

MOAQ: A Multi-Scale Outlier-Aware Framework for Extreme Quantization of Large Language Models

Jiayi Tan¹, Xudong Li¹, Yan Zhang^{1*}, Jiayi Ji¹,
Xiawu Zheng¹, Yunhang Shen², Ke Li², Liujuan Cao¹

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China

²Tencent Youtu Lab, Shanghai, China
bzhy986@gmail.com

Abstract

Low-bit quantization is crucial for deploying Large Language Models (LLMs) on resource-constrained hardware. However, existing Post-Training Quantization (PTQ) methods adopt a monolithic view of outliers, overlooking their global spatial patterns within LLM architectures and weights that are quantization-sensitive but not numerically large. To overcome these challenges, we introduce MOAQ, a novel finetuning-free weight-only PTQ framework based on a new perspective for outlier analysis. MOAQ holistically addresses the outlier problem by three core components: (1) Spatial Calibration (SC) uses a clustering-optimized Hadamard matrix to reorganize disordered outliers into both discrete and clustered patterns. (2) Local Awareness and Suppression (LAS) adopts a dual strategy guided by our outlier metric. It assigns higher bit-widths to clustered outliers via Clustered Outlier Bit Search (COBS), enabling flexible fractional-bit quantization, while further refining quantization parameters for local discrete outliers through Local Outlierness-Aware Calibration (LOAC). (3) Global Feedback and Denoising (GFD) introduces a dynamic correction mechanism informed by our outlier metric to mitigate the progressive cross-layer enrichment of outliers. Extensive experiments show that MOAQ achieves state-of-the-art performance, outperforming both finetuning-free and many finetuning-based methods with substantially lower computational cost.

1 Introduction

Recent advances in artificial intelligence have driven rapid progress in model architectures and learning methodologies (Ye et al., 2025; Li et al., 2024a; Loshchilov et al., 2025; Li et al., 2024b). Meanwhile, these models have demonstrated increasingly strong adaptability and general capabilities across diverse tasks (Behrouz et al., 2026; Li

et al., 2025a; Snell et al., 2025; Li et al., 2026a). These advances, however, are accompanied by growing computational demands.

Due to the limited memory capacity of consumer-grade devices, the community has an urgent demand for more aggressive low-bit quantization methods. However, under such configurations, the performance of quantized models degrades significantly (Frantar et al., 2022; Lin et al., 2024b). The widespread presence of outliers in weight matrices is a major factor constraining quantization performance (Zhang et al., 2024a). To address outliers, recent research (Lee et al., 2024; Huang et al., 2024a) has primarily drawn inspiration from the pruning methods (Frantar and Alistarh, 2023), identifying and mitigating outliers based on sensitivity scores of weights.

As artificial intelligence models become increasingly capable, recent studies have pursued greater computational efficiency (Bian et al., 2026; Li et al., 2026b) and more reliable evaluation across diverse application settings (Li et al., 2025c,b; Wang et al., 2026). Against this broader drive, low-bit quantization provides an important route to efficient model deployment, yet widespread outliers in weight matrices remain a key bottleneck to model performance (Zhang et al., 2024a). Although existing methods show promise in improving quantization performance, they still face three key challenges in handling outliers from local and global perspectives: **1) Local Outlier Misidentification.** Some methods (Shao et al., 2023) narrowly define outliers as large magnitude values, overlooking elements with normal magnitudes but high quantization sensitivity. Other methods (Lee et al., 2024; Huang et al., 2024a) adopt sensitivity metrics inspired by pruning (Frantar and Alistarh, 2023), combining weight magnitudes with Hessian information to identify outliers. However, these metrics do not explicitly model the difference between weight removal in pruning and weight replacement in quan-

*Corresponding author.

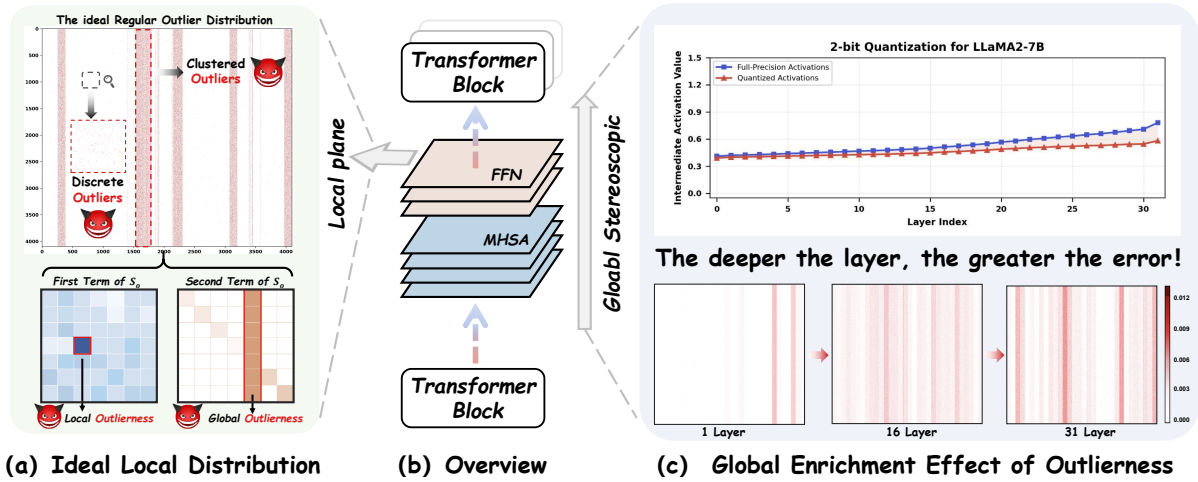


Figure 1: (a) Ideal regular local outlier distribution, where discrete and clustered outliers are reflected by the local and global terms of the outlieriness score S_o . In practice, local outlier layouts are not always regular and can also appear spatially disordered; empirical examples of regular and disordered cases are provided in Fig. 5. (b) Overview of the multi-scale analytical perspective in MOAQ, covering the local plane and the global stereoscopic space across Transformer blocks. (c) Global enrichment effect of outlieriness, where quantization errors become more pronounced in deeper layers. MHSA and FFN are the abbreviations of Multi-Head Self-Attention layer and Feed-Forward Network layer.

tization. **2) Local Inefficient Bit-Width Allocation for Identified Outliers.** Current structured mixed-precision methods (Lee et al., 2024; Huang et al., 2024b) employ rigid, coarse-grained strategies. They assign a uniform bit-width like fixed 1 or 3 bit to entire regions containing outliers. This “one-size-fits-all” approach prevents fine-grained, importance-aware bit allocation, resulting in a sub-optimal trade-off between accuracy and resource efficiency. **3) Myopic Optimization That Ignores Global Factors.** Most methods (Sun et al., 2024; Guo et al., 2024) optimize quantization at the single layer level, making this localized strategy overlook the global factor like the dynamic amplification of quantization errors.

To address these challenges, we revisit the outlier problem from a novel multi-scale analytical perspective based on our outlieriness metric, revealing two key findings (detailed proofs in Appendix B): **1) Duality in Planar Space:** From a planar perspective (tensor level), the spatial distribution of outliers exhibits a clear duality. Although Fig. 1(a) illustrates an ideal regular local distribution for clarity, real linear layers are not always in this ideal state: outliers in some layers are spatially disordered, whereas those in others form regular discrete and clustered patterns. Empirical examples of both cases are provided in Appendix 5. This finding indicates that single-pattern strategies are likely suboptimal, as they cannot efficiently handle both disordered and regularized outlier distributions. **2)**

Cumulative Effect in Stereoscopic Space: From a global layer-wise perspective, outlieriness is affected not only by local Hessian-related weight perturbations but also by the global dynamics of the quantization process, as theoretically derived in Appendix B. Dynamic outlieriness inherited from preceding layers can be transformed and enriched by the current layer, causing abnormality to become progressively stronger with depth, as shown in Fig. 1(c).

Based on these findings, we propose a finetuning-free weight-only PTQ framework MOAQ. It consists of Spatial Calibration (SC), Local Awareness and Suppression (LAS), and Global Feedback and Denoising (GFD). Firstly, **to address Challenge 1**, we propose a novel outlier metric that accounts for the fundamental differences between quantization and pruning, as well as the global dynamics of the quantization process. **To address Challenge 2**, we develop the SC and LAS modules inspired by the local duality of spatial distributions (Finding 1). SC reorganizes disordered outliers into discrete and clustered patterns, transforming the original dual outlier distributions into regularized forms, and LAS jointly optimizes bit-width allocation and quantization parameters for discrete and clustered outliers in local linear layers. **To address Challenge 3**, we create the GFD module. Inspired by the global enrichment effect of outlieriness (Finding 2), it introduces a dynamic correction mechanism based on our new metric to effectively address

the problem of local myopia in existing methods. Through the coordinated operation of these three modules, MOAQ provides a comprehensive solution to outlier-related challenges in PTQ, achieving significant improvements in both quantization accuracy and efficiency. Our contributions can be summarized as follows:

- We systematically diagnose the limitations of existing PTQ methods by pinpointing three critical challenges from both local and global perspectives: local outlier misidentification, inefficient bit-width allocation for identified outliers, and the neglect of global optimization factors.
- We introduce a novel analysis of outliers, revealing two critical findings: the duality of outlier distributions at the local level and the global cumulative effect of outlieriness.
- We propose a finetuning-free weight-only PTQ framework that addresses outliers by three core modules: the SC module reorganizes disordered outliers into discrete and clustered patterns, the LAS module enables flexible bit allocation and local parameter refinement, and the GFD module mitigates progressive enrichment of outlieriness via dynamic correction.

