KVTuner: Sensitivity-Aware Layer-Wise Mixed-Precision KV Cache Quantization for Efficient and Nearly Lossless LLM Inference

Authors: **Xing LI***, Zeyu Xing*, Yiming Li, Linping Qu, Hui-Ling Zhen, Wulong Liu, *Yiwu Yao*, Sinno Jialin Pan, Mingxuan Yuan

Affiliations: **Huawei Noah's Ark Lab**, The Chinese University of Hong Kong, *Huawei Computing Product Line*

**Equal Contribution*

li.xing2@huawei.com

ICML International Conference On Machine Learning

Paper

Code

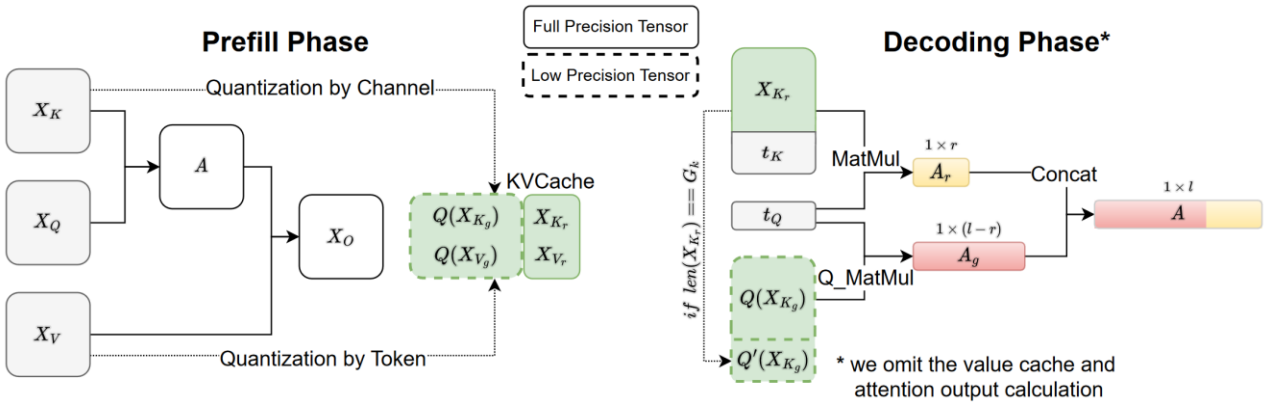
香港中文大學 The Chinese University of Hong Kong

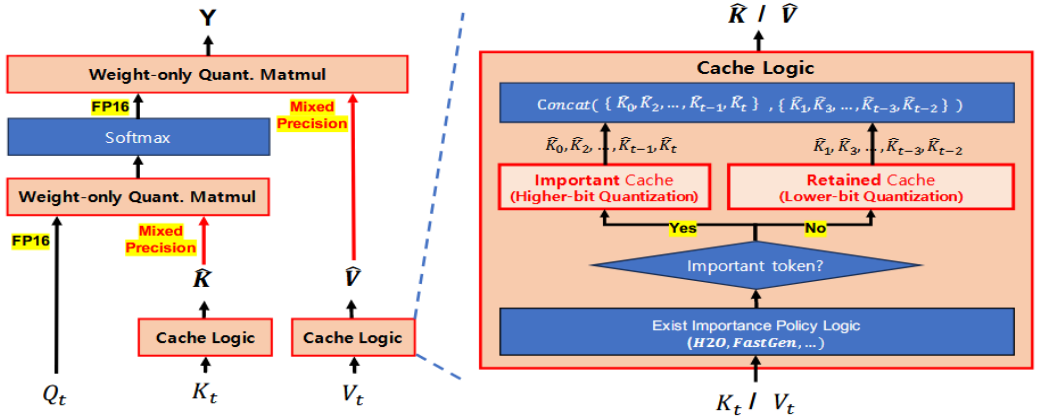
HUAWEI

KVCache Quant. Background

- KVCache properties ***Inter-layer & intra-layer coarse grain mixed precision***
  - Shape: [batchSize, ***layerNum, 2***, **seqLen**, KVHeads, hiddenDim]
  - Dynamic generation of contextual knowledge
  - Linearly scale
- Related works
  - KIVI: asymmetric 2bit KV quantization
  - MiKV: importance driven token-level

