INFORMATION-AWARE KV CACHE COMPRESSION FOR LONG REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning capability has advanced rapidly in large language models (LLMs), leading to an increasing size of key-value (KV) cache in both prefilling and decoding stages. Existing KV cache compression methods mainly rely on attention weights to estimate token importance. While attention effectively captures contextual relevance, it overlooks complementary information-theoretic signals related to predictive uncertainty and token informativeness. In this paper, we revisit token importance from a forward-looking perspective and introduce *Forward Influence*, a metric that measures how compressed tokens affect future contexts. Our analysis reveals that tokens selected by attention scores mainly influence nearby contexts, whereas tokens associated with high predictive uncertainty exhibit substantially stronger influence on distant future contexts. Based on the observation, we propose **InfoKV**, an entropy-aware KV cache compression framework that incorporates information-theoretic signals. It combines token-level predictive uncertainty with layer-wise representation evolution and integrates the resulting entropy scores with attention scores during reasoning. Experiments on long-context reasoning benchmarks with Llama-3.1, Llama-3.2, and DeepSeek-R1 demonstrate that InfoKV consistently outperforms existing attention-based KV compression methods in both long prefilling and decoding scenarios. [1]

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities in long-context understanding and reasoning (Guo et al., 2025; OpenAI et al., 2026). However, their deployment in long-sequence processing remains computationally expensive due to the quadratic growth of computing attention and the linear growth of key–value (KV) cache memory (Łańcucki et al., 2025; Song et al., 2025). This bottleneck is especially pronounced in long-form reasoning tasks, where thousands of tokens are handled as inputs or outputs for LLMs.

To address this issue, recent studies have explored KV cache compression techniques that selectively retain only a subset of past tokens. A common paradigm estimates token importance based on attention weights from a fixed observation window, e.g., the most recent tokens (Li et al., 2024; Cai. et al., 2024b; Song et al., 2025). Tokens receiving larger attention weights from recent contexts are regarded as important and preserved, while the remaining tokens are discarded. Such strategies have shown promising improvements in inference efficiency and memory reduction.

Despite their effectiveness, attention-based KV compression methods suffer from an inherent limitation: they rely on **short-term, backward-looking** signals. Specifically, importance is inferred from the extent to which recent tokens attend to past tokens, which primarily captures local dependencies. However, long-form reasoning could depend on information that may not be directly activated by recent contexts but remains crucial for future reasoning trajectories. This mismatch becomes especially problematic in long-decoding scenarios, where reasoning paths evolve dynamically over generation steps.

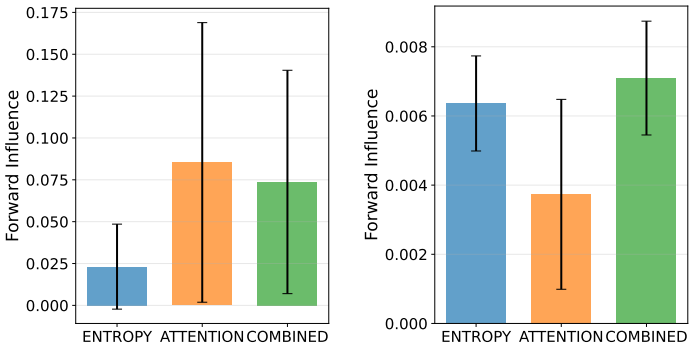
In this work, we show that effective KV cache compression should be guided by **forward-looking token utility**, namely, how much a token contributes to future generation steps rather than only its relevance to recent contexts. To investigate this phenomenon, we introduce *Forward Influence*,

[1] We will release our code for reproducibility later.

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.

which measures the divergence in future predictive distributions after removing a token from the KV cache. As shown in Figure 1, while attention emphasizes tokens that are closely relevant to recent contexts, entropy measures the informativeness of tokens, and those of high entropy exhibit substantially stronger and more persistent influence on distant future contexts.

Motivated by this observation, we propose an entropy-aware KV cache compression framework **InfoKV** that incorporates information-theoretic signals into token selection. Since entropy reflects the uncertainty of the model when predicting tokens, it naturally captures tokens carrying richer semantic information. To further characterize token importance across layers, we combine entropy with the representational evolution between intermediate and final layers, which is orthogonal to the sequence dimension.



(a) Short-range influence (128-token horizon).

(b) Long-range influence (14K-token horizon).

Figure 1: Comparison of short-range and long-range influence for top-1% tokens scored by entropy, attention weight and their combination over 100 documents from Arxiv-Summarization (Cohan et al., 2018). Short-range influence captures immediate predictive effects, whereas long-range influence reflects persistent long-context impact.

Extensive experiments on both long prefilling and long decoding benchmarks demonstrate that preserving informative tokens substantially improves reasoning performance. In long prefilling scenarios, InfoKV consistently outperforms existing attention-based KV cache compression methods on LongReason across different context lengths and cache budgets with Llama-3.1 and Llama-3.2. In long decoding scenarios, InfoKV further achieves substantial improvements on IFEval, AIME 2024, and LiveCodeBench with DeepSeek-R1, demonstrating its effectiveness for mathematical reasoning, instruction following, and code generation tasks.

2 Related Work

KV Cache Compression. A dominant line of KV cache compression research focuses on selectively evicting past tokens based on attention patterns. Recent methods such as SnapKV (Li et al., 2024), PyramidKV (Cai. et al., 2024b) and FastKV (Jo et al., 2026) propose heuristic pruning strategies that measure token importance by attention weights and discard less-attended tokens. Other works explore token merging to approximate the original attention of full cache (Zhang et al., 2024; Wang et al., 2024; Wan et al., 2025). Although these methods effectively reduce memory usage, they primarily rely on attention-based heuristics, which are inherently backward-looking and mainly take effect on long prefilling tasks with short answers.

Compression for Long-decoding. The reasoning ability of LLMs has raised increasing attention in recent years (Guo et al., 2025; OpenAI et al., 2026). With long reasoning paths to be generated, decoding latency and KV cache growth become more critical than prefilling efficiency. To address this challenge, recent studies have extended KV cache compression from the prefilling stage to the decoding stage. RPC (Song et al., 2025) generalizes SnapKV (Li et al., 2024) to online decoding by periodically compressing the KV cache throughout generation. Expected Attention (Devoto et al., 2025) further estimates the expected contribution of tokens to future attention, while LazyEviction (Zhang et al., 2026) detects token recurrence and applies a lagged KV eviction scheme. In addition, FreqKV (Kai et al., 2026) proposes an iterative frequency-domain compression framework that supports both prefilling and decoding compression, enabling efficient train-short-test-long capability.

