InQuant: In-Place Mixed-Precision KV Cache Quantization via Saliency-Aware Neighbor-Slot Reuse

Zihan Chang[1,2,3] **Shuibing He**[1,2,3] * **Bo Zhou**[1,2] **Ping Chen**[1,4] **Siling Yang**[1]

[1]Zhejiang University [2]Zhejiang Lab

[3]Zhejiang Key Laboratory of Big Data Intelligent Computing

[4]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

NeurIPS 2026

The cost of preserving outliers

KV cache grows with context length and batch size

Uniform INT4

Compact storage
Sensitive to outlier channels

Outlier-aware mixed precision

Higher precision for outliers
Extra buffers or channel permutation

47–60%

of quantization time spent moving data

ZipCache and COMET

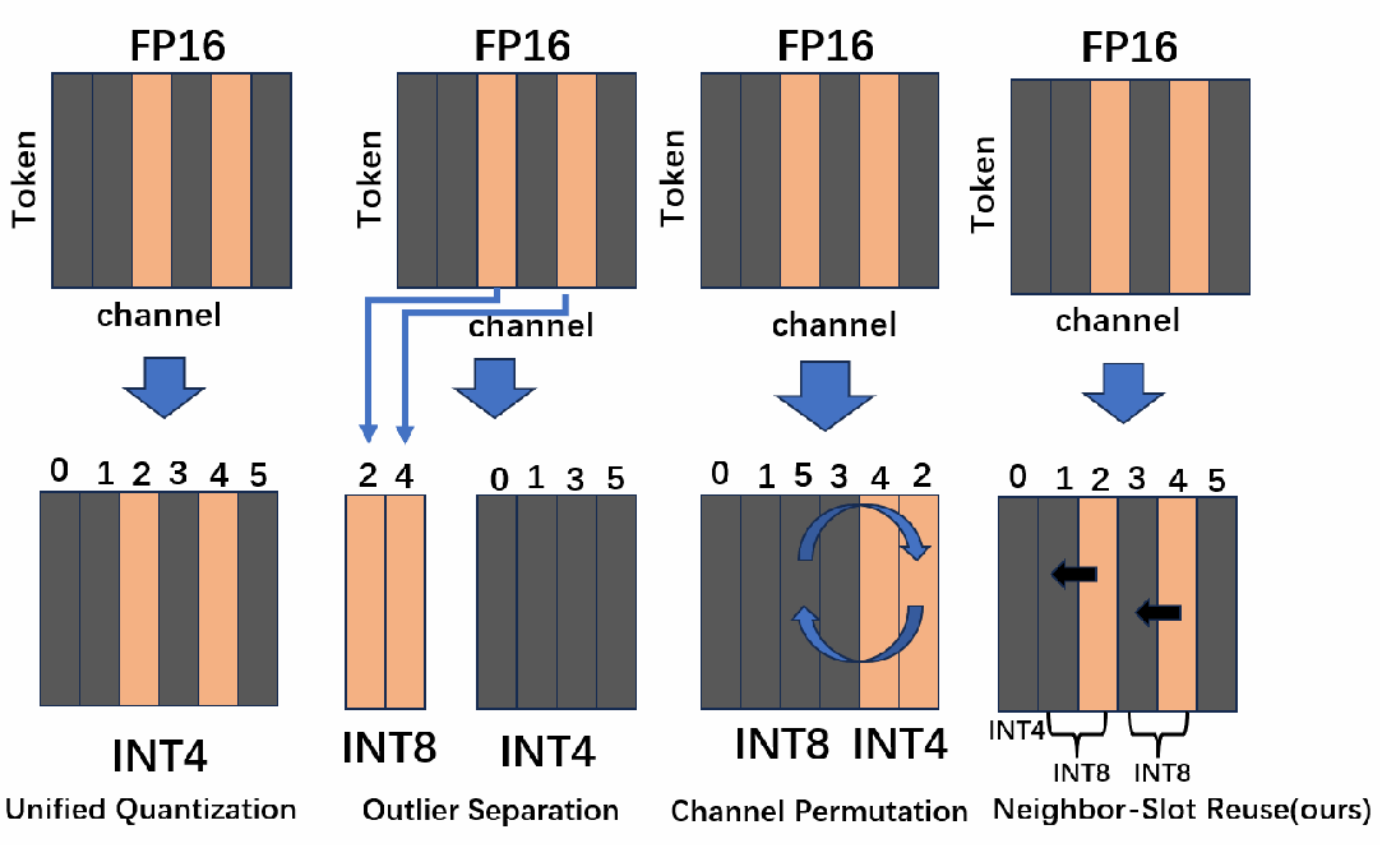
Mistral-7B profiling, varying input lengths. Paper Figure 2.