KV cache precision	Results with 15-th input prompt *In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled in hip-hop dance?*
FP16	Let X be the number of students enrolled in hip-hop dance. The number of students enrolled in contemporary dance is 20% of 20 = 4. The number of students enrolled in jazz dance is 25% of the remaining = 25% of (20 - 4) = 25% of 16 = 4. Therefore, the number of students enrolled in hip-hop dance is 20 - 4 - 4 = «20-4-4=12»12 . The percentage of the entire students enrolled in hip-hop dance is 12/20 x 100% = 60% . #### 60%.
KIVI-4	Let X be the number of students enrolled in hip-hop dance. The number of students enrolled in contemporary dance is 20% of 20 = 4. The number of students enrolled in jazz dance is 25% of the remaining = 25% of (20 - 4) = 25% of 16 = 4. Therefore, the number of students enrolled in hip-hop dance is 20 - 4 - 4 = «20-4-4=12»12 . The percentage of the entire students enrolled in hip-hop dance is 12/20 x 100% = 60% . #### 60%.
KIVI-2	Let X be the number of students who enrolled in hip-hop dance. The number of students who enrolled in contemporary dance is 20% of 20 = 4. The number of students who enrolled in jazz dance is 25% of 16 = 4. Therefore, the total number of students who enrolled in hip-hop dance is 20 + 4 + 4 = 28 . The percentage of the entire students who enrolled in hip-hop dance is 28/20 = «28/20=14»14% . #### 14.





Liu, Zirui, et al. "Kivi: A tuning-free asymmetric 2bit quantization for kv cache." *arXiv preprint arXiv:2402.02750* (2024).

Yang, June Yong, et al. "No token left behind: Reliable kv cache compression via importance-aware mixed precision quantization." *arXiv preprint arXiv:2402.18096* (2024).

