

A KV CACHE COMPRESSION ILLUSTRATION

Figures 11 to 14 show the illustration of other KV cache compression methods.

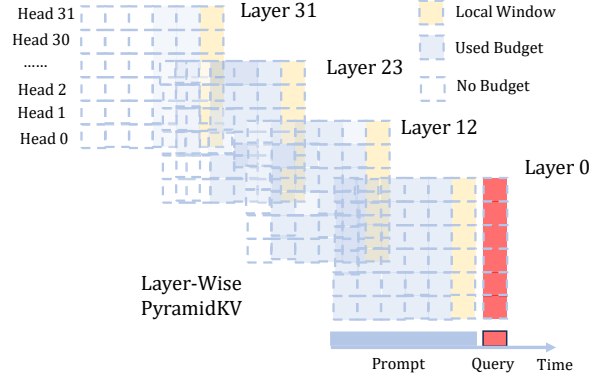


Figure 11: PyramidKV compresses the KV cache by allocating more cache in lower layers and less in higher ones.

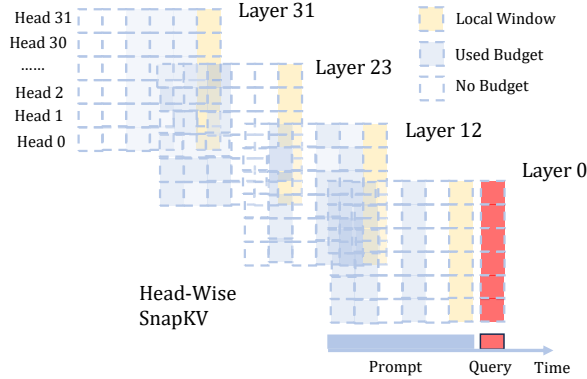


Figure 12: SnapKV compresses the KV cache by selecting grouped important positions for each attention head.

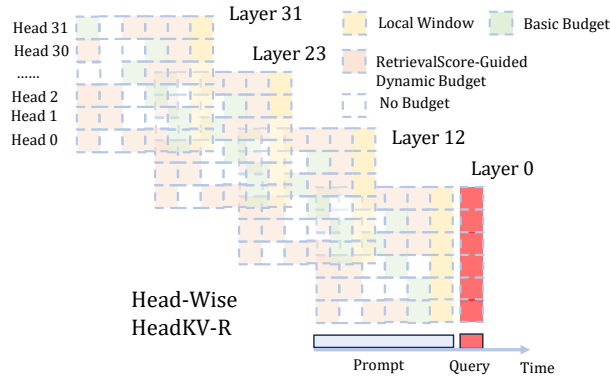


Figure 13: HeadKV-R compresses the KV cache by focusing on the retrieval-and-paste mechanism for each head.

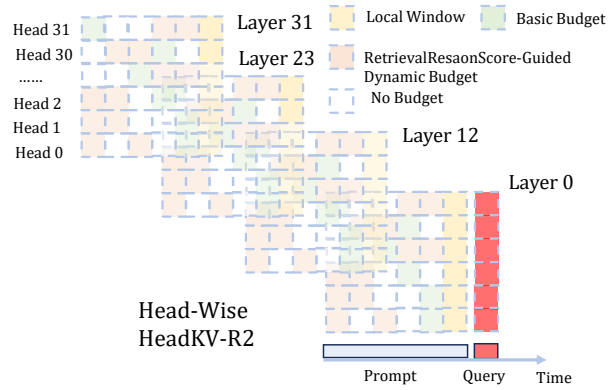


Figure 14: HeadKV-R2 compresses the KV cache by considering the contextual and reasoning skills for each head.

B NEEDLE EXAMPLE

1. **Question:** “What are good ways to spend time in campus?”
Needle: “The good ways to spend time in campus include relax and do nothing.”
Answer: “Relax and do nothing.”
2. **Question:** “What habits are beneficial for health during Ph.D. career?”
Needle: “The beneficial habits during Ph.D. career contain exercise and healthy diet.”
Answer: “Exercise and healthy diet.”
3. **Question:** “Which year did Tom start high school?”
Needle: “Tom was born in 2005. Tom started high school fifteen years after he was born.”
Answer: “Tom started high school in 2020.”