Information Signals for Token Importance. Beyond attention-based heuristics, recent work has explored information-theoretic signals to characterize token importance from a more intrinsic perspective. Unlike attention weights, which depend on contextual interactions, the information that a token carries represents its native importance. Selective Context (Li et al., 2023) leverages self-information to quantify the informativeness of tokens and prune redundant content in LLM inputs. Building on uncertainty-based measures, Kai et al. (2024) propose SH2, which utilizes prediction uncertainty to identify informative tokens and adjust the output distribution for improved factuality. In the context of long-form reasoning, SeLaR (Fu & Luo, 2026) incorporates entropy-aware contrastive regularization to encourage exploration by pushing representations away from overconfident predictions. In this paper, we introduce information signals to better reflect the token influence on future contexts and optimize KV cache compression for long-context reasoning.

3 METHODOLOGY

3.1 REVISITING TOKEN IMPORTANCE DURING INFERENCE

Existing KV cache compression methods predominantly estimate token importance according to attention scores computed from a recent observation window. Specifically, tokens receiving large attention weights from recent tokens are regarded as important and preserved in the KV cache. Although effective for maintaining short-range dependencies, such strategies implicitly assume that tokens important to recent contexts will remain important for future generation steps. However, during long-form reasoning and extended decoding, the relevance of tokens evolves continuously, and tokens with high recent attention scores may only contribute locally to nearby contexts while providing limited utility for future reasoning trajectories.

To better characterize the long-term utility of tokens, we revisit token importance from an information-theoretic perspective. Intuitively, tokens associated with high uncertainty carry more information for the language model and are therefore more likely to influence future contexts to be generated. As revealed in Kai et al. (2024), these tokens are prone to be content words such as adjectives, nouns, and conjugated verbs, which are more informative than function words like conjunctions, determiners and prepositions.

Given a sequence of tokens $\{x_0, x_1, \cdots, x_{n-1}\}$, the prediction probability of the next token x_n by an autoregressive language model θ can be formalized as:

$$\hat{p}(x_n) = p_\theta(x_n \mid x_{<n}). \tag{1}$$

For the token x_n, we measure its uncertainty using the entropy of the predictive distribution:

$$H(x_n) = -\sum_{x_n \in \mathcal{V}} \hat{p}(x_n) \log \hat{p}(x_n), \tag{2}$$

where $\mathcal{V}$ denotes the vocabulary space. A higher entropy indicates that the model is less confident when predicting the next token, implying that the corresponding context contains richer information.

3.2 INFLUENCE ESTIMATION OF COMPRESSED KV CACHE

We conduct influence estimation with Llama-3.1-8B-Instruct (Grattafiori et al., 2024) to motivate our approach. We define the *Forward Influence* of token x_i in KV cache on a future context chunk $\{x_{l_c}, \cdots, x_{r_c}\}$, as the average divergence between the original prediction distribution and the prediction distribution obtained after removing x_i from the KV cache:

$$I_{l_c:r_c}(x_i) = \frac{1}{r_c - l_c + 1} \sum_{n=l_c}^{r_c} \mathcal{D}_{\mathrm{KL}}\Big(p_\theta(x_n \mid x_{<n}) \Big\| p_\theta(x_n \mid x_{<n} \setminus \{x_i\})\Big), \tag{3}$$

where $p_\theta(x_n | x_{<n} \setminus \{x_i\})$ denotes the prediction probability of x_n after removing x_i from the KV cache. We use Kullback–Leibler (KL) divergence to quantify the difference between two predictive distributions. For simplicity, all layers share the same token choice for compression when estimating forward influence.

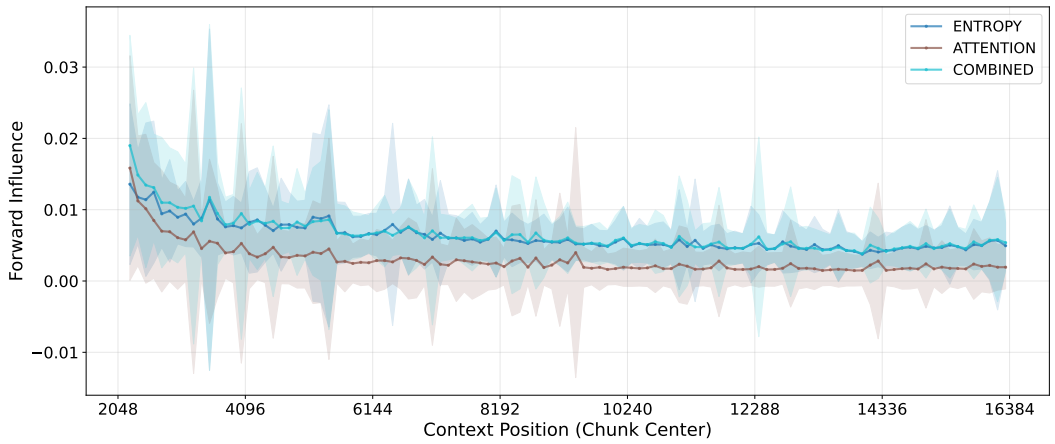


Figure 2: Forward influence of top-1% tokens selected by different scoring strategies over long generation horizons on 100 documents from Arxiv-Summarization. The first 2048 tokens are compressed using different token importance scores, and the influence is measured over future chunks of 128 tokens. The combined score achieves a better balance between short-range and long-range influence.

Based on this metric, we analyze the long-range contribution of tokens selected by different importance criteria, including attention scores, entropy, and their combinations. Specifically, we first rank tokens according to their averaged attention weights from a recent observation window $\{x_{l_o}, \cdots, x_{r_o}\}$, following prior KV cache compression methods:

$$A_i = \frac{1}{r_o - l_o + 1} \sum_{t=l_o}^{r_o} \text{Attn}(q_t, k_i), \tag{4}$$

where $\text{Attn}(q_t, k_i)$ is the attention weight from the token in the observation window to the token x_i and extracted from the last layer.

