}}$. Using equation 20, we obtain the finite-order projected eviction score by evaluating this approximation at $\widehat{m}_p = 0$. Substituting equation 33 and cancelling the signs and factorials as in equation 27 gives

$$\begin{aligned} S_{g,p}^{\mathrm{proj},(K)} &= \sum_{r=2}^{K} \frac{(0-1)^r}{r!} \left. \frac{\partial^r \mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}})}{\partial \widehat{m}_p^r} \right|_{\widehat{m}_p=1} \\ &= \sum_{i=1}^{s} \sum_{h,h' \in \mathbb{Q}_g} \left[\sum_{r=2}^{K} \sum_{\ell=1}^{r-1} (a_{i,p}^h)^\ell (a_{i,p}^{h'})^{r-\ell} \right] \\ &\quad \times (\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}) \boldsymbol{M}_{h,h'} (\boldsymbol{o}_i^{h'} - \boldsymbol{v}_{g,p})^\top. \end{aligned} \tag{34}$$

Summing all terms. As in the proof of Proposition 3.1, $0 \leq a_{i,p}^h < 1$ ensures absolute convergence at $\widehat{m}_p = 0$. For each head pair, the coefficient sum is the product of two convergent geometric series:

$$\begin{aligned} &\lim_{K\to\infty} \sum_{r=2}^{K} \sum_{\ell=1}^{r-1} (a_{i,p}^h)^\ell (a_{i,p}^{h'})^{r-\ell} \\ &= \sum_{r=2}^{\infty} \sum_{\ell=1}^{r-1} (a_{i,p}^h)^\ell (a_{i,p}^{h'})^{r-\ell} \\ &= \left[\sum_{\ell=1}^{\infty} (a_{i,p}^h)^\ell \right] \left[\sum_{r=1}^{\infty} (a_{i,p}^{h'})^r \right] \\ &= \frac{a_{i,p}^h}{1-a_{i,p}^h} \frac{a_{i,p}^{h'}}{1-a_{i,p}^{h'}}. \end{aligned} \tag{35}$$

In the double sum, ℓ and $r-\ell$ range over all pairs of positive integers, giving the product of geometric series above. Taking $K \to \infty$ in equation 34 and substituting equation 35 gives

$$\begin{aligned} \lim_{K\to\infty} S_{g,p}^{\mathrm{proj},(K)} &= \sum_{i=1}^{s} \sum_{h,h' \in \mathbb{Q}_g} \frac{a_{i,p}^h}{1-a_{i,p}^h} \frac{a_{i,p}^{h'}}{1-a_{i,p}^{h'}} (\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}) \boldsymbol{M}_{h,h'} (\boldsymbol{o}_i^{h'} - \boldsymbol{v}_{g,p})^\top \\ &= \sum_{i=1}^{s} \left\| \sum_{h \in \mathbb{Q}_g} \frac{a_{i,p}^h}{1-a_{i,p}^h} (\boldsymbol{o}_i^h - \boldsymbol{v}_{g,p}) \boldsymbol{W}_{O,h} \right\|_2^2 = S_{g,p}^{\mathrm{proj}}. \end{aligned} \tag{36}$$

The first line is the Gram-matrix expression in equation 11. Using $\boldsymbol{M}_{h,h'} = \boldsymbol{W}_{O,h} \boldsymbol{W}_{O,h'}^\top$ to collect the inner products into a squared norm gives the second line, which is the exact projected eviction score in equation 10. □

These exact scores apply to single-pair eviction. Adding scores for several pairs does not in general give their joint reconstruction error.

A.3 OBS-Inspired Compensation

We derive the compensation update from the projected reconstruction objective in equation 12. For KV head g, let $\boldsymbol{m} = \mathbf{1}_s$ be the mask before eviction and $\delta\boldsymbol{m} = \widehat{\boldsymbol{m}} - \boldsymbol{m}$ the mask change, as in Appendix A.1. We expand with respect to all s token masks. The attention denominator must remain positive for every query row i and query head $h \in \mathbb{Q}_g$.

Taylor expansion. Expanding equation 12 about $\widehat{\boldsymbol{m}} = \boldsymbol{m}$ gives

$$\begin{aligned} \mathcal{L}_g^{\mathrm{proj}}(\boldsymbol{m} + \delta\boldsymbol{m}) = \mathcal{L}_g^{\mathrm{proj}}(\boldsymbol{m}) &+ \sum_{j=1}^{s} \left. \frac{\partial \mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}})}{\partial \widehat{m}_j} \right|_{\widehat{\boldsymbol{m}}=\boldsymbol{m}} \delta m_j \\ &+ \frac{1}{2} \delta\boldsymbol{m}^\top \boldsymbol{H}_O \delta\boldsymbol{m} + \mathcal{O}\big(\|\delta\boldsymbol{m}\|_2^3\big). \end{aligned} \tag{37}$$

Here $\boldsymbol{H}_O \in \mathbb{R}^{s\times s}$ is the Hessian of $\mathcal{L}_g^{\mathrm{proj}}$ with respect to $\widehat{\boldsymbol{m}}$, evaluated at $\widehat{\boldsymbol{m}} = \boldsymbol{m}$. The remainder contains terms of order three and higher as $\delta\boldsymbol{m} \to \mathbf{0}$.

Hessian. To compute $\boldsymbol{H}_O$, we differentiate equation 12 twice. Expanding its squared norm using $\boldsymbol{M}_{h,h'} = \boldsymbol{W}_{O,h}\boldsymbol{W}_{O,h'}^\top$ gives

$$\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}}) = \sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}\big(\boldsymbol{o}_i^h(\widehat{\boldsymbol{m}}) - \boldsymbol{o}_i^h\big)\boldsymbol{M}_{h,h'}\big(\boldsymbol{o}_i^{h'}(\widehat{\boldsymbol{m}}) - \boldsymbol{o}_i^{h'}\big)^\top. \tag{38}$$

Differentiating with respect to $\widehat{m}_j$ gives two sums by the product rule. These sums are equal after exchanging h and h' and using $\boldsymbol{M}_{h',h} = \boldsymbol{M}_{h,h'}^\top$. Thus

$$\frac{\partial\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j} = 2\sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}\frac{\partial\boldsymbol{o}_i^h(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j}\boldsymbol{M}_{h,h'}\big(\boldsymbol{o}_i^{h'}(\widehat{\boldsymbol{m}}) - \boldsymbol{o}_i^{h'}\big)^\top. \tag{39}$$

Before eviction, $\widehat{\boldsymbol{m}} = \boldsymbol{m} = \mathbf{1}_s$ and $\boldsymbol{o}_i^h(\boldsymbol{m}) - \boldsymbol{o}_i^h = \mathbf{0}$ for every query row and query head. The reconstruction error and its gradient are therefore zero.

Differentiating equation 39 with respect to $\widehat{m}_\ell$ gives

$$\frac{\partial^2\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j\,\partial\widehat{m}_\ell} = 2\sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}\left[\begin{array}{l}\dfrac{\partial\boldsymbol{o}_i^h(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j}\boldsymbol{M}_{h,h'}\left(\dfrac{\partial\boldsymbol{o}_i^{h'}(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_\ell}\right)^\top \\ +\dfrac{\partial^2\boldsymbol{o}_i^h(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j\,\partial\widehat{m}_\ell}\boldsymbol{M}_{h,h'}\big(\boldsymbol{o}_i^{h'}(\widehat{\boldsymbol{m}}) - \boldsymbol{o}_i^{h'}\big)^\top\end{array}\right]. \tag{40}$$

At $\widehat{\boldsymbol{m}} = \boldsymbol{m}$, the zero output differences eliminate the second term in the brackets in equation 40.

To evaluate the output derivatives, we use equation 16. At $\widehat{\boldsymbol{m}} = \boldsymbol{m}$, $\delta\boldsymbol{m} = \mathbf{0}$, so the numerator of the output perturbation is zero and the denominator is one. The quotient rule gives

$$\left.\frac{\partial\boldsymbol{o}_i^h(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j}\right|_{\widehat{\boldsymbol{m}}=\boldsymbol{m}} = a_{i,j}^h(\boldsymbol{v}_{g,j} - \boldsymbol{o}_i^h). \tag{41}$$

Substituting equation 41 for both first derivatives in equation 40 gives

$$\begin{aligned}(\boldsymbol{H}_O)_{j,\ell} &= \left.\frac{\partial^2\mathcal{L}_g^{\mathrm{proj}}(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j\,\partial\widehat{m}_\ell}\right|_{\widehat{\boldsymbol{m}}=\boldsymbol{m}} \\ &= 2\sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}\left.\frac{\partial\boldsymbol{o}_i^h(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_j}\right|_{\widehat{\boldsymbol{m}}=\boldsymbol{m}}\boldsymbol{M}_{h,h'}\left[\left.\frac{\partial\boldsymbol{o}_i^{h'}(\widehat{\boldsymbol{m}})}{\partial\widehat{m}_\ell}\right|_{\widehat{\boldsymbol{m}}=\boldsymbol{m}}\right]^\top \\ &= 2\sum_{i=1}^{s}\sum_{h,h'\in\mathbb{Q}_g}a_{i,j}^h a_{i,\ell}^{h'}(\boldsymbol{v}_{g,j} - \boldsymbol{o}_i^h)\boldsymbol{M}_{h,h'}(\boldsymbol{v}_{g,\ell} - \boldsymbol{o}_i^{h'})^\top.\end{aligned} \tag{42}$$

For query row i and query head $h' \in \mathbb{Q}_g$, the attention weights $a_{i,\ell}^{h'}$ satisfy $\sum_{\ell=1}^{s} a_{i,\ell}^{h'} = 1$ and $\sum_{\ell=1}^{s} a_{i,\ell}^{h'}\boldsymbol{v}_{g,\ell} = \boldsymbol{o}_i^{h'}$. Combining these identities gives

$$\sum_{\ell=1}^{s} a_{i,\ell}^{h'}(\boldsymbol{v}_{g,\ell} - \boldsymbol{o}_i^{h'}) = \sum_{\ell=1}^{s} a_{i,\ell}^{h'}\boldsymbol{v}_{g,\ell} - \boldsymbol{o}_i^{h'}\sum_{\ell=1}^{s} a_{i,\ell}^{h'} = \mathbf{0}. \tag{43}$$

Using equation 43 when summing equation 42 over ℓ gives $\sum_{\ell=1}^{s}(\boldsymbol{H}_O)_{j,\ell} = 0$ for every j. Thus $\boldsymbol{H}_O\mathbf{1}_s = \mathbf{0}$, which we use below to simplify the compensation update.

Using $\boldsymbol{M}_{h,h'} = \boldsymbol{W}_{O,h}\boldsymbol{W}_{O,h'}^\top$ in equation 42, we write the Hessian's quadratic form as a sum of squared norms:

$$\delta\boldsymbol{m}^\top\boldsymbol{H}_O\delta\boldsymbol{m} = 2\sum_{i=1}^{s}\left\|\sum_{h\in\mathbb{Q}_g}\sum_{j=1}^{s}a_{i,j}^h\delta m_j(\boldsymbol{v}_{g,j} - \boldsymbol{o}_i^h)\boldsymbol{W}_{O,h}\right\|_2^2 \geq 0. \tag{44}$$

The quadratic form $\delta\boldsymbol{m}^\top\boldsymbol{H}_O\delta\boldsymbol{m}$ is nonnegative for every $\delta\boldsymbol{m}$, so $\boldsymbol{H}_O$ is positive semidefinite.

Quadratic compensation. Since the reconstruction error and its gradient are zero before eviction, dropping the higher-order remainder in equation 37 leaves $\frac{1}{2}\delta\boldsymbol{m}^\top \boldsymbol{H}_O \delta\boldsymbol{m}$. As in Section 3.5, let $\mathbb{D}$ be the nonempty set of retained token indices. Its complement $\mathbb{D}^c = \{1,\dots,s\} \setminus \mathbb{D}$ contains the evicted indices. Adding the retained-mask penalty gives the regularized quadratic problem in equation 13:

$$\min_{\delta\boldsymbol{m}} \ \frac{1}{2}\delta\boldsymbol{m}^\top \boldsymbol{H}_O \delta\boldsymbol{m} + \frac{1}{2}\delta\boldsymbol{m}_{\mathbb{D}}^\top \boldsymbol{\Lambda}\delta\boldsymbol{m}_{\mathbb{D}}, \quad \text{s.t.} \quad \delta\boldsymbol{m}_{\mathbb{D}^c} = -\mathbf{1}_{|\mathbb{D}^c|}. \tag{45}$$

The block $(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}}$ contains retained rows and retained columns; $(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}^c}$ contains retained rows and evicted columns.

To solve for $\delta\boldsymbol{m}$, we differentiate the full objective in equation 45 with respect to $\delta\boldsymbol{m}$, keeping $\boldsymbol{H}_O$ and $\boldsymbol{\Lambda}$ fixed. We set the retained components of the gradient to zero because the constraint fixes all evicted entries:

$$\begin{aligned}
&\left[\frac{\partial}{\partial(\delta\boldsymbol{m})}\left(\frac{1}{2}\delta\boldsymbol{m}^\top \boldsymbol{H}_O \delta\boldsymbol{m} + \frac{1}{2}\delta\boldsymbol{m}_{\mathbb{D}}^\top \boldsymbol{\Lambda}\delta\boldsymbol{m}_{\mathbb{D}}\right)\right]_{\mathbb{D}} \\
&\quad = \frac{1}{2}\big[(\boldsymbol{H}_O + \boldsymbol{H}_O^\top)\delta\boldsymbol{m}\big]_{\mathbb{D}} + \frac{1}{2}(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top)\delta\boldsymbol{m}_{\mathbb{D}} \\
&\quad = (\boldsymbol{H}_O\delta\boldsymbol{m})_{\mathbb{D}} + \boldsymbol{\Lambda}\delta\boldsymbol{m}_{\mathbb{D}} \\
&\quad = (\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}}\delta\boldsymbol{m}_{\mathbb{D}} + (\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}^c}\delta\boldsymbol{m}_{\mathbb{D}^c} + \boldsymbol{\Lambda}\delta\boldsymbol{m}_{\mathbb{D}} \\
&\quad = \big[(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda}\big]\delta\boldsymbol{m}_{\mathbb{D}} + (\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}^c}\delta\boldsymbol{m}_{\mathbb{D}^c} = \mathbf{0}.
\end{aligned} \tag{46}$$

The first equality in equation 46 applies the product rule to both occurrences of the mask change in each quadratic term. The second uses $\boldsymbol{H}_O^\top = \boldsymbol{H}_O$ and $\boldsymbol{\Lambda}^\top = \boldsymbol{\Lambda}$. The third splits the matrix product into the blocks indexed by $\mathbb{D}$ and $\mathbb{D}^c$, and the fourth collects the terms multiplying $\delta\boldsymbol{m}_{\mathbb{D}}$. Substituting $\delta\boldsymbol{m}_{\mathbb{D}^c} = -\mathbf{1}_{|\mathbb{D}^c|}$ into the last line and moving the term for the evicted masks to the right gives a linear system for $\delta\boldsymbol{m}_{\mathbb{D}}$:

$$\big[(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda}\big]\delta\boldsymbol{m}_{\mathbb{D}} = (\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}^c}\mathbf{1}_{|\mathbb{D}^c|}. \tag{47}$$

To simplify the right-hand side, we use $\boldsymbol{H}_O\mathbf{1}_s = \mathbf{0}$, derived above from equation 42 and equation 43. Taking the retained rows gives

$$\mathbf{0} = (\boldsymbol{H}_O\mathbf{1}_s)_{\mathbb{D}} = (\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}}\mathbf{1}_{|\mathbb{D}|} + (\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}^c}\mathbf{1}_{|\mathbb{D}^c|}. \tag{48}$$

Substituting equation 48 into equation 47 and using $\delta\boldsymbol{m}_{\mathbb{D}} = \widehat{\boldsymbol{m}}_{\mathbb{D}} - \mathbf{1}_{|\mathbb{D}|}$ gives

$$\begin{aligned}
\big[(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda}\big]\big(\widehat{\boldsymbol{m}}_{\mathbb{D}} - \mathbf{1}_{|\mathbb{D}|}\big) &= -(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}}\mathbf{1}_{|\mathbb{D}|}, \\
\big[(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda}\big]\widehat{\boldsymbol{m}}_{\mathbb{D}} &= \boldsymbol{\Lambda}\mathbf{1}_{|\mathbb{D}|}.
\end{aligned} \tag{49}$$

Solving this linear system gives the compensated masks:

$$\widehat{\boldsymbol{m}}_{\mathbb{D}}^\star = \big[(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda}\big]^{-1}\boldsymbol{\Lambda}\mathbf{1}_{|\mathbb{D}|}, \tag{50}$$

which agrees with equation 14. The matrix $(\boldsymbol{H}_O)_{\mathbb{D},\mathbb{D}} + \boldsymbol{\Lambda}$ is positive definite, so this is the unique minimizer of equation 45. The retained masks are unconstrained in the quadratic problem, so we check their validity after solving. This solution need not reduce the nonlinear reconstruction error.