2 Related Work

2.1 Post-Training Quantization of LLMs

Due to the presence of outliers, post-training quantization (PTQ) of large language models (LLMs) is a challenging task (Ma et al., 2024b). Existing PTQ methods for LLMs can be categorized into weight-only quantization (Lin et al., 2024b; Park et al., 2022) and weight-activation quantization (Lin et al., 2024a; Ma et al., 2024a; Liu et al., 2023) based on whether activations are quantized. For weight-only quantization, some methods (Guo et al., 2024; Zhang et al., 2024a; Chee et al., 2023) attempt to mitigate numerical outliers by iterative optimization and linear transformations. Others (Zhang et al., 2024b; Yu et al., 2025) have employed low-rank decomposition to compensate for quantization errors of weight matrix. Unlike weight-only quantization, weight-activation quantization needs to address outliers in both the weight matrix and activation matrix. Current weight-activation quantization methods primarily rely on finetuning-free

online transformation (Ashkboos et al., 2024; Yi et al., 2024) and finetuning-based offline transformation (Liu et al., 2024; Hu et al., 2025) to mitigate outliers. Given the massive scale of LLMs, memory access becomes a key bottleneck for deploying. Weight-only quantization can effectively overcome this issue by compressing the weight matrix to low-bit representations. Therefore, our work primarily focuses on low-bit weight-only quantization.

2.2 Fractional-Bit Quantization

Fractional-bit quantization provides finer control over model size and can improve the utilization of a fixed storage budget (Adepu et al., 2024). Currently, some mixed-precision quantization methods indirectly achieve fractional-bit quantization by isolating outliers in the weight matrix and storing them in high precision. For example, SqueezeLLM (Kim et al., 2023) and SpQR (Dettmers et al., 2023) identify elements sensitive to quantization error and store them in full-precision format, achieving a quantization bit-width of 4.27 bits. In binarization scenarios, PB-LLM (Shang et al., 2023) and BiLLM (Huang et al., 2024a) store high-error elements in INT8, achieving 1.7-bit and 1.08-bit quantization respectively. However, these element-wise unstructured methods neglect the clustered distribution of outliers in the weight matrix and are unfriendly to hardware. Considering these factors, our approach utilizes a structured block-level greedy algorithm to adaptively adjust each block’s bit-width based on its numerical fluctuations and intrinsic attributes within a given target bit-width.

3 Methods

In this section, we provide an overview of MOAQ. The overall framework is shown in Fig. 2. Specifically, in Sec. 3.2, we reorganize disordered outliers into regular discrete and clustered patterns by introducing the Spatial Calibration module. Subsequently, in Sec. 3.3, we propose the Local Awareness and Suppression module to jointly handle clustered outliers and discrete outliers. Finally, in Sec. 3.4, for the global enrichment effect of outlieriness, we introduce the Global Feedback and Denoising module to denoise during the quantization process.

3.1 Outlieriness Score Metric

Existing quantization methods mainly identify outliers using two restricted criteria. The first category narrowly defines outliers as large-magnitude

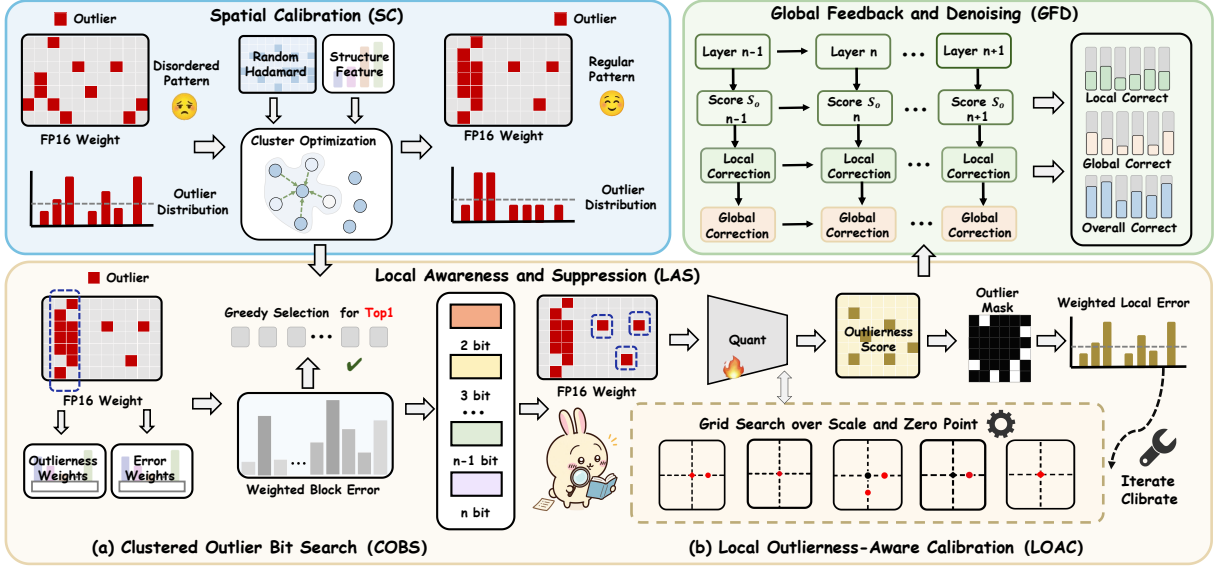


Figure 2: Overview of MOAQ. Given the FP16 weights as input, Spatial Calibration (SC) first reorganizes disordered outliers into regular discrete and clustered patterns. Local Awareness and Suppression (LAS) then allocates flexible bit-widths to clustered outliers through Clustered Outlier Bit Search (COBS) and refines quantization parameters for local discrete outliers through Local Outlierness-Aware Calibration (LOAC). Finally, Global Feedback and Denoising (GFD) mitigates cross-layer enrichment of outlierness through local and global dynamic correction, ultimately producing the quantized weights.

weights (Shao et al., 2023; Zhang et al., 2024a). This criterion is based on the intuition that weights with larger magnitudes may induce larger numerical deviations after quantization. The second category is inspired by element saliency metrics in pruning and estimates the importance of each weight element using both the weight norm and Hessian information (Frantar and Alistarh, 2022, 2023). These two types of metrics can be written as follows:

$$S_{\text{mag}}(i, j) = w_{i,j}^2, \quad S_{\text{pru}}(i, j) = \frac{w_{i,j}^2}{[H_\ell^{-1}]_{j,j}^2}. \quad (1)$$

Here, S_{mag} denotes the magnitude-based outlier score, and S_{pru} denotes the pruning-inspired Hessian-aware score. The former ignores the effect of quantization-induced activation perturbations on output errors, while the latter simplifies quantization replacement as weight removal, overlooking the fundamental difference between pruning and quantization in their perturbation forms. To address these limitations, we propose a new outlierness score metric S_o . For the ℓ -th linear layer, MOAQ defines this metric as:

$$S_o(i, j) = \frac{(w_{i,j} - Q(w_{i,j}))^2}{[H_\ell^{-1}]_{j,j}} + \|w_{i,j} (X_{\ell,j} - \hat{X}_{\ell,j})\|_2^2. \quad (2)$$

Here, $Q(w_{i,j})$ denotes the quantized value of

$w_{i,j}$, and H_ℓ denotes the Hessian matrix of the ℓ -th layer. The first term measures the local error caused by the weight quantization residual, while the second term measures the global influence of preceding-layer quantization errors on the current activation. Thus, S_o integrates local weight perturbation and preceding-layer-induced global activation discrepancy into a unified outlierness metric. The detailed derivation is provided in Appendix A.

3.2 Spatial Calibration

Outliers in weight matrices exhibit disordered spatial layouts in some linear layers, while forming relatively regular discrete or clustered patterns in others. Directly applying quantization to such heterogeneous layouts can disturb the subsequent optimization process. SC performs this reorganization with a clustering-optimized Hadamard transform (Ashkboos et al., 2024). Firstly, SC extracts structural features from each Hadamard basis vector, including local pattern coherence, positive-negative balance, and pattern complexity. Basis vectors with similar structures are grouped by local K-means clustering. SC then sorts basis vectors within each cluster by their distance to the cluster centroid and orders clusters by their average pattern complexity, yielding the optimized Hadamard matrix. Given a weight matrix W and input activation X , this transformation process can be expressed

as:

$$Y = WX = (WH)(H^{-1}X) \quad (3)$$

where H denotes the optimized Hadamard transform matrix. Since the transform and its inverse are applied to the weight and activation sides, respectively, the original computation remains unchanged. Orthogonal mixing redistributes concentrated large-magnitude outliers across transformed dimensions, while the clustering-based basis order preserves more coherent local structures. After SC, the outlier layout becomes more regular and is passed to LAS for local processing. The theoretical support for this reorganization is provided in Appendix B, and more implementation details of the SC module are provided in Appendix D.

3.3 Local Awareness and Suppression

3.3.1 Clustered Outlier Bit Search

For clustered outliers, Clustered Outlier Bit Search (COBS) partitions the weight matrix into blocks with m columns and computes a quantization-sensitivity score for each block based on the outlieriness score matrix S_o :

$$B = \sum_{i=1}^n \sum_{j=1}^m S_o(i, j) \quad (4)$$

where $S_o(i, j)$ denotes the outlieriness score at row i and column j of the current block, and n and m denote the block size. Block importance B reflects the accumulated outlieriness of each weight block and its sensitivity to quantization error. The COBS module uses this block outlieriness score to weight the quantization error in the greedy optimization target, ensuring that each optimization step prioritizes weight blocks with larger accumulated outlieriness and higher quantization sensitivity. The greedy optimization target and constraints in the COBS module are defined as follows:

$$\text{Obj} : \underset{b_1, \dots, b_n}{\text{argmin}} \sum_{i=1}^n (B_i)^\alpha \cdot \|W_i - \text{quant}_{b_i}(W_i)\|_2^2 \quad (5)$$

$$\text{Constraint} : \frac{1}{n} \sum_{i=1}^n b_i \leq \text{target_bit} \quad (6)$$

where W_i , B_i , and b_i denote the weight matrix, block outlieriness score, and bit-width of the i -th block, respectively. The coefficient α controls the weight of block outlieriness in the greedy process, and we set it to 2. COBS repeats three steps until the average bit-width reaches the preset target. **1) Initialization:** all weight blocks are assigned the same low bit-width, typically 1 or 2 bits. **2)**

Greedy optimization: COBS tentatively increases each block by 1 bit and computes the reduction in weighted quantization loss based on Eq. 5. **3)**

Greedy selection: the block with the largest loss reduction receives the actual bit-width update.