We then compare them with tokens selected by entropy-based criteria. To combine attention weights and entropy, we use softmax to normalize the scale of entropy in the sequence dimension and add it to the attention score:

$$S_i = A_i + \text{Softmax}(\boldsymbol{H})_i. \tag{5}$$

We use these three scoring strategies to compress the first 2048 tokens in a document and estimate their forward influence over a short future horizon and a long future horizon. Figure 1 demonstrates that the combination score balances short-range influence and long-range influence.

Forward influence along the long sequence is presented in Figure 2. It reveals a clear distinction between attention-based and entropy-based importance estimation. Tokens with high attention scores mainly influence nearby future contexts, and their influence decays rapidly as the generation distance increases. In contrast, tokens with high entropy exhibit substantially stronger influence on distant future contexts, suggesting that entropy better captures information relevant to long-range reasoning and generation. By combining both signals, we can pick out tokens that are important for recent contexts as well as distant future contexts. Example visualizations are provided in Appendix F to further illustrate the scoring difference between attention and entropy.

3.3 Information-Aware KV Compression

We propose InfoKV to incorporate information signals into KV cache compression. As illustrated in Figure 3, InfoKV captures informativeness from two complementary perspectives: information variation along the sequence dimension and representation evolution across layers. We combine these signals to derive an entropy score for each token at each layer, which is then integrated with attention scores to determine the final token importance.

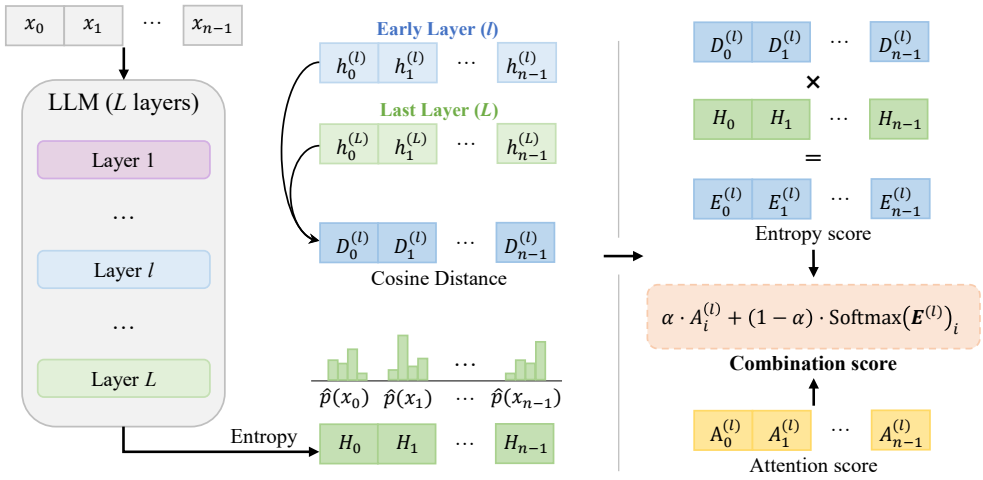


Figure 3: The overview of how to compute the importance score for KV cache compression in each layer. InfoKV combines predictive entropy, layer-wise representation evolution, and attention scores for token selection.

To quantify token importance for each layer, we measure how much the hidden representation of a token evolves from an intermediate layer to the final layer. Specifically, for token x_i, we compute the cosine distance between the hidden states from the early layer l and the final layer L:

$$D_i^{(l)} = 1 - \cos\left(h_i^{(l)}, h_i^{(L)}\right), \tag{6}$$

where $h_i^{(l)}$ and $h_i^{(L)}$ denote the hidden representations of token x_i at layer l and the final layer L, respectively.

Due to the residual connections within Transformer architectures (Vaswani et al., 2017), hidden representations evolve progressively across layers. A small distance $D_i^{(l)}$ indicates that the representation at layer l is already close to its final form, suggesting that the token's representation has undergone relatively little subsequent refinement. In contrast, tokens exhibiting larger representation shifts across layers tend to carry more unresolved semantic information and remain influential during subsequent generation. We therefore use the representation distance as a measure of the amount of information associated with a token that remains to be refined at layer l.

We combine the representation distance with the entropy computed from the final layer to estimate the entropy score for each token to capture token informativeness from both the hidden-state and output-distribution perspectives.

$$E_i^{(l)} = (D_i^{(l)} + \tau) \cdot H_i. \tag{7}$$

A bias τ will be added to $D_i^{(l)}$ so that the entropy score of the final layer will not be 0. Integrating token-level informativeness with layer-wise representation evolution, the entropy score jointly captures predictive uncertainty and the degree of representation transformation throughout layers.

Since the predictive probability mass of LLMs is typically concentrated on a small subset of highly probable tokens, these tokens dominate the model's decision-making process. Consequently, uncertainty estimated over the entire vocabulary can be heavily affected by numerous low-probability tokens that contribute little to generation behavior. Following prior work (Fu & Luo, 2026), we employ *Top-k Restricted Entropy* by using only the top-k most probable tokens in the predictive distribution, which provides a more stable and informative estimation of uncertainty.

For each layer, we compute the token importance score by combining the attention score and the entropy score:

$$S_i^{(l)} = \alpha \cdot A_i^{(l)} + (1 - \alpha) \cdot \text{Softmax}(\boldsymbol{E}^{(l)})_i, \tag{8}$$

Table 1: Accuracy (%) comparison on LongReason across different cache rates and prefill lengths (16K – 64K). We mark the best scores in bold and underline the second-best scores in the table.