C KV CACHE BUDGET ALLOCATION

Algorithms 1 and 2 present the REAL method for KV cache compression by selecting the top- k entries. The procedure first checks whether the KV length exceeds the budget. If it does not, the original K and V are returned unchanged. If it does, the full query window is retained, and the top- k most relevant historical tokens within the budget are selected to form the compressed KV.

Algorithm 1 Procedure for KV budget allocation.

Input: Total budget b_c , allocation ratio β , head (i, j) , layer count L , head count in a layer H , INFsc of j_{th} head (i, j) in the i_{th} layer $H_{j,inf}$

Output: Capacity $C_{i,j}$

- 1: Basic budget $b_{base} = b \left(1 - \frac{1}{\beta}\right)$
 - 2: Global dynamic budget pool $B_{total} = \frac{b}{\beta} L H$
 - 3: i_{th} layer's total INFScore $L_{i,\alpha} = \sum_j H_{j,inf}$
 - 4: $L_s \leftarrow \sum_{i=0}^{L-1} L_{i,\alpha}$,
 - 5: Relative Importance of head and layer $S_{j,h} \leftarrow \frac{H_{j,inf}}{L_s}, L_{i,h} = \frac{L_{i,\alpha}}{L_s}$
 - 6: Dynamic allocation $b_{i,j}^{dyn} = B_{total} \cdot (\gamma + L_{i,h}) \cdot S_{j,h}$
 - 7: Total KV cache budget $b_{i,j} = b_{base} + b_{i,j}^{dyn}$
 - 8: **return** $C_{i,j} = \lfloor \max(0, b_{i,j}) + 0.5 \rfloor$
-

Algorithm 2 Procedure for KV entry selection.

Input: key states K , query states Q , value states V , budget C

Output: Final_KV

- 1: **if** current_KV_length $\leq B$ **then**
 - 2: **return** K, V
 - 3: **end if**
 - 4: $W \leftarrow$ Last window-size tokens from KV
 - 5: $Q_H \leftarrow$ Remaining history from KV
 - 6: Attention relevance $A = \text{Softmax}\left(Q_{window} H^T / \sqrt{d_k}\right)$
 - 7: Budget for history $C_H = C - \text{window_size}$
 - 8: TopK_indices = TopK($A, k = B_H$)
 - 9: Relevant history $H' = \text{Gather}(H, \text{TopK_indices})$
 - 10: Final_KV = Concat(H', W)
 - 11: **return** Final_KV
-

D DATASET DETAILS

Table 5 shows the details of sixteen QA datasets from LongBench (Bai et al., 2024a), four QA datasets from LooGLE (Li et al., 2024a), and eleven extended-length datasets from LongBench v2 (Bai et al., 2024b).

Table 5: Dataset Details.

Label	Task	Avg Len
NrtvQA	NarrativeQA	18,409
Qasper	Qasper	3,619
MF-en	MultiFieldQA-EN	4,559
HotpotQA	HotpotQA	9,151
2WikiMultiHopQA	2WikiMultiHopQA	4,887
Musique	Musique	11,214
QMSum	QMSum	10614
MultiNews	MultiNews	2113
TREC	TREC	5177
TriviaQA	TriviaQA	8209
SAMSum	SAMSum	6258
PCount	PassageCount	11141
Pre	PassageRetrieval-en	9289
LCC	LCC	1235
RB-P	RepoBench-P	4206
GovReport	Government report	8734
Doc.QA	Comprehension & reasoning	15,498
Info.Retrieval	Multiple information retrieval	14,808
Timeline	Timeline reorder	15,425
Computation	Computation	17,001
Literary	Literary	72K
Legal	Legal	28K
Detective	Detective	70K
Detective	Academic	27K
Financial	Financial	49K
Govern	Government reports	20K
UserGuide	User guide QA	61K
Many-Shot	Many-Shot	71K
Dialogue History	Dialogue history QA	77K
Table	Table QA	42K
KnGraph	Knowledge graph reasoning	52K

E COMPREHENSIVE EXPERIMENTAL RESULTS ON LONGBENCH

Table 6 shows that the REAL method achieves substantial improvements over the baselines across different settings.

Table 6: Results for varying KV cache sizes (64, 128, 256, 512, 1024) on Llama-3-8B-Instruct and Mistral-7B-Instruct, with baseline results reported by Fu et al. (2025).

