ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference

ByteDance | Seed

Carnegie Mellon University

Hanshi Sun, Li-Wen Chang, Wenlei Bao, Size Zheng, Ningxin Zheng, Xin Liu,
Harry Dong, Yuejie Chi, and Beidi Chen

ICML International Conference On Machine Learning

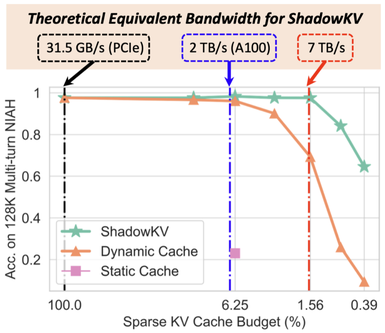
Code | ArXiv | Website

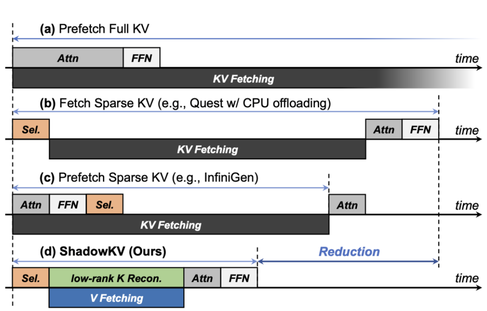
Overview

With the widespread deployment of **long-context LLMs**, KV cache has emerged as a critical bottleneck by expanding linearly in size with the sequence length.

We present **ShadowKV**, a high-throughput long-context LLM inference system that **stores the low-rank key cache** and **offloads the value cache** to **reduce the memory footprint for larger batch sizes and longer sequences**.

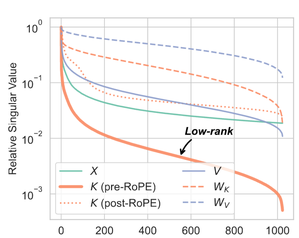
- ✓ supports up to **6x larger** batch sizes or longer sequences
- ✓ boosts throughput by **up to 3.04x** on an A100 GPU without sacrificing accuracy
- ✓ even surpassing the performance achievable with **infinite batch size** under the assumption of infinite GPU memory

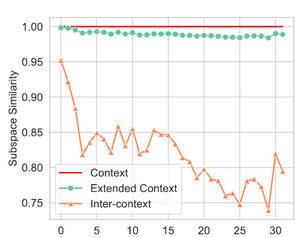


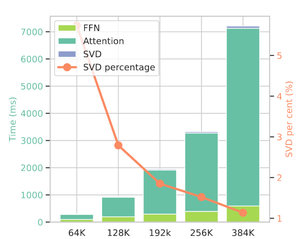


Motivation of ShadowKV

- ✦ **Pre-RoPE key caches** are **exceptionally low-rank**
- ✦ **Pre-RoPE key caches** share low-rank subspaces for a sequence and its continuation, but **NOT** across different sequences
- ✦ Attention's **quadratic** scaling makes the **linear** cost of SVD decomposition negligible







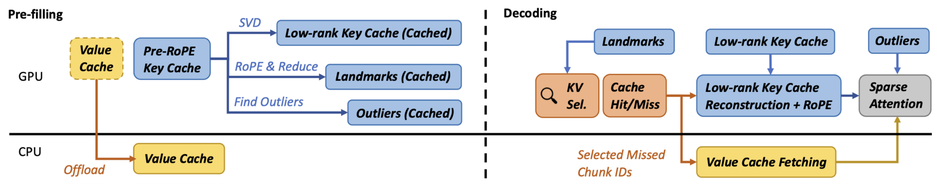
Method

Memory Efficient Pre-filling:

- ➡ Store the low-rank keys and offload the values to **reduce the memory footprint**
- ➡ Build landmarks and outliers for **accurate sparse attention**

High-throughput Decoding:

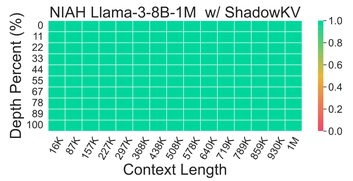
- ➡ Approximate attention with landmarks → TopK chunk IDs → Sparse Attention
- ➡ **Overlapping** Low-rank key cache reconstruction + offloaded value cache fetching



Accuracy Evaluation

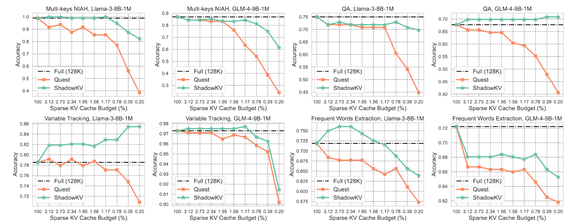
We evaluate our approach on **RULER, LongBench, and Needle In A Haystack (NIAH)**.

- ★ ShadowKV demonstrates excellent performance with a fixed sparse KV cache budget of **1.56%**.
- ★ ShadowKV can process information at any position in context windows **from 16K to 1M tokens**.



