HYBRIDKV: Exploiting Head-Dominant Reconstruction for Efficient Query-Agnostic KV Cache Compression

Changwoo Baek [1] **Kyeongbo Kong** [1]

Abstract

Efficient key–value (KV) cache compression is crucial for large language models with long contexts. While context-reconstruction attention enables query-agnostic KV compression, its practical use is limited by large compression overhead, i.e., additional prefill-time computation required for reconstruction-based importance scoring beyond standard prefill. We show that reconstruction-based KV importance consistently concentrates on a subset of attention heads, largely independent of the input context. Based on this observation, we propose a hybrid KV cache compression method that combines context-independent head pre-pruning with token-level reconstruction-based pruning. By restricting expensive reconstruction scoring to selected heads, our method significantly reduces compression overhead. Experiments on long-context benchmarks demonstrate up to a 36% overhead reduction while largely preserving inference accuracy.

1. Introduction

Large language models (Yang et al., 2025b; Grattafiori et al., 2024; Liu et al., 2024; Achiam et al., 2023; Comanici et al., 2025) are increasingly pushed to longer contexts to support complex reasoning and retrieval. During inference, the key–value (KV) cache enables efficient autoregressive decoding, but its footprint grows linearly with context length and quickly becomes a dominant bottleneck in both memory and runtime. As a result, KV cache compression—selectively discarding less useful entries while maintaining accuracy—has become essential for scalable long-context inference.

[1]Pusan National University. Correspondence to: Changwoo Baek <higok18@pusan.ac.kr>, Kyeongbo Kong <kbkong@pusan.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning Workshop AdaptFM, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Most existing KV compression methods aim to reduce memory and attention compute under minimal accuracy loss (Xiao et al.; Liu et al., 2023; Li et al., 2024; Cai et al., 2024; Zhang et al., 2023), leveraging approximated attention structures, compact KV representations, or restricted effective context ranges. Many of these techniques estimate token importance in a query-dependent manner (Li et al., 2024; Cai et al., 2024). While effective for single-query use, query-dependent compression becomes problematic when multiple queries are issued sequentially over the same context: the cache must be recompressed per query, or a fixed compressed cache is reused and can be biased toward the initial query, degrading downstream performance.

As an alternative, context-reconstruction attention has recently been proposed as a query-agnostic criterion for KV cache compression (Kim et al., 2025). Instead of estimating importance with respect to a particular downstream query, it measures how well the original input context can be reconstructed from the encoded KV cache. KV entries that contribute most to this reconstruction can be interpreted as a compact, context-level representation—capturing information intrinsic to the context rather than features specialized to any single query. This view yields a natural query-agnostic property, allowing one compressed KV cache to be reused across diverse downstream queries without per-query recompression.

Despite these advantages, context-reconstruction attention incurs substantial compression overhead, since it requires additional prefill-time computation to simulate context reconstruction (Fig. 1(a)). Profiling in Fig. 1(c) shows that for a 124K-token context, the standard prefill on an RTX PRO 6000 96GB (bfloat16) takes approximately 14 seconds, whereas enabling context-reconstruction-based compression adds more than 34 seconds of extra computation. This large overhead severely limits the practicality of context-reconstruction attention for real-world long-context deployment.

Against this backdrop, we ask a simple question: how context-dependent is reconstruction-based KV importance in practice, and where does its structure come from? Through an empirical analysis of context-reconstruction attention, we find a consistent head-dominant pat-