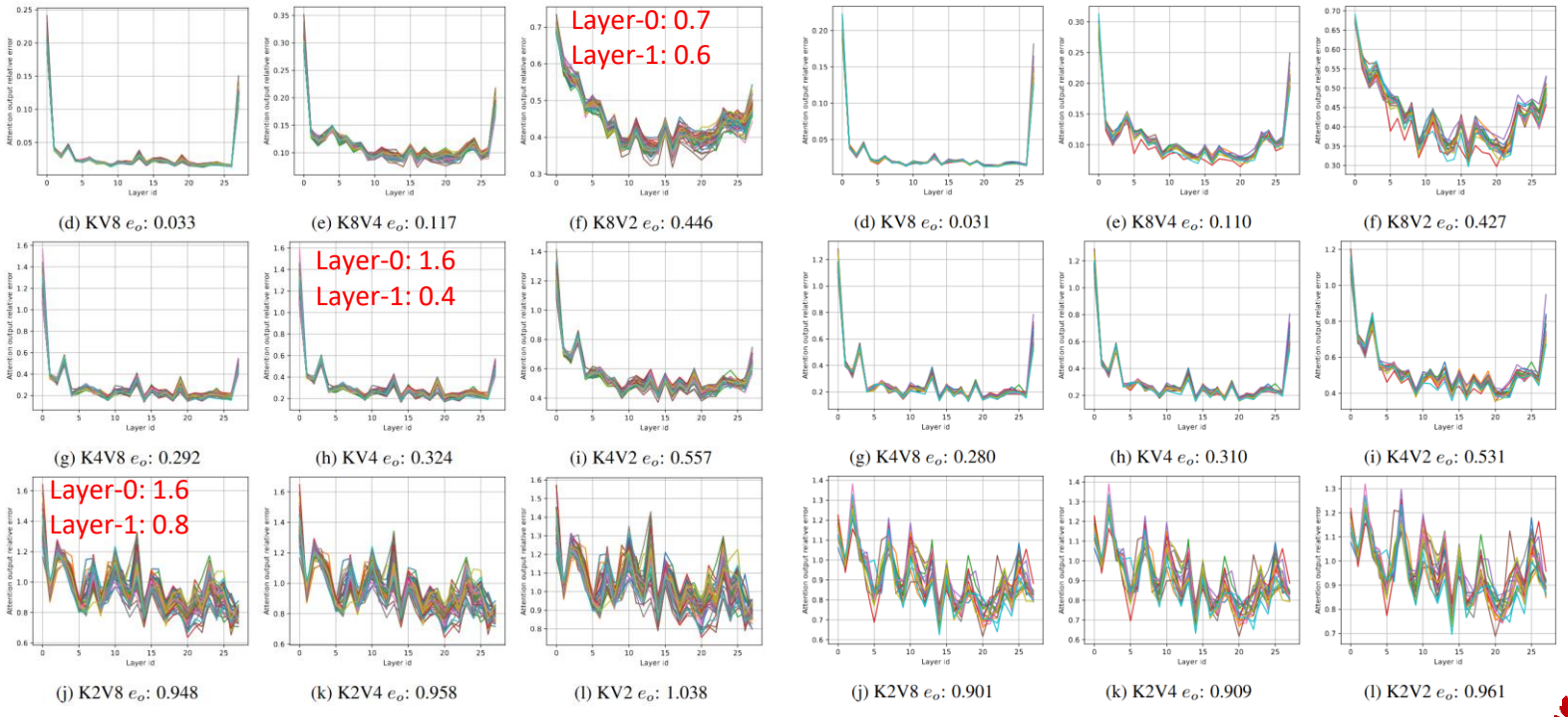
HUAWEI

KVTuner Motivation KVCache Quant Sensitivity Analysis

- Generalization: key is generally more important than value for model accuracy (perplexity)
  - 9 models
  - 9 precision pairs
  - Lower key precision
    - -> perplexity degradation

Model	KV8	K8V4	K8V2	K4V8	KV4	K4V2	K2V8	K2V4	KV2
Llama3-8B-Instruct	9.95	9.94	10.04	9.99	9.99	10.11	31.92	31.48	37.29
Llama2-7B-chat-hf	11.60	11.60	11.67	11.61	11.62	11.67	13.86	13.92	14.92
Llama2-13B-chat-hf	10.04	10.05	10.08	10.06	10.07	10.11	13.30	13.37	14.25
Mistral-7B-Instruct-v0.3	8.28	8.27	8.35	8.31	8.29	8.44	12.61	12.71	15.18
Qwen2.5-3B-Instruct	10.60	10.59	11.36	11.11	11.11	12.28	147.03	151.30	251.89
Qwen2.5-7B-Instruct	9.56	9.39	9.45	220.83	235.03	149.15	1866.33	1831.33	4016.10
Qwen2.5-Math-7B-Instruct	168.92	169.60	175.34	588.34	599.02	725.10	1746.07	1760.31	1829.26
Qwen2.5-14B-Instruct	6.65	6.67	7.19	6.81	6.83	7.32	16.05	16.37	18.22
Qwen2.5-32B-Instruct	6.68	6.85	6.34	6.47	6.52	6.43	9.13	9.20	9.56

- Intra-layer and inter-layer sensitivities are the inherent model property and independent of inputs
  - Layer-0: K8V2<KV4
  - Layer-1: K8V2>KV4



Qwen2.5-7B-Instruct, math GSM8K

Qwen2.5-7B-Instruct, creative generation AIGC softage

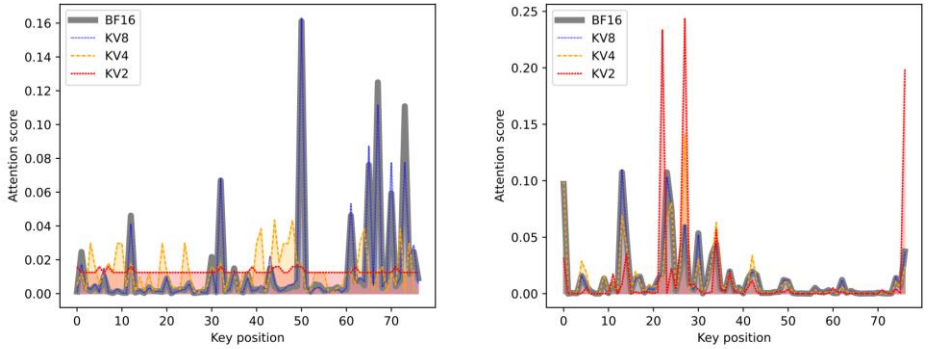
HUAWEI

KVTuner Observation Empirical and Theoretical Analysis

- Sensitivities strongly correlate with attention patterns
  - Key errors may lead to attention distribution shift
  - Sparse/concentrated v.s. random/retrieval