B Experimental Setup

Eviction results. Tables 5–16 report scores for each task using different methods at different KV budgets. H2O, SnapKV, and AdaKV are the attention-based baselines. The "+" rows show results after replacing each baseline's attention-based eviction scores with those from another scoring method. For example, "+ ProjCache" under SnapKV uses ProjCache scores in place of SnapKV's attention-based scores, with all other settings unchanged. ProjCache uses ALL-FULL scoring, and ProjCache-SelfDiag uses SELF-DIAG. For LLaMA-3.1-8B, the Full KV, H2O, SnapKV, and AdaKV results are taken from Gu et al. (2026, arXiv v4, Appendix Tables 3, 5, and 7).

Compensation. Tables 17–20 compare compensation methods on each task. For OBS, we set the diagonal damping factor to 50. We skip compensation for a KV head if its solved mask contains any negative value. For CaM, we follow the implementation in KVCache-Factory. For D2O, we follow the authors' LongBench implementation.

Reconstruction MSE. For observation query i, let $\widehat{\boldsymbol{o}}_i^h$ be the output of query head h using the evaluated cache, and let $\boldsymbol{o}_i^h$ be its output using the full cache. Let w be the number of observation queries and d the output dimension after projection. The set $\mathbb{Q}_g$ contains the query heads that share KV head g. The reconstruction MSE for this group is

$$\mathrm{MSE}_g = \frac{1}{wd} \sum_{i=1}^{w} \left\| \sum_{h \in \mathbb{Q}_g} (\widehat{\boldsymbol{o}}_i^h - \boldsymbol{o}_i^h) \boldsymbol{W}_{O,h} \right\|_2^2 . \tag{51}$$

At each evaluated layer, we average MSE_g over batch elements and KV groups. We denote this average by $\mathrm{MSE}_{\text{before}}$ for ProjCache without compensation and $\mathrm{MSE}_{\text{after}}$ for the same cache after compensation. We compute the percentage change in MSE for each sample and layer, average within each task, and then average across tasks:

$$\Delta\mathrm{MSE} = \operatorname*{mean}_{\text{tasks}} \left[\operatorname*{mean}_{\text{samples, layers}} \left[100 \left(\frac{\mathrm{MSE}_{\text{after}}}{\mathrm{MSE}_{\text{before}}} - 1 \right) \right] \right]. \tag{52}$$

We exclude evaluations with $\mathrm{MSE}_{\text{before}} = 0$. A negative $\Delta\mathrm{MSE}$ means that compensation reduces reconstruction MSE.

B.1 Inference Efficiency

Table 4: Scoring variants and per-layer costs. ALL-FULL uses all blocks $\boldsymbol{M}_{h,h'}$, $h, h' \in \mathbb{Q}_g$, to compute the exact projected eviction score in equation 10. SELF-FULL retains only $\boldsymbol{M}_{h,h}$, omitting cross-head terms; SELF-DIAG retains only their diagonal entries. Required blocks or entries are precomputed from $\boldsymbol{W}_O$.

	Head pairs			Within block		
Variant	All	Self	Sparse	Full	Diagonal	Score computational cost
Attention-based	–	–	–	–	–	$O(BHs)$
CriticalKV	–	–	–	–	–	$O(BHsd_h \cdot d)$
ALL-FULL	✓	–	–	✓	–	$O(BHsd_h \cdot \lvert\mathbb{Q}_g\rvert d_h)$
SELF-FULL	–	✓	–	✓	–	$O(BHsd_h \cdot d_h)$
SELF-DIAG	–	✓	–	–	✓	$O(BHsd_h)$
SPARSE-FULL	–	–	✓	✓	–	$O(BHsd_h \cdot rd_h)$
ALL-DIAG	✓	–	–	–	✓	$O(BHsd_h \cdot \lvert\mathbb{Q}_g\rvert)$

Score computational costs. Table 4 reports the computational costs of different variants of the exact projected eviction score. Here B, H, s, d_h, d denote batch size, query-head count, cache length, head dimension, and model dimension; all KV groups have size $|\mathbb{Q}_g|$. SPARSE-FULL uses disjoint subgroups of r heads within each KV group ($r \mid |\mathbb{Q}_g|$); ALL-DIAG keeps only the diagonal entries of every block. Neither variant is evaluated. CriticalKV lists value-projection cost; its attention score adds $O(BHs)$.

Setup. We evaluate LLaMA-3.1-8B on one H100 in BF16, using synthetic inputs of 24K, 32K, 64K, and 96K tokens with batch sizes one and four. Compressed methods use SnapKV selection and retain approximately 10% of prompt KV pairs; eviction occurs once per layer after prefill, and decoding only appends KV pairs. We compare Full KV, ProjCache (ALL-FULL, SELF-FULL, SELF-DIAG), OBCache-VK, and CriticalKV.

Timing and memory. We report means and sample standard deviations over three runs with fresh caches after one warmup. With inputs on the GPU, time to first token (TTFT) includes prefill, scoring, selection, KV compaction, and first-token selection; Gram matrices are constructed during warmup. Decoding latency is the average time per step over 512 greedy steps. Peak allocated GPU memory includes weights, KV states, Gram matrices, and temporary tensors, measured separately for prefill and decoding.

Results. At batch size one and 96K tokens (Figures 4 and 5), SELF-DIAG reduces TTFT from 11.534 to 10.547 s relative to ALL-FULL. Relative to Full KV, it reduces decoding latency from 29.532 to 14.426 ms per step and decoding memory from 27.241 to 16.273 GiB.

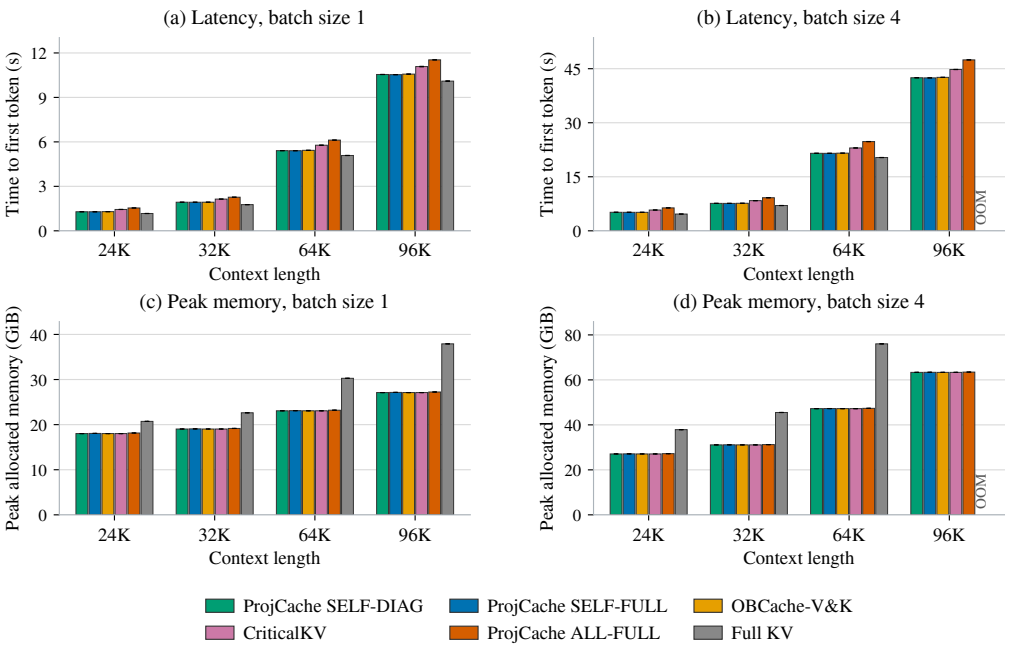

Figure 4: Prefill TTFT and peak GPU memory for LLaMA-3.1-8B on one H100. Error bars show one standard deviation over three runs.

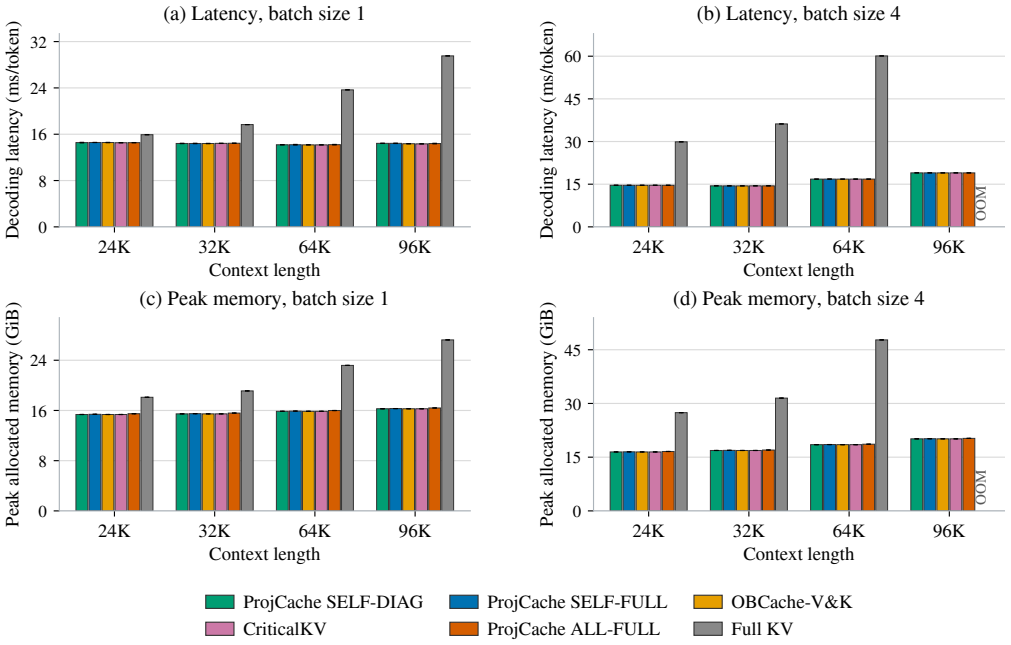

Figure 5: Decoding latency and peak GPU memory for LLaMA-3.1-8B on one H100. Latency averages 512 steps; other settings follow Figure 4.

B.2 RESULTS

Table 5: LongBench scores on LLaMA-3.1-8B (Q-Aware), without compensation.

	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Synthetic		Code		
Method	NQA	Qasper	MF-en	HQA	2WQA	MSQ	GovR	QMSum	MNews	TREC	TQA	SAMSum	PCount	PRe	LCC	RB-P	Avg.
Full KV	30.7	47.7	54.6	59.5	51.0	32.9	35.3	25.2	27.0	72.5	89.5	44.8	10.2	100	52.6	49.2	48.9
10% KV budget																	
H2O	31.2	30.3	50.7	55.8	47.7	27.8	26.4	23.9	20.7	43.5	89.5	43.3	13.1	99.5	52.3	52.2	44.2
+ DropKV	30.3	35.1	52.0	55.5	47.7	31.4	27.1	24.3	20.9	47.5	90.1	43.9	12.7	100	51.6	52.5	45.1
+ OBCache-VK	30.6	33.8	49.4	55.5	48.8	29.5	26.8	23.7	20.7	48.5	89.7	43.1	12.6	100	52.2	52.8	44.9
+ CriticalKV	30.4	32.6	51.8	55.8	47.2	31.1	26.8	23.9	20.8	45.0	89.4	43.0	11.6	100	51.6	52.8	44.6
+ ProjCache-SelfDiag	30.4	34.5	51.7	55.9	50.6	29.9	26.9	24.2	21.1	48.0	90.0	43.3	11.7	100	51.7	52.7	45.2
+ ProjCache	29.9	34.1	52.5	56.0	48.3	31.2	27.1	24.3	21.1	48.5	90.0	43.6	12.6	100	50.4	52.6	45.1
SnapKV	30.9	37.1	53.9	57.1	44.4	27.9	25.7	23.9	19.6	48.0	87.8	42.5	10.1	99.5	53.0	52.2	44.6
+ DropKV	31.0	37.4	53.2	55.9	48.4	29.5	26.5	24.3	20.0	48.0	90.0	43.8	11.1	99.5	52.2	51.4	45.1
+ OBCache-VK	31.1	38.5	51.8	57.0	45.9	28.5	26.3	24.4	20.3	48.5	89.0	43.6	11.0	99.5	52.4	51.8	45.0
+ CriticalKV	31.1	39.9	53.1	57.9	47.1	31.8	26.7	24.1	20.4	46.0	89.5	42.9	10.6	100	53.0	52.1	45.4
+ ProjCache-SelfDiag	31.1	37.9	53.8	56.6	48.3	29.6	26.5	24.3	19.9	46.5	89.7	43.6	10.6	100	52.1	51.9	45.1
+ ProjCache	30.8	39.2	53.1	56.0	46.6	29.6	26.5	24.0	20.2	47.0	90.2	43.5	11.1	100	51.5	52.2	45.1
AdaKV	30.9	38.9	53.2	57.8	47.1	30.7	26.1	24.3	20.0	52.0	89.6	43.2	10.1	99.5	54.0	52.7	45.6
+ DropKV	31.0	41.5	54.7	57.6	48.5	32.2	27.4	24.2	20.9	58.5	88.6	44.1	10.6	100	53.2	54.0	46.7
+ OBCache-VK	30.5	41.6	55.5	58.0	47.5	32.5	26.9	24.0	21.0	58.0	90.3	43.7	11.5	100	54.2	54.0	46.8
+ CriticalKV	30.2	40.3	53.8	58.0	47.5	32.8	26.8	24.5	21.1	53.0	89.0	43.1	11.6	100	54.0	53.3	46.2
+ ProjCache-SelfDiag	31.1	41.5	54.6	57.5	48.8	31.9	27.4	24.5	20.9	59.0	89.9	44.6	11.6	100	53.1	54.1	46.9
+ ProjCache	30.1	41.6	54.8	57.8	49.6	32.2	27.6	25.2	20.8	58.5	88.9	44.1	12.1	100	52.7	54.0	46.9
20% KV budget																	
H2O	31.8	37.5	52.8	57.3	50.9	29.8	28.5	24.3	22.8	53.0	90.2	43.1	13.7	99.5	52.4	52.2	46.2
+ DropKV	30.8	41.3	54.8	57.2	50.7	31.8	29.1	24.2	23.4	55.0	90.2	43.5	12.2	100	52.3	50.8	46.7
+ OBCache-VK	29.6	39.7	52.8	56.9	49.3	32.6	28.8	24.3	23.3	56.5	90.1	43.2	12.2	100	51.7	50.8	46.4
+ CriticalKV	30.0	40.0	52.1	56.6	49.3	30.8	29.1	24.2	23.3	56.0	90.0	43.7	11.7	100	52.0	51.3	46.3
+ ProjCache-SelfDiag	30.7	41.7	54.2	57.0	50.0	31.3	29.6	24.1	23.5	54.5	90.3	43.6	12.2	100	51.3	51.4	46.6
+ ProjCache	30.9	40.1	54.8	57.2	50.0	33.0	29.2	24.2	24.0	54.5	89.9	43.8	11.2	100	51.3	51.1	46.6
SnapKV	29.8	44.4	55.0	58.8	50.2	32.3	28.6	24.6	22.8	59.0	89.5	44.2	11.1	99.5	53.7	51.5	47.2
+ DropKV	29.6	44.5	56.1	58.8	50.1	31.8	29.2	24.6	23.0	59.5	89.8	44.3	12.6	100	53.5	50.8	47.4
+ OBCache-VK	30.1	43.6	55.6	59.1	50.6	33.2	28.9	24.9	23.0	60.5	89.4	43.8	11.1	100	52.9	51.5	47.4
+ CriticalKV	30.4	44.7	55.2	58.4	49.9	32.7	29.8	24.6	23.5	61.5	89.4	43.9	11.4	100	53.7	51.9	47.5
+ ProjCache-SelfDiag	30.0	45.7	55.9	59.1	50.4	31.0	29.2	24.6	22.7	59.5	89.6	44.3	11.1	100	53.9	51.2	47.4
+ ProjCache	29.9	45.7	55.2	58.9	50.2	31.9	29.1	24.6	23.2	58.0	89.4	44.3	12.1	100	53.5	52.2	47.4
AdaKV	29.6	43.1	55.1	58.5	49.3	32.3	28.6	24.5	23.5	62.5	89.6	43.5	11.1	100	53.1	51.6	47.2
+ DropKV	29.1	44.9	56.2	59.4	50.5	33.8	29.7	25.1	23.8	68.5	89.5	44.3	12.1	100	52.5	51.9	48.2
+ OBCache-VK	29.3	43.9	56.7	57.9	51.5	33.7	29.2	24.6	23.7	66.5	89.1	44.3	11.6	100	53.9	51.8	48.0
+ CriticalKV	30.3	45.1	55.3	57.9	50.2	32.9	29.5	24.3	23.5	62.5	89.6	44.2	11.1	100	53.3	51.7	47.6
+ ProjCache-SelfDiag	29.7	44.3	55.9	58.8	51.3	32.9	29.9	24.9	24.0	69.0	90.0	44.6	11.1	100	53.0	52.4	48.2
+ ProjCache	29.8	45.9	56.5	58.6	50.6	33.5	30.1	25.0	23.9	68.5	90.0	44.4	11.6	100	52.8	53.1	48.4
30% KV budget																	
H2O	30.7	41.1	54.1	55.9	49.6	29.8	30.0	24.9	24.1	56.0	90.0	43.6	11.7	99.5	52.5	51.4	46.6
+ DropKV	30.8	43.1	55.5	56.7	50.9	31.1	30.7	24.8	24.3	58.0	90.7	43.8	12.2	100	52.5	50.5	47.2
+ OBCache-VK	30.9	43.1	54.5	57.6	50.3	29.7	30.6	24.6	24.4	61.0	90.7	43.8	12.2	100	51.7	50.5	47.2
+ CriticalKV	30.4	44.1	55.0	57.0	50.2	31.1	30.8	24.9	24.5	58.0	89.5	43.7	12.2	100	52.3	50.6	47.2
+ ProjCache-SelfDiag	30.8	43.5	54.9	56.6	49.3	31.6	30.5	24.4	24.4	58.5	90.7	43.6	12.7	100	51.5	50.3	47.1
+ ProjCache	30.4	44.2	55.2	57.8	49.5	32.1	31.0	24.3	24.5	58.5	90.2	43.8	12.7	100	51.5	51.0	47.3
SnapKV	29.8	46.9	56.1	58.9	51.5	32.6	30.5	24.4	24.0	64.0	89.4	43.1	10.1	99.5	53.3	51.0	47.8
+ DropKV	30.4	45.6	55.3	59.8	50.5	32.3	30.8	24.5	24.2	68.0	89.7	44.5	11.6	100	53.8	50.7	48.2
+ OBCache-VK	30.1	45.6	55.4	58.8	50.6	32.7	30.6	24.8	24.2	66.5	89.5	44.2	12.1	100	53.5	50.5	48.1
+ CriticalKV	30.5	46.5	54.3	58.9	52.5	34.0	31.6	24.4	24.8	66.0	89.1	44.1	10.6	100	52.0	50.2	48.1
+ ProjCache-SelfDiag	30.1	46.3	55.5	59.4	51.1	33.1	30.6	24.7	24.2	67.5	90.0	44.0	11.1	100	53.1	51.0	48.2
+ ProjCache	30.0	45.3	55.4	59.2	51.4	32.3	30.8	24.7	24.5	66.0	89.7	44.1	11.1	100	53.0	51.1	48.0
AdaKV	30.1	46.3	56.3	58.7	51.7	32.2	30.4	24.8	24.1	66.0	89.9	44.2	11.1	100	52.1	51.0	48.1
+ DropKV	29.9	46.4	56.3	58.8	51.7	32.9	31.5	24.6	24.9	70.0	89.7	44.0	10.6	100	52.5	50.9	48.4
+ OBCache-VK	29.5	46.5	55.1	58.6	52.5	32.6	30.8	25.0	24.8	70.5	89.8	44.2	10.6	100	52.6	51.0	48.4
+ CriticalKV	30.1	46.9	55.0	58.5	53.0	33.4	31.7	24.8	25.0	68.5	90.0	44.4	12.1	100	52.2	50.1	48.5
+ ProjCache-SelfDiag	30.4	46.1	55.3	58.0	50.1	32.5	31.4	25.2	24.9	71.0	89.8	44.0	10.1	100	51.7	50.4	48.2
+ ProjCache	29.6	46.1	56.1	58.4	51.0	32.8	31.4	24.7	25.0	70.5	89.7	44.2	11.7	100	52.2	50.8	48.4