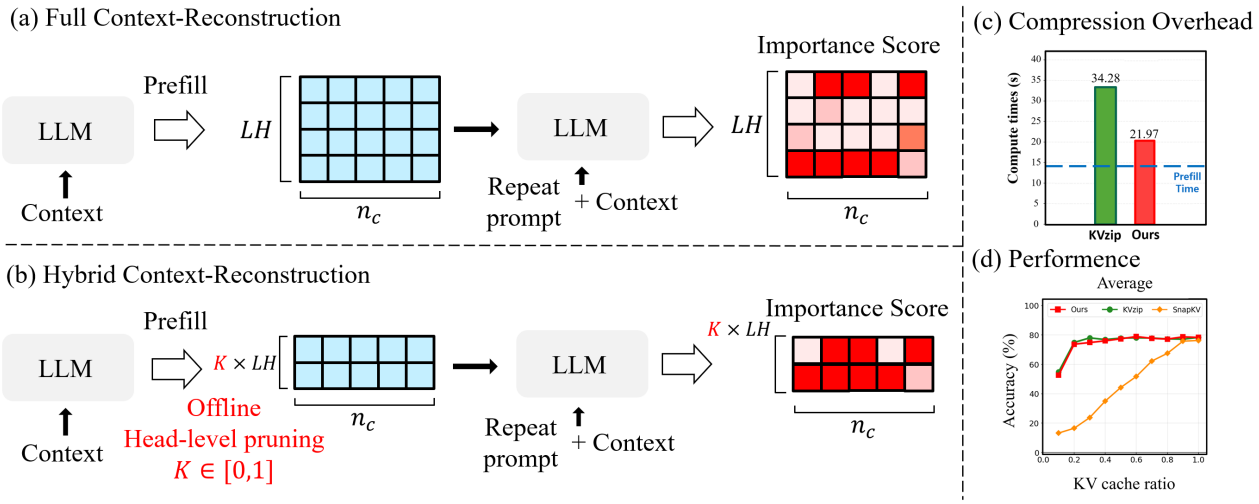

Figure 1. **Comparison between full and hybrid context-reconstruction attention.** (a) Full context-reconstruction performs importance scoring over all attention heads, incurring high compression overhead. (b) The proposed hybrid scheme reduces the search space by pre-selecting important heads in a context-independent manner, followed by token-level reconstruction-based scoring. (c) Computational overhead comparison between KVzip and the proposed method. (d) Performance comparison under different KV cache ratios on Qwen2.5-7B-1M.

tern—importance concentrates on a small, fixed subset of attention heads and remains largely stable across diverse input contexts. This suggests that reconstruction-based importance is not purely driven by fine-grained contextual variation, but instead reflects a model-dependent, head-level structure that can be exploited to reduce redundant computation.

Based on these findings, we make the following contributions:

- **Structural finding.** We show that context-reconstruction attention induces a *head-dominant* importance pattern: KV importance consistently concentrates on a small subset of attention heads largely independent of the input context, and we find that these heads are closely tied to long-range retrieval behavior.
- **Method.** Building on this observation, we propose a *hybrid* KV cache compression scheme that (i) pre-selects important heads in a *context-independent* offline step to shrink the reconstruction scoring space, and (ii) applies token-level context-reconstruction-based pruning within the selected heads at deployment time (Fig.1(b)).
- **Results.** Across long-context benchmarks and model settings, our method reduces *compression overhead* (additional prefill-time computation) by up to 36% while largely preserving inference accuracy (Fig.1(c,d)).

2. Related Work

KV Cache Compression. Efficient context management, such as KV cache compression, has emerged as a crucial strategy for long-context inference in LLMs and multimodal LLMs (Xiao et al.; Zhang et al., 2023; Baek et al., 2026; Tu et al., 2025). Early approaches rely on position-based policies: StreamingLLM (Xiao et al.) retains attention sinks together with a sliding window of recent tokens, achieving stable generation but discarding potentially important middle-context information. Subsequent work estimates token importance from attention patterns: H2O (Zhang et al., 2023) identifies Heavy Hitters based on accumulated attention scores during decoding, SnapKV (Li et al., 2024) leverages attention from recent queries to select important KV entries, and PyramidKV (Cai et al., 2024) allocates cache budgets pyramidally across layers. Because these methods estimate importance in a query-dependent manner, the resulting caches cannot be reused across different queries over the same context. KVzip (Kim et al., 2025) removes this restriction with a query-agnostic criterion based on context-reconstruction attention, but incurs substantial prefill-time overhead. We extend the query-agnostic perspective by exploiting the head-dominant structure of reconstruction attention, substantially reducing this cost.

Specialized Heads in LLMs. Recent studies have revealed that attention heads in LLMs exhibit strong functional specialization. (Wu et al., 2025) identify *retrieval heads* responsible for retrieving information from long contexts, and (Olsson et al., 2022) discover *induction heads*

