NEURAL INFORMATION PROCESSING SYSTEMS

ADKV: A Low-Overhead Adaptive Delta Quantization for KV Cache in LLM Inference

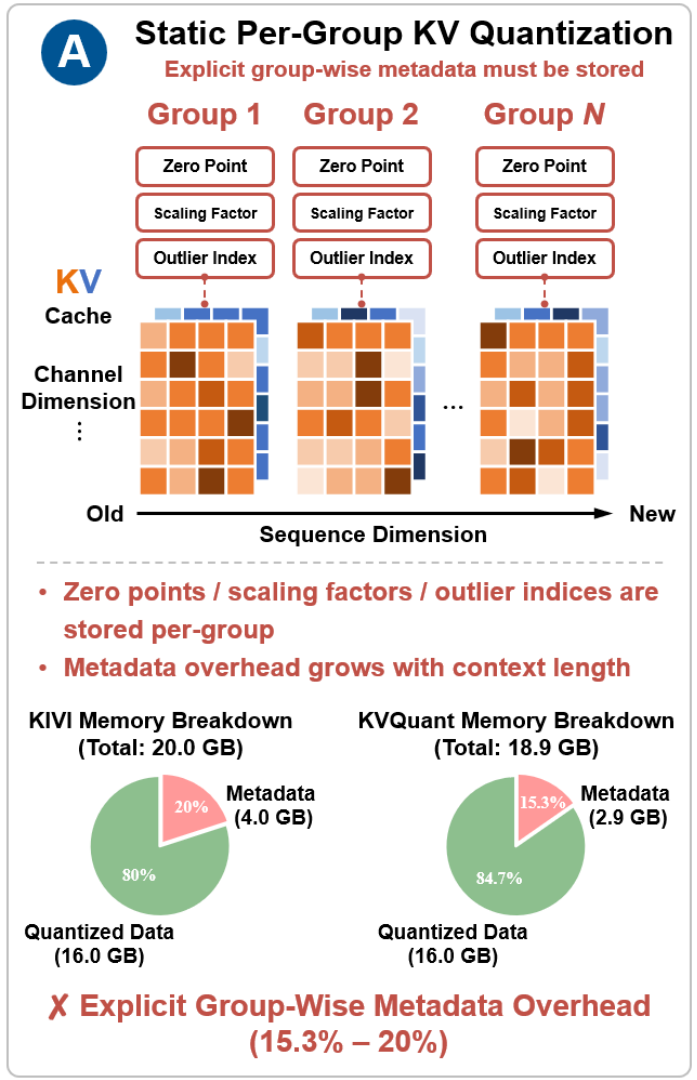
Honghao Jia[1], Zhenxing Li[2], Jialiang Guo[3], Jiacheng Gan[2]

The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS), 2026.

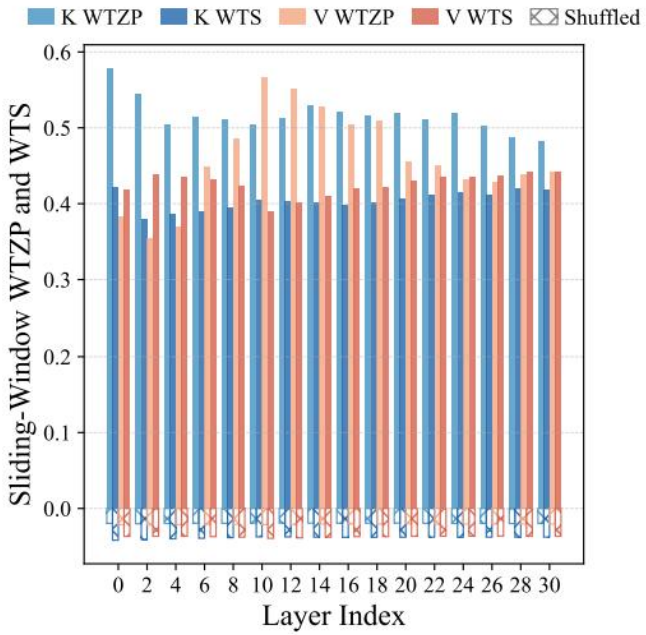
[1]University of Science and Technology of China, [2]Harbin Engineering University, [3]Tianjin University