It is worth noting that in the COBS module, the preset target bit-width can be either integer or fractional. This unique advantage enables users to flexibly adjust the quantization bit-width such as 2.1-bit based on the actual storage space of hardware devices, so as to maximize the storage space utilization and the performance of the quantized model.

3.3.2 Local Outlierness-Aware Calibration

Local Outlierness-Aware Calibration (LOAC) uses outlieriness score S_o to weight the quantization error when searching for the scaling factor and zero point. Given the original weight W and quantized weight $\hat{W}$, LOAC computes the outlieriness-aware weighted error matrix as:

$$\mathcal{L} = S_o^P \odot \left(\text{quant}(\hat{W}, s, z) - W \right)^2 \quad (7)$$

where P controls the sharpness of the outlieriness weighting. The scaling factor and zero point are denoted by s and z . LOAC applies the 3- σ rule to $\mathcal{L}$ and obtains masks M_o and M_u for outliers and normal elements, which are used to compute weighted errors for the two regions separately.

To avoid the time and memory cost of gradient-based optimization, LOAC searches for s and z over a constrained grid:

$$\underset{s, z}{\text{argmin}} \lambda_1 \mathcal{L}_u(s, z) + \lambda_2 \mathcal{L}_o(s, z) \quad (8)$$

$$s = \frac{\alpha_t (\max(W) - \min(W))}{2^b - 1} \quad (9)$$

$$z = - \left\lceil \frac{\alpha_t \min(W)}{s} \right\rceil \quad (10)$$

where $\mathcal{L}_u(s, z)$ and $\mathcal{L}_o(s, z)$ are obtained by multiplying $\mathcal{L}$ with M_u and M_o , respectively. The coefficients λ_1 and λ_2 balance the normal and outlier regions, and we set them to 1 and 1.2. The grid coefficient $\alpha_t = l + \frac{tR}{T}$ controls the candidate scaling range at step t , where R , l , and T denote the search range, left endpoint, and grid resolution. We set them to 0.8, 0.2 and 100.

3.4 Global Feedback and Denoising

SC and LAS regularize local outliers within each layer, but they remain focused on layer-local distributions and quantization parameters rather than the cross-layer enrichment phenomenon. In sequential

quantization, the error inherited from preceding layers can be amplified by high-outlierness weights in the current layer, causing outlierness to be progressively enriched with depth. Global Feedback and Denoising (GFD) addresses this enrichment path through a dynamic correction mechanism guided by our outlier metric.

For the ℓ -th layer, let W_ℓ and X_ℓ be the full-precision weight matrix and input activation, and let $\widehat{W}_\ell$ and $\widehat{X}_\ell$ be their quantized counterparts. GFD defines an outlierness-aware correction objective using the score matrix S_o :

$$\underset{\widehat{W}_\ell}{\operatorname{argmin}} \|(S_o \odot W_\ell)X_\ell - (S_o \odot \widehat{W}_\ell)\widehat{X}_\ell\|_2^2 \quad (11)$$

This objective uses S_o to reweight output discrepancies, allowing GFD to focus on deviations at high-outlierness positions rather than treating layer residuals without explicit outlierness weighting, as in prior layer-wise error-compensation methods (Li et al., 2025d; Arai and Ichikawa, 2025). We further derive the final denoising correction δ_i for the i -th weight column using the Lagrangian function.

$$\delta_i^{local} = \frac{S_{o,i}(quant(w_i) - w_i)}{H_{i,i}^{-1}} H_{i,i}^{-1} \quad (12)$$

$$\delta_i^{global} = (S_o \odot W_\ell^{row}) \Delta X_\ell \widehat{X}_\ell^\top H_{-i}^{-1} \quad (13)$$

$$\delta_i = \delta_i^{local} + \alpha_\ell \delta_i^{global} \quad (14)$$

where δ_i^{local} corrects the local perturbation from quantizing w_i , and δ_i^{global} provides the global feedback for the cross-layer enrichment path. The final denoising correction δ_i injects δ_i^{global} through the global correction factor α_ℓ , allowing GFD to correct both local quantization error and the cross-layer enrichment path. We define α_ℓ using the inter-block change of quantized activation error, with the ℓ -th block error computed as:

$$D_\ell = \frac{1}{NL} \sum_{n,l} \|X_\ell^{(n,l)} - \widehat{X}_\ell^{(n,l)}\|_2 \quad (15)$$

where N and L denote the number of samples and the sequence length, respectively. D_ℓ measures the deviation between the full-precision input activation and the quantized input activation of the ℓ -th block. The first Transformer block has no inherited error from preceding quantized blocks. We set $\alpha_1 = 0.2$ and initialize the first nontrivial feedback factor as $\alpha_2 = 0.2$. For each block after the second, GFD derives the global correction factor from the relative change of quantized activation error across adjacent blocks:

$$\alpha_\ell = \frac{D_\ell}{D_{\ell-1}} \alpha_{\ell-1}, \quad \ell > 2 \quad (16)$$

This mechanism adjusts the feedback strength according to the relative change of quantized activation error across adjacent blocks. Thus, GFD responds to how the inherited error is enriched along depth, rather than using a uniform correction coefficient for all blocks. As the quantization proceeds layer by layer, each corrected layer provides a less perturbed quantized input for the following layer. This progressively weakens the chain in which quantization errors are repeatedly transformed by high-outlierness weights and accumulated with depth.

4 Experiments

4.1 Experimental Setup

Models and Baselines. We apply MOAQ to pre-trained LLMs including LLaMA (Touvron et al., 2023a), LLaMA-2 (Touvron et al., 2023b) and LLaMA-3 (Dubey et al., 2024) family. These models are representative open-source LLMs, enabling research across diverse model scales and architectures. We compared our method against both finetuning-free and finetuning-based weight-only PTQ methods. For finetuning-free PTQ methods, we compare MOAQ with Round-To-Nearest(RTN), GPTQ (Frantar et al., 2022), AWQ (Lin et al., 2024b), LeanQuant (Zhang and Shrivastava, 2024), Slim-LLM (Huang et al., 2024b), QuIP (Chee et al., 2023), PB-LLM (Shang et al., 2023). For finetuning-based PTQ methods, we compare MOAQ with OmniQuant (Shao et al., 2023), AffineQuant (Ma et al., 2024a) and decoupleQ (Guo et al., 2024).

Datasets and Evaluation. First, we evaluate the language modeling capabilities of quantized models on language generation tasks based on the WikiText2 datasets (Merity et al., 2016). Then we conduct experiments on zero-shot evaluation tasks including WinoGrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020), ARC-Challenge and ARC-Easy (Clark et al., 2018), and HellaSwag (Zellers et al., 2019).

Experimental Details. We implement MOAQ based on Hugging Face (Wolf et al., 2019) and PyTorch (Paszke et al., 2019). All experiments were conducted on NVIDIA A800 80GB GPUs.

Bits	Method	Tuning-free	WikiText2 perplexity ↓									Zero-shot avg. acc. ↑						
			7B	13B	30B	65B	2-7B	2-13B	2-70B	3-8B	3-70B	7B	13B	30B	2-7B	2-13B	3-8B	3-70B
16	FP16	-	5.68	5.09	4.10	3.53	5.47	4.88	3.31	5.75	2.9	68.6	70.96	74.52	69.18	71.75	72.87	79.94
3.0	RTN	✓	7.01	5.88	4.87	4.24	6.66	5.51	3.97	29.91	11.84	47.99	45.99	51.77	62.06	65.77	58.72	65.29
3.0	GPTQ	✓	6.55	5.62	4.80	4.17	6.29	5.42	3.85	8.19	5.22	39.65	54.63	57.76	62.48	66.18	60.58	71.28
3.0	AWQ	✓	6.46	5.51	4.63	3.99	6.24	5.32	3.74	8.22	4.81	32.26	35.44	35.07	62.82	66.14	64.82	73.65
3.0	SliM-LLM	✓	6.40	5.48	4.61	3.99	6.24	5.26	3.67	7.16	4.08	48.36	59.02	64.21	66.42	65.10	65.25	73.35
3.0	LeanQuant	✓	6.62	5.76	4.72	4.10	6.61	5.66	3.91	7.88	5.02	63.85	64.88	63.14	61.06	65.18	54.54	72.20
3.0	OmniQuant	✗	6.15	5.44	4.56	3.94	6.03	5.28	3.78	7.12	4.77	63.89	67.91	72.04	62.42	66.18	64.09	71.90
3.0	AffineQuant	✗	6.14	5.45	4.59	4.01	6.08	5.28	3.85	7.20	4.74	63.82	67.58	72.45	64.02	67.68	60.26	72.61
3.0	decoupleQ	✗	6.38	5.60	4.67	6.05	6.22	5.72	3.84	7.22	4.51	62.85	66.72	70.02	63.17	66.82	58.20	69.04
3.0	MOAQ	✓	5.97	5.34	4.43	3.77	5.78	5.11	3.60	6.99	3.99	66.65	69.09	73.20	67.26	69.76	70.27	78.58
2.0	RTN	✓	1.9e3	781.2	68.04	15.08	4.2e3	122.08	27.27	1.9e3	4.6e5	35.97	35.12	35.98	35.60	35.56	35.87	39.14
2.0	GPTQ	✓	152.31	20.44	13.01	9.51	60.45	28.14	8.78	210	11.9	36.88	55.24	56.90	46.31	40.49	40.12	45.42
2.0	AWQ	✓	2.6e5	2.8e5	2.4e5	7.4e4	2.2e5	1.2e5	4.6e4	1.7e6	1.7e6	32.47	36.35	35.28	37.30	34.89	35.62	44.11
2.0	QuIP	✓	29.74	12.48	11.57	7.83	39.73	13.48	6.64	84.97	13.03	36.12	47.37	54.15	45.24	49.73	43.06	49.43
2.0	PB-LLM	✓	24.61	17.73	12.65	7.85	25.37	49.81	NAN	44.12	11.68	39.06	44.23	52.19	44.12	43.16	48.75	50.18
2.0	SliM-LLM	✓	14.58	8.87	7.33	5.90	16.01	9.41	6.28	39.66	9.46	41.14	56.45	59.25	46.71	52.80	48.02	53.77
2.0	LeanQuant	✓	18.53	14.42	12.68	8.84	25.59	24.43	7.92	41.78	10.32	49.86	53.38	58.70	42.43	48.00	38.98	48.21
2.0	OmniQuant	✗	9.71	7.93	7.12	5.95	11.06	8.26	6.55	31.40	9.02	47.32	56.68	59.87	46.98	53.56	52.66	60.06
2.0	AffineQuant	✗	13.51	7.22	6.49	5.88	10.87	7.64	6.62	37.27	9.13	48.55	52.56	62.00	37.20	49.50	36.03	58.87
2.0	decoupleQ	✗	9.49	7.86	6.37	5.59	9.74	13.03	5.23	35.32	9.18	48.20	55.41	63.29	47.69	54.62	53.86	62.75
2.0	MOAQ	✓	7.52	6.51	5.63	4.86	7.40	6.30	4.93	10.76	7.35	57.94	62.15	67.64	57.69	62.29	54.58	71.61
2.1	MOAQ	✓	7.24	6.28	5.44	4.75	7.08	6.10	4.66	9.91	6.98	58.44	63.82	68.69	58.95	63.47	58.24	73.68

Table 1: Main results of evaluation experiment. We report the perplexity and zero-shot accuracy. The best score is **bolded**.