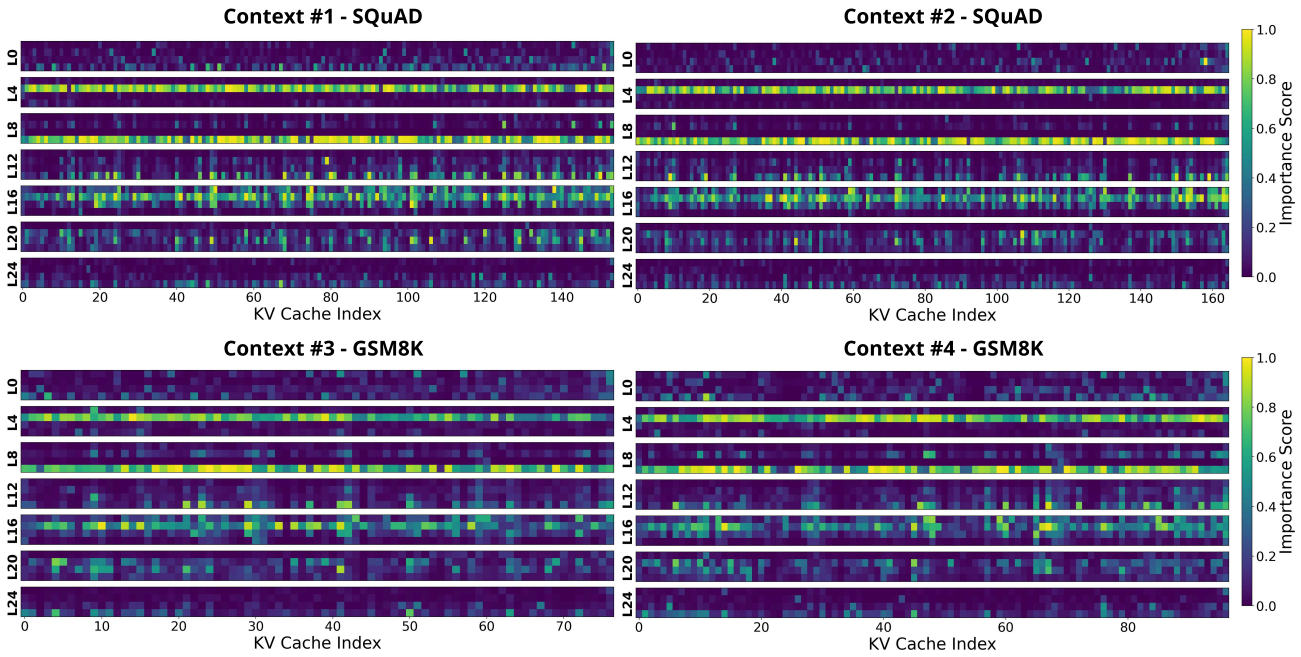

Figure 2. **Head-dominant patterns in context-reconstruction attention** Across different contexts, attention score visualizations show that the same layer–head indices consistently exhibit higher scores, while other heads maintain lower score distributions.

that perform in-context pattern matching. Building on these findings, recent works exploit head-level specialization to reduce KV cache cost: DuoAttention (Xiao et al., 2025) preserves the full KV cache for retrieval heads while aggressively compressing the cache of streaming heads, and RazorAttention (Tang et al., 2025) keeps retrieval-head caches intact while compressing non-retrieval heads via "compensation tokens". In contrast, our method identifies a head-dominant pattern directly from context-reconstruction attention, and combines context-independent head pre-pruning with token-level reconstruction-based pruning within the selected heads.

3. Preliminary

Context reconstruction attention. Context reconstruction attention was introduced in KVzip (Kim et al., 2025), which quantifies the importance of each KV pair based on its contribution to reconstructing the original context. Specifically, reconstruction is simulated through a single teacher-forced forward pass using an input sequence formed by concatenating a repeat prompt with the original context. The importance of a KV pair is then defined as the maximum attention it receives during this reconstruction process.

Formally, given a context of length n_c, we construct an input sequence of length $n_{\text{in}} = n_{\text{prompt}} + n_c$. Forwarding this sequence through the language model with the prefilled KV cache produces grouped-query features $\mathbf{Q}_{l,h} \in \mathbb{R}^{G \times n_{\text{in}} \times d}$ and key features $\mathbf{K}_{l,h} \in \mathbb{R}^{(n_c+n_{\text{in}}) \times d}$ at layer l and head h. The resulting cross-attention matrix is

$$\mathbf{A}_{l,h} = \text{Softmax}\big(\mathbf{Q}_{l,h}\mathbf{K}_{l,h}^{\top}\big), \quad \mathbf{A}_{l,h} \in \mathbb{R}^{G \times n_{\text{in}} \times (n_c+n_{\text{in}})}.$$

from which attention entries corresponding to the context keys are extracted as $\bar{\mathbf{A}}_{l,h} \in \mathbb{R}^{G \times n_{\text{in}} \times n_c}$. The importance score for the i-th KV pair is then computed as

$$S_{l,h}(i) = \max_{g=1,\ldots,G;\ j=1,\ldots,n_{\text{in}}} \bar{\mathbf{A}}_{l,h}[g, j, i].$$

We refer to these values as the maximum cross-attention scores.

4. Methods

Observation We analyze the structural characteristics of context-reconstruction attention scores across diverse input contexts using Qwen2.5-7B-1M (Yang et al., 2025b). As shown in Fig. 2, reconstruction-based importance exhibits a pronounced *head-dominant* structure: across distinct contexts, the *same* small subset of layer–head indices repeatedly produces disproportionately high importance scores, while most other heads remain consistently low. This reveals a stable head-level separation between structurally salient heads and largely inactive heads, suggesting that reconstruction-based scoring over all heads may be computationally redundant. Importantly, even within the salient heads, importance is not uniform across tokens; instead, token-level variation