Table 6: LongBench scores on LLaMA-3.1-8B (Q-Agnostic), without compensation.

	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Synthetic		Code		
Method	NQA	Qasper	MF-en	HQA	2WQA	MSQ	GovR	QMSum	MNews	TREC	TQA	SAMSum	PCount	PRe	LCC	RB-P	Avg.
Full KV	30.7	47.7	54.6	59.5	51.0	32.9	35.3	25.2	27.0	72.5	89.5	44.8	10.2	100	52.6	49.2	48.9
10% KV budget																	
H2O	25.8	15.3	26.1	41.8	24.1	18.0	25.4	19.9	20.9	44.0	90.9	43.4	9.50	50.0	52.3	47.9	34.7
+ DropKV	27.0	16.5	24.9	44.6	27.6	20.8	26.2	20.4	21.1	44.0	90.6	43.5	10.1	63.5	51.6	47.8	36.3
+ OBCache-VK	26.3	18.3	25.9	42.1	28.6	20.1	25.9	20.7	21.2	45.5	90.6	43.7	10.1	57.0	52.2	47.8	36.0
+ CriticalKV	26.9	17.2	25.4	44.6	25.9	18.6	25.9	20.1	21.1	45.0	89.6	43.0	9.60	47.0	51.6	47.1	34.9
+ ProjCache-SelfDiag	27.7	16.9	27.4	45.2	26.1	19.8	25.9	20.6	20.9	45.5	90.3	43.6	9.10	65.5	51.7	47.4	36.5
+ ProjCache	28.0	17.5	26.5	45.9	26.6	19.7	26.1	20.1	21.3	45.0	90.8	43.6	11.0	72.0	50.4	48.0	37.0
SnapKV	24.2	20.5	23.4	43.2	23.4	19.9	25.3	19.8	20.0	41.5	89.6	42.8	7.00	54.5	53.0	48.6	34.8
+ DropKV	23.3	22.1	23.5	45.9	25.0	19.4	25.8	21.3	20.4	43.5	89.6	43.0	8.00	59.5	52.2	47.6	35.6
+ OBCache-VK	26.4	21.4	24.6	44.3	25.7	20.4	25.6	21.0	20.6	44.5	90.0	43.2	8.00	55.0	52.4	48.6	35.7
+ CriticalKV	27.5	23.6	25.0	48.1	26.5	20.1	26.0	20.5	20.7	46.5	90.5	43.6	8.00	61.5	53.0	48.4	36.8
+ ProjCache-SelfDiag	25.6	22.0	23.2	45.1	25.4	19.0	25.5	20.9	20.4	41.5	89.3	43.0	7.00	64.5	52.1	48.1	35.8
+ ProjCache	25.0	22.0	24.0	45.1	25.1	19.5	25.6	21.3	20.6	42.5	89.6	43.5	8.00	63.0	51.5	48.8	35.9
AdaKV	25.2	21.6	26.7	47.0	26.0	19.4	25.5	20.6	20.5	49.0	89.7	43.0	6.00	55.5	54.0	50.4	36.3
+ DropKV	26.9	23.9	29.0	50.7	33.7	23.1	26.8	21.9	21.1	51.0	89.3	44.0	8.55	81.5	53.2	48.6	39.6
+ OBCache-VK	27.4	22.4	32.2	51.5	31.3	20.9	25.9	21.6	21.1	51.5	89.1	44.0	8.00	71.5	54.2	49.9	38.9
+ CriticalKV	27.7	24.0	27.9	50.6	27.6	22.5	26.4	21.6	21.3	47.5	90.1	43.8	7.71	67.0	54.0	49.6	38.1
+ ProjCache-SelfDiag	28.3	25.1	29.7	50.6	32.6	23.1	26.4	21.6	21.4	52.5	89.7	43.8	8.50	85.0	53.1	49.8	40.1
+ ProjCache	27.5	23.8	27.5	51.2	32.1	23.0	26.9	22.1	21.2	50.5	89.4	44.5	10.0	82.5	52.7	49.8	39.7
20% KV budget																	
H2O	28.7	23.4	35.6	49.6	36.1	21.2	28.1	21.2	22.5	48.5	90.5	44.2	10.0	84.0	52.4	48.8	40.3
+ DropKV	29.1	26.3	38.8	52.4	40.0	27.6	28.4	21.6	23.1	51.0	90.6	44.4	10.5	90.0	51.9	48.0	42.1
+ OBCache-VK	28.1	27.2	37.2	51.7	35.5	26.0	28.8	21.5	22.9	53.0	90.2	44.1	11.5	89.0	51.7	48.4	41.7
+ CriticalKV	29.3	24.0	35.6	51.3	38.3	24.5	28.7	21.4	23.3	51.0	90.4	44.4	10.5	82.5	52.0	48.2	41.0
+ ProjCache-SelfDiag	28.9	27.1	38.6	51.5	37.1	26.7	28.6	21.4	23.0	50.5	89.7	45.1	10.5	91.0	51.3	47.9	41.8
+ ProjCache	29.8	28.5	39.5	52.4	39.7	29.5	28.5	21.7	23.5	50.0	91.3	45.1	11.5	94.0	51.3	48.0	42.8
SnapKV	27.9	25.3	30.2	53.4	36.3	26.0	27.7	21.5	22.6	50.0	89.3	44.1	7.10	86.0	53.7	49.0	40.6
+ DropKV	27.7	26.7	35.8	53.4	37.2	26.0	28.2	22.3	23.1	48.0	89.6	44.5	8.10	89.5	53.5	49.2	41.4
+ OBCache-VK	27.8	26.6	36.3	57.0	36.9	27.0	28.2	21.9	22.6	53.5	89.6	44.0	10.1	89.0	52.9	48.5	42.0
+ CriticalKV	29.8	31.6	37.9	53.6	39.7	25.5	29.5	22.6	23.4	55.5	89.1	44.4	9.67	96.5	53.7	49.0	43.2
+ ProjCache-SelfDiag	28.4	26.7	34.4	51.1	36.8	26.8	28.4	22.4	22.5	48.5	89.4	44.2	8.58	91.0	53.9	49.5	41.4
+ ProjCache	27.7	27.2	32.7	51.9	36.8	27.6	28.3	22.2	22.7	50.0	89.5	44.0	9.60	93.0	53.5	48.6	41.6
AdaKV	27.7	29.2	33.4	55.6	39.8	26.1	27.7	22.2	23.1	56.0	89.8	44.3	9.60	86.5	53.1	49.5	42.1
+ DropKV	28.6	34.9	39.5	54.2	43.8	30.5	29.1	23.0	23.4	62.5	89.3	44.5	10.0	97.0	53.0	50.3	44.6
+ OBCache-VK	27.6	30.6	38.3	54.6	42.1	25.9	28.5	22.2	23.4	61.0	90.1	44.2	9.10	96.5	53.9	50.4	43.7
+ CriticalKV	30.1	35.7	38.1	55.4	41.1	28.4	29.2	23.1	23.6	58.0	89.5	44.1	9.10	94.5	53.3	50.7	44.0
+ ProjCache-SelfDiag	29.7	34.0	38.9	53.6	44.1	30.6	29.2	22.7	23.4	62.5	90.1	44.7	8.55	96.0	53.0	49.9	44.4
+ ProjCache	28.6	34.4	40.0	54.0	43.9	30.9	29.0	22.9	23.9	63.0	90.4	44.7	9.55	95.5	52.8	49.7	44.6
30% KV budget																	
H2O	29.0	30.6	42.5	54.4	43.5	24.5	29.5	22.3	24.5	51.0	89.8	44.9	12.6	94.5	52.5	49.2	43.5
+ DropKV	29.4	35.4	43.0	55.8	45.9	28.7	30.2	22.8	24.5	55.0	90.2	45.0	11.7	98.0	52.5	50.0	44.9
+ OBCache-VK	29.9	34.1	41.5	55.4	43.2	28.5	30.3	23.0	24.6	54.0	89.8	44.6	12.1	97.5	51.7	49.7	44.4
+ CriticalKV	29.7	32.7	42.0	55.9	43.6	26.4	30.3	22.6	24.7	53.0	90.0	44.2	11.6	96.5	52.3	49.0	44.0
+ ProjCache-SelfDiag	28.4	34.4	42.7	54.4	45.7	28.3	30.2	22.8	24.4	54.5	90.2	44.6	11.1	98.5	51.5	50.1	44.5
+ ProjCache	29.5	35.1	42.4	55.4	47.0	29.0	30.4	22.5	24.6	55.0	90.0	44.8	11.1	99.0	51.5	49.5	44.8
SnapKV	27.0	31.4	39.5	55.5	46.5	27.2	29.8	23.0	23.8	56.5	90.5	44.0	9.05	92.5	53.3	49.2	43.7
+ DropKV	29.9	35.5	41.4	55.7	46.4	30.4	30.1	22.8	24.0	56.0	90.4	44.5	8.55	98.0	53.8	48.8	44.8
+ OBCache-VK	30.2	35.3	41.3	55.8	43.1	28.6	30.1	23.0	24.2	58.0	90.0	44.6	7.05	96.5	53.5	49.6	44.4
+ CriticalKV	29.9	37.6	45.5	55.9	49.2	28.1	30.8	23.4	24.7	58.0	90.7	44.4	9.55	99.0	52.0	50.0	45.6
+ ProjCache-SelfDiag	29.6	34.2	42.0	55.8	45.5	31.3	30.3	22.4	24.1	55.0	90.2	44.2	9.55	97.5	53.1	49.1	44.6
+ ProjCache	29.5	35.2	41.9	54.8	45.0	30.3	30.1	23.0	24.3	56.0	89.8	44.2	8.50	98.0	53.0	49.2	44.6
AdaKV	28.8	35.8	42.0	55.1	46.1	29.5	29.7	23.4	24.4	60.5	89.7	43.6	10.1	96.0	52.1	49.7	44.8
+ DropKV	28.6	39.7	44.1	57.0	48.9	29.0	31.3	23.2	25.1	65.5	90.3	44.5	9.55	100	51.7	50.0	46.1
+ OBCache-VK	28.8	36.8	46.3	55.5	47.2	28.3	30.1	23.3	24.7	65.5	89.9	44.3	10.1	99.5	52.6	50.3	45.8
+ CriticalKV	31.3	41.2	48.7	55.8	51.0	31.1	30.9	23.6	25.3	64.5	89.7	44.4	10.1	99.0	52.2	51.1	46.9
+ ProjCache-SelfDiag	28.9	39.6	46.2	56.4	47.4	31.1	30.7	23.6	24.8	65.5	89.6	44.0	10.1	100	51.7	49.8	46.2
+ ProjCache	28.6	40.2	45.5	56.3	48.0	29.0	31.2	23.5	25.0	66.5	90.6	44.9	11.1	99.5	52.2	50.1	46.4