SC	COBS	LOAC	GFD	WikiText2 perplexity ↓			
				1-7B	1-13B	2-7B	2-13B
✗	✓	✓	✓	9.75	8.84	10.08	8.37
✓	✗	✓	✓	9.32	7.25	8.56	7.44
✓	✓	✗	✓	9.09	8.67	8.84	8.68
✓	✓	✓	✗	9.98	7.83	9.67	8.33
✓	✓	✓	✓	7.52	6.51	7.40	6.30

Table 2: Ablation study of main components in MOAQ. The evaluation was conducted by removing individual components and measuring the WikiText2 perplexity.

Method	2-7B		2-13B		2-70B	
	Time	Mem	Time	Mem	Time	Mem
OmniQuant	7.27	14.35	10.68	19.91	24.88	39.69
AffineQuant	7.11	23.49	11.17	30.60	23.92	130.26
Ours	0.65	11.27	1.11	14.67	5.26	33.23
Ours(2.1bit)	0.68	11.32	1.16	14.69	5.81	33.24

Table 3: Comparison of quantization time (hours) and memory cost (GB) with finetuning-based PTQ methods. Our approach consistently achieves the lowest computational cost.

4.2 Main Results

We firstly test the performance of MOAQ on the LLaMA family. As shown in Table 1, our method consistently outperforms previous state-of-the-art (SOTA) methods across all configurations and models. Specifically, in language generation tasks, compared with the best baseline method, our method achieves average perplexity reductions of 3.38 and 0.14 points under 2-bit and 3-bit quantization configurations, respectively. On the most challenging LLaMA-3-8B model, MOAQ reduces perplexity by 20.64 points compared with the best baseline method under the 2-bit quantization config-

Method	Bits	1-7B	1-13B	1-30B	1-65B
OWQ	3.1	6.41	5.56	4.63	4.09
SliM-LLM	3.0	6.40	5.48	4.61	3.99
COBS	3.0	6.23	5.45	4.54	3.92

Table 4: Perplexity performance on WikiText2 datasets of different structured bit allocation strategies on LLaMA model families.

Method	Qwen1.5-7B		Qwen2-7B		Qwen2.5-7B	
	2-bit	3-bit	2-bit	3-bit	2-bit	3-bit
GPTQ	37.34	9.28	23.61	7.85	25.8	7.98
OmniQuant	29.02	9.18	14.39	7.96	14.73	7.87
Ours	11.04	8.44	9.28	7.48	9.46	7.35

Table 5: Perplexity on WikiText2 datasets of Qwen model families when quantized to 2-bit and 3-bit by different methods.

uration. In zero-shot tasks, our method achieves average accuracy improvements of 6.45 and 2.51 percentage points under 2-bit and 3-bit quantization configurations, respectively, compared with the best baseline method. These significant performance improvements demonstrate the effectiveness of our method. It is worth noting that MOAQ is a finetuning-free PTQ method, which requires fewer computational resources to achieve better quantization performance compared with the finetuning-based PTQ methods in Table 1.

Furthermore, we found that in low-bit quantization scenarios, for quantized models, a slight increase in the quantization bit-width can further improve their performance. For example, the 2.1-bit quantized LLaMA-3-70B version of MOAQ reduces perplexity by 5.0% and improves zero-shot

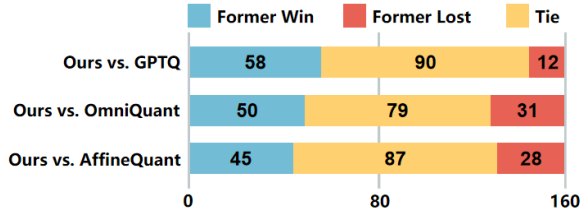


Figure 3: GPT-4 evaluation on the Vicuna.

accuracy by 2.07 percentage points compared with its 2-bit quantized version. This indicates that in low-bit quantization deployment scenarios, if hardware devices have available storage capacity, appropriately increasing the bit-width of the quantized model using MOAQ can enhance the model performance.

4.3 Ablation Study

Effectiveness of Individual Components. In this section, we conduct a detailed ablation study to evaluate the impact of each component on the performance of MOAQ. As shown in Table 2, we observe that when removing the SC, COBS, and LOAC modules, the perplexity of our method on different LLMs increases by 2.07-2.68 points, 0.74-1.80 points, and 1.44-2.38 points respectively. This indicates that effectively focusing on the spatial distribution patterns and intrinsic attributes of outliers can enhance quantization performance. Meanwhile, we note that when the GFD module is removed, the perplexity of our method rises by 1.32-2.46 points. This further demonstrates the significance of addressing cross-layer enrichment of quantization errors from a global perspective.

Results of Instruction-Following Models. To evaluate its instruction-following capability, we quantized the LLaMA-3-8B-Chat model and evaluated its performance using the Vicuna benchmark (Chiang et al., 2023) and GPT-4 evaluation protocol (Chiang et al., 2023). Previous study (Zheng et al., 2023) has identified an order bias in GPT-4 scoring, where models presented first tend to receive higher ratings. To mitigate this bias in the GPT-4-based evaluation protocol, we conduct 160 pairwise comparisons using two permutation orders and aggregate the results. As shown in Fig. 3, under 2-bit quantization scenarios, MOAQ achieved win-tie rates of 92.5%, 80.6% and 82.5% against GPTQ, OmniQuant and AffineQuant, respectively. These results demonstrate that MOAQ exhibits strong capability in quantizing instruction-finetuned LLMs.

Results of Different Structured Bit Allocation

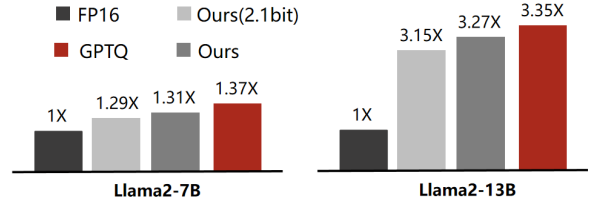


Figure 4: Inference speed comparison.

Strategies. We compared two sensitivity-based structured bit allocation algorithms with the COBS module on the LLaMA model, using the GPTQ quantization framework. As shown in Table 4, COBS consistently achieves optimal performance. This can be attributed to two key factors: 1) It comprehensively considers both outlieriness scores and dynamic quantization errors during bit-width allocation. 2) It adaptively assigns the most suitable bit-width to each region, without being constrained by predefined bit-widths.

4.4 Efficiency and Generalization Analysis

Quantization Speed and Peak Memory. One of the significant advantages of MOAQ is that it is a finetuning-free method, which requires less computational resources compared to finetuning-based methods. To further demonstrate this advantage, we compare the quantization time and peak memory between MOAQ and existing finetuning-based quantization methods in Table 3. Combined with results from Table 1, we can conclude that MOAQ can achieve superior quantization performance at the cost of lower resource overhead.

Results of Inference Efficiency. Due to the potential inference latency that may be introduced by the mixed-precision quantization method in MOAQ, we compared the inference speed differences among MOAQ, GPTQ, and the full-precision model in Fig. 4. The results show that the inference speeds of the 2-bit LLaMA-2-7B and LLaMA-2-13B models quantized by MOAQ are 1.31x and 3.27x those of the full-precision models. Moreover, the inference speeds of our method on these two models are only 4.4% and 2.4% slower than those of the lighter GPTQ method. In addition, we also tested the inference speed of MOAQ in the 2.1-bit quantization scenario, and the results show that the additionally allocated fractional-level bit-widths have nearly negligible impact on the inference performance.

Results of Other Model Families. To evaluate the effectiveness of MOAQ in quantizing other model families, we selected the Qwen series of LLMs

as new quantization targets. Table 5 presents the performance of Qwen1.5-7B (Bai et al., 2023), Qwen2-7B (Yang et al., 2024a), and Qwen2.5-7B (Yang et al., 2024b) after being quantized by MOAQ. These results indicate that the quantization capability of MOAQ is not limited to LLMs with a specific architecture.

5 Conclusion

In this paper, we propose MOAQ, a finetuning-free weight-only PTQ framework for low-bit quantization of LLMs. MOAQ addresses the outlier bottleneck by translating a multi-scale understanding of outliers into a unified framework with local calibration and global feedback. Specifically, the SC module reorganizes disordered outliers into regular discrete and clustered patterns. LAS module uses an outlier-aware strategy to allocate flexible bit-widths to clustered outliers and refine quantization parameters for local discrete outliers. Additionally, GFD module mitigates the progressive enrichment of outliers through dynamic correction. Extensive experiments demonstrate that MOAQ achieves state-of-the-art performance, outperforming finetuning-free methods and many finetuning-based methods with substantially lower computational cost.