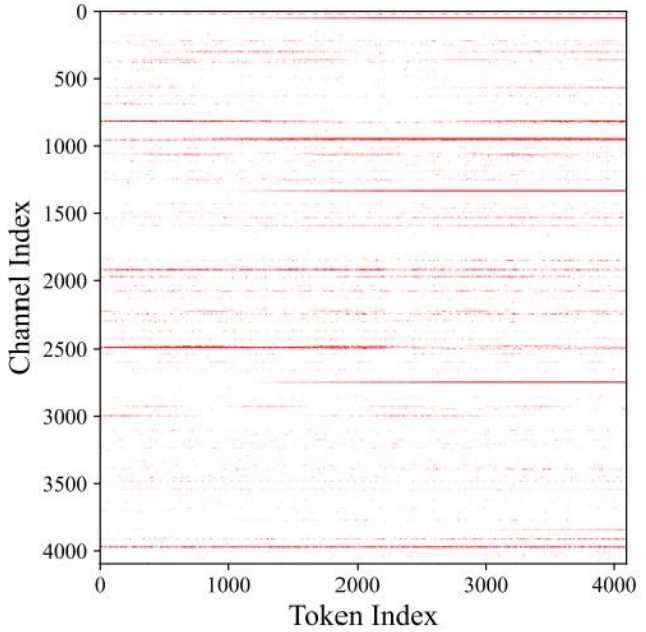
Challenges

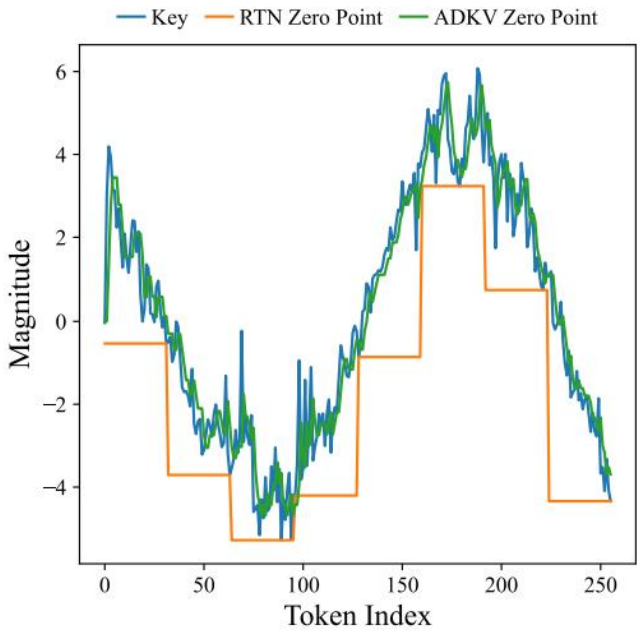
- The **KV cache**, which stores intermediate attention activations, becomes a major memory bottleneck.
- **KV cache quantization** is particularly practical: it directly reduces KV cache bit-width, achieving a 2–4× memory footprint reduction.
- Existing **static per-group quantization** methods rely on fixed-window statistics, incurring non-negligible metadata overhead (15.3–20%) for storing zero points, scaling factors, and outlier indices.
- Extending the statistical range reduces the metadata overhead but degrades accuracy due to distributional shifts in long contexts.