Table 7: RULER-4K accuracy (%) on LLaMA-3.1-8B (Q-Aware), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	99.6	85.1	100	100	100	100	100	100	100	100	88.0	65.6	100	95.3
10% KV budget														
H2O	65.2	42.7	29.6	8.80	0.00	34.6	7.00	44.8	44.8	1.60	79.2	59.2	17.9	33.5
+ DropKV	65.0	42.7	64.8	20.8	0.00	50.8	8.80	77.6	72.0	1.60	83.2	60.0	29.6	44.4
+ OBCache-VK	61.3	43.2	56.0	12.0	0.00	51.8	9.80	69.6	67.2	1.60	81.6	59.2	15.8	40.7
+ CriticalKV	69.0	42.9	55.2	14.4	0.00	47.8	8.60	50.4	68.0	1.60	80.8	60.0	18.6	39.8
+ ProjCache	67.0	42.1	68.8	24.8	0.00	48.6	9.60	80.0	74.4	1.60	84.0	62.4	36.5	46.1
SnapKV	41.0	51.5	96.0	44.8	0.00	92.2	24.2	98.4	91.2	1.60	84.0	60.8	47.2	56.4
+ DropKV	28.2	50.9	98.4	59.2	0.00	81.0	33.8	97.6	89.6	1.60	87.2	62.4	65.9	58.1
+ OBCache-VK	29.0	52.0	96.8	53.6	0.00	72.6	29.2	100	92.0	1.60	82.4	60.8	42.2	54.8
+ CriticalKV	48.0	51.2	98.4	59.2	0.800	94.6	42.4	99.2	93.6	1.60	87.2	59.2	61.0	61.3
+ ProjCache	27.0	49.3	97.6	60.8	0.00	77.8	33.2	98.4	92.0	1.60	85.6	60.0	66.6	57.7
AdaKV	61.0	54.1	90.4	62.4	5.60	81.0	15.0	100	89.6	1.60	84.8	60.0	60.6	58.9
+ DropKV	73.3	56.3	99.2	84.0	15.2	86.4	42.2	99.2	95.2	1.60	85.6	62.4	86.7	68.3
+ OBCache-VK	69.4	57.3	93.6	75.2	10.4	66.2	33.4	100	88.0	1.60	86.4	61.6	80.5	63.4
+ CriticalKV	64.8	55.2	96.0	69.6	4.00	83.2	28.0	99.2	92.8	1.60	83.2	59.2	74.7	62.4
+ ProjCache	71.7	58.7	97.6	78.4	22.4	86.8	36.4	100	96.8	1.60	86.4	64.0	90.4	68.5
20% KV budget														
H2O	86.3	48.8	85.6	28.0	16.8	74.6	17.2	75.2	79.2	1.60	88.0	63.2	40.2	54.2
+ DropKV	87.8	50.7	96.8	60.8	14.4	83.6	36.2	92.8	94.4	1.60	91.2	63.2	75.8	65.3
+ OBCache-VK	87.8	50.1	98.4	53.6	8.80	89.8	35.8	98.4	93.6	1.60	88.8	64.8	40.2	62.4
+ CriticalKV	89.4	48.8	95.2	45.6	16.0	87.0	26.8	84.0	89.6	1.60	88.0	64.8	46.6	60.3
+ ProjCache	87.6	49.9	97.6	60.0	13.6	85.0	39.8	95.2	92.0	1.60	88.0	63.2	79.7	65.6
SnapKV	73.4	59.2	100	72.0	24.0	96.8	72.4	100	96.8	1.60	89.6	63.2	80.5	71.5
+ DropKV	69.6	60.3	100	84.0	18.4	96.0	81.8	100	95.2	1.60	87.2	64.8	95.8	73.4
+ OBCache-VK	64.5	62.9	100	77.6	16.0	95.0	82.6	100	100	1.60	87.2	63.2	85.8	72.0
+ CriticalKV	75.2	60.8	100	90.4	21.6	98.0	89.4	100	98.4	1.60	88.0	63.2	92.0	75.3
+ ProjCache	70.7	60.3	100	85.6	19.2	96.2	76.2	100	96.8	1.60	88.0	63.2	96.3	73.4
AdaKV	89.1	66.1	97.6	84.8	36.8	94.8	71.2	100	96.8	1.60	87.2	64.0	90.7	75.4
+ DropKV	93.3	66.9	99.2	96.0	73.6	96.4	91.4	100	99.2	1.60	90.4	63.2	98.7	82.3
+ OBCache-VK	92.4	65.6	100	96.8	62.4	95.8	89.2	100	99.2	1.60	89.6	65.6	96.0	81.1
+ CriticalKV	91.2	67.5	100	94.4	37.6	97.2	81.8	100	99.2	1.60	88.0	64.0	96.2	78.4
+ ProjCache	93.1	67.5	99.2	96.0	76.8	97.0	88.0	100	98.4	1.60	89.6	64.8	98.7	82.4
30% KV budget														
H2O	95.0	56.8	99.2	57.6	34.4	91.0	41.0	83.2	94.4	1.60	88.8	64.0	65.8	67.1
+ DropKV	96.9	60.3	99.2	80.8	36.0	94.8	67.2	96.0	96.8	1.60	91.2	63.2	89.4	74.9
+ OBCache-VK	96.9	64.3	100	82.4	31.2	94.4	68.2	100	96.0	1.60	90.4	64.8	81.1	74.7
+ CriticalKV	96.5	56.5	100	71.2	39.2	95.0	57.6	98.4	96.8	1.60	91.2	62.4	69.9	72.0
+ ProjCache	97.2	58.7	99.2	81.6	35.2	93.8	70.4	96.0	99.2	1.60	91.2	63.2	92.6	75.4
SnapKV	87.1	64.8	98.4	87.2	47.2	97.8	90.2	100	99.2	1.60	92.0	65.6	92.8	78.8
+ DropKV	86.3	65.9	99.2	94.4	41.6	94.4	91.6	100	97.6	1.60	90.4	68.0	98.4	79.2
+ OBCache-VK	80.9	65.3	100	91.2	40.0	96.6	93.4	100	99.2	1.60	91.2	64.8	92.8	78.2
+ CriticalKV	90.1	66.4	100	96.8	43.2	98.6	99.2	100	100	8.80	90.4	64.8	97.0	81.2
+ ProjCache	86.3	65.9	99.2	92.8	42.4	94.4	90.6	100	97.6	1.60	91.2	64.8	99.5	78.9
AdaKV	95.9	68.8	99.2	96.0	68.8	99.4	89.4	100	99.2	1.60	89.6	66.4	97.1	82.4
+ DropKV	97.9	73.1	100	99.2	94.4	99.6	97.8	100	99.2	2.40	90.4	66.4	100	86.2
+ OBCache-VK	97.7	72.5	100	99.2	93.6	98.4	96.4	100	100	3.20	89.6	67.2	99.7	86.0
+ CriticalKV	97.2	75.2	100	100	62.4	98.2	99.0	100	100	6.40	89.6	67.2	99.4	84.2
+ ProjCache	97.6	72.5	100	99.2	97.6	99.0	97.2	100	99.2	2.40	90.4	65.6	100	86.2

Table 8: RULER-4K accuracy (%) on LLaMA-3.1-8B (Q-Agnostic), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	99.6	85.1	100	100	100	100	100	100	100	100	88.0	65.6	100	95.3
10% KV budget														
H2O	72.6	60.5	12.0	3.20	0.800	9.20	6.80	13.6	0.800	1.60	16.8	29.6	13.0	18.5
+ DropKV	72.9	60.5	12.8	4.00	0.800	9.80	8.00	37.6	4.80	1.60	19.2	34.4	25.1	22.4
+ OBCache-VK	72.9	60.8	13.6	3.20	0.00	9.80	7.80	36.8	2.40	1.60	16.8	33.6	16.8	21.2
+ CriticalKV	73.8	61.1	12.8	4.00	0.800	9.20	7.20	21.6	2.40	1.60	19.2	32.0	16.5	20.2
+ ProjCache	72.6	61.6	12.8	4.80	0.800	9.80	8.00	43.2	2.40	1.60	16.8	35.2	27.2	22.8
SnapKV	10.2	69.6	19.2	6.40	1.60	15.0	10.0	44.8	6.40	1.60	14.4	31.2	21.4	19.4
+ DropKV	13.7	61.6	13.6	4.80	0.800	11.6	9.40	63.2	16.8	1.60	17.6	30.4	41.8	22.1
+ OBCache-VK	14.0	64.5	13.6	5.60	0.800	12.2	9.80	64.0	9.60	1.60	13.6	31.2	24.2	20.4
+ CriticalKV	31.1	71.5	26.4	9.60	1.60	18.0	12.2	54.4	26.4	1.60	15.2	28.8	28.5	25.0
+ ProjCache	12.7	61.6	14.4	3.20	0.800	11.4	9.20	66.4	12.8	1.60	14.4	27.2	39.8	21.2
AdaKV	36.8	72.5	12.8	8.00	1.60	11.0	9.00	45.6	4.00	1.60	16.0	32.8	28.0	21.5
+ DropKV	49.8	69.1	16.0	14.4	2.40	12.6	9.60	72.8	12.8	1.60	18.4	35.2	61.4	28.9
+ OBCache-VK	44.3	71.7	13.6	16.0	1.60	10.8	9.80	75.2	8.00	1.60	20.8	33.6	41.4	26.8
+ CriticalKV	57.0	73.9	22.4	8.80	1.60	11.6	10.4	62.4	8.80	1.60	16.8	36.8	38.6	27.0
+ ProjCache	51.4	68.3	17.6	15.2	2.40	13.8	9.40	77.6	20.0	1.60	18.4	36.0	69.3	30.8
20% KV budget														
H2O	89.8	69.6	22.4	10.4	1.60	11.2	8.80	32.8	4.80	1.60	32.8	44.0	28.8	27.6
+ DropKV	89.0	71.7	35.2	21.6	0.800	14.2	9.20	65.6	16.8	1.60	36.0	44.8	47.5	34.9
+ OBCache-VK	88.3	74.1	32.0	15.2	0.800	12.4	8.80	61.6	20.8	1.60	32.8	44.8	28.6	32.5
+ CriticalKV	88.5	71.2	29.6	15.2	1.60	12.8	9.20	41.6	14.4	1.60	36.0	44.8	29.1	30.4
+ ProjCache	89.6	72.5	34.4	24.0	2.40	15.0	9.80	68.8	20.0	1.60	36.8	44.8	56.8	36.7
SnapKV	68.9	72.5	42.4	16.8	3.20	23.2	14.6	51.2	32.8	1.60	26.4	43.2	48.2	34.2
+ DropKV	71.4	67.7	39.2	16.0	1.60	20.0	11.6	80.8	40.0	1.60	24.8	41.6	76.2	37.9
+ OBCache-VK	66.1	69.3	40.8	17.6	1.60	23.2	13.6	69.6	48.0	1.60	27.2	40.8	51.0	36.2
+ CriticalKV	75.4	73.1	66.4	24.8	4.00	36.8	21.4	62.4	60.8	1.60	30.4	43.2	59.4	43.1
+ ProjCache	70.6	66.1	40.8	17.6	2.40	22.6	11.2	84.8	40.8	1.60	27.2	40.0	77.8	38.7
AdaKV	91.2	76.0	39.2	23.2	6.40	18.4	12.8	69.6	20.8	1.60	38.4	44.8	66.1	39.1
+ DropKV	96.2	73.3	57.6	52.0	26.4	26.0	12.8	95.2	47.2	1.60	40.8	43.2	92.0	51.1
+ OBCache-VK	95.1	76.5	51.2	48.0	19.2	28.8	14.8	100	42.4	1.60	38.4	48.8	81.6	49.7
+ CriticalKV	94.2	78.4	56.8	38.4	8.00	25.6	15.2	84.0	48.8	1.60	43.2	48.8	81.6	48.0
+ ProjCache	95.6	72.8	59.2	56.0	20.8	28.6	12.0	95.2	46.4	1.60	40.8	46.4	94.4	51.5
30% KV budget														
H2O	97.2	75.2	42.4	24.0	2.40	15.0	10.0	51.2	19.2	1.60	50.4	48.8	44.2	37.0
+ DropKV	96.4	77.6	49.6	37.6	1.60	18.6	11.6	80.0	34.4	1.60	46.4	56.0	65.3	44.4
+ OBCache-VK	98.2	80.0	45.6	34.4	3.20	20.8	12.0	84.0	39.2	1.60	45.6	52.8	56.2	44.1
+ CriticalKV	96.6	75.2	48.8	36.8	3.20	22.0	11.4	52.0	31.2	1.60	49.6	52.0	46.6	40.5
+ ProjCache	97.1	78.1	50.4	45.6	3.20	20.8	12.0	80.0	39.2	1.60	46.4	56.8	75.2	46.7
SnapKV	88.2	77.1	60.0	28.8	4.80	31.0	16.0	67.2	49.6	1.60	40.8	48.8	64.3	44.5
+ DropKV	89.7	70.9	62.4	32.8	6.40	32.2	13.0	88.8	54.4	1.60	40.8	50.4	87.5	48.5
+ OBCache-VK	86.2	73.9	65.6	32.0	5.60	33.8	17.8	88.0	60.0	1.60	41.6	50.4	65.0	47.8
+ CriticalKV	93.2	77.1	86.4	40.8	5.60	63.0	67.0	72.8	90.4	3.20	60.0	52.8	73.9	60.5
+ ProjCache	89.8	69.9	60.8	36.0	4.00	31.2	13.6	93.6	52.0	1.60	39.2	49.6	87.0	48.3
AdaKV	97.8	79.5	61.6	47.2	18.4	28.4	15.8	78.4	40.0	1.60	51.2	53.6	83.7	50.5
+ DropKV	98.9	78.7	83.2	77.6	55.2	40.8	15.8	100	68.8	1.60	56.0	49.6	97.0	63.3
+ OBCache-VK	99.0	79.7	80.8	78.4	49.6	45.4	21.4	100	72.0	1.60	59.2	52.0	94.7	64.1
+ CriticalKV	98.1	82.7	86.4	64.8	26.4	55.4	42.4	89.6	86.4	1.60	63.2	53.6	92.3	64.8
+ ProjCache	98.9	76.5	84.0	79.2	53.6	37.6	13.4	99.2	66.4	1.60	55.2	56.0	97.8	63.0

Table 9: RULER-32K accuracy (%) on LLaMA-3.1-8B (Q-Aware), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	31.4	88.3	100	98.4	99.2	99.8	99.4	100	100	100	84.0	52.0	98.9	88.6
10% KV budget														
H2O	22.7	69.6	92.8	20.0	15.2	91.8	51.6	98.4	96.8	4.00	81.6	50.4	84.6	60.0
+ DropKV	29.8	72.3	99.2	36.8	14.4	92.4	69.4	99.2	96.8	5.60	83.2	50.4	94.7	64.9
+ OBCache-VK	30.5	71.7	96.8	44.0	12.0	95.4	73.8	99.2	96.8	6.40	81.6	50.4	96.8	65.8
+ CriticalKV	24.0	69.3	98.4	27.2	20.8	93.8	65.4	98.4	96.8	7.20	82.4	48.8	87.7	63.1
+ ProjCache	31.1	70.4	99.2	38.4	18.4	93.8	70.0	99.2	96.8	7.20	82.4	48.8	94.7	65.4
SnapKV	5.84	61.3	99.2	72.8	36.8	97.0	92.6	100	97.6	6.40	83.2	51.2	95.4	69.2
+ DropKV	8.00	57.9	99.2	88.8	28.8	94.6	92.2	100	98.4	6.40	83.2	52.0	98.2	69.8
+ OBCache-VK	8.40	58.4	99.2	83.2	27.2	94.6	92.0	100	97.6	13.6	84.8	52.8	97.9	70.0
+ CriticalKV	14.9	54.9	99.2	87.2	12.8	98.8	92.6	100	99.2	11.2	83.2	51.2	97.3	69.4
+ ProjCache	8.64	56.0	99.2	88.0	34.4	93.2	89.0	100	98.4	7.20	84.0	52.8	98.2	69.9
AdaKV	5.76	68.5	99.2	81.6	52.0	97.4	95.2	100	97.6	10.4	82.4	51.2	97.3	72.2
+ DropKV	18.1	72.8	98.4	91.2	80.0	98.6	96.2	100	97.6	19.2	83.2	52.8	99.5	77.5
+ OBCache-VK	13.6	69.6	98.4	96.0	71.2	98.0	97.6	100	97.6	39.2	84.0	52.8	98.4	78.2
+ CriticalKV	20.0	66.4	100	94.4	26.4	98.4	93.6	100	99.2	9.60	83.2	51.2	99.4	72.4
+ ProjCache	20.2	73.3	98.4	92.8	80.0	98.6	97.8	100	96.8	18.4	83.2	51.2	99.0	77.7
20% KV budget														
H2O	31.4	79.7	99.2	43.2	55.2	99.0	91.6	100	96.8	20.0	81.6	51.2	94.2	72.5
+ DropKV	44.2	80.3	100	62.4	54.4	98.4	94.2	99.2	97.6	27.2	80.8	48.8	98.4	75.8
+ OBCache-VK	39.4	79.5	100	71.2	51.2	98.6	94.4	98.4	98.4	32.8	82.4	52.0	99.2	76.7
+ CriticalKV	34.1	81.1	99.2	48.0	56.0	98.4	93.2	100	97.6	22.4	82.4	51.2	95.2	73.8
+ ProjCache	44.9	81.6	100	67.2	54.4	97.4	94.4	100	98.4	23.2	83.2	48.8	97.8	76.2
SnapKV	14.6	70.4	98.4	94.4	72.8	98.8	98.0	100	97.6	48.0	82.4	52.0	97.9	78.9
+ DropKV	17.8	70.1	98.4	95.2	68.8	98.4	97.8	100	96.8	40.8	84.0	53.6	99.2	78.5
+ OBCache-VK	21.4	68.5	98.4	98.4	63.2	98.6	97.6	100	96.8	56.0	83.2	53.6	99.0	79.6
+ CriticalKV	27.4	66.9	99.2	96.8	49.6	100	98.8	100	100	51.2	84.0	50.4	99.2	78.7
+ ProjCache	18.6	69.6	97.6	92.0	72.0	98.8	97.2	100	96.8	40.8	84.0	51.2	99.0	78.3
AdaKV	19.2	75.2	100	93.6	88.0	99.4	99.4	100	99.2	56.0	84.0	51.2	99.5	81.9
+ DropKV	29.8	79.7	100	98.4	98.4	99.4	99.0	100	97.6	75.2	84.0	51.2	99.8	85.6
+ OBCache-VK	30.9	77.3	100	97.6	96.8	99.4	99.4	100	99.2	88.0	84.0	52.0	99.0	86.4
+ CriticalKV	30.6	78.9	100	97.6	70.4	100	100	100	100	63.2	82.4	51.2	99.0	82.6
+ ProjCache	32.3	80.5	100	97.6	98.4	99.8	99.2	100	97.6	79.2	84.8	51.2	99.8	86.2
30% KV budget														
H2O	31.7	83.2	99.2	58.4	76.8	100	97.0	99.2	98.4	47.2	82.4	51.2	96.2	78.5
+ DropKV	47.4	83.7	100	80.0	81.6	99.4	97.4	98.4	98.4	51.2	80.0	50.4	99.0	82.1
+ OBCache-VK	39.7	82.4	99.2	81.6	77.6	99.4	98.0	100	98.4	59.2	83.2	51.2	99.2	82.2
+ CriticalKV	36.3	81.6	100	69.6	79.2	99.8	97.6	100	98.4	56.8	81.6	50.4	97.3	80.7
+ ProjCache	48.2	83.7	100	81.6	83.2	99.2	97.6	99.2	98.4	54.4	81.6	50.4	99.4	82.8
SnapKV	22.4	75.5	99.2	95.2	88.8	99.8	98.2	100	97.6	71.2	84.0	51.2	98.2	83.2
+ DropKV	29.5	75.2	99.2	97.6	88.0	99.8	98.6	100	97.6	74.4	84.0	51.2	99.5	84.2
+ OBCache-VK	29.0	73.9	99.2	97.6	84.8	99.0	98.8	100	97.6	76.0	84.0	51.2	99.2	83.9
+ CriticalKV	30.8	76.0	99.2	97.6	75.2	99.8	99.6	100	100	83.2	84.8	52.0	99.2	84.4
+ ProjCache	27.8	75.5	99.2	98.4	88.8	99.2	97.8	100	97.6	69.6	84.8	52.0	99.2	83.8
AdaKV	25.4	78.9	100	99.2	96.0	99.6	99.8	100	100	86.4	84.0	50.4	99.0	86.1
+ DropKV	34.6	84.5	100	98.4	99.2	99.4	99.6	100	100	96.0	84.0	51.2	99.0	88.1
+ OBCache-VK	32.8	82.7	100	97.6	98.4	99.8	99.4	100	100	96.0	83.2	51.2	99.0	87.7
+ CriticalKV	39.0	81.9	100	98.4	88.0	99.6	99.8	100	100	98.4	83.2	51.2	99.0	87.6
+ ProjCache	39.0	84.5	100	98.4	99.2	100	99.4	100	100	95.2	84.8	50.4	99.5	88.5