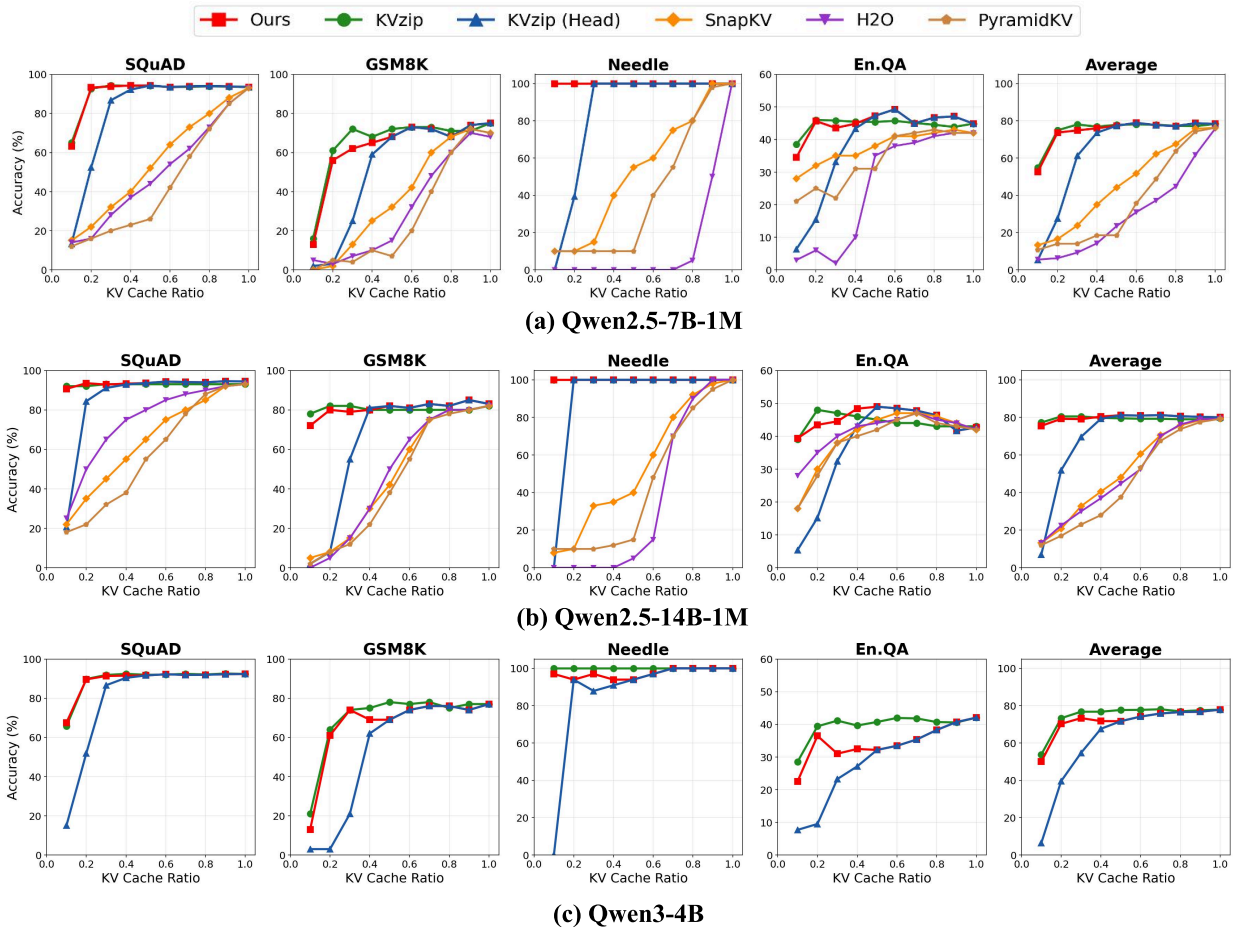

Figure 3. **Effect of KV cache ratio on accuracy across tasks for Qwen2.5-7B-1M, Qwen2.5-14B-1M, and Qwen3-4B** with pre-defined k=0.5. For cache ratios above 0.5, only head pruning is applied with k matched to the target cache ratio.

persists, indicating that context dependence is primarily expressed *within* a limited set of heads rather than across the entire head space.

This head-level differentiation is consistently observed across samples from SQuAD (Rajpurkar et al., 2016) and GSM8K (Cobbe et al., 2021), suggesting that reconstruction-induced importance is not primarily driven by fine-grained contextual differences, but instead reflects a model-dependent property of the attention structure. In particular, the heads that repeatedly carry high importance exhibit behaviors consistent with retrieval-oriented heads that attend to semantically salient tokens. However, high-importance heads do not assign uniformly high scores to all KV entries, nor do low-importance heads contain exclusively low-scoring entries. Rather, substantial *token-level* variation persists within heads, indicating that context-reconstruction attention preserves context-dependent signals while concentrating most of the mass on a small subset of structurally salient heads.

Hybrid Context-Reconstruction Compression. Motivated by the head-dominant structure of reconstruction-based importance, we propose HYBRIDKV (Algorithm 1), a hybrid KV cache compression method that removes *structurally redundant* head-level computation while retaining *context-adaptive* token-level decisions. Our idea is to restrict expensive context-reconstruction scoring to a small subset of consistently salient heads, since most heads contribute little to high-importance KV selection across contexts.

We first identify a subset of important attention heads by aggregating context-reconstruction attention scores over a small calibration set. For each head, we compute the maximum score across calibration samples and rank heads accordingly, then pre-select a fraction k of heads ($0 < k \leq 1$), where k denotes the *head-retention ratio* (equivalently, $1-k$ heads are pruned).