Limitations

Although MOAQ demonstrates strong performance on representative open-source LLM families, its evaluation is still constrained by the availability of model weights and internal activations. Recent frontier proprietary LLMs, such as the Claude and GPT series, are typically accessible only through restricted interfaces rather than publicly released full-precision weights. Therefore, MOAQ cannot be directly applied to these highly capable closed-weight models to examine whether the observed quantization benefits transfer to the most advanced industrial-scale systems. This limitation stems from the current accessibility boundary of proprietary LLMs and leaves broader validation on future open frontier models as an important direction.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122701), the National Key Research and Development Program of China (No. 2025YFE0113500), the National Science Fund for Distinguished Young Scholars

(No. 62525605) and the National Natural Science Foundation of China (Nos. 625B2156, U25B2066, 62506313, and U22B2051).

References

- Harshavardhan Adepu, Zhanpeng Zeng, Li Zhang, and Vikas Singh. 2024. Framequant: Flexible low-bit quantization for transformers. *arXiv preprint arXiv:2403.06082*.
- Yamato Arai and Yuma Ichikawa. 2025. Quantization error propagation: Revisiting layer-wise post-training quantization. *arXiv preprint arXiv:2504.09629*.
- Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoeffler, and James Hensman. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. 2026. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543.
- Song Bian, Tao Yu, Shivaram Venkataraman, and Youngsuk Park. 2026. Scaling laws meet model architecture: Toward inference-efficient llms. In *International Conference on Learning Representations*, volume 2026, pages 136664–136695.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.
- Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. 2023. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36:4396–4429.
- Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

- Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefer, and Dan Alistarh. 2023. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Elias Frantar and Dan Alistarh. 2022. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488.
- Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.
- Yi Guo, Fanliu Kong, Xiaoyang Li, Hui Li, Wei Chen, Xiaogang Tian, Jinping Cai, Yang Zhang, and Shouda Liu. 2024. decoupleq: Towards 2-bit post-training uniform quantization via decoupling parameters into integer and floating points. *arXiv preprint arXiv:2404.12759*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Zhihang Yuan, Jiangyong Yu, Chen Xu, Zhe Jiang, and Sifan Zhou. 2025. Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. *arXiv preprint arXiv:2501.13987*.
- Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. 2024a. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*.
- Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. 2024b. Slim-llm: Saliency-driven mixed-precision quantization for large language models. *arXiv preprint arXiv:2405.14917*.
- Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. 2023. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*.
- Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. 2024. Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13355–13364.
- Xudong Li, Timin Gao, Runze Hu, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Jingyuan Zheng, Yunhang Shen, Ke Li, Yutao Liu, and 1 others. 2024a. Adaptive feature selection for no-reference image quality assessment by mitigating semantic noise sensitivity. In *International Conference on Machine Learning*, pages 27808–27821. PMLR.
- Xudong Li, Runze Hu, Jingyuan Zheng, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Ke Li, Yunhang Shen, Yutao Liu, Pingyang Dai, and 1 others. 2024b. Integrating global context contrast and local sensitivity for blind image quality assessment. In *Forty-first International Conference on Machine Learning*.
- Xudong Li, Zihao Huang, Yan Zhang, Yunhang Shen, Ke Li, Xiawu Zheng, Liujuan Cao, and Rongrong Ji. 2025a. Few-shot image quality assessment via adaptation of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10442–10452.
- Xudong Li, Wenjie Nie, Yan Zhang, Runze Hu, Ke Li, Xiawu Zheng, and Liujuan Cao. 2025b. Distilling spatially-heterogeneous distortion perception for blind image quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2344–2354.
- Xudong Li, Mengdan Zhang, Peixian Chen, Xiawu Zheng, Yan Zhang, Jingyuan Zheng, Yunhang Shen, Ke Li, Chaoyou Fu, Xing Sun, and 1 others. 2026a. Zooming from context to cue: Hierarchical preference optimization for multi-image mllms. *Advances in Neural Information Processing Systems*, 38:179–214.
- Xudong Li, Yan Zhang, Yunhang Shen, Ke Li, Runze Hu, Xiawu Zheng, and Sicheng Zhao. 2025c. Feature denoising diffusion model for blind image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5004–5012.
- Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. 2025d. Gptaq: Efficient finetuning-free quantization for asymmetric calibration. *arXiv preprint arXiv:2504.02692*.
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2026b. Eagle-3: Scaling up inference acceleration of large language models via training-time test. *Advances in Neural Information Processing Systems*, 38:136737–136756.
- Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. 2024a. Duquant: Distributing outliers via dual transformation makes stronger quantized

- llms. *Advances in Neural Information Processing Systems*, 37:87766–87800.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024b. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.
- Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. 2023. Qllm: Accuracy and efficient low-bitwidth quantization for large language models. *arXiv preprint arXiv:2310.08041*.
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2024. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.
- Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. 2025. ngpt: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, volume 2025, pages 74014–74038.
- Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. 2024a. Affinequant: Affine transformation quantization for large language models. *arXiv preprint arXiv:2403.12544*.
- Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. 2024b. Outlier-aware slicing for post-training quantization in vision transformer. In *Forty-first International Conference on Machine Learning*.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Gunho Park, Baeseong Park, Minsub Kim, Sungjae Lee, Jeonghoon Kim, Beomseok Kwon, Se Jung Kwon, Byeongwook Kim, Youngjoo Lee, and Dongsoo Lee. 2022. Lut-gemm: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models. *arXiv preprint arXiv:2206.09557*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Yuzhang Shang, Zhihang Yuan, Qiang Wu, and Zhen Dong. 2023. Pb-llm: Partially binarized large language models. *arXiv preprint arXiv:2310.00034*.
- Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2023. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, volume 2025, pages 10131–10165.
- Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, and 1 others. 2024. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Yuyang Wang, Shikai Guo, Zijian Tao, Hui Li, Jiayi Tan, Liying Hao, Fa Zhu, Xingchi Chen, and Moustafa Youssef. 2026. Correvator: Evaluating patch correctness in automated program repair from online community discussions. *IEEE Transactions on Consumer Electronics*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024a. [Qwen2 technical report](#). *ArXiv*, abs/2407.10671.
- Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan

- Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao-ran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024b. [Qwen2.5 technical report](#). *ArXiv*, abs/2412.15115.
- Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2025. Differential transformer. In *International Conference on Learning Representations*, volume 2025, pages 144–164.
- Ke Yi, Zengke Liu, Jianwei Zhang, Chengyuan Li, Tong Zhang, Junyang Lin, and Jingren Zhou. 2024. Rotated runtime smooth: Training-free activation smoother for accurate int4 inference. *arXiv preprint arXiv:2409.20361*.
- Hao Yu, Yang Zhou, Bohua Chen, Zelan Yang, Shen Li, Yong Li, and Jianxin Wu. 2025. Treasures in discarded weights for llm quantization.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Aozhong Zhang, Naigang Wang, Yanxia Deng, Xin Li, Zi Yang, and Penghang Yin. 2024a. Magr: Weight magnitude reduction for enhancing post-training quantization. *arXiv preprint arXiv:2406.00800*.
- Cheng Zhang, Jianyi Cheng, George A Constantinides, and Yiren Zhao. 2024b. Lqer: Low-rank quantization error reconstruction for llms. *arXiv preprint arXiv:2402.02446*.
- Tianyi Zhang and Anshumali Shrivastava. 2024. Leanquant: Accurate and scalable large language model quantization with loss-error-aware grid. *arXiv preprint arXiv:2407.10032*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Derivation of the Outlierness Score Metric

For the ℓ -th linear layer, let W_ℓ and $\widehat{W}_\ell$ denote the full-precision and quantized weight matrices, and let X_ℓ and $\widehat{X}_\ell$ denote the corresponding full-precision and quantized-state input activations. The full-precision output and quantized output are

$$Y_\ell = W_\ell X_\ell, \quad \widehat{Y}_\ell = \widehat{W}_\ell \widehat{X}_\ell. \quad (17)$$

Their output discrepancy can be decomposed as

$$Y_\ell - \widehat{Y}_\ell = W_\ell (X_\ell - \widehat{X}_\ell) + (W_\ell - \widehat{W}_\ell) \widehat{X}_\ell. \quad (18)$$

The first term corresponds to activation perturbations inherited from preceding quantized layers, while the second term is introduced by quantizing the current-layer weights.

We first derive the local cost of the current weight perturbation. For the i -th row, define

$$d_i = (W_{\ell,i:} - \widehat{W}_{\ell,i:})^\top. \quad (19)$$

With $\widehat{X}_\ell$ fixed, consider the regularized mean reconstruction objective

$$\begin{aligned} \overline{\mathcal{L}}_i^w(d_i) &= \frac{1}{2N} \|d_i^\top \widehat{X}_\ell\|_2^2 + \frac{\lambda_\ell}{2} \|d_i\|_2^2 \\ &= \frac{1}{2} d_i^\top \left(\frac{1}{N} \widehat{X}_\ell \widehat{X}_\ell^\top + \lambda_\ell I \right) d_i. \end{aligned} \quad (20)$$

where N denotes the number of calibration tokens or samples and $\lambda_\ell \geq 0$ is the damping coefficient in the normalized curvature. To put both components of the final score on the same unnormalized calibration scale, define

$$\begin{aligned} H_\ell &= N \left(\frac{1}{N} \widehat{X}_\ell \widehat{X}_\ell^\top + \lambda_\ell I \right) \\ &= \widehat{X}_\ell \widehat{X}_\ell^\top + N\lambda_\ell I. \end{aligned} \quad (21)$$

We assume either $\lambda_\ell > 0$ or $\widehat{X}_\ell$ has full row rank. The rescaled local objective is then

$$\mathcal{L}_i^w(d_i) = N \overline{\mathcal{L}}_i^w(d_i) = \frac{1}{2} d_i^\top H_\ell d_i. \quad (22)$$

For a single element $w_{i,j}$, its quantization residual is

$$r_{i,j} = w_{i,j} - Q(w_{i,j}). \quad (23)$$

When this residual is fixed and the remaining elements in the same row are allowed to compensate locally, the constrained optimization problem is

$$\min_{d_i} \frac{1}{2} d_i^\top H_\ell d_i, \quad \text{s.t.} \quad e_j^\top d_i = r_{i,j}, \quad (24)$$

where e_j is the one-hot vector at position j . The Lagrangian is

$$\mathcal{J}(d_i, \mu) = \frac{1}{2} d_i^\top H_\ell d_i + \mu (e_j^\top d_i - r_{i,j}). \quad (25)$$