Rate	Method	16K		32K		64K		Ave.	
		w. CoT	w/o. CoT	w. CoT	w/o. CoT	w. CoT	w/o. CoT	w. CoT	w/o. CoT
Llama-3.1-8B-Instruct									
100%	Full	55.67	52.08	53.90	49.75	53.02	48.61	54.20	50.15
40%	HeadKV	52.64	47.73	49.87	45.97	46.35	**46.35**	49.62	46.68
	AdaKV	54.03	48.11	51.01	46.60	47.36	46.10	50.80	46.94
	SnapKV	53.15	50.13	51.13	45.72	48.99	45.59	51.09	47.15
	PyramidKV	53.67	47.36	51.01	46.35	47.61	45.04	50.76	46.25
	FastKV	51.89	48.74	47.73	46.47	47.10	45.72	48.91	46.98
	Expected	54.16	48.36	50.50	46.35	48.74	45.72	51.13	46.81
	InfoKV	**55.32**	**51.80**	**52.39**	**48.61**	**49.87**	46.22	**52.53**	**48.88**
20%	HeadKV	49.37	45.72	47.23	45.09	45.47	44.96	47.36	45.26
	AdaKV	50.38	46.22	48.11	44.96	45.84	45.59	48.11	45.59
	SnapKV	**52.39**	47.23	47.98	45.59	47.73	44.28	49.37	45.70
	PyramidKV	50.88	**49.12**	48.36	44.84	46.47	43.53	48.57	45.83
	FastKV	50.38	48.24	47.98	**46.47**	46.85	45.97	48.40	**46.89**
	Expected	50.88	45.72	47.74	45.97	47.48	45.34	48.70	45.68
	InfoKV	51.77	48.74	**49.50**	46.22	**48.11**	**45.72**	**49.79**	**46.89**
Llama-3.2-3B-Instruct									
100%	Full	48.23	45.59	46.47	44.96	44.96	42.70	46.55	44.42
40%	SnapKV	45.34	42.95	42.81	**42.57**	40.93	39.42	43.03	41.65
	PyramidKV	44.71	**43.20**	43.07	42.44	40.81	41.06	42.86	**42.23**
	Expected	45.09	42.82	43.07	42.07	41.18	39.29	43.11	41.39
	InfoKV	**46.47**	42.44	**43.82**	41.69	**41.56**	**41.31**	**43.95**	41.81
20%	SnapKV	43.19	**41.06**	40.30	39.92	38.53	37.28	40.67	39.42
	PyramidKV	43.42	40.18	40.42	39.67	38.41	**38.28**	40.75	39.38
	Expected	42.44	40.55	40.93	40.05	38.79	37.66	40.72	39.42
	InfoKV	**43.83**	40.55	**41.69**	**40.43**	**39.17**	38.04	**41.56**	**39.67**

where $A_i^{(l)}$ denotes the attention score of token x_i at layer l, and $\mathbf{E}^{(l)}$ represents the entropy scores of all tokens in layer l. Given the importance scores $S^{(l)}$, we retain the top-ranked tokens in KV cache for each layer.

4 Experiments

4.1 Setup

We assess InfoKV on both long prefilling and decoding scenarios. For long prefilling, we evaluate Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct (Grattafiori et al., 2024) on the long-context reasoning benchmark LongReason (Ling et al., 2025). Models will process the entire input prompt in parallel and compress KV cache of the prompt for the following decoding stage.

As for long decoding, we employ InfoKV on reasoning models DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025). Models are evaluated on IFEval (Zhou et al., 2023), American Invitational Mathematics Examination (AIME) 2024 (Mathematical Association of America, 2025), and LiveCodeBench (Jain et al., 2025). Models will compress KV cache for the generated tokens periodically during the decoding phase.

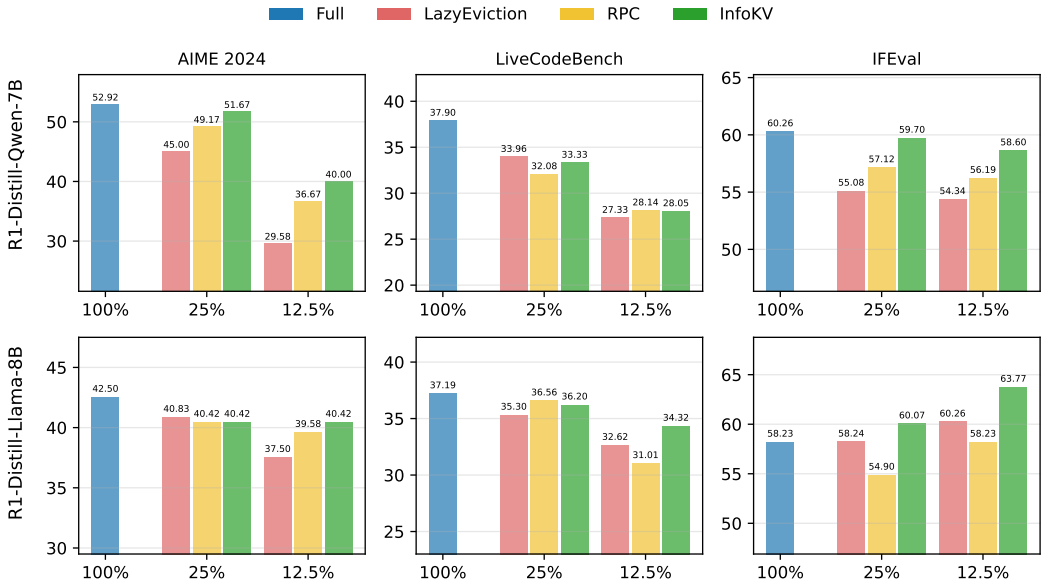


Figure 4: Performance of different KV cache compression methods on long decoding benchmarks.

4.2 Long Prefilling

We evaluate InfoKV on the long-context reasoning benchmark LongReason, which expands original reasoning tasks into long-context inputs containing extensive supporting evidence and distractor information. Thereby, it stresses the ability of KV cache compression methods to preserve reasoning-critical information under limited KV cache budgets. We compare InfoKV against representative attention-based methods: SnapKV (Li et al., 2024), PyramidKV (Cai. et al., 2024b), FastKV (Jo et al., 2026), and Expected Attention (Devoto et al., 2025); as well as two head-wise KV cache compression methods, HeadKV (Fu et al., 2025) and AdaKV (Feng et al., 2026), which implement adaptive budget allocation using attention scores for compression.

Experiments are conducted on Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct under different cache retaining ratios. Models are evaluated in chain-of-thought (w. CoT) and direct-answer (w/o. CoT) settings. For fair comparison, all baselines are implemented following the official implementations and share the same evaluation configurations. More details can be referred to in Appendix A.1.