At deployment time, we discard KV entries associated with the pruned heads, thereby shrinking the search space for reconstruction-based scoring. We then apply token-

Algorithm 1 Overview of the proposed HYBRIDKV

Require: Pre-trained LLM, calibration set $\mathcal{D}_{\text{cal}}$, head-retention ratio k, cache ratio r, input context c
Ensure: Compressed KV cache
Stage 1: Offline Head Selection
Initialize per-head score $s_{l,h} \leftarrow 0$ for all heads (l, h)
for each $c' \in \mathcal{D}_{\text{cal}}$ **do**
  Compute $S_{l,h}(i)$ on c'
  $s_{l,h} \leftarrow \max\big(s_{l,h},\ \max_i S_{l,h}(i)\big)$
end for
Select retained heads $\mathcal{H}_{\text{keep}} \leftarrow$ top-k heads ranked by $s_{l,h}$
Stage 2: Online Per-context Compression
Prefill KV cache with input context c
Discard KV entries of pruned heads $(l, h) \notin \mathcal{H}_{\text{keep}}$
Compute reconstruction scores $S_{l,h}(i)$ on $\mathcal{H}_{\text{keep}}$
Evict lowest-score KV entries until ratio $\leq r$
return Compressed KV cache

level context-reconstruction attention only within the retained heads to perform fine-grained KV eviction, following KVzip, our compression operates between the prefill and decoding stages. By limiting reconstruction scoring to the selected heads, our method substantially reduces *compression overhead* (additional prefill-time computation) while preserving the token-level adaptivity of reconstruction attention. Overall, the proposed design offers an effective trade-off between computational cost and compression fidelity.

5. Experiments

Baselines. We compare against representative state-of-the-art KV cache eviction/compression methods, including H2O (Zhang et al., 2023), SnapKV (Li et al., 2024), PyramidKV (Cai et al., 2024), and KVzip (Kim et al., 2025). To disentangle the effect of *context-reconstruction attention* from its *online* reconstruction cost, we additionally report **KVzip (head)**, a head-level variant that uses *offline* head importance scores for head selection without performing token-level reconstruction scoring at deployment time. Unless otherwise noted, all experiments use greedy decoding with bfloat16 precision and are conducted on a single NVIDIA RTX PRO 6000 96GB GPU.

Dataset and Models. We evaluate long-document question answering on Needle-in-a-Haystack (NIAH) (Kamradt, 2023) and the long-context En.QA subset of SCBench (LI et al., 2025), which stress retrieval and multi-hop reasoning under long contexts. We use instruction-tuned long-context LLMs, including Qwen2.5-7B-1M (Yang et al., 2025b), Qwen2.5-14B-1M (Yang et al., 2025b), and Qwen3-4B (Yang et al., 2025a), all of which employ grouped-query

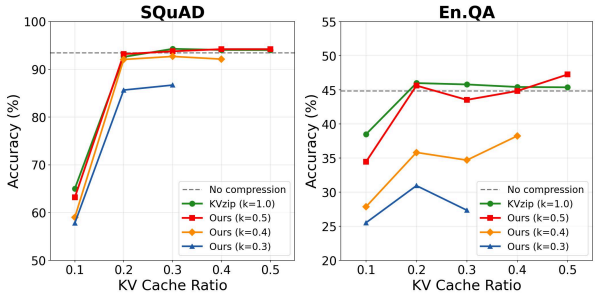

Figure 4. **Ablation on head-retention ratio k on Qwen2.5-7B-Instruct-1M.** Trade-off between retained heads and compression fidelity.

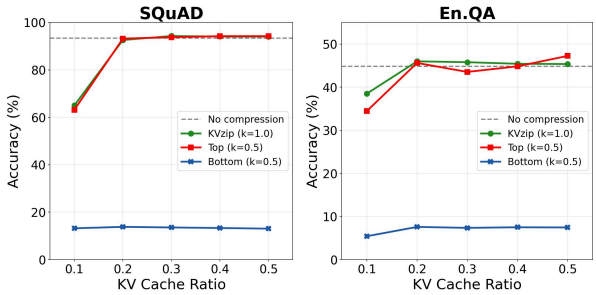

Figure 5. **Top vs Bottom head selection on Qwen2.5-7B-Instruct-1M.** Top retains the highest-scoring heads (our default); Bottom retains the lowest-scoring ones.

attention with their original configurations.

Experiment Setup We set the context-independent head-pruning ratio k to 0.5 by default. For calibration, we compute head importance scores using only 10 randomly sampled contexts (average $\sim$10K tokens) from LongAlpaca-12k (Chen et al., 2024). For cache ratios above 0.5, we set k equal to the target cache ratio and apply only head pruning without token-level reconstruction. All evaluations follow the evaluation setting of KVzip. Specifically, KV caches are prefilled and compressed using only the input context in a query-agnostic manner, and the resulting compressed cache is reused to answer one or multiple queries per context.