Setting $\frac{\partial \mathcal{J}}{\partial d_i} = 0$ gives

$$H_\ell d_i + \mu e_j = 0, \quad d_i = -\mu H_\ell^{-1} e_j. \quad (26)$$

Substituting this solution into the constraint yields

$$d_i^* = \frac{r_{i,j}}{[H_\ell^{-1}]_{j,j}} H_\ell^{-1} e_j. \quad (27)$$

The corresponding minimum local cost is

$$\varepsilon_{i,j}^w = \frac{1}{2} \frac{(w_{i,j} - Q(w_{i,j}))^2}{[H_\ell^{-1}]_{j,j}}. \quad (28)$$

Both error sources below use the same rescaling by N . Removing their common factor $1/2$, we obtain the local quantization term:

$$s_{i,j}^w = \frac{(w_{i,j} - Q(w_{i,j}))^2}{[H_\ell^{-1}]_{j,j}}. \quad (29)$$

Next, we consider the activation perturbation term in Eq. 18. This term is not caused by the current weight residual, but by the discrepancy between X_ℓ and $\widehat{X}_\ell$ inherited from preceding quantized layers. Under a diagonal attribution over input dimensions, its mean-square contribution is

$$\bar{\varepsilon}_{i,j}^a = \frac{1}{2N} \|w_{i,j} (X_{\ell,j} - \widehat{X}_{\ell,j})\|_2^2. \quad (30)$$

Multiplying by the same factor N used in Eq. 22 gives $\varepsilon_{i,j}^a = N \bar{\varepsilon}_{i,j}^a$. Removing the same factor $1/2$ as in the local term yields

$$s_{i,j}^a = \|w_{i,j} (X_{\ell,j} - \widehat{X}_{\ell,j})\|_2^2. \quad (31)$$

Combining the local weight quantization term and the global activation perturbation term gives the final outlierness score:

$$S_o(i, j) = \frac{(w_{i,j} - Q(w_{i,j}))^2}{[H_\ell^{-1}]_{j,j}} + \|w_{i,j} (X_{\ell,j} - \widehat{X}_{\ell,j})\|_2^2. \quad (32)$$

This unnormalized convention is equivalent to averaging both components over calibration samples. In particular, if every calibration sample is repeated c times, then $N' = cN$, $H'_\ell = cH_\ell$, and the inherited squared norm is also multiplied by c . Consequently,

$$S'_o(i, j) = c S_o(i, j), \quad (33)$$

so calibration-set replication changes neither the relative weighting of the two components nor the element ranking.

This formulation supports the outlier definition used in MOAQ. The first term uses the actual quantization residual, so it measures the perturbation caused by quantization replacement rather than pruning removal. The second term incorporates the preceding-layer-induced activation discrepancy, allowing the score to account for global effects during sequential quantization. Therefore, S_o identifies abnormal weights according to the actual PTQ trajectory, where local weight perturbations and global activation discrepancies jointly determine outlieriness.

B Spatial Distribution and Enrichment of Outlieriness

Theorem 1 (Two-Scale Support Characterization of Outlieriness). For an LLM linear layer, write the score in Eq. 2 as

$$S_o(i, j) = \frac{q_{i,j}}{h_j} + w_{i,j}^2 a_j, \quad (34)$$

where $q_{i,j} = (w_{i,j} - Q(w_{i,j}))^2$, $h_j = [H_\ell^{-1}]_{j,j}$, and $a_j = \|X_{\ell,j} - \widehat{X}_{\ell,j}\|_2^2$. By Eq. 21, for every nonzero vector v ,

$$v^\top H_\ell v = \|\widehat{X}_\ell^\top v\|_2^2 + N\lambda_\ell \|v\|_2^2 > 0, \quad (35)$$

where strict positivity follows from either $\lambda_\ell > 0$ or the full-row-rank condition on $\widehat{X}_\ell$. Hence $H_\ell \succ 0$, $H_\ell^{-1} \succ 0$, and $h_j > 0$. The two components of S_o have different deterministic spatial responses: increasing one local residual statistic $q_{i,j}$ affects only the score at (i, j) , whereas increasing one inherited-discrepancy statistic a_j changes the scores along column j .

Proof 1. Fix W_ℓ and H_ℓ . If $q_{i,j}$ is increased by any $u \geq 0$ while all other quantities are fixed, direct substitution into Eq. 34 gives

$$\Delta_q S_o(m, n) = \begin{cases} u/h_j, & (m, n) = (i, j), \\ 0, & \text{otherwise.} \end{cases} \quad (36)$$

Thus, a local quantization-residual elevation has point support. If a_j is increased by any $v \geq 0$, then

$$\Delta_a S_o(m, n) = \begin{cases} vw_{m,j}^2, & n = j, \\ 0, & n \neq j. \end{cases} \quad (37)$$

Hence, an inherited activation-discrepancy elevation forms a column-confined carrier for clustered support and raises every nonzero-weight score in that column. We next make the resulting two-scale support statement explicit. Let

$$L_{i,j} = \frac{q_{i,j}}{h_j}, \quad A_{i,j} = w_{i,j}^2 a_j, \quad (38)$$

$$\Omega_\tau(Z) = \{(i, j) : Z_{i,j} > \tau\}.$$

Because L and A are nonnegative, for every $\rho > 0$ and $\theta \in (0, 1)$,

$$\Omega_\rho(L) \cup \Omega_\rho(A) \subseteq \Omega_\rho(S_o) \subseteq \Omega_{\theta\rho}(L) \cup \Omega_{(1-\theta)\rho}(A). \quad (39)$$

The second inclusion follows because $L_{i,j} \leq \theta\rho$ and $A_{i,j} \leq (1-\theta)\rho$ together imply $S_o(i, j) \leq \rho$. If $M_w = \max_{i,j} |w_{i,j}|$ is zero, then $A = 0$ and the inherited-support claim is immediate. Otherwise, $M_w > 0$; define $\mathcal{J}_a(\tau) = \{j : a_j > \tau\}$. Then

$$\Omega_{(1-\theta)\rho}(A) \subseteq \left\{ (i, j) : j \in \mathcal{J}_a\left(\frac{(1-\theta)\rho}{M_w^2}\right) \right\}. \quad (40)$$

If d_{out} is the number of weight rows, Eqs. 39–40 further give

$$|\Omega_\rho(S_o)| \leq |\Omega_{\theta\rho}(L)| + d_{\text{out}} \left| \mathcal{J}_a\left(\frac{(1-\theta)\rho}{M_w^2}\right) \right|. \quad (41)$$

Thus, whenever the local superlevel set is sparse and only a small number of inherited-discrepancy columns cross the indicated threshold, the high-outlieriness set is confined to sparse coordinate-level positions together with a bounded collection of column carriers. This establishes the claimed co-existence of discrete pointwise support and column-confined clustered support under explicit, verifiable conditions.

Theoretical Justification of Spatial Calibration. Let $H_0 = [b_1, \dots, b_d]$ be the Hadamard matrix before SC's two-level ordering, and let $\phi(b_j)$ be the seven-dimensional structural feature vector used by the method. K-means partitions the basis indices into K nonempty sets $\mathcal{C}_1, \dots, \mathcal{C}_K$, with label $z_j = k$ for $j \in \mathcal{C}_k$. Let $\pi_{\text{SC}}(1), \dots, \pi_{\text{SC}}(d)$ be the returned order, let $P_{\pi_{\text{SC}}}$ place basis $b_{\pi_{\text{SC}}(t)}$ at coordinate t , and write $H_{\text{SC}} = H_0 P_{\pi_{\text{SC}}}$.

For a threshold $\rho > 0$, define the high-outlieriness mask after SC by

$$M_{\rho,i,t}^{\text{SC}} = \mathbf{1}[S_o^{\text{SC}}(i, t) > \rho]. \quad (42)$$

For each weight row i and structural cluster k , let $p_{\rho,i,k}$ be the binary cluster representative that minimizes disagreement with this mask. Its cluster projection and total mismatch are

$$p_{\rho,i,k} \in \arg \min_{a \in \{0,1\}} \sum_{t: \pi_{\text{SC}}(t) \in \mathcal{C}_k} |M_{\rho,i,t}^{\text{SC}} - a|, \\ \overline{M}_{\rho,i,t}^{\text{SC}} = p_{\rho,i,z_{\pi_{\text{SC}}(t)}}, \quad D_\rho = \sum_{i,t} |M_{\rho,i,t}^{\text{SC}} - \overline{M}_{\rho,i,t}^{\text{SC}}|. \quad (43)$$

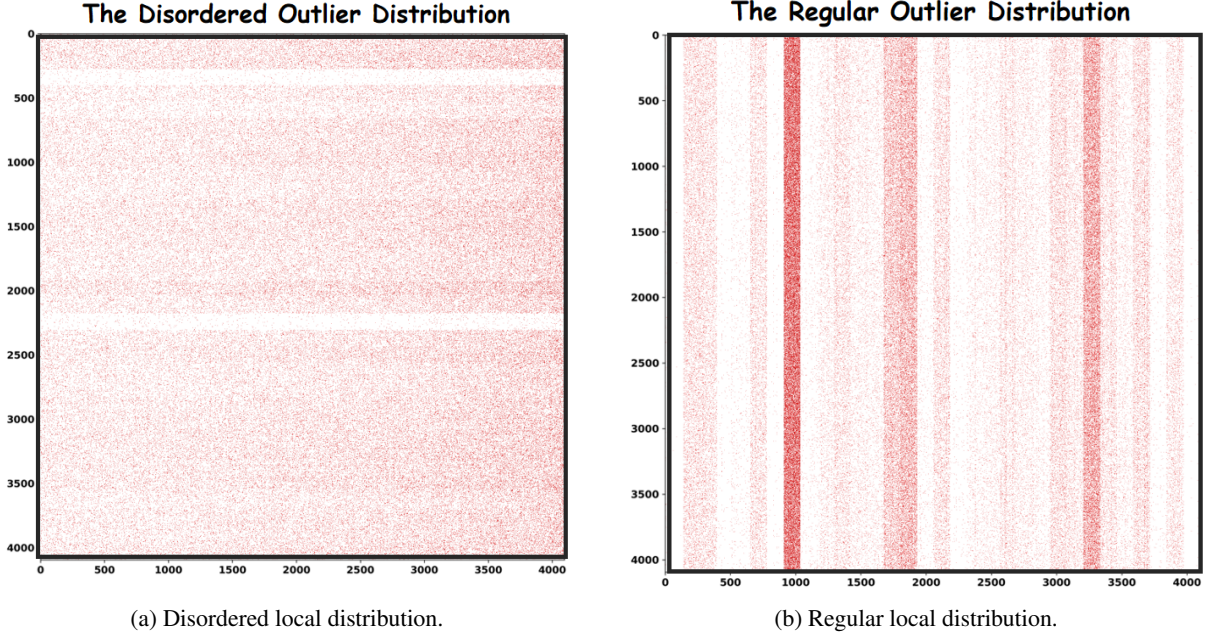


Figure 5: Empirical examples of local outlier distributions from 4096-dimensional projection matrices in LLaMA-2-7B. The disordered case shows high-outlierness elements scattered over the weight plane, while the regular case shows structured column-wise concentration. These examples complement the idealized regular illustration in Fig. 1(a) and motivate MOAQ to handle both disordered and regular local layouts.