Methods	S1	S2	MK1	MK2	MQ	MV	QA-1	QA-2	VT	FWE	Avg.
Llama-3-8B-1M	100.00	100.00	98.96	98.96	98.96	95.57	75.00	48.96	78.54	71.85	86.68
Loki	18.75	1.04	2.08	0.00	1.56	0.78	4.17	13.54	26.04	25.35	9.33
Loki (V)	41.67	6.25	37.50	1.04	8.07	30.73	10.42	19.79	51.67	37.50	24.46
InfiniGen	**100.00**	98.96	84.38	53.13	63.28	54.95	65.63	48.96	**81.67**	50.35	70.13
InfiniGen (V)	**100.00**	98.96	96.88	76.04	81.25	77.08	67.71	50.00	**81.67**	53.47	78.31
Quest	**100.00**	**100.00**	**98.96**	77.08	97.65	93.49	60.42	50.00	77.08	65.63	82.03
Quest (V)	**100.00**	**100.00**	**98.96**	85.42	**97.92**	95.49	70.83	46.88	78.75	65.63	83.99
SHADOWKV	**100.00**	**100.00**	97.92	**98.96**	96.88	**95.83**	**72.92**	**52.08**	**81.67**	**72.57**	**86.88**
GLM-4-9B-1M	100.00	100.00	94.79	87.50	99.74	93.75	67.71	55.21	97.29	72.22	86.82
Loki	71.88	27.08	22.92	2.08	9.90	11.46	28.13	27.08	31.04	54.17	28.57
Loki (V)	96.88	55.21	56.25	18.75	51.04	50.52	45.83	39.58	72.71	59.72	54.65
InfiniGen	**100.00**	93.75	82.29	0.00	79.43	60.16	57.29	53.13	92.71	57.29	67.60
InfiniGen (V)	**100.00**	96.88	87.50	7.29	95.31	75.26	56.25	54.17	95.63	60.76	72.91
Quest	**100.00**	95.83	90.62	54.17	94.01	76.30	55.21	52.08	95.83	64.58	77.86
Quest (V)	**100.00**	96.88	93.75	72.92	95.83	83.07	56.25	53.13	96.88	65.97	81.47
SHADOWKV	**100.00**	**100.00**	**95.83**	**83.33**	**98.70**	**87.76**	**69.79**	**55.21**	**97.50**	**68.06**	**85.62**

★ ShadowKV consistently surpasses Quest under the same sparse budgets and achieves higher throughput.



Efficiency Evaluation

ShadowKV supports **6x** larger batch sizes and boosts generation throughput up to **3.04x**

Table 3 Generation throughput (tokens/s) on an A100. The gray text in brackets denotes batch size.

Model	Context	Full Attn	SHADOWKV	Gain	Full Attn (Inf)
Llama-3-8B-1M (8 KV heads)	60K	160.62 (8)	**455.14** (48)	2.83×	168.72 (48) / 273.07 (Inf)
	122K	80.77 (4)	**239.51** (24)	2.97×	83.05 (24) / 134.30 (Inf)
	244K	40.37 (2)	**119.01** (12)	2.95×	52.00 (12) / 67.15 (Inf)
Llama-3.1-8B (8 KV heads)	60K	160.93 (8)	**472.77** (48)	2.94×	168.72 (48) / 273.07 (Inf)
	122K	80.78 (4)	**245.90** (24)	3.04×	83.05 (24) / 134.30 (Inf)
GLM-4-9B-1M (4 KV heads)	60K	241.05 (12)	**615.89** (50)	2.56×	266.24 (50) / 436.91 (Inf)
	122K	122.67 (6)	**293.40** (25)	2.39×	158.83 (25) / 214.87 (Inf)
	244K	61.13 (3)	**136.51** (12)	2.23×	78.84 (12) / 107.44 (Inf)
Yi-9B-200K (4 KV heads)	60K	204.81 (10)	**544.36** (42)	2.66×	271.21 (42) / 364.09 (Inf)
	122K	101.44 (5)	**260.03** (21)	2.56×	133.53 (21) / 179.06 (Inf)
	244K	46.74 (2)	**118.55** (10)	2.54×	65.79 (10) / 89.53 (Inf)

ShadowKV supports **6x** longer contexts **without OOM.**

Table 4 Generation throughput (tokens/s) under varying batch sizes and sequence lengths on Llama-3-8B-1M.

Context	2	3	4	5	6	8	12	16	24	32	48
Full KV											
60K	89.19	111.44	126.73	142.62	147.40	160.62	OOM	OOM	OOM	OOM	OOM
122K	65.12	75.16	80.77	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
244K	40.37	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
488K	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
ShadowKV											
60K	89.69	126.61	159.41	184.92	205.41	244.20	306.09	346.34	399.65	428.63	**455.14**
122K	65.61	94.28	115.01	132.23	143.77	166.72	196.73	217.24	**239.51**	OOM	OOM
244K	48.39	65.95	78.92	87.83	94.07	104.73	**119.01**	OOM	OOM	OOM	OOM
488K	29.82	41.01	47.13	50.85	**53.46**	OOM	OOM	OOM	OOM	OOM	OOM