OSCAR

Offline Spectral Covariance-Aware Rotation for 2-bit KV Cache Quantization



Zhongzhu Zhou, **Donglin Zhuang**, Jisen Li, Ziyan Chen
Shuaiwen Leon Song, Ben Athiwaratkun, Xiaoxia Wu

Together AI · University of Sydney · University of Illinois Urbana-Champaign

Key quantization and attention logits

At 2 bits, the direction of a key error matters for attention.

STANDARD TARGET · KEY RECONSTRUCTION

$$\|K - \mathrm{Reconstruct}(K)\|_F^2$$

OSCAR TARGET · ATTENTION LOGITS

$$\|QK^\top - Q\,\mathrm{Reconstruct}(K)^\top\|_F^2$$

$$= \mathrm{tr}\left(\Delta K\, Q^\top Q\, \Delta K^\top\right)$$

$$\Delta K = K - \mathrm{Reconstruct}(K)$$

Q^TQ weights each key-error direction.

Value quantization and attention output

Hold the attention weights S fixed to isolate the effect of value quantization.

SQUARED ERROR IN THE WEIGHTED OUTPUT

$$\left\| SV - S\widehat{V} \right\|_F^2 = \mathrm{tr}\left(\Delta V^\top S^\top S \, \Delta V \right)$$

OSCAR TARGET · SECOND MOMENT OF SCORE-WEIGHTED VALUES

$$\frac{1}{N}(SV)^\top (SV) = \frac{1}{N} V^\top S^\top S V$$

For long calibration sequences, estimate the target with streaming score-energy weights.

$$\Delta V = V - \widehat{V}, \qquad S = \mathrm{softmax}_{\mathrm{row}}(QK^\top / \sqrt{d})$$

Key and value rotations

KEYS · QUERY MOMENT

$$\frac{1}{N}Q^\top Q = U_Q \Lambda_Q U_Q^\top$$

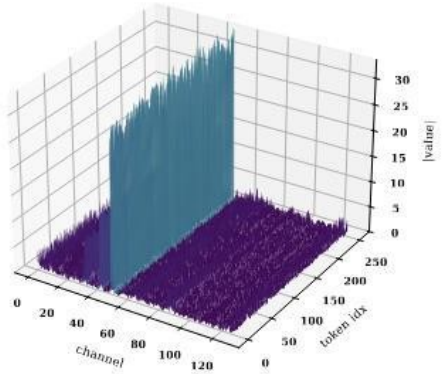
$$R_K = U_Q H_{\mathrm{Had}} P_{\mathrm{br}},$$

VALUES · SCORE-WEIGHTED MOMENT

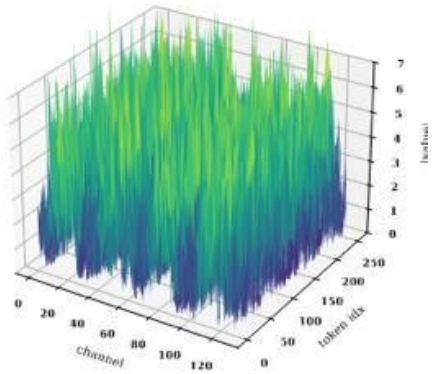
$$\frac{1}{N}(SV)^\top (SV) = U_S \Lambda_S U_S^\top$$

$$R_V = U_S H_{\mathrm{Had}} P_{\mathrm{br}}$$

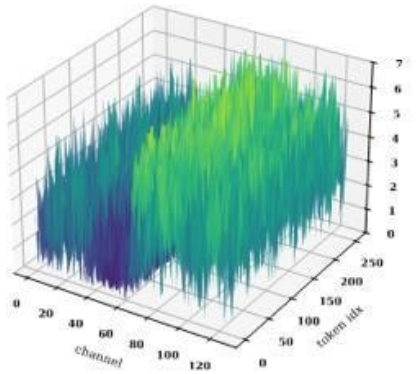
PCA gives each basis. Hadamard spreads outliers. Bit reversal balances groups.



Original keys



Hadamard only



OSCAR rotation

Illustrative key activations from Figure 1

INT2 history in a paged KV cache

SGLang stores the long history in INT2 and protects two short windows in BF16.

BF16 SINK	INT2 HISTORY	BF16 RECENT
64 tokens	**Grouped INT2 KV**	256 tokens

Offline calibration

Fit fixed per-layer rotations and clipping thresholds on calibration data.

Online serving

Fused Triton kernels handle writes and decoding with paging and prefix reuse.

Shown with the evaluation’s default sink and recent windows.

Accuracy on reasoning and coding

Mean accuracy across five tasks, with up to 32K generated tokens

Model	BF16	OSCAR	Gap (points)
Qwen3-4B-Thinking	75.64	**71.86**	−3.78
Qwen3-8B	70.84	**69.63**	−1.21
Qwen3-32B	74.19	**74.17**	−0.02
GLM-4.7-FP8	77.89	**78.16**	+0.27

Qwen3-32B and GLM-4.7-FP8 are within 0.3 points of BF16.

Five seeds per Qwen3 model and three GLM runs. Effective storage at 128K is 2.28–2.53 bits per KV element.