For a binary mask M , measure its column-wise fragmentation by

$$\mathfrak{F}(M) = \sum_{i=1}^{d_{\text{out}}} \sum_{t=1}^{d-1} |M_{i,t+1} - M_{i,t}|. \quad (44)$$

Then SC produces a K -block cluster projection of the actual outlierness mask and gives the deterministic regularity bound

$$\mathfrak{F}(M_{\rho}^{\text{SC}}) \leq d_{\text{out}}(K-1) + 2D_{\rho}. \quad (45)$$

Proof. Because the two-level sorting lists all members of one K-means cluster before moving to the next cluster, there is a contiguous interval I_k such that

$$\{t : \pi_{\text{SC}}(t) \in \mathcal{C}_k\} = I_k. \quad (46)$$

Consequently, $\overline{M}_{\rho,i,t}^{\text{SC}}$ is constant on every I_k . Each row of this projected mask can change only at the $K-1$ boundaries between consecutive cluster intervals, and hence

$$\mathfrak{F}(\overline{M}_{\rho}^{\text{SC}}) \leq d_{\text{out}}(K-1). \quad (47)$$

For every row i and adjacent pair $(t, t+1)$, the triangle inequality gives

$$\begin{aligned} |M_{\rho,i,t+1}^{\text{SC}} - M_{\rho,i,t}^{\text{SC}}| &\leq |\overline{M}_{\rho,i,t+1}^{\text{SC}} - \overline{M}_{\rho,i,t}^{\text{SC}}| \\ &\quad + |M_{\rho,i,t+1}^{\text{SC}} - \overline{M}_{\rho,i,t+1}^{\text{SC}}| \\ &\quad + |M_{\rho,i,t}^{\text{SC}} - \overline{M}_{\rho,i,t}^{\text{SC}}|. \end{aligned} \quad (48)$$

Summing over all rows and adjacent columns counts every mismatch at most twice. Combining the result with Eq. 47 proves Eq. 45. The within-cluster sorting further gives

$$\|\phi(b_{\pi_{\text{SC}}(t)}) - \mu_k\|_2 \leq \|\phi(b_{\pi_{\text{SC}}(t+1)}) - \mu_k\|_2 \quad (49)$$

whenever the two adjacent indices are in cluster k , where μ_k is its centroid.

For any ordering π , define the number of structural-cluster transitions by

$$T(\pi) = \sum_{t=1}^{d-1} \mathbf{1}[z_{\pi(t)} \neq z_{\pi(t+1)}]. \quad (50)$$

An ordering that visits all K nonempty clusters must contain at least $K-1$ transitions. Since SC represents every cluster by exactly one contiguous block, it attains this lower bound:

$$T(\pi_{\text{SC}}) = K-1 = \min_{\pi} T(\pi). \quad (51)$$

Finally, since $P_{\pi_{\text{SC}}} P_{\pi_{\text{SC}}}^{\top} = I$,

$$\begin{aligned} (WH_{\text{SC}})(H_{\text{SC}}^{-1}X) &= (WH_0 P_{\pi_{\text{SC}}}) (P_{\pi_{\text{SC}}}^{\top} H_0^{-1}X) \\ &= WX. \end{aligned} \quad (52)$$

Equations 46–45 establish the regularity tendency induced by SC: the cluster-explainable component of the actual high-outlierness mask occupies at most K contiguous structural carriers,

while within-cluster deviations contribute additively through D_ρ . Equations 51 and 52 further show that this organization attains the minimum structural transition count and preserves the full-precision layer function. The resulting layout is aligned with the contiguous blocks used by COBS and the local calibration performed by LOAC.

Theorem 2 (Cross-Layer Enrichment Tendency of Outlierness Exposure). For the l -th layer, let $\delta_l = X_l - \hat{X}_l$ and define

$$\begin{aligned} \mathcal{A}_l &= \sum_{i,j} \|w_{l,i,j} \delta_{l,j}\|_2^2 \\ &= \sum_j c_{l,j} \|\delta_{l,j}\|_2^2, \quad c_{l,j} = \sum_i w_{l,i,j}^2. \end{aligned} \quad (53)$$

Thus, $\mathcal{A}_l$ is exactly the global component of the aggregate score $\mathcal{O}_l = \sum_{i,j} S_{o,l}(i,j)$. Let $C_l = \text{diag}(c_{l,1}, \dots, c_{l,d})$. At a fixed target layer, let $e_1, \dots, e_k$ denote perturbations introduced by k preceding quantized layers, and let $J_{l,r}$ be the Jacobian mapping e_r to its first-order contribution to δ_l . Define the first-order inherited discrepancy and its outlierness-weighted unit-budget envelope as

$$\begin{aligned} \delta_{l,k}^{(1)}(e) &= \sum_{r=1}^k J_{l,r} e_r, \\ \mathfrak{E}_{l,k} &= \sup_{\sum_{r=1}^k \|e_r\|_F^2 \leq 1} \left\| C_l^{1/2} \delta_{l,k}^{(1)}(e) \right\|_F^2. \end{aligned} \quad (54)$$

To index the same accumulation directly by network depth, define the prefix exposure through layer L as

$$\mathfrak{E}_L = \sup_{\sum_{r=1}^{L-1} \|e_r\|_F^2 \leq 1} \left\| C_l^{1/2} \sum_{r=1}^{L-1} J_{l,r} e_r \right\|_F^2. \quad (55)$$

As the preceding-layer path is extended from k to $k+1$ sources, both the maximum attainable first-order global outlierness $\mathfrak{E}_{l,k}$ and its total directional exposure are monotone non-decreasing. Across network prefixes, $\mathfrak{E}_L$ is likewise monotone non-decreasing in depth L .

Proof 2. Suppressing the identity factor over calibration tokens, Eq. 53 is

$$\mathcal{A}_l = \left\| C_l^{1/2} \delta_l \right\|_F^2. \quad (56)$$

After vectorizing the calibration-token dimensions, define the block sensitivity operator and its Gramian by

$$\begin{aligned} B_{l,k} &= \left[C_l^{1/2} J_{l,1}, \dots, C_l^{1/2} J_{l,k} \right], \\ \mathcal{G}_{l,k} &= B_{l,k} B_{l,k}^\top = \sum_{r=1}^k \left(C_l^{1/2} J_{l,r} \right) \left(C_l^{1/2} J_{l,r} \right)^\top. \end{aligned} \quad (57)$$

The variational characterization of the largest singular value yields

$$\mathfrak{E}_{l,k} = \|B_{l,k}\|_2^2 = \lambda_{\max}(\mathcal{G}_{l,k}). \quad (58)$$

Every summand in Eq. 57 is positive semidefinite. Adding the $(k+1)$ -st preceding quantization source therefore gives

$$\begin{aligned} \mathcal{G}_{l,k+1} - \mathcal{G}_{l,k} &= \left(C_l^{1/2} J_{l,k+1} \right) \left(C_l^{1/2} J_{l,k+1} \right)^\top \\ &\succeq 0. \end{aligned} \quad (59)$$

Eigenvalue monotonicity under the Loewner order and the trace identity now give

$$\begin{aligned} \mathfrak{E}_{l,k+1} &\geq \mathfrak{E}_{l,k}, \\ \text{tr}(\mathcal{G}_{l,k+1}) &= \text{tr}(\mathcal{G}_{l,k}) + \left\| C_l^{1/2} J_{l,k+1} \right\|_F^2 \\ &\geq \text{tr}(\mathcal{G}_{l,k}). \end{aligned} \quad (60)$$

It remains to verify the depth-indexed statement. Write $D_{l,r} = C_l^{1/2} J_{l,r}$ and define the block lower-triangular prefix operator $\mathbb{B}_L$ by

$$(\mathbb{B}_L e)_l = \sum_{r=1}^{l-1} D_{l,r} e_r, \quad l = 2, \dots, L. \quad (61)$$

Equation 55 gives $\mathfrak{E}_L = \|\mathbb{B}_L\|_2^2$. For any unit-budget prefix perturbation $e = (e_1, \dots, e_{L-1})$, embed it into depth $L+1$ as $\tilde{e} = (e_1, \dots, e_{L-1}, 0)$. Then

$$\begin{aligned} \|\mathbb{B}_{L+1} \tilde{e}\|_F^2 &= \|\mathbb{B}_L e\|_F^2 + \left\| \sum_{r=1}^{L-1} D_{L+1,r} e_r \right\|_F^2 \\ &\geq \|\mathbb{B}_L e\|_F^2. \end{aligned} \quad (62)$$

Taking the supremum yields $\mathfrak{E}_{L+1} \geq \mathfrak{E}_L$. The corresponding total directional exposure obeys

$$\|\mathbb{B}_{L+1}\|_F^2 = \|\mathbb{B}_L\|_F^2 + \sum_{r=1}^L \|D_{L+1,r}\|_F^2 \geq \|\mathbb{B}_L\|_F^2. \quad (63)$$

Therefore, extending a fixed target's preceding-layer path cannot decrease either its strongest attainable outlierness response or its total sensitivity, and extending the network prefix cannot decrease the cumulative response across depths. These monotone quantities formally characterize the cross-layer enrichment tendency of outlierness.