Table 10: RULER-32K accuracy (%) on LLaMA-3.1-8B (Q-Agnostic), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	31.4	88.3	100	98.4	99.2	99.8	99.4	100	100	100	84.0	52.0	98.9	88.6
10% KV budget														
H2O	0.00	72.5	10.4	3.20	0.00	11.2	7.00	76.0	6.40	2.40	21.6	32.0	31.5	21.1
+ DropKV	0.320	66.9	12.0	4.80	0.00	11.6	8.00	89.6	7.20	2.40	25.6	31.2	62.1	24.7
+ OBCache-VK	0.0800	68.5	13.6	4.00	0.00	12.6	9.40	92.8	11.2	2.40	24.8	32.8	65.0	25.9
+ CriticalKV	0.0800	73.9	12.8	4.80	0.00	12.6	8.60	80.8	10.4	2.40	21.6	34.4	36.0	23.0
+ ProjCache	0.480	67.7	14.4	4.80	0.00	12.2	8.80	94.4	8.00	2.40	24.8	30.4	62.6	25.5
SnapKV	0.00	65.1	12.0	4.80	0.00	12.8	8.20	89.6	8.80	2.40	25.6	31.2	57.4	24.5
+ DropKV	0.00	53.3	12.0	4.80	0.00	11.4	7.60	96.0	7.20	2.40	24.0	34.4	83.5	25.9
+ OBCache-VK	0.00	55.5	12.8	4.80	0.00	12.8	8.20	99.2	8.80	2.40	25.6	28.8	81.3	26.2
+ CriticalKV	0.00	55.7	17.6	4.80	0.00	18.4	17.0	99.2	32.0	2.40	24.0	36.0	77.6	29.6
+ ProjCache	0.00	52.5	11.2	4.80	0.00	11.4	7.40	98.4	8.00	2.40	25.6	33.6	84.2	26.1
AdaKV	0.00	72.0	12.0	8.00	0.800	13.2	8.60	92.0	13.6	2.40	24.8	30.4	72.2	26.9
+ DropKV	0.560	70.9	15.2	16.8	6.40	15.4	9.80	99.2	17.6	2.40	28.8	37.6	92.0	31.7
+ OBCache-VK	0.160	67.2	17.6	12.8	4.80	16.0	13.0	100	29.6	3.20	24.8	36.0	91.7	32.1
+ CriticalKV	0.320	68.5	19.2	7.20	1.60	23.2	24.6	100	41.6	2.40	28.0	35.2	92.0	34.1
+ ProjCache	0.400	71.2	16.0	16.8	4.80	14.6	10.4	100	16.8	2.40	32.0	36.8	93.0	31.9
20% KV budget														
H2O	0.560	80.3	20.0	4.80	0.00	15.2	11.6	86.4	18.4	4.00	26.4	37.6	51.7	27.5
+ DropKV	1.52	78.4	20.8	11.2	0.00	21.2	15.4	96.8	15.2	4.00	32.8	39.2	80.5	32.1
+ OBCache-VK	3.76	79.5	24.0	8.80	0.00	22.6	17.2	96.8	17.6	4.80	33.6	38.4	88.3	33.5
+ CriticalKV	0.0800	80.5	25.6	6.40	0.00	21.0	13.4	90.4	24.8	4.00	29.6	39.2	56.6	30.1
+ ProjCache	1.76	78.1	21.6	10.4	0.00	20.8	16.0	96.8	16.8	2.40	33.6	36.8	84.5	32.3
SnapKV	0.320	70.4	23.2	9.60	1.60	16.6	12.8	97.6	28.8	4.80	27.2	36.0	75.0	31.1
+ DropKV	0.320	64.3	18.4	11.2	1.60	15.2	11.4	99.2	22.4	2.40	32.8	40.8	92.3	31.7
+ OBCache-VK	0.560	64.5	27.2	10.4	0.800	18.2	14.8	100	37.6	3.20	32.0	39.2	92.2	33.9
+ CriticalKV	0.960	67.7	58.4	8.80	1.60	68.8	71.0	100	88.0	3.20	40.8	37.6	86.9	48.8
+ ProjCache	0.640	64.3	18.4	9.60	1.60	16.0	11.6	99.2	21.6	2.40	33.6	39.2	92.2	31.6
AdaKV	1.12	75.5	25.6	21.6	12.0	20.0	18.2	99.2	52.8	8.00	32.0	41.6	89.4	38.2
+ DropKV	2.32	79.7	35.2	47.2	32.0	28.8	29.6	100	72.8	8.00	34.4	43.2	98.9	47.1
+ OBCache-VK	1.68	75.5	38.4	44.8	30.4	35.2	42.6	100	83.2	14.4	35.2	41.6	99.0	49.4
+ CriticalKV	4.24	78.9	66.4	32.0	8.00	86.0	82.8	100	85.6	8.80	47.2	41.6	98.1	56.9
+ ProjCache	2.16	79.7	34.4	45.6	32.0	30.4	28.4	100	72.0	7.20	39.2	43.2	98.6	47.1
30% KV budget														
H2O	1.28	83.5	31.2	12.8	0.800	23.6	22.8	90.4	35.2	8.00	32.8	40.0	65.4	34.4
+ DropKV	6.00	84.5	38.4	18.4	0.800	31.0	32.8	100	39.2	8.00	39.2	41.6	89.6	40.7
+ OBCache-VK	6.64	83.5	37.6	24.0	1.60	33.6	33.8	100	40.8	8.80	40.8	43.2	91.7	42.0
+ CriticalKV	1.12	82.9	40.0	16.0	0.800	35.6	34.2	94.4	50.4	9.60	36.8	40.0	71.2	39.5
+ ProjCache	4.64	83.5	40.0	21.6	1.60	32.6	35.8	100	40.8	8.80	43.2	44.0	91.4	42.1
SnapKV	0.880	73.3	32.8	15.2	4.80	23.8	20.2	100	62.4	11.2	32.8	43.2	86.4	39.0
+ DropKV	1.12	70.1	32.0	21.6	6.40	21.0	20.2	100	56.8	6.40	34.4	40.8	95.5	39.0
+ OBCache-VK	1.04	72.0	32.0	21.6	6.40	27.4	28.2	100	63.2	9.60	33.6	45.6	95.5	41.2
+ CriticalKV	3.92	75.7	83.2	23.2	4.00	93.4	93.0	100	96.0	16.8	50.4	44.0	91.5	59.6
+ ProjCache	1.20	72.5	33.6	22.4	5.60	22.6	19.6	100	56.8	5.60	36.0	42.4	95.5	39.5
AdaKV	2.24	80.0	52.8	45.6	31.2	38.0	43.6	100	81.6	24.0	39.2	47.2	97.9	52.6
+ DropKV	6.00	83.5	64.0	71.2	59.2	48.6	66.8	100	88.8	28.8	51.2	48.8	99.7	62.8
+ OBCache-VK	3.12	81.3	72.8	74.4	60.8	53.6	71.2	100	91.2	35.2	48.0	46.4	98.9	64.4
+ CriticalKV	14.3	82.7	91.2	78.4	24.0	94.4	93.2	100	94.4	45.6	61.6	47.2	99.8	71.3
+ ProjCache	6.16	84.0	63.2	72.8	55.2	44.6	66.8	100	89.6	28.0	52.8	45.6	99.4	62.2

Table 11: LongBench scores on Qwen3-8B (Q-Aware), without compensation.

	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Synthetic		Code		
Method	NQA	Qasper	MF-en	HQA	2WQA	MSQ	GovR	QMSum	MNews	TREC	TQA	SAMSum	PCount	PRe	LCC	RB-P	Avg.
Full KV	28.8	44.1	55.0	62.9	49.2	35.3	33.5	24.5	24.7	71.5	89.3	44.6	9.50	91.8	55.8	54.9	48.5
10% KV budget																	
H2O	28.0	33.7	49.8	58.9	41.9	30.5	24.6	23.4	17.4	47.0	88.5	42.5	9.11	93.5	54.5	55.7	43.7
+ DropKV	28.4	35.7	50.3	60.9	45.0	32.3	25.9	23.5	18.5	50.0	88.6	43.4	9.25	93.8	54.9	55.7	44.8
+ OBCache-VK	28.1	36.1	50.2	59.1	45.7	34.5	25.8	23.4	18.6	47.5	88.6	43.6	8.55	92.4	54.5	55.7	44.5
+ CriticalKV	28.5	33.9	50.4	59.4	43.3	32.6	25.6	23.4	18.0	49.0	88.9	42.6	8.16	95.0	54.5	55.9	44.3
+ ProjCache-SelfDiag	28.3	36.5	49.9	60.1	45.3	33.3	26.1	23.8	18.3	50.5	88.9	43.0	8.55	93.6	55.0	55.5	44.8
+ ProjCache	28.4	37.0	48.5	59.4	45.3	33.8	26.6	23.4	18.8	49.0	89.0	43.2	8.85	97.4	54.6	55.0	44.9
SnapKV	28.4	35.7	49.1	61.0	41.6	32.7	25.5	22.9	17.0	52.0	89.5	41.4	7.10	93.2	54.9	57.0	44.3
+ DropKV	29.9	37.2	50.4	60.9	42.3	33.7	26.2	23.4	17.3	49.5	89.0	41.9	6.55	93.7	55.0	56.4	44.6
+ OBCache-VK	29.9	37.6	51.2	61.2	43.4	33.8	26.7	23.1	17.4	50.5	89.5	42.4	7.00	93.5	55.0	56.5	44.9
+ CriticalKV	28.7	39.2	50.6	61.2	44.4	32.9	27.0	23.2	17.6	54.0	88.7	41.7	6.39	94.7	56.0	56.7	45.2
+ ProjCache-SelfDiag	29.5	36.7	49.4	61.4	43.7	33.0	26.4	23.5	17.1	50.0	89.0	41.7	6.00	93.5	55.2	57.0	44.6
+ ProjCache	29.9	35.9	50.5	61.1	42.4	36.0	26.4	23.5	18.0	49.0	89.6	41.5	7.10	93.2	55.4	56.8	44.8
AdaKV	29.3	38.0	49.6	61.2	43.1	32.9	25.7	23.2	17.6	60.5	89.7	42.1	6.55	94.2	56.3	56.9	45.4
+ DropKV	29.1	38.6	50.7	61.3	45.3	33.4	27.0	23.6	18.1	61.0	90.1	43.0	5.60	93.8	57.0	57.5	45.9
+ OBCache-VK	28.9	39.3	50.5	61.1	46.0	33.8	27.5	23.7	18.7	68.0	89.2	43.1	8.05	93.5	56.8	57.7	46.6
+ CriticalKV	28.4	38.8	50.0	60.5	44.2	32.2	26.9	23.3	18.0	60.0	89.4	42.5	7.64	94.3	56.8	56.8	45.6
+ ProjCache-SelfDiag	28.1	39.7	51.0	61.0	45.5	33.6	26.9	23.5	18.1	59.5	90.0	43.1	6.10	94.1	56.8	57.3	45.9
+ ProjCache	28.7	39.8	52.4	60.6	44.3	33.3	26.9	23.3	18.4	62.0	89.7	42.8	7.66	94.2	57.0	56.9	46.1
20% KV budget																	
H2O	29.7	38.4	51.6	60.0	47.0	35.4	27.9	23.9	20.6	53.5	89.0	43.5	10.1	93.0	56.4	56.0	46.0
+ DropKV	28.9	40.1	53.1	59.7	47.1	34.8	29.1	24.3	20.9	57.5	89.2	43.8	9.55	94.0	56.8	55.8	46.5
+ OBCache-VK	28.0	40.0	51.8	59.7	47.6	34.5	29.2	24.0	20.6	53.0	88.7	44.0	9.50	92.7	56.6	56.0	46.0
+ CriticalKV	29.1	39.8	52.2	59.8	47.5	36.0	28.8	24.2	20.5	56.5	88.9	43.3	10.1	93.9	56.5	56.1	46.4
+ ProjCache-SelfDiag	29.1	40.2	52.8	60.6	47.2	34.5	29.4	24.0	20.9	59.5	88.8	44.0	9.05	94.5	56.4	56.1	46.7
+ ProjCache	28.6	40.3	52.2	61.1	47.0	33.6	29.8	23.9	21.1	61.0	89.0	43.3	8.61	96.8	56.7	55.9	46.8
SnapKV	28.9	39.9	52.3	61.2	47.2	34.9	28.8	23.3	20.1	60.0	89.3	42.6	6.57	93.1	56.5	56.1	46.3
+ DropKV	29.2	41.8	52.9	60.6	46.5	35.1	29.5	23.8	20.1	60.5	88.9	43.4	7.55	93.1	56.5	56.9	46.6
+ OBCache-VK	29.4	42.2	53.0	61.0	46.9	35.7	29.5	23.9	20.8	62.0	88.8	43.5	7.55	93.2	56.5	56.4	46.9
+ CriticalKV	28.7	42.8	52.2	61.7	46.2	34.6	30.4	23.7	20.5	65.0	88.7	43.2	9.07	94.9	56.7	56.2	47.2
+ ProjCache-SelfDiag	29.1	42.6	52.2	61.4	47.2	34.8	29.6	24.0	20.5	60.5	88.9	43.4	8.00	94.1	55.5	56.6	46.8
+ ProjCache	28.9	42.8	52.6	61.6	47.6	34.4	29.5	23.8	20.6	63.0	88.8	43.5	8.00	92.6	56.7	56.3	46.9
AdaKV	29.2	41.1	53.6	62.4	48.8	34.3	28.6	23.9	20.2	65.0	89.0	43.3	7.07	94.2	56.5	56.5	47.1
+ DropKV	29.5	41.9	52.8	61.8	48.7	35.2	30.1	24.2	20.9	68.5	88.8	43.8	7.55	94.6	57.1	57.2	47.7
+ OBCache-VK	29.4	41.9	53.6	61.9	48.1	35.1	30.7	23.8	21.1	69.5	89.5	43.7	7.55	94.5	56.6	56.9	47.7
+ CriticalKV	29.2	42.8	53.1	62.5	47.4	34.2	30.0	23.9	21.0	67.0	89.0	43.9	8.82	94.2	56.5	56.7	47.5
+ ProjCache-SelfDiag	29.7	42.2	53.1	62.0	48.2	35.4	30.0	24.0	20.8	67.5	88.8	44.2	7.55	94.3	56.9	56.9	47.6
+ ProjCache	29.2	42.3	53.1	61.7	47.8	36.0	30.1	24.1	20.8	67.5	88.8	43.9	8.55	94.1	57.4	57.2	47.7
30% KV budget																	
H2O	28.8	41.3	52.8	61.3	48.3	35.2	30.5	23.8	21.7	60.5	88.8	43.6	10.6	92.7	56.2	55.6	47.0
+ DropKV	28.6	41.7	53.2	61.9	48.4	36.1	31.4	23.9	22.3	64.0	88.8	44.1	9.05	93.2	56.5	55.3	47.4
+ OBCache-VK	28.8	43.0	53.0	61.9	48.5	35.2	31.1	24.0	21.8	60.5	89.5	44.0	9.55	93.5	56.3	55.4	47.3
+ CriticalKV	28.9	42.5	53.3	62.0	47.0	36.5	30.9	23.8	22.1	63.5	88.8	43.5	9.05	94.7	55.9	55.3	47.4
+ ProjCache-SelfDiag	28.4	42.9	53.7	61.6	49.1	36.4	31.5	24.3	22.1	65.0	88.8	43.7	8.50	93.5	56.0	55.4	47.6
+ ProjCache	29.1	43.2	53.0	62.1	48.0	34.4	31.4	24.1	22.7	65.0	88.8	43.5	8.59	96.2	56.3	55.8	47.6
SnapKV	28.1	42.0	51.8	62.3	48.6	35.9	30.5	23.9	21.2	64.5	88.8	43.4	9.05	92.7	56.6	56.4	47.2
+ DropKV	30.0	42.3	53.1	61.8	47.8	36.8	30.8	23.9	21.8	65.5	88.8	44.3	9.05	92.6	56.1	56.4	47.6
+ OBCache-VK	29.7	42.1	53.8	61.8	47.0	35.6	31.4	24.0	22.0	66.5	88.8	44.0	8.55	92.5	56.2	56.7	47.5
+ CriticalKV	30.1	43.0	53.8	61.7	48.5	35.2	31.6	24.2	22.2	68.0	88.8	43.8	8.57	93.9	56.5	55.8	47.9
+ ProjCache-SelfDiag	30.1	42.4	53.0	61.4	47.5	36.2	31.0	23.8	21.6	65.5	88.8	43.9	9.00	92.7	56.4	56.2	47.5
+ ProjCache	28.9	42.0	53.3	61.4	48.4	36.2	31.1	23.9	22.0	67.0	88.8	44.1	8.50	92.3	56.4	56.3	47.5
AdaKV	28.5	42.5	53.5	61.8	49.3	35.5	30.4	24.0	21.4	67.5	89.0	43.9	9.05	92.6	56.6	56.5	47.6
+ DropKV	29.1	43.1	53.2	62.0	47.4	36.0	31.6	24.1	22.2	70.0	89.0	44.6	8.55	93.7	57.0	56.7	48.0
+ OBCache-VK	29.1	43.8	54.1	62.4	48.3	36.0	32.0	24.3	22.5	71.5	89.0	44.5	10.1	94.3	56.6	57.4	48.5
+ CriticalKV	29.3	43.7	54.0	62.8	47.7	34.7	31.6	24.1	22.5	68.5	89.0	44.2	8.36	93.9	56.4	56.4	47.9
+ ProjCache-SelfDiag	29.7	43.2	53.3	62.1	47.9	35.6	31.6	24.2	22.2	70.5	89.0	44.5	9.00	94.0	56.8	57.0	48.2
+ ProjCache	29.0	43.1	53.7	62.4	48.8	35.2	31.7	24.2	22.3	70.5	89.0	44.2	8.50	94.1	57.2	56.6	48.2