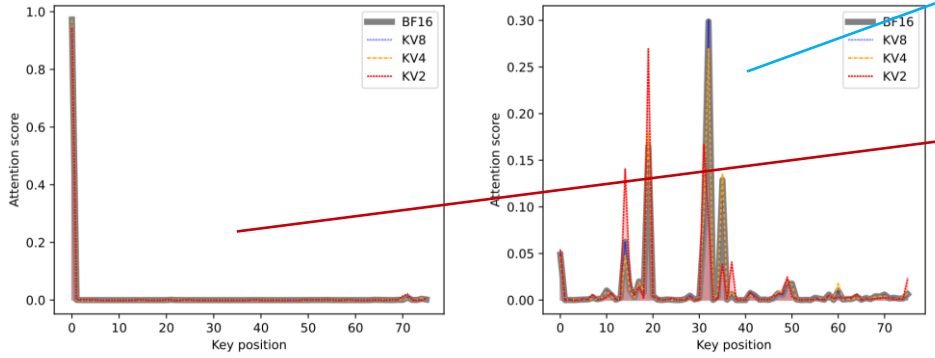
Lemma 1. *Only attention heads with sparse and concentrated patterns demonstrate consistent robustness to low-precision KV cache quantization.*

  - Noticeable model-wise and layer-wise difference



(a) Layer-0 query head-2 (b) Layer-21 query head-4

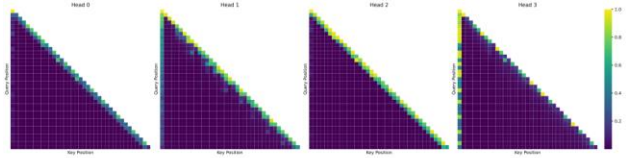
Qwen2.5-7B-Instruct



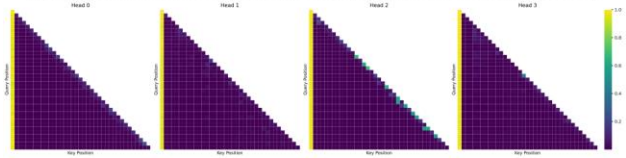
(a) Layer-2 streaming head (b) Layer-13 retrieval head

Llama-3.1-8B-Instruct

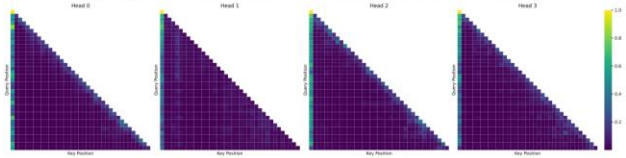
Llama-3.1-8B-Instruct



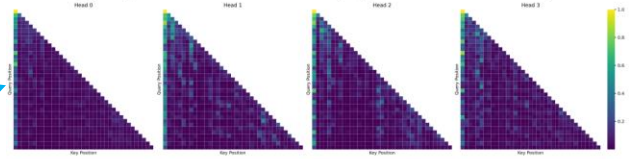
(a) Layer-0 with recent attention patterns (medium attention score errors)



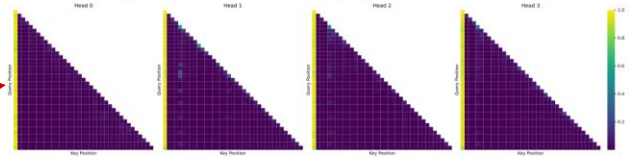
(b) Layer-2 with attention sinks (low attention score errors)



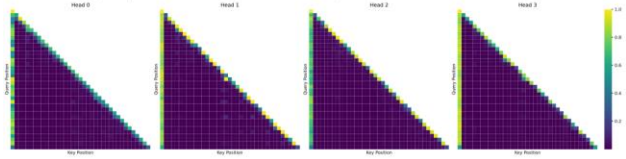
(c) Layer-12 with retrieval heads (high attention score errors)