Experimental results across multiple context lengths are reported in Table 1. Overall, InfoKV obtains highly competitive performance across most settings, demonstrating the effectiveness of incorporating entropy-aware information signals into KV cache compression. InfoKV consistently outperforms attention-based baselines under both 40% and 20% cache budgets on Llama-3.1-8B-Instruct. The advantage becomes more evident as the sequence length increases. It suggests that entropy-aware token selection can better retain globally informative tokens that remain useful throughout long reasoning trajectories, whereas recent-attention heuristics tend to emphasize short-range dependencies and may discard information important for future reasoning steps. Besides, the gains are particularly pronounced under CoT evaluation, also indicating that incorporating information-theoretic signals is especially beneficial for long-context reasoning. More evaluation results on a general benchmark LongBench (Bai et al., 2024) are presented in Appendix B.

4.3 Long Decoding

For long decoding, we consider AIME 2024 for mathematical reasoning, LiveCodeBench for coding evaluation, and IFEval for instruction following. Models are required to generate reasoning steps and derive final answers with a maximum output length of 32768 tokens. We compare InfoKV against Full KV Cache, RPC (Song et al., 2025), and LazyEviction (Zhang et al., 2026). Following RPC, which periodically compresses KV cache in the decoding stage, we sample 8 completions

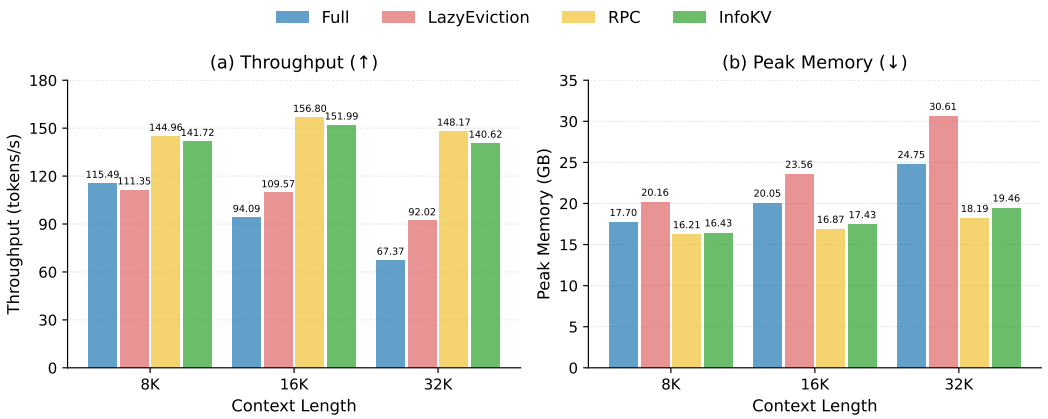


Figure 5: Throughput (tokens/s) and peak memory usage (GB) of different methods.

per instance for AIME 2024, 4 completions for LiveCodeBench, and 1 completion for IFEval to compute pass@1 scores. KV cache compression is triggered every 1024 tokens during decoding. Settings of hyperparameters are summarized in Appendix A.2.

Performance on DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B is shown in Figure 4. InfoKV achieves better performance on the three task categories compared to RPC and LazyEviction. Notably, on IFEval, InfoKV with retaining ratios of 25% and 12.5% even surpasses the full cache performance of R1-Distill-Llama-8B. To verify that this improvement is not caused by sampling variance, we repeat the evaluation using three random seeds and report statistics as in Appendix C. It suggests that long reasoning trajectories contain substantial redundancy, and retaining all historical tokens may introduce distracting or less informative contexts during generation. By selectively compressing tokens associated with high predictive certainty and lower information content, InfoKV enables the model to focus more effectively on informative reasoning contexts and improves generation quality as well as efficiency. More evaluation results on the larger DeepSeek-R1-Distill-Qwen-14B are provided in Appendix D.

We also investigate an adaptive compression strategy that dynamically allocates KV cache budgets according to layer-wise entropy statistics in Appendix E. Since the adaptive strategy leads to greater performance degradation, we adopt the uniform strategy as the default setting throughout the paper for better robustness and simplicity.

5 Analysis

In this section, we analyze the efficiency of InfoKV in terms of throughput and peak memory usage. Furthermore, we conduct ablation studies regarding the choice of τ, top-k restricted entropy, and the balance between entropy and attention in InfoKV.

5.1 Efficiency Analysis

We measure the generation throughput and peak memory usage to evaluate the efficiency of InfoKV as in Figure 5. Experiments are conducted on DeepSeek-R1-Distill-Qwen-7B using a single NVIDIA H100 SXM GPU with a batch size of 4. Each input prompt contains 128 tokens, while the generated sequence lengths are 8K, 16K, and 32K tokens. All compression methods use the same configuration, including a compression interval of 1024 tokens and a retaining ratio of 25%.

Compared with RPC, InfoKV introduces a modest computational overhead for entropy and representation distance computation, resulting in slightly lower throughput and higher memory consumption. Nevertheless, InfoKV remains substantially more efficient than the full KV cache and LazyEviction, achieving a favorable trade-off between compression effectiveness and inference efficiency.

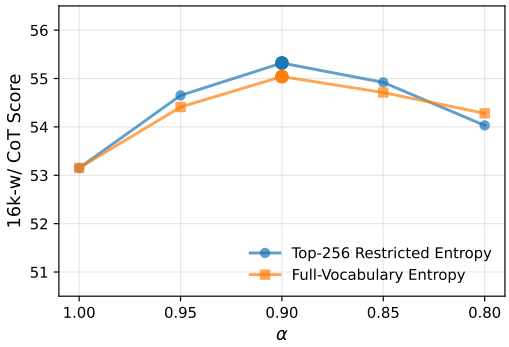


Figure 6: Performance of InfoKV with top-256 restricted entropy and full-vocabulary entropy across a range of α.

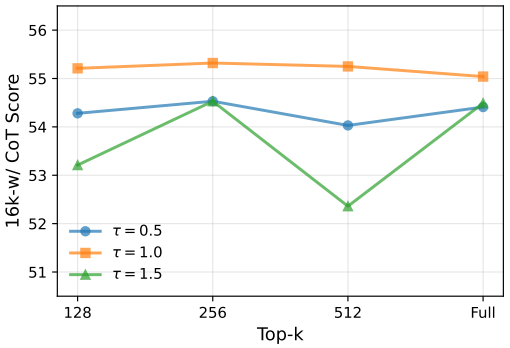


Figure 7: Performance of InfoKV with different τ and top-k. "Full" means that it computes entropy on the original full vocabulary.