Main results Figure 3 summarizes the performance of different KV cache compression methods under varying cache ratios. Across all benchmarks, and all evaluated models (Qwen2.5-7B-1M, Qwen2.5-14B-1M, and Qwen3-4B), our method consistently matches or outperforms existing approaches at the same KV cache ratio, particularly in low- and mid-cache regimes. Compared to fully context-dependent reconstruction (KVzip), the proposed hybrid method achieves similar accuracy while substantially reducing compression overhead. Notably, it outperforms context-dependent baselines such as SnapKV, H2O, and PyramidKV on retrieval-

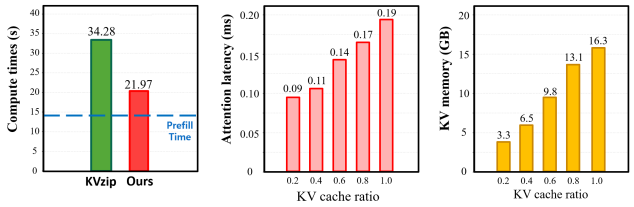

Figure 6. **Efficiency analysis of KV cache compression on Qwen2.5-7B-1M.**

and reasoning-intensive tasks. We further compare against the head-level eviction variant of KVzip, which relies on maximum reconstruction scores for offline head pruning, and observe that while such head-only methods suffer from severe performance degradation at low cache ratios, our method consistently preserves accuracy by retaining token-level adaptivity within structurally important heads.

Effect of head-retention ratio. We study the sensitivity of the proposed hybrid compression to the head-retention ratio k, which controls the fraction of attention heads on which token-level context-reconstruction scoring is performed. Figure 4 reports accuracy on SQuAD and En.QA with Qwen2.5-7B-1M as k varies from 0.3 to 0.5, alongside KVzip (k=1.0) as a full-head reference that performs reconstruction scoring over all heads. At k=0.5, our method matches KVzip across the entire KV cache ratio range on both datasets despite restricting reconstruction scoring to only half of the heads. This suggests that our calibration-based ranking reliably identifies the heads that contribute to reconstruction. Reducing k to 0.4 leads to a small drop on SQuAD (roughly 2%) but a more pronounced degradation on En.QA (around 9% at a cache ratio of 0.2, and k=0.3 causes a substantial drop on both benchmarks. The sharp transition between k=0.5 and k=0.4 suggests that the boundary of the heads that need to be retained is relatively well-defined. The difference in degradation across datasets further indicates that sensitivity to head pruning may vary with task characteristics.

Based on these observations, we adopt k=0.5 as the default configuration. We further note that the performance drop at very low cache ratios (e.g., 0.1) is largely independent of k, indicating that this regime is limited by the token-level cache budget rather than by head selection.

Effect of head selection strategy. To verify that our calibration-based ranking captures meaningful structural importance, we compare two opposite selection strategies at k=0.5 on Qwen2.5-7B-1M: *Top* retains the highest-scoring heads (our default), while *Bottom* retains the lowest-scoring ones. Both variants share the same head budget and apply token-level reconstruction scoring within their respective head subsets, isolating the effect of head selection itself. As shown in Fig. 5, Top matches KVzip across all cache ratios on both SQuAD and En.QA, indicating that the top-ranked half of heads alone is sufficient to recover full-head reconstruction performance. In contrast, Bottom collapses across all cache ratios on both benchmarks, with the gap to KVzip remaining wide even at high cache ratios where token-level eviction removes only a small fraction of entries. This dramatic gap shows that the top-ranked heads play a decisive role in reconstruction, while the bottom-ranked heads carry little useful signal regardless of how many tokens are retained within them. These results confirm that our calibration-based ranking accurately identifies these critical heads, supporting it as an effective method for selecting structurally important heads.

Efficiency Analysis Figure 6 demonstrates the efficiency of the proposed method under practical inference settings. With KV cache compression at k=0.5 on a 124K-context input, the compression overhead is reduced from 34.26 seconds with KVzip to 21.97 seconds—a 36% reduction—substantially lowering the cost of context-reconstruction-based compression by restricting reconstruction scoring to a subset of attention heads. As a reference, the standard prefill itself takes approximately 14 seconds, meaning that KVzip more than triples the total prefill cost while our method limits this inflation to roughly $2.5\times$. Moreover, reducing the KV cache ratio consistently decreases both decoding latency and memory consumption: attention latency drops from 0.19 ms to 0.09 ms, and KV cache memory shrinks from 16.3 GB to 3.3 GB as the cache ratio decreases from 1.0 to 0.2, both scaling near-linearly with the cache ratio. These results confirm that the proposed approach achieves significant memory and latency savings while preserving stable decoding performance.

6. Conclusion

We presented HYBRIDKV, an efficient query-agnostic KV cache compression method that exploits the head-dominant structure of context-reconstruction attention. Through empirical analysis, we showed that reconstruction-based KV importance consistently concentrates on a small, fixed subset of attention heads largely independent of the input context. Building on this finding, we proposed a hybrid scheme that combines context-independent head pre-pruning with token-level reconstruction-based pruning, restricting expensive reconstruction scoring to structurally salient heads. Experiments on long-context benchmarks across three model scales (Qwen2.5-7B-1M, Qwen2.5-14B-1M, and Qwen3-4B) demonstrate that our method reduces compression overhead by up to 36% while preserving inference accuracy comparable to fully context-dependent reconstruction. We believe this head-dominant perspective opens a promising direction for designing efficient query-agnostic KV cache compression methods that better exploit the inherent structure of attention.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT) (No. RS-2024-00456152) and the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea(Ministry of Science and ICT), and the authors gratefully acknowledge the Cluster Server for Computational Science at Pusan National University for providing computational resources.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Baek, C., Song, J., Kim, S., and Kong, K. Agilepruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2NLkhPex1M.