Neighbor-slot reuse

8-bit outlier code

One 4-bit piece stays in the original slot

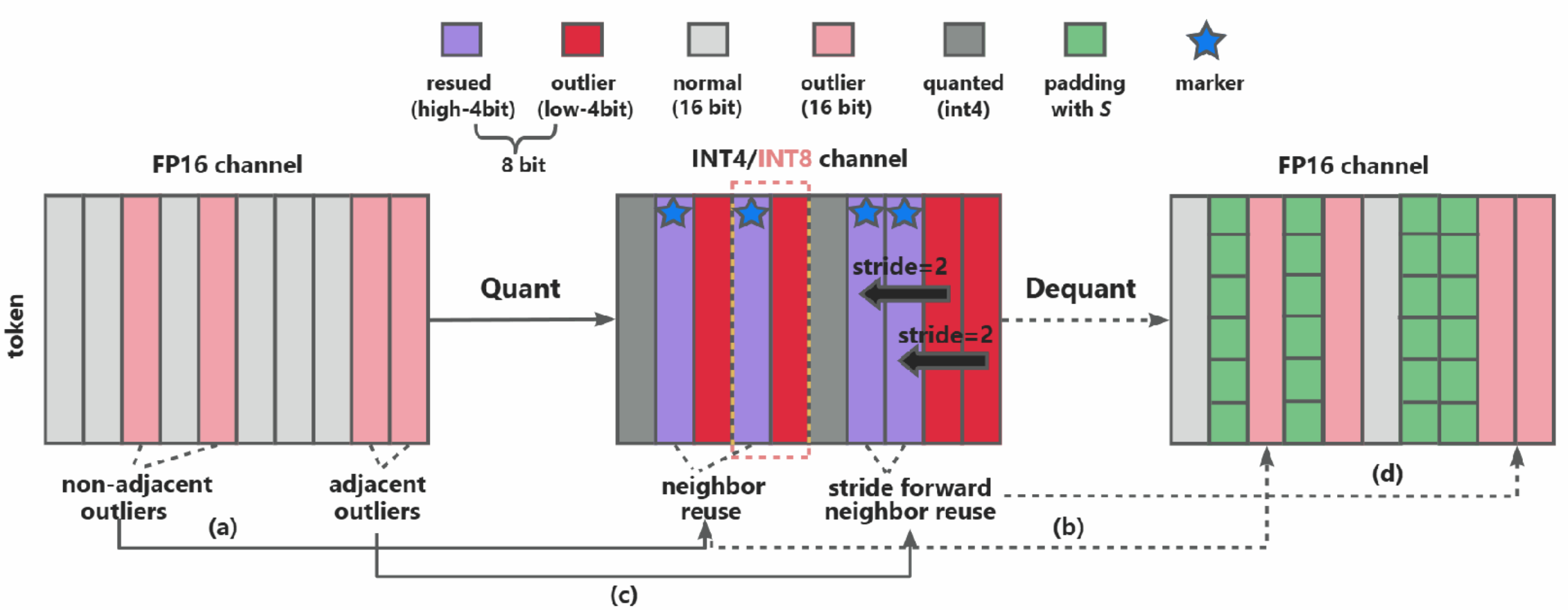
One 4-bit piece reuses a low-saliency slot



Selected donor values need approximate reconstruction

Regular packed layout, with compact metadata. Original diagram: paper Figure 1.

Adjacent outliers and recovery



Stride search finds low-saliency donors beyond adjacent outliers

Descriptors guide recovery. Sampled padding approximates overwritten values.

Lightweight channel selection

1 / 8

token positions sampled
for channel saliency

Top 20%

salient channels receive
8-bit precision

Spearman rank correlation > 0.98 with full-cache scoring

Fused CUDA kernels reduce intermediate memory traffic and synchronization

Default paper configuration. Sections 3.1 and 4.1, Figure 6.