(d) Layer-13 with retrieval heads (high attention score errors)

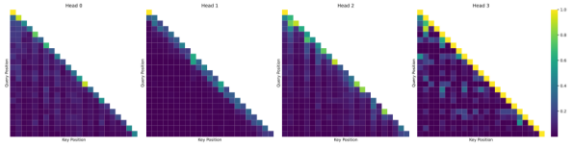


(e) Layer-23 with attention sink (low attention score errors)

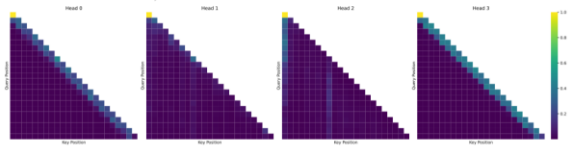


(f) Layer-31 with mixture of retrieval and streaming heads (medium attention score errors)

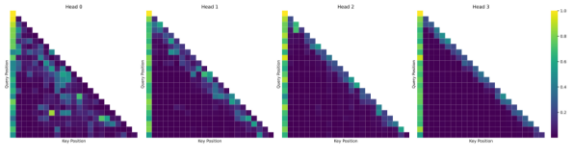
Qwen2.5-7B-Instruct



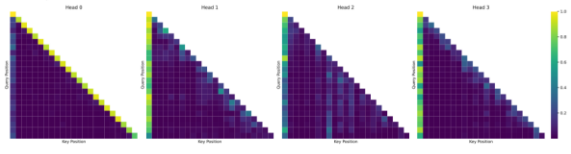
(a) Layer-0 with mixture of recent window, re-access, and retrieval heads (high attention score errors)



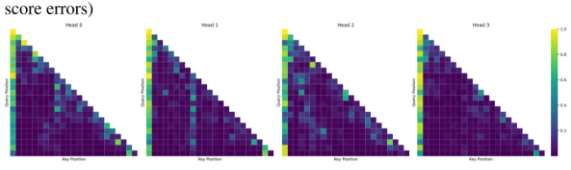
(b) Layer-1 with mixture of recent window and re-access patterns (medium attention score errors)



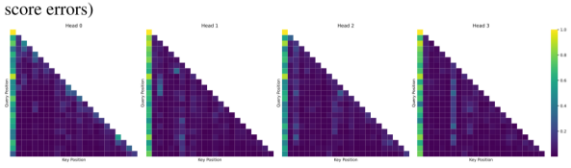
(c) Layer-5 with mixture of retrieval and streaming heads (low attention score errors)



(d) Layer-12 with mixture of retrieval and streaming heads (medium attention score errors)



(e) Layer-21 with mixture of retrieval heads and attention sinks (medium attention score errors)