Cai, Z., Zhang, Y., Gao, B., Liu, Y., Li, Y., Liu, T., Lu, K., Xiong, W., Dong, Y., Hu, J., et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., and Jia, J. Longlora: Efficient fine-tuning of long-context large language models. In *International Conference on Learning Representations*, volume 2024, pp. 8220–8238, 2024.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kamradt, G. Needle in a haystack-pressure testing llms. *GitHub repository*, pp. 28, 2023.

Kim, J.-H., Kim, J., Kwon, S., Lee, J. W., Yun, S., and Song, H. O. Kvzip: Query-agnostic kv cache compression with context reconstruction. *Advances in Neural Information Processing Systems*, 2025.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., and Chen, D. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

LI, Y., Jiang, H., Wu, Q., Luo, X., Ahn, S., Zhang, C., Abdi, A. H., Li, D., Gao, J., Yang, Y., and Qiu, L. SCBench: A KV cache-centric analysis of long-context methods. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gkUyYcY1W9.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Liu, Z., Desai, A., Liao, F., Wang, W., Xie, V., Xu, Z., Kyrillidis, A., and Shrivastava, A. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Tang, H., Lin, Y., Lin, J., Han, Q., Ke, D., Hong, S., Yao, Y., and Wang, G. Razorattention: Efficient kv cache compression through retrieval heads. 2025.

Tu, D., Vashchilenko, D., Lu, Y., and Xu, P. Vl-cache: Sparsity and modality-aware kv cache compression for vision-language model inference acceleration. In *International Conference on Learning Representations*, volume 2025, pp. 219–239, 2025.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. Retrieval head mechanistically explains long-context factuality. 2025.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Xiao, G., Tang, J., Zuo, J., Yang, S., Tang, H., Fu, Y., Han, S., et al. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yang, A., Yu, B., Li, C., Liu, D., Huang, F., Huang, H., Jiang, J., Tu, J., Zhang, J., Zhou, J., et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025b.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.

A. Head-Dominant Pattern on Other Models

In the main paper, we observed that context-reconstruction attention exhibits a pronounced head-dominant structure on Qwen2.5-7B-1M. To verify that this observation generalizes beyond a single model, we visualize the same importance score distributions on Qwen2.5-14B-1M (Figure 7) and Qwen3-4B (Figure 8). In both cases, the same head-dominant pattern consistently appears, confirming that this structure is a model-general property of context-reconstruction attention.