Accuracy and quantization cost

Mistral-7B, paper main experiment

Method	Physical CR	MMLU (%)	Quant. (ms)	Dequant. (ms)
FP16	1.0×	61.9	—	—
ZipCache	2.47×	58.9	242.8	129.2
COMET	2.31×	58.8	189.7	157.7
InQuant	**4.0×**	**58.4**	**78.4**	**58.4**

2.72× faster combined Q/DQ than ZipCache
372.0 ms vs. 136.8 ms

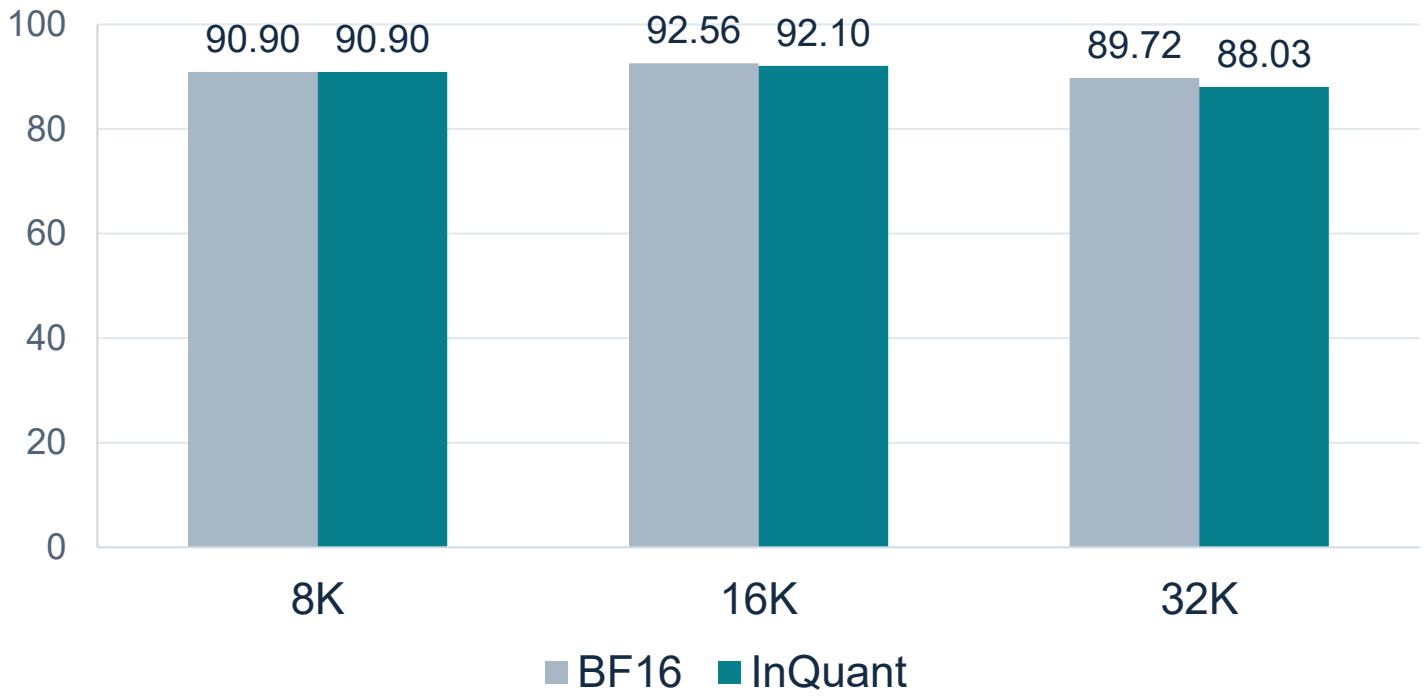
At 4-bit footprint: MMLU 58.4 vs. 52.8 for uniform INT4 (+5.6 points)

Tables 1–2. CR is physical KV-cache compression vs. FP16. Q/DQ speedup excludes other inference work.

Long-context quality and storage

Qwen2.5-7B-Instruct, K4/V2, batch size 1

RULER v1 score



78.21%

lower persistent KV and decode-state storage at 32K (mean paired reduction)

BF16 1784.78 MiB
InQuant 386.77 MiB

RULER pilot: 13 tasks, 5 samples per task at each length. Small score differences need caution.

Paper Table 4. Storage values are per-request medians. This K4/V2 experiment uses a separate configuration.

InQuant

Mixed precision within a regular packed layout

Low-saliency slots preserve outlier precision with less data movement

Evaluated on six models, up to 13B parameters and 32K context

Code and implementation

github.com/ISCS-ZJU/InQuant-open-source

Thank you