Table 12: LongBench scores on Qwen3-8B (Q-Agnostic), without compensation.

	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Synthetic		Code		
Method	NQA	Qasper	MF-en	HQA	2WQA	MSQ	GovR	QMSum	MNews	TREC	TQA	SAMSum	PCount	PRe	LCC	RB-P	Avg.
Full KV	28.8	44.1	55.0	62.9	49.2	35.3	33.5	24.5	24.7	71.5	89.3	44.6	9.50	91.8	55.8	54.9	48.5
10% KV budget																	
H2O	24.1	17.4	27.1	39.2	31.8	20.0	24.8	18.6	18.3	42.5	90.5	41.9	7.55	48.8	54.6	51.5	34.9
+ DropKV	24.5	20.4	29.1	45.1	30.5	21.1	26.0	19.9	18.8	44.0	90.7	42.0	10.1	70.5	54.9	51.0	37.4
+ OBCache-VK	22.9	19.5	28.9	42.4	31.3	22.1	25.7	19.7	18.7	45.0	90.5	41.7	9.63	68.8	54.5	51.5	37.1
+ CriticalKV	23.5	19.2	27.8	42.2	30.3	23.0	25.6	19.0	18.5	44.5	90.5	42.3	7.05	60.4	54.5	50.8	36.2
+ ProjCache-SelfDiag	25.7	19.7	29.1	44.0	31.6	22.2	25.9	19.7	18.7	45.0	90.2	41.3	8.10	68.0	55.0	51.4	37.2
+ ProjCache	24.3	21.6	28.3	45.6	33.4	24.7	26.2	20.1	19.2	45.0	90.8	42.1	7.65	70.9	54.7	51.4	37.9
SnapKV	21.1	20.7	25.9	42.4	30.3	19.0	25.4	19.1	17.8	46.0	90.7	41.0	9.05	48.0	55.2	52.8	35.3
+ DropKV	22.5	20.8	26.4	41.1	29.3	23.0	26.0	19.5	18.0	44.0	89.8	41.3	9.00	57.5	55.0	53.0	36.0
+ OBCache-VK	23.3	20.1	26.6	44.1	32.1	21.3	26.0	19.5	18.5	47.5	90.3	41.6	7.61	56.5	55.0	52.7	36.4
+ CriticalKV	24.2	22.7	29.0	44.1	33.3	23.0	26.8	20.4	18.2	48.5	90.8	40.7	7.50	69.0	56.0	53.8	38.0
+ ProjCache-SelfDiag	24.9	20.6	27.4	41.4	29.7	23.8	25.8	19.2	17.9	44.5	90.0	41.2	9.50	63.3	55.2	52.8	36.7
+ ProjCache	22.2	20.2	27.2	42.2	30.8	23.5	26.0	19.6	18.1	45.5	90.2	41.6	6.25	65.0	55.4	52.9	36.7
AdaKV	21.5	23.1	26.3	41.0	31.0	21.5	25.2	19.3	18.2	51.0	90.7	41.6	8.15	50.6	56.3	53.5	36.2
+ DropKV	23.8	23.9	30.3	45.9	33.2	24.4	26.3	20.1	18.7	52.5	90.2	42.1	8.00	79.1	57.0	54.0	39.3
+ OBCache-VK	23.7	25.9	30.3	45.5	30.9	23.6	27.1	20.7	19.1	58.0	90.3	42.6	8.50	79.2	56.8	54.3	39.8
+ CriticalKV	24.7	24.6	31.0	47.2	33.7	26.2	26.7	20.5	18.4	54.0	90.3	41.2	7.50	79.2	56.8	54.0	39.7
+ ProjCache-SelfDiag	23.7	24.9	29.0	45.4	31.8	23.7	26.2	20.2	18.6	53.0	90.5	41.5	6.50	77.5	56.8	53.9	38.9
+ ProjCache	24.4	25.6	30.9	45.3	31.6	24.3	26.2	20.3	18.8	52.0	90.7	42.4	7.59	79.1	56.8	53.7	39.4
20% KV budget																	
H2O	25.4	24.8	32.9	49.9	36.6	25.1	27.5	19.9	20.6	47.5	89.9	42.3	8.55	82.6	56.5	51.9	40.1
+ DropKV	26.0	25.9	34.1	51.0	39.8	26.1	28.5	21.0	21.1	53.0	89.3	42.5	8.05	86.7	56.8	52.5	41.4
+ OBCache-VK	25.5	26.1	35.1	49.8	38.7	26.4	28.6	20.9	20.9	50.0	89.9	43.0	7.54	87.9	56.6	52.5	41.2
+ CriticalKV	25.6	27.1	33.7	51.2	37.8	25.9	28.0	20.6	21.1	53.0	89.8	42.5	7.05	86.3	56.5	51.7	41.1
+ ProjCache-SelfDiag	26.1	26.7	33.9	52.7	39.0	26.0	28.4	21.2	21.3	55.0	88.8	42.0	7.50	87.8	56.4	52.1	41.6
+ ProjCache	26.5	26.4	35.8	55.2	40.3	28.5	28.5	21.3	21.8	56.0	91.2	42.9	7.44	91.2	56.7	52.3	42.6
SnapKV	23.3	26.8	34.0	48.1	36.2	25.0	28.4	20.7	20.3	52.0	90.7	42.3	6.80	83.2	56.6	52.4	40.4
+ DropKV	23.9	29.1	33.9	49.0	38.2	26.3	29.2	20.7	20.9	52.0	89.5	42.8	7.05	87.2	56.5	52.9	41.2
+ OBCache-VK	24.6	28.2	34.8	47.7	38.3	24.7	28.9	20.9	20.8	56.0	89.7	43.0	6.55	85.6	56.5	52.0	41.1
+ CriticalKV	25.7	32.9	36.8	52.0	40.4	27.8	30.3	22.1	21.0	60.5	89.5	42.5	8.04	88.5	56.7	54.5	43.1
+ ProjCache-SelfDiag	24.4	28.8	33.0	49.5	38.4	26.6	29.2	20.9	20.9	52.0	89.5	43.1	7.05	87.0	55.5	52.6	41.2
+ ProjCache	24.0	30.2	33.1	48.5	38.1	26.3	29.1	20.7	21.2	51.5	90.7	42.5	6.05	87.6	56.7	52.7	41.2
AdaKV	25.5	27.3	35.7	48.4	37.6	27.6	28.2	20.9	20.5	57.0	89.0	42.8	6.73	83.4	56.5	53.5	41.3
+ DropKV	25.3	31.8	37.1	53.8	39.8	28.0	29.3	21.5	21.1	61.5	89.4	43.6	6.05	89.2	57.1	54.1	43.0
+ OBCache-VK	26.0	31.6	37.8	53.0	42.8	29.1	29.8	22.0	21.4	65.0	89.4	43.4	6.09	89.6	56.6	54.6	43.6
+ CriticalKV	26.9	32.3	37.5	54.8	41.3	29.0	29.8	22.1	21.3	60.5	89.7	43.0	7.25	88.3	56.5	55.0	43.5
+ ProjCache-SelfDiag	25.6	31.6	36.7	53.4	40.8	26.9	29.4	21.6	21.4	61.5	89.8	43.1	7.05	89.9	56.9	54.2	43.1
+ ProjCache	24.8	31.1	37.7	53.6	39.0	29.3	29.3	21.7	21.1	61.5	90.7	43.4	8.19	91.5	57.4	54.1	43.4
30% KV budget																	
H2O	27.0	31.4	37.5	55.1	39.9	27.5	29.5	21.1	21.9	59.0	89.7	42.8	9.50	89.2	56.2	53.5	43.2
+ DropKV	25.7	32.2	38.4	55.7	39.6	28.6	30.3	22.6	22.5	61.5	89.7	43.3	9.50	89.8	56.5	52.5	43.6
+ OBCache-VK	27.1	32.7	37.6	54.4	41.3	29.1	30.1	22.0	22.6	58.0	88.8	42.7	10.5	89.7	56.3	53.5	43.5
+ CriticalKV	26.9	30.9	38.4	55.6	37.9	28.6	30.3	21.9	22.4	60.5	89.8	43.4	8.00	90.8	55.9	53.4	43.4
+ ProjCache-SelfDiag	26.6	31.9	40.3	55.9	38.4	27.8	30.6	22.5	22.3	63.0	89.8	43.1	9.50	91.8	56.0	52.6	43.9
+ ProjCache	26.3	32.5	42.3	56.7	41.5	30.7	30.8	22.7	22.4	63.0	90.5	43.3	8.94	97.9	56.1	53.5	44.9
SnapKV	25.9	29.7	36.7	52.9	42.6	27.9	29.9	21.2	21.9	60.0	89.7	43.4	9.05	89.4	56.1	53.6	43.1
+ DropKV	25.2	33.0	37.5	53.1	41.4	27.9	30.8	22.1	22.2	58.5	90.2	43.7	9.50	91.6	56.1	54.4	43.6
+ OBCache-VK	26.6	32.4	37.5	53.3	41.8	27.8	30.6	21.8	22.3	61.0	89.7	44.0	10.1	90.8	56.2	53.5	43.7
+ CriticalKV	26.4	35.5	43.2	57.3	44.6	29.1	31.6	22.8	22.5	62.5	89.5	43.4	8.55	89.7	56.5	55.0	44.9
+ ProjCache-SelfDiag	26.1	33.1	37.6	53.1	41.6	27.4	30.5	21.9	22.0	59.0	89.7	43.8	9.00	91.0	56.4	54.1	43.5
+ ProjCache	25.6	32.5	38.2	54.2	42.3	28.8	30.6	21.9	22.3	58.0	89.7	43.4	9.12	91.5	55.9	53.8	43.6
AdaKV	26.3	32.6	38.0	55.8	43.9	28.7	30.1	21.8	22.0	62.5	89.2	43.4	10.1	91.3	56.6	54.0	44.1
+ DropKV	27.0	35.3	40.4	58.1	44.0	31.5	31.3	22.3	22.4	68.0	89.8	43.8	10.1	91.4	57.0	54.6	45.4
+ OBCache-VK	27.3	35.3	42.3	59.5	45.6	33.9	31.4	22.8	22.8	69.0	89.0	44.0	9.05	92.3	56.6	55.0	46.0
+ CriticalKV	28.3	37.4	43.4	58.4	45.1	31.6	31.5	23.3	22.7	68.0	89.7	43.6	8.34	89.7	56.4	55.0	45.8
+ ProjCache-SelfDiag	26.9	34.5	41.2	56.5	43.4	30.7	31.1	22.5	22.3	65.5	89.9	43.8	8.05	91.5	56.8	53.9	44.9
+ ProjCache	27.9	35.0	40.6	58.3	44.5	31.6	31.0	22.8	22.6	64.5	89.3	43.8	9.62	91.3	56.5	54.5	45.2

Table 13: RULER-4K accuracy (%) on Qwen3-8B (Q-Aware), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	99.0	84.8	100	100	100	100	100	100	100	100	82.4	67.2	100	94.9
10% KV budget														
H2O	59.3	42.9	18.4	6.40	3.20	47.6	9.00	1.60	23.2	1.60	68.0	55.2	25.3	27.8
+ CriticalKV	64.6	43.5	33.6	10.4	7.20	71.0	8.80	19.2	50.4	1.60	68.8	55.2	38.9	36.4
+ ProjCache	67.5	44.0	63.2	11.2	7.20	73.6	12.8	92.0	79.2	1.60	72.8	55.2	60.5	49.3
SnapKV	56.8	44.3	24.0	12.8	13.6	66.4	8.80	37.6	51.2	1.60	73.6	55.2	37.6	37.2
+ CriticalKV	71.8	44.3	47.2	19.2	9.60	90.4	14.8	80.0	89.6	1.60	73.6	56.8	66.7	51.2
+ ProjCache	44.1	47.7	56.0	14.4	9.60	63.6	13.0	96.0	83.2	1.60	75.2	55.2	73.0	48.7
AdaKV	68.9	45.9	20.8	26.4	20.8	40.6	8.80	65.6	19.2	1.60	74.4	55.2	44.2	37.9
+ CriticalKV	82.1	47.5	44.8	30.4	22.4	71.8	9.80	99.2	56.8	1.60	77.6	60.0	83.5	52.9
+ ProjCache	81.8	50.7	71.2	49.6	60.0	83.6	14.2	100	88.8	1.60	76.8	59.2	88.0	63.5
20% KV budget														
H2O	77.4	45.6	76.0	22.4	54.4	97.4	21.0	50.4	88.0	1.60	76.8	60.0	43.8	55.0
+ CriticalKV	83.6	46.9	94.4	32.0	60.0	99.6	29.4	91.2	98.4	1.60	78.4	58.4	69.0	64.8
+ ProjCache	87.9	53.1	96.8	40.8	57.6	99.6	35.6	100	99.2	1.60	77.6	59.2	89.8	69.1
SnapKV	88.0	50.7	87.2	37.6	69.6	99.2	26.2	81.6	98.4	1.60	77.6	62.4	70.6	65.4
+ CriticalKV	93.4	55.2	97.6	44.8	64.0	100	67.6	100	100	1.60	81.6	64.0	95.2	74.2
+ ProjCache	87.7	58.7	96.8	44.0	68.8	99.4	33.6	100	99.2	1.60	79.2	61.6	94.6	71.2
AdaKV	93.2	57.6	76.0	51.2	89.6	98.2	19.6	100	92.8	1.60	80.0	63.2	85.8	69.9
+ CriticalKV	96.7	68.0	100	66.4	82.4	100	46.0	100	100	2.40	80.8	62.4	99.4	77.3
+ ProjCache	95.8	66.1	100	84.0	96.8	99.8	38.4	100	100	1.60	80.0	64.0	98.9	78.9
30% KV budget														
H2O	89.1	49.6	97.6	45.6	86.4	100	35.6	88.0	99.2	2.40	81.6	60.0	62.7	69.1
+ CriticalKV	93.3	53.6	99.2	52.8	88.8	100	47.2	100	100	4.00	79.2	60.8	86.4	74.3
+ ProjCache	95.9	66.4	100	64.0	89.6	100	55.2	100	100	5.60	80.0	61.6	96.3	78.0
SnapKV	95.4	59.2	99.2	48.0	93.6	100	41.8	99.2	100	2.40	81.6	61.6	85.1	74.4
+ CriticalKV	97.3	69.1	100	60.0	92.8	100	92.2	100	100	12.8	82.4	64.8	99.2	82.3
+ ProjCache	94.1	65.3	100	65.6	90.4	100	61.0	100	100	2.40	81.6	62.4	98.4	78.6
AdaKV	97.4	68.8	100	72.8	96.8	100	34.2	100	100	1.60	81.6	62.4	97.9	78.0
+ CriticalKV	98.5	76.8	100	85.6	99.2	100	94.2	100	100	24.8	81.6	65.6	100	86.6
+ ProjCache	98.2	72.8	100	93.6	99.2	100	68.8	100	100	4.80	82.4	65.6	99.8	83.5