Method	Single-Doc QA			Multi-Doc QA			Avg.	Long Dependency QA				Avg.	β
	NartQA Qasper MF-en HotpotQA 2WikiMQA Musique							DocQA Info. Retrieval Timeline Computation					
Llama-3-8B-Instruct, KV Size = Full													
FullKV	25.56	32.07	39.71	43.57	35.28	21.18	32.90	8.73	11.21	0.67	7.43	7.01	-
Llama-3-8B-Instruct, KV Size = 64													
PyramidKV	21.17	13.66	29.34	34.86	23.46	15.88	23.06	8.27	9.31	0.63	6.86	6.27	-
SnapKV	20.51	12.80	31.69	37.02	25.91	17.02	24.16	8.84	9.43	0.66	6.18	6.28	-
HeadKV-R	22.67	23.54	37.51	37.45	29.76	19.01	28.32	8.80	10.51	0.58	6.68	6.64	2
HeadKV-R2	23.21	25.33	38.71	40.64	31.33	19.35	29.76	9.46	10.66	0.61	6.92	6.91	1.2
REAL	26.22	26.30	38.05	43.89	31.06	20.75	31.05	9.23	10.67	0.63	7.42	6.99	1.351
Llama-3-8B-Instruct, KV Size = 128													
PyramidKV	22.01	17.05	31.52	39.27	28.99	18.34	26.20	8.89	9.63	0.61	6.72	6.46	-
SnapKV	22.11	15.79	31.01	41.12	29.20	19.35	26.43	8.36	9.46	0.79	6.56	6.29	-
HeadKV-R	23.49	25.39	38.15	42.45	32.84	19.95	30.38	8.87	10.35	0.78	7.52	6.88	1.5
HeadKV-R2	21.80	29.19	41.89	43.73	35.01	20.40	32.00	9.60	11.13	0.67	7.22	7.16	1.01
REAL	25.47	29.95	38.02	44.67	34.28	20.66	32.18	9.20	11.32	0.62	7.83	7.24	1.351
Llama-3-8B-Instruct, KV Size = 256													
PyramidKV	23.94	20.27	36.27	42.51	31.44	19.99	29.07	8.66	10.61	0.53	6.98	6.70	-
SnapKV	23.38	20.18	37.65	42.80	33.23	20.01	29.54	9.04	10.59	0.53	7.53	6.92	-
HeadKV-R	23.83	29.04	39.90	42.36	33.58	20.57	31.54	9.05	11.15	0.52	7.22	6.99	1.1
HeadKV-R2	24.68	30.49	38.59	44.32	36.41	20.54	32.51	9.47	11.56	0.54	7.65	7.31	1.1
REAL	26.10	30.26	37.14	44.61	38.53	22.40	33.17	10.01	11.71	0.53	7.23	7.37	1.351
Llama-3-8B-Instruct, KV Size = 512													
PyramidKV	24.69	23.65	35.10	43.25	31.16	20.06	29.65	8.90	10.62	0.74	7.57	6.96	-
SnapKV	25.47	23.75	38.64	43.66	33.98	19.83	30.89	9.00	11.07	0.63	7.34	7.01	-
HeadKV-R	23.84	29.21	39.79	44.41	36.09	20.59	32.32	9.13	11.61	0.56	7.12	7.11	1.2
HeadKV-R2	24.75	29.75	38.03	44.43	36.45	21.67	32.51	9.34	11.26	0.56	7.54	7.18	1.1
REAL	26.13	31.58	38.94	44.23	36.76	22.82	33.41	9.32	11.97	0.55	7.34	7.30	1.351
Llama-3-8B-Instruct, KV Size = 1024													
PyramidKV	25.38	26.83	36.90	44.09	34.24	21.49	31.49	8.98	11.41	0.53	6.96	6.97	-
SnapKV	25.76	27.50	38.38	43.40	34.81	20.07	31.65	9.61	11.34	0.53	7.22	7.18	-
HeadKV-R	24.85	30.94	39.82	43.52	36.58	20.37	32.68	9.20	11.67	0.55	7.71	7.28	1.2
HeadKV-R2	24.66	30.82	39.56	43.97	36.47	22.24	32.95	9.02	11.51	0.47	7.85	7.21	1.2
REAL	25.57	32.37	39.50	44.72	36.87	22.45	33.58	9.68	11.48	0.49	7.12	7.19	1.351