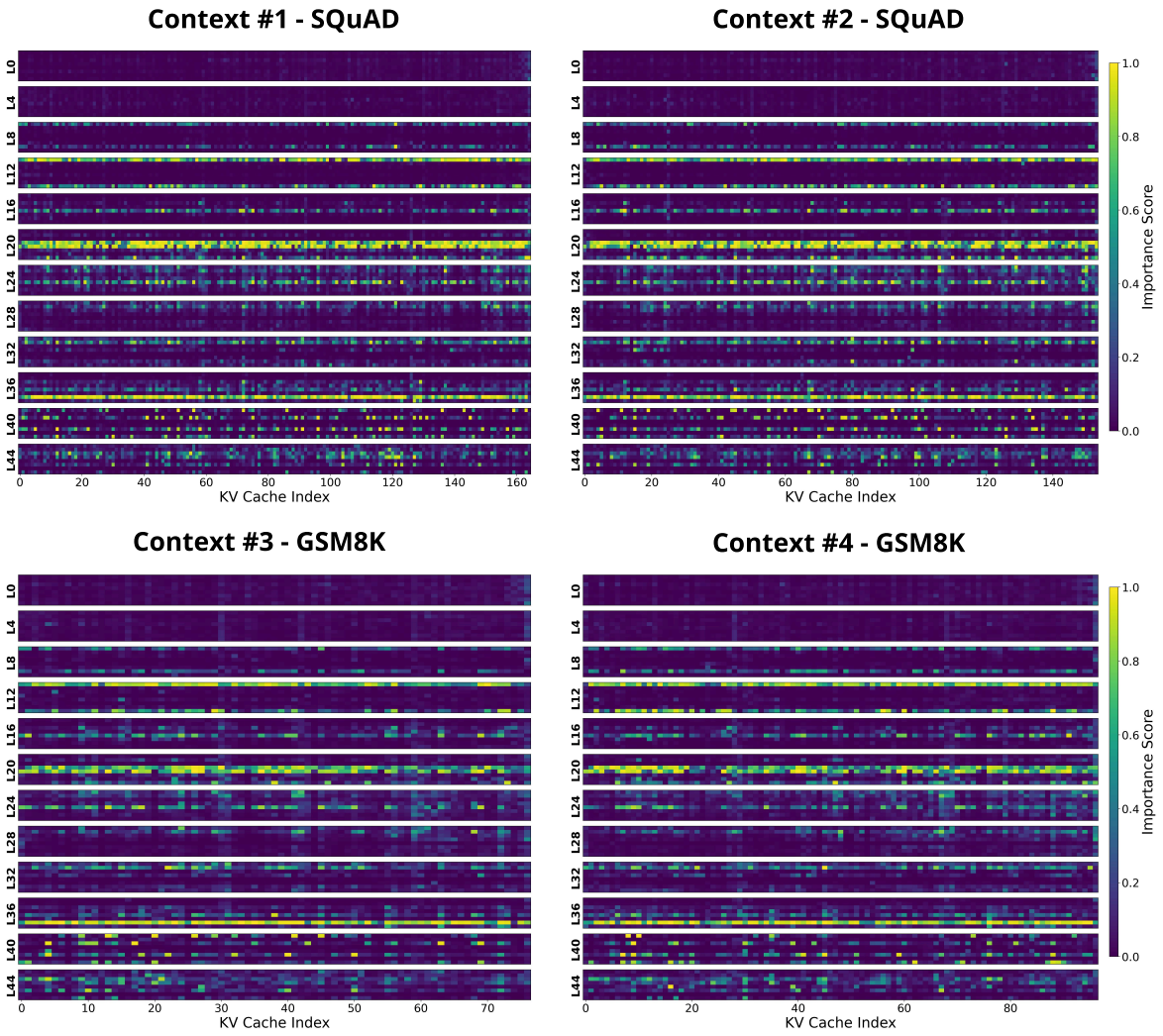

Figure 7. **Head-dominant patterns in context-reconstruction attention on Qwen2.5-14B-1M.**

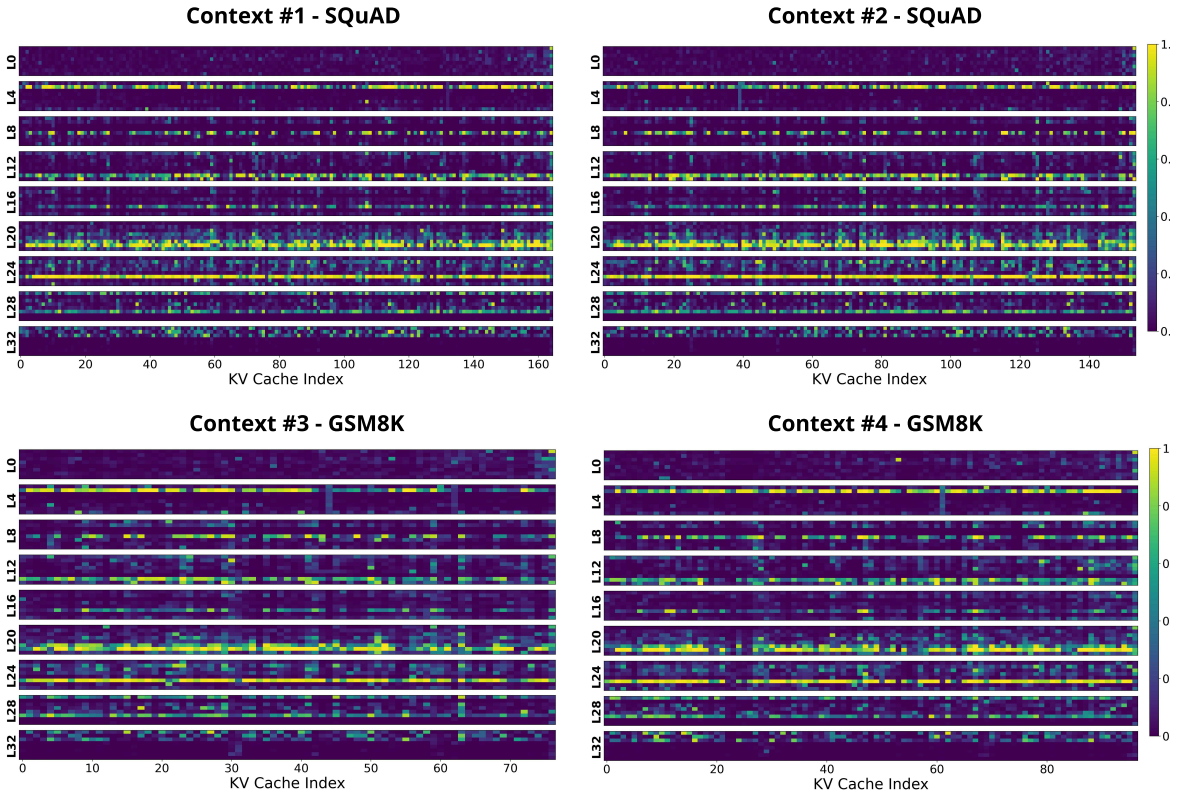

Figure 8. **Head-dominant patterns in context-reconstruction attention on Qwen3-4B.**