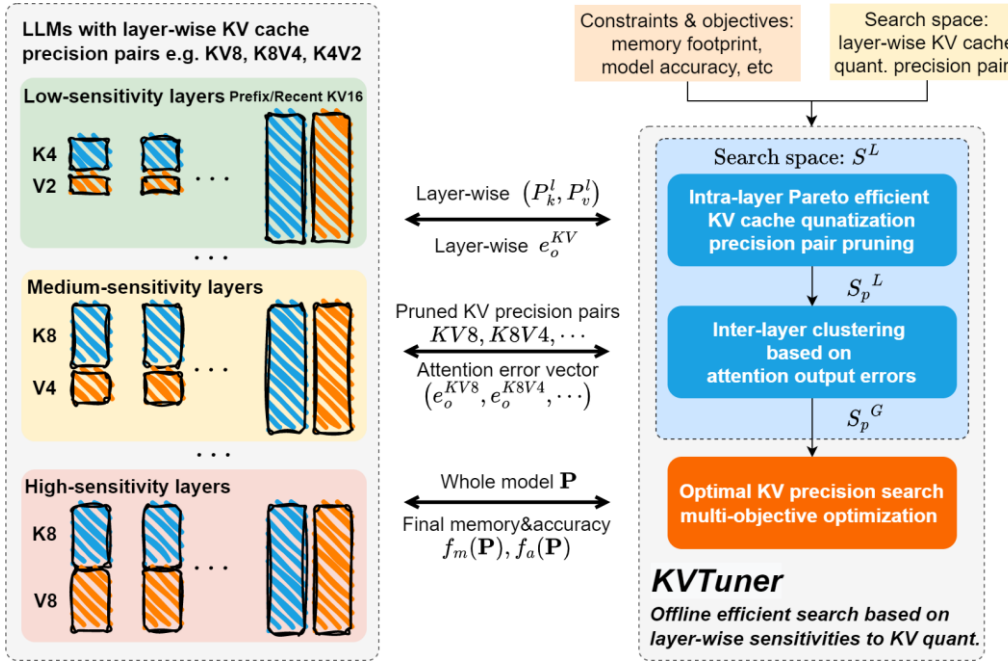


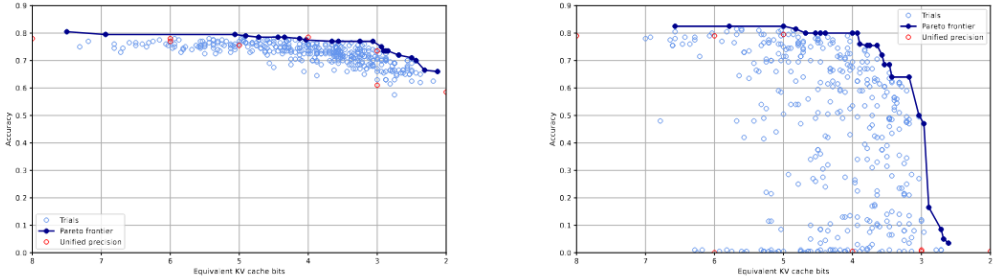
(f) Layer-27 with mixture of retrieval heads and attention sinks (high attention score errors)

HUAWEI

KVTuner Method

- Intra-layer and inter-layer coarse-grain **KV**Cache mixed precision quantization **tuner**
  - KV sensitivity theoretical analysis driven & hardware friendly
  - Generalize to different quant. modes and inference frameworks
- *Two-stage search space pruning* $3.4 \times 10^{30} \rightarrow 15625$
  - Efficient sampling
- *Offline multi-objective KVCache quant. precision tuning*
  - ✓ Automatically adapt to different models and constraints
  - ✓ Zero online overhead with calibrated config.
- Gain
  - ✓ Nearly lossless **3.25bit KVCache quant**. on math./scientific datasets
  - ✓ **21% throughput improvement** on GQA models





(a) Llama-3.1-8B-Instruct with KIVI

(b) Qwen2.5-7B-Instruct with per-token-asym

Paper

Code

HUAWEI

KVTuner Experimental Results

- Accuracy and performance gain
  - Nearly lossless 3.25bit KVCache quant. on math./scientific datasets
  - **Nearly lossless 4bit KVCache quant. on the sensitive Qwen2.5-7B-Instruct**
    - Uniform per-token-asym KV4 and KIVI-4 lead to >67% accuracy loss
  - 21% throughput improvement on GQA models
  - **Simple per-token-asym matches the accuracy level of complex KIVI**