Mistral-7B-Instruct, KV Size = Full													
FullKV	26.63	32.99	49.34	42.77	27.35	18.78	32.98	12.17	15.52	0.49	10.03	9.55	-
Mistral-7B-Instruct, KV Size = 64													
PyramidKV	20.91	19.61	38.05	32.18	22.87	15.26	24.81	10.64	11.69	0.56	9.06	7.99	-
SnapKV	19.95	18.63	38.16	31.24	21.39	13.81	23.86	10.41	11.49	0.46	9.38	7.94	-
HeadKV-R	24.23	25.22	46.02	38.82	26.05	17.41	29.63	10.94	13.14	0.63	9.11	8.46	1.5
HeadKV-R2	21.77	26.57	48.39	40.12	26.76	16.21	29.97	11.19	13.94	0.48	9.87	8.87	1.2
REAL	24.34	27.45	47.31	40.04	25.14	17.34	30.27	12.47	13.48	0.66	10.17	9.20	1.351
Mistral-7B-Instruct, KV Size = 128													
PyramidKV	21.76	21.98	43.72	32.76	22.73	15.59	26.42	10.64	11.90	0.47	8.69	7.93	-
SnapKV	21.47	21.95	45.24	33.88	21.83	15.53	26.65	10.86	12.24	0.57	8.81	8.12	-
HeadKV-R	23.97	29.60	48.40	39.66	26.31	18.13	31.01	11.43	13.04	0.53	10.26	8.82	1.5
HeadKV-R2	25.04	27.95	48.48	41.28	27.65	18.05	31.41	11.44	13.08	0.63	10.20	8.84	1.1
REAL	25.99	30.22	48.45	40.78	28.34	18.38	32.03	11.86	14.08	0.49	10.85	9.32	1.351
Mistral-7B-Instruct, KV Size = 256													
PyramidKV	21.42	25.36	47.94	38.75	25.82	15.30	29.10	11.57	12.35	0.56	9.51	8.50	-
SnapKV	22.26	24.94	48.30	36.76	25.16	14.93	28.72	11.07	12.39	0.53	9.13	8.28	-
HeadKV-R	24.98	29.31	49.01	41.36	27.16	17.34	31.53	11.94	13.30	0.63	10.95	9.21	2
HeadKV-R2	24.24	31.02	50.76	42.11	26.14	18.47	32.24	12.37	13.88	0.48	9.86	9.15	1.005
REAL	26.98	30.60	49.07	42.32	27.36	19.04	32.56	12.48	13.51	0.53	11.26	9.45	1.351
Mistral-7B-Instruct, KV Size = 512													
PyramidKV	23.07	28.97	48.37	39.54	25.63	16.59	30.36	11.34	13.32	0.65	10.81	9.03	-
SnapKV	24.18	28.87	48.74	38.84	25.48	15.04	30.19	11.96	13.47	0.52	10.50	9.11	-
HeadKV-R	24.97	30.94	49.45	42.25	26.34	18.54	32.08	12.09	13.88	0.62	10.94	9.38	1.01
HeadKV-R2	25.59	31.33	50.26	42.66	27.20	19.37	32.74	11.62	15.61	0.50	9.97	9.43	1.005
REAL	27.87	30.87	49.23	42.36	28.06	19.14	32.92	12.53	13.76	0.53	10.84	9.42	1.351
Mistral-7B-Instruct, KV Size = 1024													
PyramidKV	24.28	30.05	49.17	40.49	26.43	18.80	31.54	11.77	14.51	0.51	10.19	9.25	-
SnapKV	25.38	30.22	49.29	41.84	26.60	18.08	31.90	11.69	13.89	0.52	10.54	9.16	-
HeadKV-R	25.87	31.44	49.55	41.95	27.09	19.88	32.63	12.21	14.17	0.50	10.58	9.37	1.5
HeadKV-R2	25.64	32.54	50.49	41.80	27.88	18.89	32.87	11.94	14.93	0.50	10.49	9.47	1.01
REAL	26.77	31.99	48.98	42.61	28.04	19.35	32.96	12.82	14.82	0.50	10.33	9.62	1.351

F THE USE OF LLMs

We used LLMs to assist with grammar checking and correction. All ideas and technical content were entirely developed by the authors.