5.2 Ablation Studies

Choice of τ. We study the effect of the bias term τ on Llama-3.1-8B-Instruct under the 40% retaining ratio. Results on the CoT setting of LongReason are presented in Figure 7. As τ increases, the contribution of layer-wise representation distance is gradually reduced, making the entropy score for each layer rely more on predictive uncertainty from the final layer. We observe that $\tau = 1$ achieves the best overall performance while also providing more stable behavior across different settings. Therefore, we adopt $\tau = 1$ as the default configuration in experiments.

Top-k Restricted Entropy. Moreover, we investigate the effect of top-k restricted entropy by varying the value of k in Figure 7. Overall, restricting entropy computation to the most probable tokens consistently improves performance compared with computing entropy over the entire vocabulary. In particular, $k = 256$ achieves the best performance under different values of τ. This result suggests that low-probability tokens contribute limited useful information to uncertainty estimation and may introduce noise into entropy-based token importance measurement.

Balance between Entropy and Attention. We further study the balance between entropy-based uncertainty and attention-based relevance by varying the coefficient α in Equation 8. Results are shown in Figure 6. When $\alpha = 1$, the importance score degenerates to pure attention-based selection, which leads to inferior performance compared with using a moderate combination of entropy and attention. This observation indicates that attention alone is insufficient to fully characterize the long-range utility of tokens in KV cache.

Introducing entropy consistently improves performance, and $\alpha = 0.9$ obtains the best scores under both the full-vocabulary and top-256 restricted entropy settings. However, further reducing α leads to performance degradation, suggesting the importance of short-range dependency from the attention perspective. Therefore, the results demonstrate that entropy and attention provide complementary signals, and a moderate integration achieves the best balance for long reasoning.

6 Conclusion

In this paper, we revisit KV cache compression from a forward-looking perspective and introduce forward influence to measure the effect of compressed tokens on future predictive distributions. Our analysis reveals that attention weights mainly capture short-range dependencies, whereas tokens associated with high predictive uncertainty exhibit substantially stronger influence on distant future contexts. Motivated by this observation, we propose InfoKV to combine predictive entropy, layer-wise representation evolution, and attention scores for token selection during long-context reasoning. Extensive experiments on long prefilling and long decoding benchmarks demonstrate that our information-aware KV cache compression framework consistently achieves better performance than existing attention-based compression methods across multiple models and reasoning tasks.

AI USE STATEMENT

In this work, we used generative AI tools to polish grammar, improve clarity, and refine the academic style of the manuscript. All research ideas, methods, experiments, and analyses were independently developed and conducted by the authors. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work pertains to key-value compression of large language models for long reasoning scenarios. In this work, we use only publicly available data and artifacts. There are no ethical issues in our paper, including its motivation and experiments.

REPRODUCIBILITY STATEMENT

We have provided detailed implementations of our method throughout the paper. Our method is comprehensively elaborated in Section 3. Detailed settings of our experiments and analysis are given in Section 4, 5, and Appendix A.1, A.2, B, C, and D. Our code will be released for reproducibility.

REFERENCES

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, 2024. URL https://arxiv.org/abs/2308.14508.

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, 2024b. URL https://arxiv.org/abs/2406.02069.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2097. URL https://aclanthology.org/N18-2097.

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

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. Advances in Neural Information Processing Systems, 38:113152–113188, 2026.

Renyu Fu and Guibo Luo. Selar: Selective latent reasoning in large language models, 2026. URL https://arxiv.org/abs/2604.08299.

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. In International Conference on Learning Representations, volume 2025, pp. 99269–99290, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,

Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,

Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature, 645(8081):633–638, 2025.

Naman Jain, Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), International Conference on Learning Representations, volume 2025, pp. 58791–58831, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/94074dd5a072d28ff75a76dabed43767-Paper-Conference.pdf.

Dongwon Jo, Jiwon Song, Yulhwa Kim, and Jae-Joon Kim. Fastkv: Decoupling of context reduction and kv cache compression for prefill-decoding acceleration. In Findings of the Association for Computational Linguistics: ACL 2026, pp. 32167–32186, 2026.

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. Sh2: Self-highlighted hesitation helps you decode more truthfully, 2024. URL https://arxiv.org/abs/2401.05930.

Jushi Kai, Yixuan Wang, Boyi Zeng, Haoli Bai, Bo Jiang, Ziwei He, and Zhouhan Lin. Freqkv: Key-value compression in frequency domain for context window extension, 2026. URL https://arxiv.org/abs/2505.00570.

Adrian Ł ańcucki, Konrad Staniszewski, Piotr Nawrot, and Edoardo Maria Ponti. Inference-time hyper-scaling with kv cache compression. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), Advances in Neural Information Processing Systems, volume 38, pp. 9365–9397. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0d781fa5f639bf2caf728a68e9678362-Paper-Conference.pdf.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. Compressing context to enhance inference efficiency of large language models, 2023. URL https://arxiv.org/abs/2310.06201.

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

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion. arXiv preprint arXiv:2501.15089, 2025.

Mathematical Association of America. American invitational mathematics examination. https://maa.org/maa-invitational-competitions/, 2025. Accessed: 2025-05-15.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Bohan Zhang, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wenting Zhan, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

Jiwon Song, Dongwon Jo, Yulhwa Kim, and Jae-Joon Kim. Reasoning path compression: Compressing generation trajectories for efficient llm reasoning, 2025. URL https://arxiv.org/abs/2505.13866.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

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, 2025. URL https://arxiv.org/abs/2406.13035.

Zheng Wang, Boxiao Jin, Zhongzhi Yu, and Minjia Zhang. Model tells you where to merge: Adaptive kv cache merging for llms on long-context tasks, 2024. URL https://arxiv.org/abs/2407.08454.

Haoyue Zhang, Hualei Zhang, Xiaosong Ma, Jie Zhang, and Song Guo. Lazyeviction: Lagged kv eviction with attention pattern observation for efficient long reasoning. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 36335–36352, 2026.

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. CaM: Cache merging for memory-efficient LLMs inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 58840–58850. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24n.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Experiment Details

A.1 Long Prefill

For a fair comparison, all common configurations adopt the official implementations of previous KV cache compression methods, and all public hyperparameter settings are kept consistent. Specifically, the observation window size is set to 64, the pooling function adopts average pooling (`avgpool`), and the pooling kernel size is set to 9.