Table 14: RULER-4K accuracy (%) on Qwen3-8B (Q-Agnostic), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	99.0	84.8	100	100	100	100	100	100	100	100	82.4	67.2	100	94.9
10% KV budget														
H2O	63.8	62.9	8.80	1.60	0.00	8.00	8.80	1.60	0.800	1.60	24.8	24.0	5.44	16.3
+ CriticalKV	66.7	65.1	8.80	1.60	0.800	8.40	8.80	1.60	1.60	1.60	27.2	28.8	7.36	17.6
+ ProjCache	69.8	63.2	10.4	2.40	0.00	9.60	8.80	8.00	2.40	1.60	31.2	28.8	15.4	19.3
SnapKV	57.4	54.9	11.2	2.40	0.800	9.60	8.80	2.40	2.40	1.60	24.0	24.8	8.80	16.1
+ CriticalKV	69.4	57.9	11.2	3.20	0.800	9.80	8.80	14.4	6.40	1.60	28.0	33.6	18.4	20.3
+ ProjCache	48.9	58.1	11.2	2.40	0.00	9.60	8.80	40.8	1.60	1.60	26.4	29.6	38.1	21.3
AdaKV	71.3	55.5	11.2	3.20	1.60	9.60	8.80	9.60	1.60	1.60	25.6	30.4	11.5	18.6
+ CriticalKV	79.4	59.5	11.2	6.40	1.60	9.60	8.80	35.2	1.60	1.60	28.8	32.8	33.6	23.9
+ ProjCache	78.1	57.3	12.8	10.4	4.00	9.80	8.80	65.6	4.00	1.60	25.6	28.8	55.4	27.9
20% KV budget														
H2O	75.5	70.1	13.6	3.20	0.800	8.80	8.80	4.00	4.80	1.60	34.4	40.0	8.80	21.1
+ CriticalKV	78.8	69.6	25.6	7.20	0.800	11.0	9.00	7.20	8.00	1.60	37.6	42.4	11.7	23.9
+ ProjCache	86.3	71.7	23.2	6.40	0.800	12.8	10.0	54.4	9.60	1.60	35.2	42.4	35.0	30.0
SnapKV	86.6	66.1	18.4	7.20	1.60	10.8	8.80	23.2	11.2	1.60	30.4	40.8	22.2	25.3
+ CriticalKV	90.7	68.3	35.2	14.4	1.60	30.2	21.2	59.2	28.8	1.60	37.6	45.6	64.5	38.4
+ ProjCache	88.0	64.5	22.4	8.80	2.40	12.4	9.40	77.6	12.0	1.60	32.0	40.8	75.8	34.4
AdaKV	92.0	67.2	16.0	12.0	2.40	10.2	9.20	45.6	7.20	1.60	36.0	41.6	45.3	29.7
+ CriticalKV	95.1	73.6	37.6	19.2	2.40	20.8	11.4	92.8	23.2	1.60	40.8	46.4	92.0	42.8
+ ProjCache	93.8	69.6	37.6	23.2	16.0	20.6	10.4	92.8	18.4	1.60	38.4	42.4	89.8	42.7
30% KV budget														
H2O	87.0	71.5	28.0	8.80	1.60	14.2	9.20	8.80	9.60	1.60	43.2	46.4	11.8	26.3
+ CriticalKV	89.3	72.5	36.8	12.0	1.60	23.4	10.6	23.2	20.0	1.60	49.6	48.0	21.0	31.5
+ ProjCache	96.4	80.0	37.6	22.4	3.20	23.0	11.4	72.0	16.8	1.60	48.8	47.2	55.2	39.7
SnapKV	93.2	71.5	32.0	15.2	3.20	19.8	10.2	40.0	20.8	1.60	42.4	44.8	43.7	33.7
+ CriticalKV	96.1	77.3	72.0	26.4	2.40	56.4	54.4	87.2	68.0	4.00	54.4	50.4	85.9	56.5
+ ProjCache	93.8	72.5	40.8	20.0	4.00	25.6	12.0	94.4	26.4	2.40	40.8	50.4	88.3	44.0
AdaKV	96.2	74.1	29.6	21.6	6.40	17.2	10.4	79.2	19.2	1.60	45.6	48.8	71.2	40.1
+ CriticalKV	98.0	79.5	78.4	39.2	12.8	50.0	40.2	99.2	65.6	6.40	58.4	52.0	99.8	60.0
+ ProjCache	97.4	78.9	57.6	40.8	37.6	34.0	16.0	99.2	36.0	2.40	50.4	52.8	98.2	53.9

Table 15: RULER-32K accuracy (%) on Qwen3-8B (Q-Aware), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	67.8	88.3	100	92.8	80.8	100	99.0	100	100	100	72.8	60.0	99.8	89.3
10% KV budget														
H2O	16.3	78.7	92.8	17.6	26.4	99.6	26.2	100	96.0	3.20	68.8	52.0	98.4	59.7
+ CriticalKV	17.8	78.7	96.8	24.0	24.0	99.6	37.0	100	99.2	3.20	70.4	52.8	99.5	61.8
+ ProjCache	23.9	79.5	99.2	27.2	28.8	99.6	38.6	100	100	3.20	68.8	54.4	98.4	63.2
SnapKV	28.6	75.5	97.6	35.2	31.2	99.8	37.8	100	100	2.40	70.4	56.0	98.7	64.1
+ CriticalKV	25.0	77.1	100	27.2	26.4	100	96.8	100	100	7.20	72.8	55.2	99.8	68.3
+ ProjCache	26.7	73.6	99.2	48.0	30.4	99.8	45.2	100	100	2.40	71.2	57.6	99.5	65.7
AdaKV	30.0	78.9	97.6	46.4	33.6	99.6	29.2	100	100	2.40	70.4	56.8	99.4	64.9
+ CriticalKV	28.6	79.7	100	35.2	28.0	100	97.4	100	100	24.0	72.0	58.4	100	71.0
+ ProjCache	41.9	77.9	100	75.2	57.6	99.8	63.8	100	100	7.20	72.0	60.0	99.5	73.5
20% KV budget														
H2O	27.2	80.5	100	33.6	45.6	99.6	57.6	100	100	8.00	71.2	54.4	99.4	67.5
+ CriticalKV	31.0	83.7	100	40.8	47.2	99.8	79.6	100	100	16.8	72.0	55.2	99.5	71.2
+ ProjCache	38.6	82.9	100	47.2	45.6	99.8	78.2	100	100	16.0	72.0	56.8	99.5	72.1
SnapKV	48.9	81.6	100	56.8	50.4	99.8	78.4	100	100	8.00	72.0	58.4	99.4	73.4
+ CriticalKV	48.2	82.4	100	49.6	51.2	100	99.4	100	100	54.4	72.8	56.8	99.8	78.0
+ ProjCache	45.7	80.3	100	69.6	60.0	99.8	84.4	100	100	14.4	72.8	60.0	99.7	75.9
AdaKV	50.4	83.5	100	73.6	44.0	99.8	71.6	100	100	13.6	72.8	57.6	99.8	74.4
+ CriticalKV	50.6	85.9	100	74.4	57.6	100	99.6	100	100	82.4	71.2	58.4	99.8	83.1
+ ProjCache	59.6	84.0	100	87.2	66.4	100	95.4	100	100	37.6	71.2	60.0	99.7	81.6
30% KV budget														
H2O	37.9	84.8	100	53.6	58.4	99.8	83.8	100	100	24.0	72.8	57.6	99.4	74.8
+ CriticalKV	42.4	85.9	100	60.0	60.8	100	97.2	100	100	62.4	73.6	57.6	99.7	80.0
+ ProjCache	49.4	86.1	100	69.6	59.2	99.8	96.0	100	100	42.4	71.2	58.4	99.8	79.4
SnapKV	57.9	83.2	100	68.0	60.8	99.8	93.6	100	100	27.2	74.4	57.6	99.7	78.6
+ CriticalKV	62.4	85.1	100	67.2	64.8	100	99.6	100	100	77.6	70.4	58.4	99.8	83.5
+ ProjCache	58.9	83.2	100	77.6	64.8	100	96.6	100	100	47.2	74.4	60.0	99.7	81.7
AdaKV	59.8	86.7	100	84.0	57.6	99.8	94.2	100	100	44.8	72.0	58.4	99.8	81.3
+ CriticalKV	63.1	88.0	100	86.4	67.2	100	99.6	100	100	96.8	71.2	58.4	99.8	87.0
+ ProjCache	65.2	87.2	100	89.6	72.0	100	99.4	100	100	77.6	72.0	59.2	99.8	86.3

Table 16: RULER-32K accuracy (%) on Qwen3-8B (Q-Agnostic), without compensation.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	Avg.
Full KV	67.8	88.3	100	92.8	80.8	100	99.0	100	100	100	72.8	60.0	99.8	89.3
10% KV budget														
H2O	13.0	77.6	8.80	1.60	0.00	9.80	6.40	53.6	2.40	2.40	30.4	32.0	43.2	21.6
+ CriticalKV	15.1	76.5	9.60	2.40	0.00	10.0	6.40	78.4	4.00	2.40	29.6	32.8	78.2	26.6
+ ProjCache	19.8	75.7	9.60	2.40	0.00	10.4	6.40	98.4	4.00	2.40	30.4	34.4	97.9	30.1
SnapKV	23.4	74.9	9.60	4.00	0.00	10.6	6.40	85.6	5.60	2.40	31.2	33.6	90.1	29.0
+ CriticalKV	20.6	75.2	56.0	4.80	0.00	73.0	63.6	100	72.8	4.00	30.4	32.0	100	48.6
+ ProjCache	20.0	68.3	9.60	4.00	0.00	10.8	6.40	98.4	5.60	2.40	32.8	33.6	99.5	30.1
AdaKV	21.5	76.5	10.4	5.60	0.800	10.4	6.40	96.0	6.40	2.40	29.6	32.8	94.7	30.3
+ CriticalKV	22.2	79.5	58.4	4.80	0.00	71.2	65.6	100	80.0	7.20	29.6	31.2	99.7	50.0
+ ProjCache	37.2	76.0	11.2	6.40	1.60	10.8	7.00	99.2	11.2	2.40	27.2	38.4	99.4	32.9
20% KV budget														
H2O	24.4	80.3	11.2	4.00	0.800	11.4	6.80	84.0	7.20	4.00	31.2	38.4	81.9	29.7
+ CriticalKV	27.7	80.3	16.8	8.00	0.800	13.4	11.6	96.8	14.4	4.80	30.4	41.6	95.0	34.0
+ ProjCache	36.1	83.2	13.6	8.00	0.00	12.0	7.80	100	16.0	4.80	32.0	40.0	99.2	34.8
SnapKV	40.6	81.6	12.8	7.20	1.60	11.8	7.40	96.8	14.4	4.00	32.0	40.0	99.8	34.6
+ CriticalKV	43.5	84.0	88.8	12.0	0.800	93.8	96.4	100	93.6	37.6	34.4	46.4	99.7	63.9
+ ProjCache	40.6	77.3	13.6	6.40	0.800	12.0	8.40	99.2	16.8	5.60	28.0	39.2	100	34.5
AdaKV	44.6	83.5	12.8	9.60	1.60	11.8	7.20	99.2	16.0	5.60	28.8	43.2	99.5	35.6
+ CriticalKV	43.3	85.1	92.0	18.4	4.80	97.4	96.2	100	99.2	65.6	37.6	44.0	100	68.0
+ ProjCache	56.0	84.5	16.8	21.6	8.80	14.6	9.60	100	38.4	8.00	29.6	45.6	99.5	41.0
30% KV budget														
H2O	34.2	84.0	16.0	8.00	1.60	13.6	11.4	95.2	19.2	8.00	28.8	41.6	96.8	35.3
+ CriticalKV	38.6	84.3	27.2	13.6	1.60	23.8	25.6	99.2	60.8	18.4	31.2	44.8	99.7	43.7
+ ProjCache	46.0	87.5	20.0	15.2	1.60	17.6	11.8	99.2	44.8	8.00	36.0	46.4	99.7	41.1
SnapKV	53.2	84.0	16.8	11.2	3.20	14.2	10.6	100	24.0	8.00	28.0	44.0	100	38.2
+ CriticalKV	58.0	86.1	97.6	22.4	4.00	96.6	97.8	100	99.2	72.0	43.2	48.0	99.8	71.1
+ ProjCache	55.0	83.2	18.4	15.2	3.20	15.6	11.4	100	27.2	6.40	29.6	43.2	99.8	39.1
AdaKV	57.1	85.6	18.4	18.4	6.40	17.8	13.6	99.2	38.4	10.4	34.4	44.0	100	41.8
+ CriticalKV	60.8	88.5	98.4	50.4	11.2	98.8	98.2	100	100	89.6	49.6	52.8	99.8	76.8
+ ProjCache	65.8	86.9	23.2	43.2	20.0	23.6	19.8	100	76.0	28.0	43.2	48.0	99.8	52.1

Table 17: Compensation on LongBench with LLaMA-3.1-8B at a 20% KV budget; compensation rows use fixed ProjCache eviction sets.