Motivations

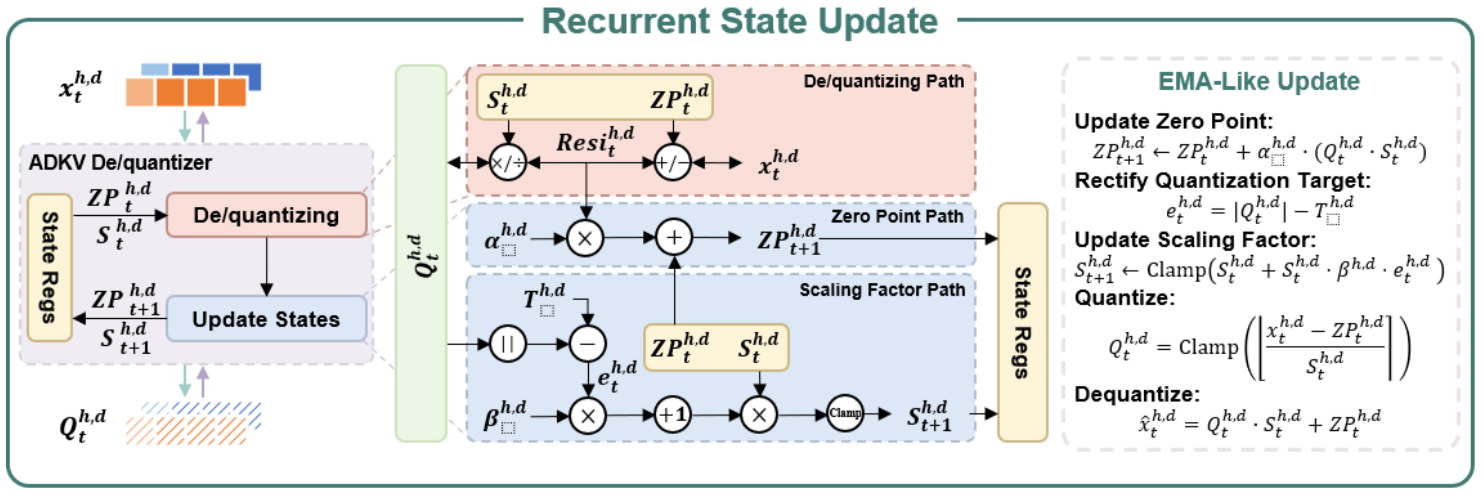


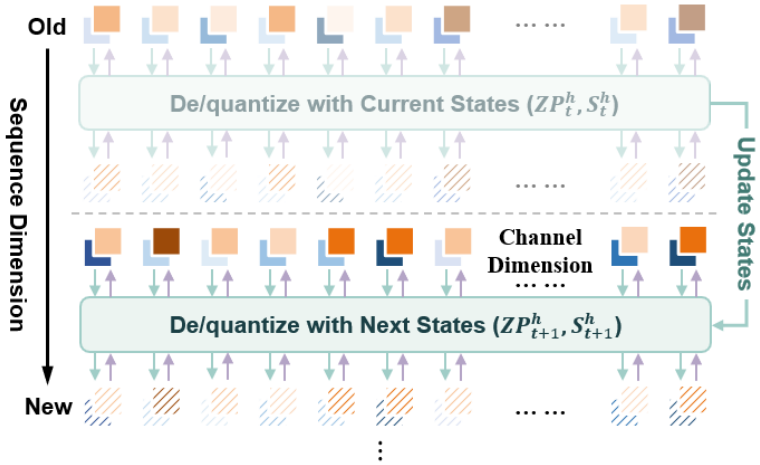




- **Temporal Trackability (WTZP & WTS):** The KV cache grows in a sequential, time-ordered manner, exhibiting an inherent temporal structure that static per-group quantization fails to exploit.
- To effectively track and compress the trackable components, we propose **ADKV**, an adaptive delta quantization method for KV cache.

ADKV: Low-Overhead Adaptive Delta Quantization





- **De/Quantization Phase:** $Resi_t^{h,d} = x_t^{h,d} - ZP_t^{h,d}$, $Q_t^{h,d} = \mathrm{Clamp}\left(\left\lfloor \frac{Resi_t^{h,d}}{S_t^{h,d}} \right\rceil, Q_\mathrm{L}, Q_\mathrm{U}\right)$, $\hat{x}_t^{h,d} = Q_t^{h,d} \cdot S_t^{h,d} + ZP_t^{h,d}$
- **Update Phase:** The ADKV quantizer updates all state variables it maintains online.

Static per-group metadata

No Explicit Group-Wise Metadata!

Online Quantization states

$$ZP_{t+1}^{h,d} \leftarrow ZP_t^{h,d} + \alpha^{h,d} \cdot \left(Q_t^{h,d} \cdot S_t^{h,d}\right)$$

$$e_t^{h,d} = \left|Q_t^{h,d}\right| - T^{h,d}$$

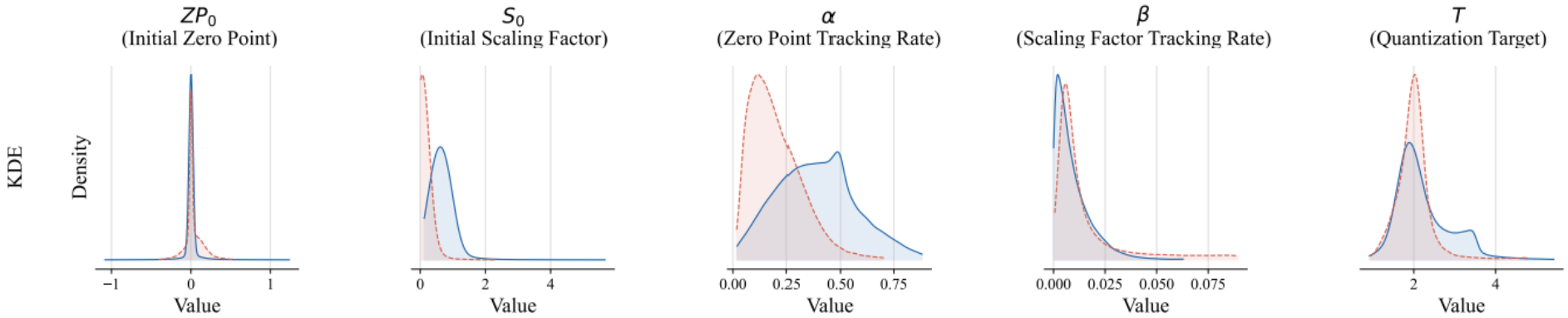
$$S_{t+1}^{h,d} \leftarrow \mathrm{Clamp}(S_t^{h,d} + S_t^{h,d} \cdot \beta^{h,d} \cdot e_t^{h,d}, S_{\min}, S_{\max})$$