For InfoKV, we set the bias term τ added to $D_i^{(l)}$ to 1.0 and compute top-k restricted entropy using the top 256 predicted tokens. To balance the weights of the entropy score and the attention score α in Eq. (8) is set to 0.9. We adopt fixed prompting templates for both direct-answer and Chain-of-Thought (CoT) reasoning settings in LongReason. Same prompts are used for all compared methods to ensure that the performance differences mainly arise from KV cache compression strategies rather than prompting variations.

A.2 Long Decoding

Table 2: Hyper-parameter settings on three long decoding tasks.

Model	Parameter	IFEval	AIME 2024	LiveCodeBench
R1-Distill-Qwen-7B	τ	1.5	0.5	0.5
	k	256	128	512
	α	0.9	0.95	0.95
R1-Distill-Llama-8B	τ	1	0.5	1
	k	256	256	256
	α	0.9	0.95	0.95

The settings of hyperparameters on the three long decoding benchmarks are summarized in Table 2. The weight of attention scores α is set to 0.95 for AIME 2024 and LiveCodeBench, whose samples contain a large number of mathematical notations and symbolic reasoning steps. In such cases, attention scores provide more reliable structural signals for preserving locally important contexts. For IFEval, we adopt a slightly smaller value $\alpha = 0.9$ to introduce stronger entropy guidance, which better captures informative tokens relevant to instruction following and long-range code generation.

Table 3: Performance on LongBench.

Method	NtrvQA	Qasper	MF-en	MF-zh	HotpotQA	2WikiMQA	Musique	PR-en	PR-zh	P-cnt	Avg
Full	29.69	45.67	55.21	62.12	55.99	44.97	31.34	99.50	93.00	9.41	52.69
AdaKV	26.77	43.89	**55.05**	61.08	55.57	43.66	30.80	99.50	**92.50**	6.81	51.56
SnapKV	26.71	**45.17**	54.23	61.73	**55.80**	43.74	29.50	99.50	**92.50**	6.60	51.55
PyramidKV	26.93	43.26	55.03	61.56	55.26	44.08	30.38	99.50	92.00	6.76	51.48
Expected	23.74	38.54	45.39	**61.85**	54.57	43.14	28.42	**97.00**	68.50	**7.08**	46.82
InfoKV	**30.27**	43.58	54.86	61.40	55.79	**44.52**	**31.00**	99.50	92.29	6.76	**52.00**

B Evaluation on LongBench

InfoKV is primarily motivated by long reasoning, where preserving tokens with long-range influence is particularly important yet not well captured by attention alone. We introduce entropy as a complementary signal because it better reflects token informativeness for future reasoning.

To further evaluate the generality of InfoKV, we additionally conduct experiments on LongBench, a widely used benchmark covering diverse long-context tasks. Results on Llama-3.1-8B-Instruct with a retaining budget of 1024 are shown in Table 3. Although InfoKV is motivated by long reasoning, it achieves comparable performance across these general long context tasks. These results suggest that InfoKV generalizes beyond reasoning-oriented scenarios.

C Significance Test on IFEval

We also noticed that InfoKV surpasses full cache of R1-Distill-Llama-8B on IFEval. To verify that this observation is not caused by sampling variance, we repeat the evaluation using three random seeds and report statistics in Table 4.

The improvements remain consistent across independent runs. It suggests that long reasoning trajectories contain considerable redundancy, and retaining all historical tokens is not always beneficial. By preferentially preserving informative tokens and compressing tokens with high predictive certainty, InfoKV can improve both memory efficiency and reasoning quality.

D Scaling to a Larger Model

To further evaluate the scalability of InfoKV, we additionally conduct experiments on DeepSeek-R1-Distill-Qwen-14B. Evaluation results with a retaining ratio of 12.5% are reported in Table 5. The improvements are preserved on a larger reasoning model.

E Adaptive Compression

Although the uniform strategy is effective, we observe that different Transformer layers exhibit substantially different entropy distributions. Early and middle layers generally contain richer uncertainty and broader contextual information, whereas higher layers become increasingly confident and redundant.

Table 4: Evaluation results on IFEval with three random seeds.

Rate	Method	Run 1	Run 2	Run 3	Avg
100%	Full	58.23	57.74	58.60	58.19±0.35
25%	RPC	54.90	56.19	57.86	56.32±1.21
	InfoKV	**60.07**	**61.55**	**61.00**	**60.87±0.61**
12.5%	RPC	58.23	58.04	57.86	58.04±0.15
	InfoKV	**63.77**	**60.63**	**60.07**	**61.49±1.63**

Table 5: Evaluation results on DeepSeek-R1-Distill-Qwen-14B.

Method	AIME 2024	LiveCodeBench	IFEval
Full	67.08	55.36	73.38
RPC	67.08	49.46	72.83
InfoKV	**68.75**	**49.73**	**75.23**

Table 6: Comparison of uniform budget and adaptive budget on IFEval.

Model	Rate	Uniform	Adaptive
R1-Distill-Qwen-7B	25%	**59.70**	55.82
	12.5%	**58.60**	57.67
R1-Distill-Llama-8B	25%	60.07	**61.18**
	12.5%	**63.77**	60.81

Motivated by this observation, we further introduce an adaptive compression strategy that dynamically allocates KV cache budgets according to layer-wise entropy statistics. Specifically, we compute the accumulated entropy score for each layer:

$$\bar{E}^{(l)} = \sum_{i=0}^{n-1} E_i^{(l)}. \tag{9}$$

The retaining budget for layer l is then allocated proportionally:

$$k_l = \frac{\bar{E}^{(l)}}{\sum_{m=1}^{L} \bar{E}^{(m)}} \cdot B, \tag{10}$$

where B denotes the total KV cache budget and L is the number of Transformer layers.

Layers with larger entropy scores receive larger KV budgets, enabling the model to preserve more informative contexts in uncertainty-rich layers while aggressively compressing more redundant layers.

Table 6 presents the results on IFEval. Under an overall retaining ratio of 25%, the adaptive strategy improves performance on R1-Distill-Llama-8B compared with the uniform setting. However, the gains are less consistent on R1-Distill-Qwen-7B, where adaptive allocation introduces larger performance degradation. We conjecture that excessively imbalanced layer-wise budgets may overcompress certain layers and harm the stability of long-range reasoning. Therefore, we adopt the uniform strategy as the default setting throughout the paper for better robustness and simplicity.