	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Synthetic		Code			
Method	NQA	Qasper	MF-en	HQA	2WQA	MSQ	GovR	QMSum	MNews	TREC	TQA	SAMSum	PCount	PRe	LCC	RB-P	ΔMSE (%)	Avg.
Full KV	30.7	47.7	54.6	59.5	51.0	32.9	35.3	25.2	27.0	72.5	89.5	44.8	10.2	100	52.6	49.2	–	48.9
Q-Aware, 20% KV budget																		
H2O w/ ProjCache	30.9	40.1	54.8	57.2	50.0	33.0	29.2	24.2	24.0	54.5	89.9	43.8	11.2	100	51.3	51.1	–	46.6
+ OBS	31.3	41.1	54.7	56.6	48.8	32.1	29.4	24.2	23.7	57.0	90.0	44.5	13.2	100	51.1	51.0	–42.15	46.8
+ CaM	31.0	40.5	55.0	56.7	49.8	32.2	29.4	24.3	23.6	54.5	89.8	43.4	10.7	100	50.9	50.7	+1.78	46.4
+ D2O	30.1	36.9	53.3	56.7	49.8	30.7	26.4	24.3	22.2	51.5	89.2	43.3	11.6	100	51.1	51.6	>+100	45.5
SnapKV w/ ProjCache	29.9	45.7	55.2	58.9	50.2	31.9	29.1	24.6	23.2	58.0	89.4	44.3	12.1	100	53.5	52.2	–	47.4
+ OBS	30.0	45.7	55.0	59.2	49.9	32.4	29.6	24.7	23.7	59.5	89.4	45.0	11.6	100	52.5	51.2	–29.97	47.5
+ CaM	30.5	46.5	55.4	58.9	50.0	32.3	29.1	24.4	23.0	59.5	89.3	44.2	11.6	100	53.3	51.8	+1.14	47.5
+ D2O	30.9	39.8	54.2	57.8	47.4	30.8	26.8	24.2	21.4	49.0	89.8	44.5	14.1	100	53.0	52.9	>+100	46.0
AdaKV w/ ProjCache	29.8	45.9	56.5	58.6	50.6	33.5	30.1	25.0	23.9	68.5	90.0	44.4	11.6	100	52.8	53.1	–	48.4
+ OBS	29.8	47.5	54.5	59.4	52.5	34.0	30.5	24.7	24.2	69.5	89.5	44.9	10.1	100	52.5	51.5	–37.63	48.4
+ CaM	29.4	45.6	56.4	58.8	51.4	33.4	30.0	24.6	23.9	69.0	89.7	44.1	9.55	100	52.8	52.6	+1.43	48.2
+ D2O	29.2	43.2	55.2	58.2	49.2	32.8	27.5	24.4	22.8	64.0	90.1	43.5	11.7	100	49.2	49.2	>+100	46.9
Q-Agnostic, 20% KV budget																		
H2O w/ ProjCache	29.8	28.5	39.5	52.4	39.7	29.5	28.5	21.7	23.5	50.0	91.3	45.1	11.5	94.0	51.3	48.0	–	42.8
+ OBS	29.2	27.9	38.8	52.1	38.4	28.1	28.9	22.1	23.6	51.0	91.2	44.4	11.0	94.0	51.1	48.5	–43.75	42.5
+ CaM	29.4	28.2	39.6	51.8	39.6	28.0	28.7	21.6	23.5	50.0	91.3	45.0	11.5	93.5	50.9	48.3	+2.37	42.6
+ D2O	28.8	18.6	26.6	45.1	26.0	20.5	26.2	20.5	22.0	48.0	90.6	43.4	10.5	75.5	51.1	47.9	>+100	37.6
SnapKV w/ ProjCache	27.7	27.2	32.7	51.9	36.8	27.6	28.3	22.2	22.7	50.0	89.5	44.0	9.60	93.0	53.5	48.6	–	41.6
+ OBS	26.9	30.1	34.1	51.5	36.5	26.8	29.1	22.3	23.2	49.0	89.5	44.1	10.1	94.0	52.5	47.0	–31.37	41.7
+ CaM	28.0	27.5	33.9	51.9	37.5	27.9	28.5	22.6	22.9	48.0	89.5	43.7	9.10	93.0	53.3	49.0	+1.50	41.6
+ D2O	26.2	22.1	29.1	47.9	27.9	21.4	25.9	20.5	21.9	46.0	89.3	43.3	10.6	71.5	53.3	48.7	>+100	37.9
AdaKV w/ ProjCache	28.6	34.4	40.0	54.0	43.9	30.9	29.0	22.9	23.9	63.0	90.4	44.7	9.55	95.5	52.8	49.7	–	44.6
+ OBS	29.3	32.3	40.7	53.3	43.3	29.6	29.8	22.9	23.7	63.0	90.3	44.0	9.50	98.0	52.5	49.2	–39.25	44.5
+ CaM	28.9	34.5	41.1	53.3	43.6	30.1	28.9	23.0	23.8	63.5	90.8	44.7	8.50	96.0	52.8	49.7	+1.92	44.6
+ D2O	27.0	28.5	34.9	52.9	39.7	22.8	27.1	21.7	22.3	61.0	89.4	43.9	9.10	90.5	49.4	50.1	>+100	41.9

Table 18: Compensation on RULER-4K with LLaMA-3.1-8B at a 20% KV budget; compensation rows use fixed ProjCache eviction sets.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	ΔMSE (%)	Avg.
Full KV	99.6	85.1	100	100	100	100	100	100	100	100	88.0	65.6	100	–	95.3
Q-Aware, 20% KV budget															
H2O w/ ProjCache	87.6	49.9	97.6	60.0	13.6	85.0	39.8	95.2	92.0	1.60	88.0	63.2	79.7	–	65.6
+ OBS	89.8	50.4	98.4	61.6	12.8	86.2	38.0	93.6	92.0	1.60	89.6	63.2	80.8	−28.32	66.0
+ CaM	87.5	50.1	98.4	63.2	12.8	86.8	39.0	95.2	92.0	1.60	88.0	64.0	81.8	+0.51	66.2
+ D2O	76.8	47.2	94.4	44.8	1.60	77.4	30.2	96.0	88.0	1.60	85.6	64.0	71.2	>+100	59.9
SnapKV w/ ProjCache	70.7	60.3	100	85.6	19.2	96.2	76.2	100	96.8	1.60	88.0	63.2	96.3	–	73.4
+ OBS	72.2	66.1	100	86.4	20.0	94.8	76.2	100	96.8	1.60	87.2	63.2	95.8	−23.63	73.9
+ CaM	69.4	60.0	100	88.8	19.2	96.0	76.6	100	96.8	1.60	88.0	64.0	95.8	+0.43	73.6
+ D2O	54.9	55.7	97.6	27.2	0.00	93.0	66.4	100	94.4	1.60	85.6	61.6	94.2	>+100	64.0
AdaKV w/ ProjCache	93.1	67.5	99.2	96.0	76.8	97.0	88.0	100	98.4	1.60	89.6	64.8	98.7	–	82.4
+ OBS	94.4	72.8	99.2	96.0	76.0	95.0	87.6	100	98.4	1.60	88.8	64.0	98.2	−29.16	82.5
+ CaM	93.1	66.9	99.2	96.0	76.8	96.6	88.4	100	98.4	1.60	88.8	65.6	98.4	+0.42	82.3
+ D2O	92.8	64.3	97.6	90.4	59.2	98.0	88.0	100	99.2	1.60	87.2	64.8	98.4	>+100	80.1
Q-Agnostic, 20% KV budget															
H2O w/ ProjCache	89.6	72.5	34.4	24.0	2.40	15.0	9.80	68.8	20.0	1.60	36.8	44.8	56.8	–	36.7
+ OBS	89.5	75.7	34.4	21.6	2.40	14.8	9.60	68.0	20.0	0.800	40.0	45.6	54.7	−30.63	36.7
+ CaM	89.0	73.3	36.0	22.4	2.40	14.6	9.80	68.0	19.2	1.60	38.4	44.0	55.2	+0.90	36.5
+ D2O	76.8	66.4	18.4	6.40	0.800	10.4	9.00	70.4	16.8	0.800	16.0	36.8	42.1	>+100	28.5
SnapKV w/ ProjCache	70.6	66.1	40.8	17.6	2.40	22.6	11.2	84.8	40.8	1.60	27.2	40.0	77.8	–	38.7
+ OBS	68.5	66.7	40.8	16.8	1.60	20.6	10.6	85.6	40.8	0.800	30.4	40.0	78.9	−21.55	38.6
+ CaM	67.8	65.9	40.8	18.4	2.40	21.2	11.2	84.8	37.6	1.60	28.0	39.2	80.0	+0.67	38.4
+ D2O	28.6	64.5	24.0	9.60	0.800	14.0	9.80	88.0	32.8	1.60	17.6	37.6	66.1	>+100	30.4
AdaKV w/ ProjCache	95.6	72.8	59.2	56.0	20.8	28.6	12.0	95.2	46.4	1.60	40.8	46.4	94.4	–	51.5
+ OBS	95.8	73.3	60.8	56.0	20.8	27.4	13.0	95.2	48.8	1.60	42.4	44.8	92.6	−27.34	51.7
+ CaM	96.4	72.5	58.4	56.0	20.8	28.4	11.8	96.0	48.8	1.60	40.8	48.0	93.1	+0.75	51.7
+ D2O	89.4	73.1	48.8	47.2	15.2	23.0	10.8	96.8	44.0	1.60	31.2	39.2	90.4	>+100	47.0

Table 19: Compensation on LongBench with Qwen3-8B at a 20% KV budget; compensation rows use fixed ProjCache eviction sets.

	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Synthetic		Code			
Method	NQA	Qasper	MF-en	HQA	2WQA	MSQ	GovR	QMSum	MNews	TREC	TQA	SAMSum	PCount	PRe	LCC	RB-P	ΔMSE (%)	Avg.
Full KV	28.8	44.1	55.0	62.9	49.2	35.3	33.5	24.5	24.7	71.5	89.3	44.6	9.50	91.8	55.8	54.9	–	48.5
Q-Aware, 20% KV budget																		
H2O w/ ProjCache	28.6	40.3	52.2	61.1	47.0	33.6	29.8	23.9	21.1	61.0	89.0	43.3	8.61	96.8	56.7	55.9	–	46.8
+ OBS	28.8	41.4	51.8	60.6	47.0	34.0	29.9	24.1	21.4	60.5	89.0	43.3	8.11	96.4	56.3	55.8	−32.37	46.8
+ CaM	28.1	41.4	52.2	61.2	47.7	34.4	29.4	23.9	21.0	61.5	89.5	43.2	9.56	95.9	56.5	55.4	+31.34	46.9
+ D2O	27.9	36.4	49.0	60.1	43.2	31.3	25.4	23.1	18.8	48.5	89.1	42.8	9.17	97.0	50.4	53.9	>+100	44.1
SnapKV w/ ProjCache	28.9	42.8	52.6	61.6	47.6	34.4	29.5	23.8	20.6	63.0	88.8	43.5	8.00	92.6	56.7	56.3	–	46.9
+ OBS	28.6	42.6	51.5	61.3	48.1	35.7	29.9	23.5	21.0	62.5	89.3	43.0	7.55	93.1	57.0	56.1	−24.37	46.9
+ CaM	28.5	42.2	52.4	61.3	46.1	33.9	29.4	24.3	20.5	62.0	88.8	43.1	7.00	92.1	56.2	56.5	+26.20	46.5
+ D2O	28.1	35.6	48.4	59.5	42.5	29.1	25.1	22.8	17.8	43.0	88.9	42.4	6.19	94.3	52.0	53.6	>+100	43.1
AdaKV w/ ProjCache	29.2	42.3	53.1	61.7	47.8	36.0	30.1	24.1	20.8	67.5	88.8	43.9	8.55	94.1	57.4	57.2	–	47.7
+ OBS	29.6	42.8	53.2	62.2	48.4	35.7	30.2	24.3	21.2	68.5	88.8	43.7	7.05	93.1	57.5	56.8	−30.82	47.7
+ CaM	28.8	42.5	53.0	61.7	47.0	33.9	29.9	23.9	21.3	68.0	88.8	43.8	8.50	93.2	57.0	56.6	+35.28	47.4
+ D2O	28.7	39.0	51.7	60.6	44.8	31.8	25.9	23.6	19.0	57.5	88.5	43.5	7.05	93.6	52.8	53.7	>+100	45.1
Q-Agnostic, 20% KV budget																		
H2O w/ ProjCache	26.5	26.4	35.8	55.2	40.3	28.5	28.5	21.3	21.8	56.0	91.2	42.9	7.44	91.2	56.7	52.3	–	42.6
+ OBS	26.4	25.9	35.8	54.1	40.0	28.8	28.7	20.9	21.7	55.0	91.2	42.2	7.78	90.5	56.3	52.6	−33.00	42.4
+ CaM	26.6	26.4	35.5	53.6	40.4	27.4	28.4	21.1	21.7	56.0	91.7	42.5	8.50	90.9	56.1	53.0	+36.66	42.5
+ D2O	24.2	17.4	26.5	42.2	28.4	20.1	25.0	19.2	19.3	40.1	92.0	42.7	6.82	60.2	50.5	50.6	>+100	35.3
SnapKV w/ ProjCache	24.0	30.2	33.1	48.5	38.1	26.3	29.1	20.7	21.2	51.5	90.7	42.5	6.05	87.6	56.7	52.7	–	41.2
+ OBS	25.5	29.3	32.5	48.5	39.2	26.7	29.5	21.4	21.4	53.0	90.9	43.5	7.60	88.0	57.0	53.2	−24.46	41.7
+ CaM	24.1	30.1	33.3	48.2	38.5	25.6	29.4	20.8	21.1	50.5	90.2	42.5	7.55	87.4	56.2	52.6	+29.76	41.1
+ D2O	19.0	18.0	26.1	40.2	27.4	17.4	24.9	18.9	18.7	39.0	91.0	41.9	6.55	57.6	52.1	51.5	>+100	34.4
AdaKV w/ ProjCache	24.8	31.1	37.7	53.6	39.0	29.3	29.3	21.7	21.1	61.5	90.7	43.4	8.19	91.5	57.4	54.1	–	43.4
+ OBS	25.0	31.1	36.8	54.1	42.0	27.7	29.7	21.8	21.7	61.0	89.3	43.9	9.85	92.8	57.5	54.7	−30.98	43.7
+ CaM	25.2	30.5	36.6	52.7	39.6	27.7	29.4	21.0	21.4	60.5	89.2	42.9	7.60	89.8	57.0	54.2	+40.51	42.8
+ D2O	23.1	23.4	30.2	47.2	33.8	22.3	25.6	19.2	19.6	51.0	90.2	42.9	6.50	77.1	52.6	51.8	>+100	38.5

Table 20: Compensation on RULER-4K with Qwen3-8B at a 20% KV budget; compensation rows use fixed ProjCache eviction sets.

Method	CWE	FWE	MK-1	MK-2	MK-3	MQ	MV	S-1	S-2	S-3	QA-1	QA-2	VT	ΔMSE (%)	Avg.
Full KV	99.0	84.8	100	100	100	100	100	100	100	100	82.4	67.2	100	–	94.9
Q-Aware, 20% KV budget															
H2O w/ ProjCache	87.9	53.1	96.8	40.8	57.6	99.6	35.6	100	99.2	1.60	77.6	59.2	89.8	–	69.1
+ OBS	86.5	54.1	96.0	42.4	58.4	99.6	33.0	100	100	1.60	77.6	58.4	87.5	−23.06	68.9
+ CaM	85.5	52.5	96.8	43.2	56.0	99.6	33.6	100	100	1.60	77.6	60.0	88.2	+16.47	68.8
+ D2O	74.0	43.7	84.8	13.6	0.00	99.2	24.0	100	96.8	1.60	68.8	54.4	68.8	>+100	56.1
SnapKV w/ ProjCache	87.7	58.7	96.8	44.0	68.8	99.4	33.6	100	99.2	1.60	79.2	61.6	94.6	–	71.2
+ OBS	87.8	61.9	96.8	44.8	71.2	99.4	34.4	100	100	1.60	79.2	61.6	94.1	−19.20	71.8
+ CaM	86.2	58.4	96.8	44.8	68.0	99.4	33.4	100	100	1.60	79.2	60.0	94.9	+13.71	71.0
+ D2O	52.2	45.1	67.2	6.40	0.00	99.4	26.6	100	96.8	1.60	72.8	54.4	92.6	>+100	55.0
AdaKV w/ ProjCache	95.8	66.1	100	84.0	96.8	99.8	38.4	100	100	1.60	80.0	64.0	98.9	–	78.9
+ OBS	96.8	70.4	100	83.2	96.8	100	36.4	100	100	1.60	80.8	63.2	98.7	−23.29	79.1
+ CaM	95.5	65.6	100	84.0	96.0	99.8	36.4	100	100	1.60	80.0	61.6	98.6	+19.12	78.4
+ D2O	88.9	54.4	95.2	55.2	25.6	100	32.4	100	98.4	1.60	76.0	56.8	98.4	>+100	67.9
Q-Agnostic, 20% KV budget															
H2O w/ ProjCache	86.3	71.7	23.2	6.40	0.800	12.8	10.0	54.4	9.60	1.60	35.2	42.4	35.0	–	30.0
+ OBS	83.0	71.5	23.2	6.40	0.800	12.0	9.80	57.6	10.4	1.60	39.2	41.6	32.0	−24.87	29.9
+ CaM	83.6	71.2	22.4	6.40	0.800	11.8	9.80	57.6	11.2	1.60	40.0	40.8	32.2	+20.54	30.0
+ D2O	77.7	63.2	11.2	1.60	0.00	9.40	9.20	58.4	8.00	1.60	27.2	27.2	22.1	>+100	24.4
SnapKV w/ ProjCache	88.0	64.5	22.4	8.80	2.40	12.4	9.40	77.6	12.0	1.60	32.0	40.8	75.8	–	34.4
+ OBS	88.9	65.6	24.0	8.00	1.60	13.2	10.0	76.8	13.6	1.60	31.2	39.2	76.8	−17.17	34.7
+ CaM	88.1	64.0	21.6	8.80	2.40	12.6	9.40	78.4	13.6	1.60	29.6	40.0	76.5	+16.98	34.4
+ D2O	54.1	54.4	12.8	1.60	0.00	9.80	8.80	81.6	11.2	1.60	22.4	24.0	57.4	>+100	26.1
AdaKV w/ ProjCache	93.8	69.6	37.6	23.2	16.0	20.6	10.4	92.8	18.4	1.60	38.4	42.4	89.8	–	42.7
+ OBS	94.8	71.7	37.6	22.4	17.6	21.4	10.2	93.6	19.2	1.60	40.0	44.8	89.9	−21.39	43.4
+ CaM	94.9	69.9	37.6	24.0	16.0	21.8	10.2	93.6	20.0	1.60	39.2	42.4	89.0	+23.64	43.1
+ D2O	88.7	59.7	31.2	17.6	4.80	15.0	9.00	96.8	16.8	1.60	30.4	36.8	83.5	>+100	37.8