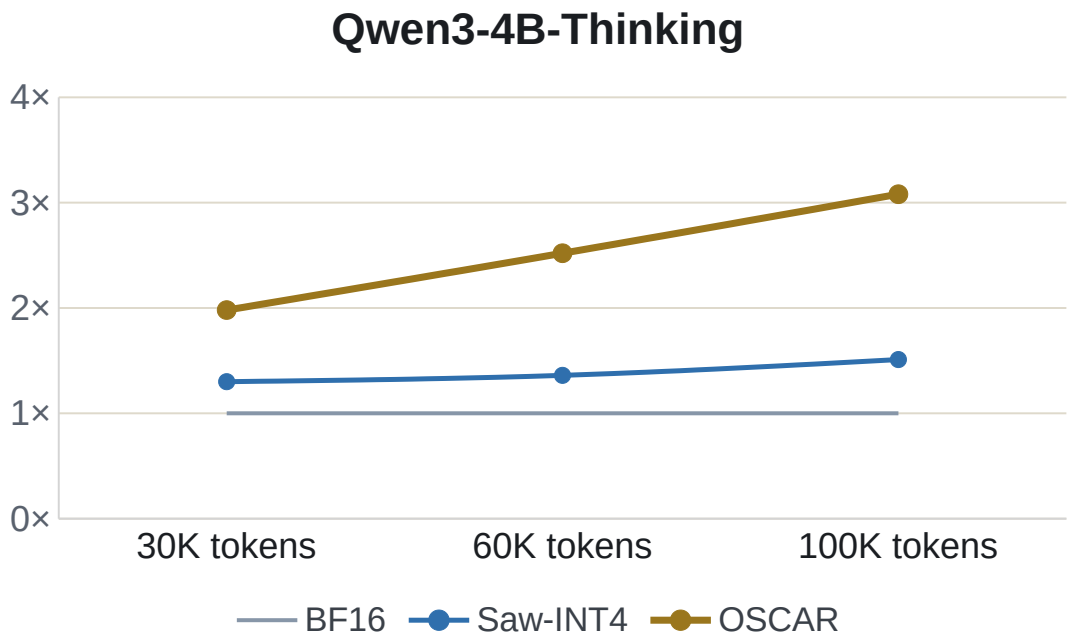
Serving speed on H100

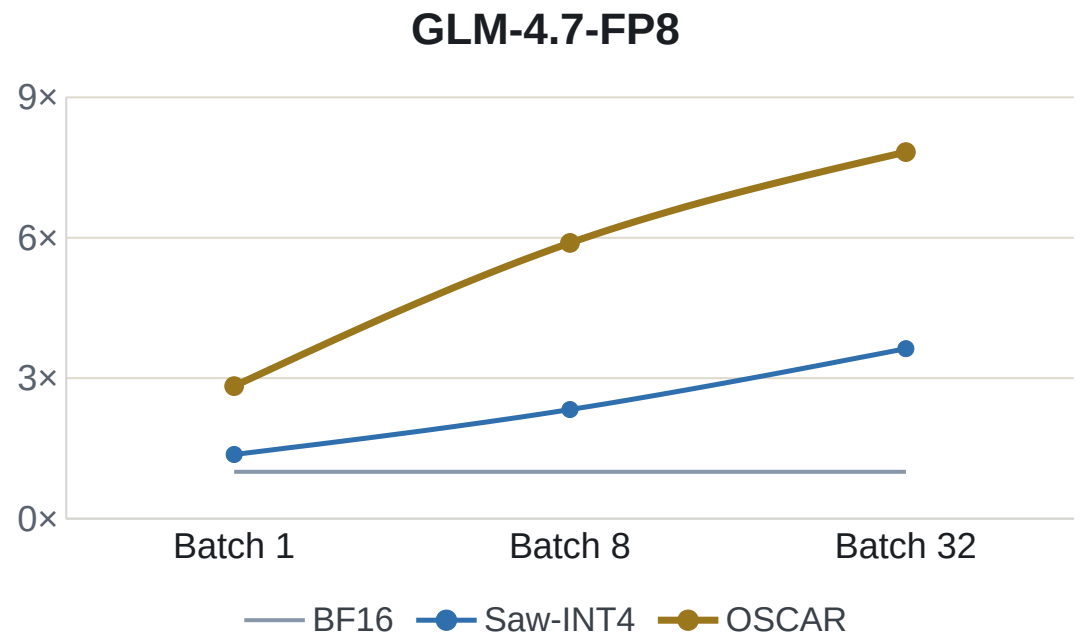
3.08×

Batch-1 decode at 100K

7.83×

Job throughput at 100K, batch 32





Relative to BF16 on H100 with prefix reuse. Decode uses 1K output tokens.

Keys are more sensitive in the 64K diagnostic

Qwen3-8B · WikiText-2 prefix at 64K · 512 scored tokens

Quantized tensors	Perplexity	Attention KL
BF16 reference	8.217	0
Values only	8.218	0.0039
Keys only	**23.318**	**1.1371**
Keys and values	21.862	1.0867

Value-only INT2 stays near BF16 in this test. Key-only INT2 shifts attention substantially.

Separate 128K RULER-NIAH test on Qwen3-8B. OSCAR scores 45.0 versus 78.2 with BF16.

Takeaway and resources

Keys preserve logits.
Values preserve the weighted output.

Two offline rotations support grouped INT2 history with small BF16 windows.



FOSCAR and OSCAR2 are cooked cooked!

Faster, more accurate KV compression

Scan for project details

oscar-quantize.github.io

Method, results and code