F Visualizations of Token Scoring

We provide visualizations of token scores from entropy and attention on two examples of reasoning tasks in Figures 8 and 9.

Prompt: ### Question
In modern society filled with choices and temptations, a study in Blackwood City found that self-control is crucial in daily life. Researchers concluded through experiments that our self-control comes from a single account used for various tasks. Therefore, it is wiser to handle one thing at a time. If you try to juggle multiple tasks simultaneously, you might manage temporarily, but as self-control depletes, the likelihood of making significant mistakes increases.
Which of the following is a necessary assumption for the above argument?
A. Your self-control is limited and is depleted with use.
B. There is a hidden connection between the various things you do throughout the day.
C. Self-control, like a muscle, can be strengthened with long-term exercise.
D. When self-control is depleted, things that usually annoy you become more irritating, and you are more likely to say things you will regret.

Please answer the above question!

Answer
Please analyze the question step by step, and analyze it only once. Do not analyze the question repeatedly. Provide the final answer on the last line using "The answer is" + option (represented by A, B, C, D, E).

Entropy (low → blue, high → red)



Attention (low → blue, high → red)

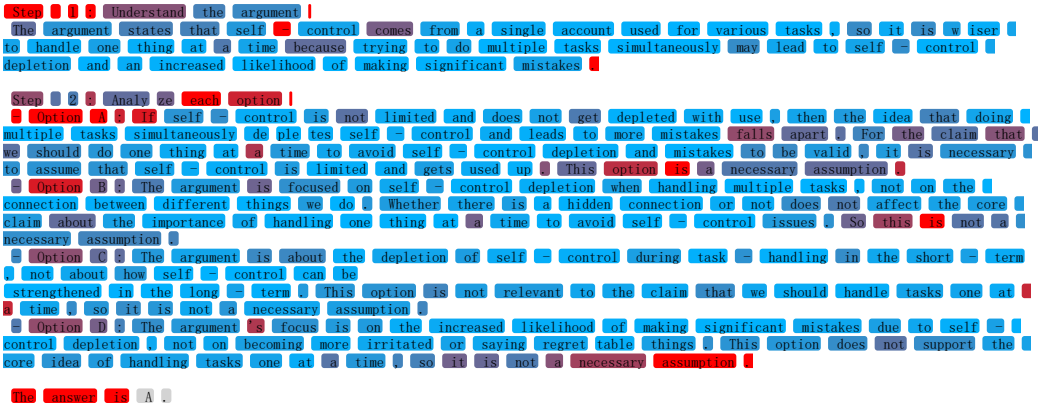


Figure 8: Visualizations of token scores from entropy and attention during the decoding stage of Llama-3.1-8B-Instruct on case 1. The example is a reasoning task that explicitly analyzes the question step by step.

Attention scores are obtained from "The answer is" in the last sequence, which is to derive the final answer. As a result, the word "Option" and the following "A", "B", "C" and "D" are all assigned high attention weights. It shows that attention tends to retrieve tokens that are closely relevant to the current query. In contrast, entropy measures the informativeness of the token itself, but does not depend on the query. It is observed to capture more content words that carry important information like "argument", "mistakes", "importance" and "depletion". Therefore, more information could be preserved for the future decoding process.

Prompt: ### Question
At the end of the 20th century, essay writing experienced unprecedented prosperity in the Western world. The most popular and active forms of essays showed a polarized developmental trend: on one hand, there was a movement towards rationalization and largeness, known as "macro essays," such as cultural essays and thought essays; on the other hand, there was a trend towards emotional, gentle, and detailed expressions, known as "micro essays," which focus on capturing the delicate feelings of urban life and the fleeting experiences of individuals. "Micro essays" bring the freedom, spontaneity, and ease of essay writing to its peak, but also, to some extent, weaken the profound exploration of life and the aesthetic nature of literature, exhibiting what is called the "consumption characteristics of essays," providing cheap comfort to people rushing in the fast-paced, utilitarian, and indifferent modern life. In the passage, "consumption characteristics of essays" refers to:
A. Focus on capturing the delicate feelings of urban life
B. Reflecting a strong humanistic spirit and aesthetic awareness
C. The emotional and fast-food nature of micro essays
D. Fast-paced, utilitarian, and indifferent to human feelings

Please answer the above question!

Answer
Please analyze the question step by step, and analyze it only once. Do not analyze the question repeatedly. Provide the final answer on the last line using "The answer is" + option (represented by A, B, C, D, E).

Entropy (low → blue, high → red)



Attention (low → blue, high → red)

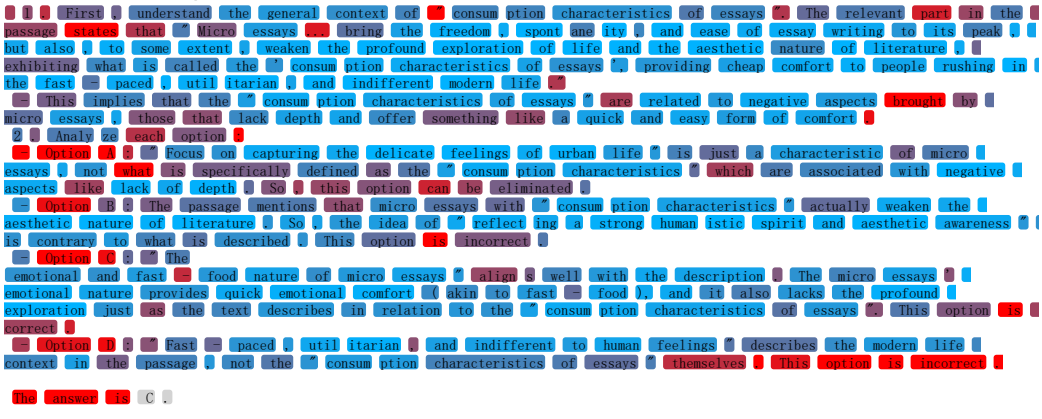


Figure 9: Visualizations of token scores from entropy and attention during the decoding stage of Llama-3.1-8B-Instruct on case 2.