Quant. method	Precision	Few-shot CoT 4-shot	Few-shot CoT 8-shot	Few-shot CoT 16-shot	Few-shot as multiturn 4-shot	Few-shot as multiturn 8-shot	Few-shot as multiturn 16-shot	Average
Llama-3.1-8B-Instruct								
BF16	BF16	0.7635	0.7741	0.7854	0.8355	0.8309	0.8332	0.8038
Per-token-asym	KV8	0.7635	0.7710	0.7908	0.8340	0.8302	0.8279	0.8029
	KV4	0.7240	0.7506	0.7354	0.8211	0.8180	0.8097	0.7765
	KV2	0.0174	0.019	0.0250	0.0167	0.019	0.0197	0.0195
	KVTuner-C5.44	0.7604	0.7726	0.7726	0.8287	0.8385	0.8309	0.8006
	KVTuner-C3.59	0.7210	0.7316	0.7407	0.8021	0.8014	0.7991	0.7660
KIVI	KIVI-8	0.7733	0.7748	0.7756	0.8347	0.8317	0.8294	0.8033
	KIVI-4	0.7566	0.7718	0.7839	0.8370	0.8241	0.8332	0.8011
	KIVI-2	0.6073	0.6080	0.5929	0.6649	0.6543	0.6687	0.6327
	KVTuner-C4.91	0.7506	0.7665	0.7657	0.8173	0.8188	0.8378	0.7928
	KVTuner-C3.25	0.7483	0.7566	0.7604	0.8362	0.8256	0.8279	0.7925
Qwen2.5-3B-Instruct								
BF16	BF16	0.6020	0.6490	0.7020	0.5679	0.6005	0.6490	0.6284
Per-token-asym	KV8	0.5959	0.6573	0.7081	0.5686	0.6080	0.6323	0.6284
	KV4	0.1888	0.1721	0.2312	0.2229	0.2616	0.2464	0.2205
	KV2	0.0099	0.0121	0.0106	0.0106	0.0091	0.0129	0.0109
	KVTuner-C5.06	0.6058	0.6664	0.6823	0.5914	0.6133	0.6490	0.6347
	KVTuner-C4.00	0.6156	0.6482	0.6672	0.5815	0.6118	0.6422	0.6278
KIVI	KIVI-8	0.5974	0.6619	0.7096	0.5648	0.5989	0.6346	0.6279
	KIVI-4	0.6156	0.6550	0.7066	0.5732	0.6073	0.6414	0.6332
	KIVI-2	0.0546	0.0576	0.0675	0.047	0.0478	0.0591	0.0556
	KVTuner-C3.44	0.5989	0.6429	0.7089	0.5701	0.5997	0.6475	0.6280
	KVTuner-C3.17	0.6065	0.6444	0.6998	0.5512	0.5891	0.6406	0.6219
Qwen2.5-7B-Instruct								
BF16	BF16	0.8059	0.8287	0.8218	0.7081	0.7339	0.7544	0.7755
Per-token-asym	KV8	0.7998	0.8203	0.8196	0.7134	0.7384	0.7354	0.7712
	KV4	0.0106	0.0121	0.0121	0.003	0.003	0.0061	0.0078
	KV2	0.0068	0.0099	0.0076	0.0083	0.0106	0.0106	0.0090
	KVTuner-C5.00	0.7885	0.8302	0.8203	0.6914	0.7445	0.7468	0.7703
	KVTuner-C4.00	0.7847	0.8112	0.7726	0.6929	0.7331	0.7407	0.7559
KIVI	KIVI-8	0.8021	0.8271	0.8302	0.7066	0.7354	0.7506	0.7753
	KIVI-4	0.0735	0.1137	0.1554	0.0667	0.0705	0.1463	0.1043
	KIVI-2	0.0379	0.0402	0.0356	0.0326	0.0258	0.0235	0.0326
	KVTuner-C5.96	0.8218	0.8309	0.8150	0.6907	0.7248	0.7513	0.7724
	KVTuner-C3.92	0.5959	0.6664	0.6558	0.5588	0.6156	0.6035	0.6160

Generalization: nearly lossless long context generation with <4bit KV

Table 7: Accuracy comparison between offline searched layer-wise KV cache precision using KVTuner in Table 5 and 6 and uniform KV precision settings of the sensitive Qwen2.5-7B-Instruct on 20 LongBench long context generation benchmarks.

KIVI					
BF16	KV8	K8V4	KV4	KVTuner-C5.96	KVTuner-C3.92
0.7956	0.7992	0.8001	0.7723	0.7956	0.7903
Per-token-asym					
BF16	KV8	K8V4	KV4	KVTuner-C5.0	KVTuner-C4.0
0.7956	0.7971	0.7953	0.6343	0.8005	0.7960

KVTuner 3.25bit memory compression and throughput gain

Table 8: Throughput comparison between offline searched layer-wise KV cache precision using KVTuner in Table 5 and uniform KV precision settings with KIVI of Llama-3.1-8B-Instruct.

BS	inputLen	KV8(baseline)	K8V4	KV4	K4V2	KVTuner-C4.91	KVTuner-C3.25
64	128	3836	4193	4567	4697	4240 +10.53%	4652 +21.25%
16	512	1102	1205	1275	1304	1239 +12.41%	1296 +17.55%
8	1024	549	597	632	645	600 +9.22%	641 +16.79%

HUAWEI