ADKV: Learnable Quantizer Parameter

- We model ADKV as an **recurrent computation graph**, using learnable quantizer parameter (LQP) and SGD to optimize the ADKV per-channel constant parameters: $\alpha^{h,d}, \beta^{h,d}, T^{h,d}, ZP_0^{h,d}, S_0^{h,d}$.

- For each sample, we first calculate the reference attention scores and attention outputs using the original query, key, and value activations. Then, the quantized keys and values produced by ADKV are fed back into the same attention operator to obtain the reconstructed outputs. The primary loss consists of an MSE term:

$$\mathcal{L}_{\text{attn}}^{(i)} = \frac{1}{H}\sum_{h=0}^{H-1}\frac{1}{|\Omega^{(i)}|}\left\|\text{Softmax}\left(\frac{Q_h^{(i)}\,\hat{K}_h^{(i)\top}}{\sqrt{d_h}}\right)\hat{V}_h^{(i)} - \text{Softmax}\left(\frac{Q_h^{(i)}\,K_h^{(i)\top}}{\sqrt{d_h}}\right)V_h^{(i)}\right\|_F^2$$



Experiments

- **Generation and Long-Context Accuracy:** Maintain near-parity with the FP16 baseline (< **1** point drop for generation, **1.02** for long-context tasks on average).

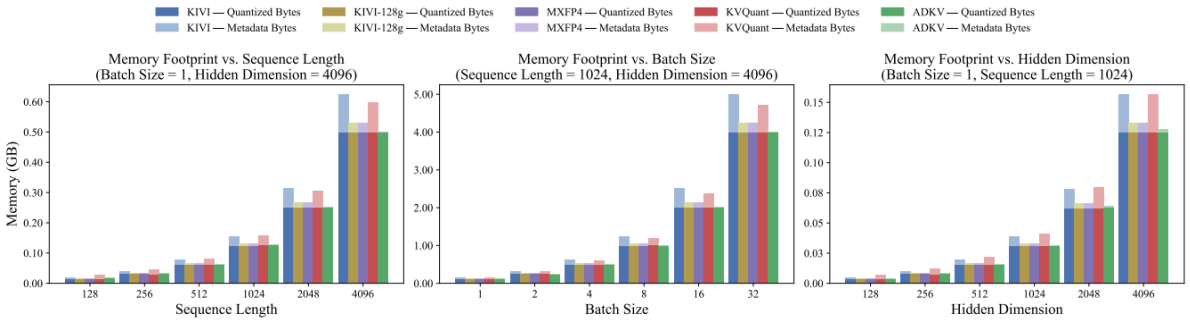
Table 1: Generation Accuracy Results

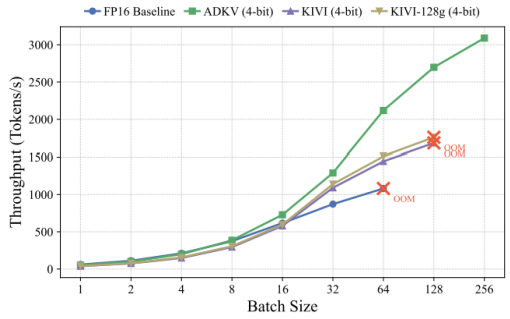
Model	Method	CoQA ↑	TruthfulQA ↑	GSM8K ↑	Avg. ↑
Llama-2 7B	FP16	64.37	32.93	13.57	36.96
	RTN	59.87	28.52	10.54	32.98
	KIVI-128g	**64.52**	31.95	12.70	36.39
	ADKV (Ours)	63.93	**34.64**	**13.50**	**37.36**
Llama-2 13B	FP16	66.26	30.60	23.43	40.10
	RTN	65.62	28.15	16.22	36.66
	KIVI-128g	66.02	30.35	23.17	39.85
	ADKV (Ours)	**67.75**	**31.21**	**23.35**	**40.77**
Falcon3 7B	FP16	65.65	45.65	76.04	62.45
	RTN	65.95	43.00	74.22	61.06
	KIVI-128g	66.10	44.21	75.89	62.07
	ADKV (Ours)	**66.37**	**44.43**	**76.02**	**62.27**
Mistral 7B	FP16	68.42	40.64	38.21	49.09
	RTN	65.45	39.78	32.22	45.82
	KIVI-128g	67.83	40.64	**36.16**	48.21
	ADKV (Ours)	**68.17**	**40.76**	35.78	**48.24**

Table 2: Long-Context Accuracy Results

Model	Method	Qasper ↑	QMSum ↑	TREC ↑	TriviaQA ↑	MultiNews ↑	SAMSum ↑	LCC ↑	RepoB-P ↑	Avg. ↑
Llama-2 7B 32K	FP16	10.53	17.18	71.00	87.79	21.96	43.90	62.08	59.86	46.79
	RTN	10.26	14.48	68.50	75.08	19.55	34.81	49.91	46.93	39.94
	KIVI-128g	10.14	17.13	71.00	**87.62**	**21.62**	43.97	61.22	58.50	46.40
	ADKV (Ours)	**10.35**	**17.16**	**71.00**	87.49	21.32	**43.98**	**61.62**	**58.98**	**46.49**
Llama-3 8B 1048K	FP16	27.84	25.13	71.00	87.70	28.38	42.42	38.34	38.45	44.91
	RTN	26.45	23.91	70.00	84.26	28.01	41.14	30.84	33.37	42.25
	KIVI-128g	27.16	24.53	71.00	**87.13**	28.04	**41.72**	35.21	35.14	43.74
	ADKV (Ours)	**27.52**	**24.63**	**72.00**	86.33	**28.31**	41.69	**35.32**	**35.35**	**43.89**
Llama-2 13B	FP16	9.28	21.45	70.00	87.37	15.96	43.47	66.64	56.45	46.33
	RTN	6.80	17.57	61.00	72.12	13.40	41.26	61.41	51.46	40.63
	KIVI-128g	8.73	21.23	69.00	86.20	15.21	**44.09**	**66.20**	57.06	45.97
	ADKV (Ours)	**8.81**	**21.48**	**70.00**	**87.17**	**15.53**	43.51	65.94	**57.10**	**46.19**
Falcon3 7B	FP16	12.33	23.37	77.00	89.38	23.49	43.77	68.29	62.10	49.97
	RTN	12.11	23.19	73.50	89.18	22.34	41.74	66.97	61.68	48.84
	KIVI-128g	**12.19**	**23.45**	76.00	89.25	22.22	43.22	67.45	**62.59**	49.55
	ADKV (Ours)	12.03	22.81	**76.50**	**89.40**	**22.72**	**43.25**	**67.79**	62.12	**49.58**
Mistral 7B	FP16	8.07	19.16	66.50	89.21	13.83	41.24	69.31	61.48	46.10
	RTN	7.43	**19.34**	64.50	86.63	13.22	40.87	66.66	58.44	44.64
	KIVI-128g	7.55	19.03	65.50	**89.10**	13.62	**40.95**	69.36	60.21	45.67
	ADKV (Ours)	**8.03**	19.25	**65.50**	88.94	**14.24**	39.76	**69.64**	**60.27**	**45.70**

- **Memory Savings & Throughput:** Reduce total KV cache memory by **5.87–19.99%** and quantization metadata memory by **99.76–99.94%** compared to the static per-group quantization baselines. Achieve a **2**× larger maximum feasible batch size, and **1.53**× end-to-end throughput gain.





Takeaways

- **Analysis of the Temporal Structure in the KV Cache.** We observe and formalize the inherent temporal structure of the KV cache.
- **Adaptive Delta Quantization via EMA.** We employ EMA to recurrently track and update zero points and scaling factors online, thereby improving robustness to distributional shifts.
- **Ultra-Low Metadata Overhead via Adaptive Delta Coding.** We decouple the metadata overhead from context length by embedding metadata implicitly into the quantized sequence via adaptive delta coding.
- **SGD-Based Calibration Algorithm.** We develop an SGD-based calibration algorithm that optimizes the quantizer parameters by minimizing the mean squared error (MSE) of attention outputs, which enables attention-aware calibration.
- **High-Performance Fused ADKV-GEMV CUDA Kernel.** We design an efficient CUDA kernel for the ADKV quantization and dequantization pipeline.