Implication for MOAQ. The cross-layer term used by GFD follows directly from the exact weighted reconstruction discrepancy. With S_o fixed during a correction step,

$$\begin{aligned} & (S_o \odot W_l)X_l - (S_o \odot \widehat{W}_l)\widehat{X}_l \\ &= \left(S_o \odot (W_l - \widehat{W}_l) \right) \widehat{X}_l + (S_o \odot W_l)(X_l - \widehat{X}_l). \end{aligned} \quad (64)$$

The first term is the current-layer weight residual, and the second is the inherited activation discrepancy. GFD addresses these two error sources through its local and global corrections, respectively.

C Additional Experiments

C.1 Detailed Zero-Shot Results

Table 6 reports detailed zero-shot performance of MOAQ across model families and bit-width settings, complementing the main results in Table 1.

C.2 MMLU Reasoning Experiment

Table 7 reports reasoning results for Qwen3-8B (Yang et al., 2025) under 2-bit quantization on the MMLU benchmark (Hendrycks et al., 2020). Our method consistently outperforms GPTQ and AffineQuant across all MMLU categories, demonstrating stronger reasoning ability after quantization.

C.3 Additional Ablation Experiments

Table 8 and Table 9 provide additional ablation experiments for two key designs in MOAQ. As discussed in the Introduction, existing PTQ methods often suffer from two limitations when handling outliers. First, they usually process outliers from a single local pattern, making it difficult to simultaneously accommodate spatially disordered outliers and regular clustered outliers. Second, many methods optimize only the current layer and ignore the global effect that activation perturbations and outlieriness can be inherited, transformed, and progressively enriched across layers during sequential quantization.

Table 8 ablates the design of SC by comparing it with QuaRot (Ashkboos et al., 2024). QuaRot uses Hadamard transformation to mix weight dimensions and suppress outliers, but it does not explicitly organize the local structural relations among Hadamard basis vectors. When high-outlieriness elements are spatially disordered in the original

weight layout, subsequent block-level bit-width allocation and local parameter calibration may still be unstable. In contrast, SC follows the planar-space duality motivation introduced in the Introduction. It uses a clustering-optimized Hadamard transform to reorganize the weight space, converting disordered high-outlieriness elements into more regular discrete and clustered patterns. This produces a more structured layout for LAS, where COBS performs block-level bit-width search and LOAC refines local quantization parameters. The results show that SC consistently improves MMLU average accuracy and reduces WikiText2 perplexity on both LLaMA-7B and LLaMA-13B. This supports the effectiveness of using outlier spatial structure to improve extreme low-bit quantization.

Table 9 ablates the design of GFD by comparing it with GPTAQ (Li et al., 2025d). GPTAQ mainly optimizes quantization from a local error-compensation perspective. However, the stereoscopic-space cumulative effect discussed in the Introduction indicates that activation perturbations introduced by previous quantized layers can be inherited by later layers and further amplified during forward propagation. Therefore, optimizing only the current-layer weight residual is insufficient for suppressing the global error trajectory. GFD addresses this issue by introducing global feedback and denoising based on our outlieriness metric, so that the correction process considers both current-layer perturbations and preceding-layer-induced activation discrepancies. The results show that GFD consistently outperforms GPTAQ on LLaMA-7B and LLaMA-13B in terms of both MMLU average accuracy and WikiText2 perplexity, validating the importance of global feedback and denoising for maintaining model performance under extreme quantization.

D Design and Implementation Details of the SC Module

As described in the main text, the SC module is designed according to the local duality of outlier distributions. High-outlieriness elements are spatially disordered in some linear layers, whereas they form more regular discrete and clustered patterns in others. If local quantization is directly performed on the original weight layout, LAS needs to handle isolated peaks and continuous clustered regions at the same time, which makes block-level bit-width allocation and local parameter calibration

Model	Method	Bits	WikiText2 ↓	PIQA	HellaSwag	ARC-c	ARC-e	WinoGrande	Avg.
LLaMA-7B	MOAQ	2.0	7.52	71.87	60.86	34.30	59.68	62.98	57.94
	MOAQ	2.1	7.24	72.74	63.00	35.07	59.05	62.35	58.44
	MOAQ	3.0	5.97	77.97	73.10	42.06	70.96	69.14	66.65
LLaMA-13B	MOAQ	2.0	6.51	75.73	67.30	37.46	61.57	68.67	62.15
	MOAQ	2.1	6.28	75.19	68.87	38.23	67.30	69.53	63.82
	MOAQ	3.0	5.34	79.16	76.68	46.67	71.42	71.51	69.09
LLaMA-30B	MOAQ	2.0	5.63	77.26	73.34	44.11	72.31	71.19	67.64
	MOAQ	2.1	5.44	77.80	74.47	46.25	73.40	71.51	68.69
	MOAQ	3.0	4.43	81.18	80.98	51.54	76.98	75.30	73.20
LLaMA-2-7B	MOAQ	2.0	7.40	72.25	59.45	32.59	60.77	63.38	57.69
	MOAQ	2.1	7.08	73.07	61.84	34.47	61.45	63.93	58.95
	MOAQ	3.0	5.78	78.78	73.31	43.69	72.47	68.03	67.26
LLaMA-2-13B	MOAQ	2.0	6.30	73.94	65.87	38.91	66.75	65.98	62.29
	MOAQ	2.1	6.10	75.41	68.42	39.16	66.41	67.96	63.47
	MOAQ	3.0	5.11	78.89	76.83	46.84	74.03	72.22	69.76
LLaMA-3-8B	MOAQ	2.0	10.76	68.28	56.22	33.53	52.69	62.19	54.58
	MOAQ	2.1	9.91	71.60	60.55	34.13	59.81	65.11	58.24
	MOAQ	3.0	6.99	78.40	75.37	49.74	75.72	72.14	70.27
LLaMA-3-70B	MOAQ	2.0	7.35	79.71	75.46	50.00	76.94	75.93	71.61
	MOAQ	2.1	6.98	80.52	76.89	54.27	80.47	76.24	73.68
	MOAQ	3.0	3.99	83.19	83.17	62.97	84.43	79.16	78.58

Table 6: WikiText2 perplexity and zero-shot performance of MOAQ on five common tasks.

Bit	Model	Method	MMLU_STEM	MMLU_SOCIAL_SCIENCES	MMLU_HUMANITIES	MMLU_OTHER
2-bit	Qwen3-8B	GPTQ	32.1	40.7	34.6	40.4
2-bit	Qwen3-8B	AffineQuant	34.0	41.8	36.2	42.2
2-bit	Qwen3-8B	Ours	42.4	51.8	41.3	49.9

Table 7: MMLU reasoning results on Qwen3-8B under 2-bit quantization.

Bits	Model	Method	MMLU Avg.	WikiText2 perplexity
2-bit	LLaMA-7B	QuaRot	23.76	18.07
2-bit	LLaMA-7B	SC	25.61	13.28
2-bit	LLaMA-13B	QuaRot	23.94	16.15
2-bit	LLaMA-13B	SC	26.12	11.79

Table 8: Comparison of QuaRot and SC on LLaMA-7B and LLaMA-13B.

Bits	Model	Method	MMLU Avg.	WikiText2 perplexity
2-bit	LLaMA-7B	GPTAQ	22.76	12.91
2-bit	LLaMA-7B	GFD	24.53	11.49
2-bit	LLaMA-13B	GPTAQ	23.08	10.81
2-bit	LLaMA-13B	GFD	25.97	10.24

Table 9: Comparison of GPTAQ and GFD on LLaMA-7B and LLaMA-13B.

less stable. Therefore, SC is placed before LAS to reorganize the weight space with a clustering-optimized Hadamard transform, so that subsequent local optimization is applied to a more regular outlier layout.

Hadamard transforms have been widely used in quantization to suppress outliers in weights (Ashkboos et al., 2024; Liu et al., 2024). Existing Hadamard transforms mainly rely on recursive or random constructions and focus on orthogonal mix-

ing, but they do not explicitly organize the local structural relations among Hadamard basis vectors. To make the transformed dimensions more suitable for block-level search and local calibration, SC reorganizes the Hadamard basis vectors according to their sign patterns. SC consists of three main components, namely feature engineering, local K-means clustering, and two-level sorting.

1) **Feature Engineering:** In this stage, we construct a 7-dimensional feature vector for each Hadamard basis vector to describe its structural pattern. The seven features are summarized in Table 10.

2) **Local K-Means Clustering:** Based on the feature vectors constructed above, all basis vectors are grouped using K-means clustering. The algorithm partitions them into different clusters by iteratively minimizing within-cluster variance, so that basis vectors with similar sign structures are assigned to the same cluster.

3) **Two-Level Sorting:** In this stage, we generate the final ordering of basis vectors using a two-level strategy that proceeds from local to global organization. We first perform fine-grained optimization within each cluster by sorting the basis vectors in ascending order according to the Euclidean distance

Feature	Definition
Global Pattern Complexity	The total number of sign transitions between adjacent elements.
Maximum Coherence Length	The length of the longest consecutive same-sign segment.
Average Coherence Length	The mean length of all same-sign segments.
Spatial Coherence	The variance of pattern complexity across several equal-length subsegments.
Sign Balance	The proportion of positive elements.
Prefix Local Complexity	The pattern complexity within the first quarter of the vector.
Suffix Local Complexity	The pattern complexity within the last quarter of the vector.

Table 10: Structural features used in the SC module for Hadamard basis vector reorganization.

between their feature vectors and the corresponding cluster center. Next, at the inter-cluster level, we apply coarse-grained sorting based on the average pattern complexity of each cluster. The original Hadamard matrix is then reorganized according to the resulting two-stage ordering.

After SC, originally disordered outliers are reorganized into more regular discrete and clustered patterns. This provides a more structured input for LAS, where COBS performs block-level bit-width allocation and LOAC calibrates local quantization parameters.

E Efficient Implementation Details of the GFD Module

E.1 Grouped Activation Storage

Using full-precision activations for quantization-error compensation requires accessing the corresponding activation tensors for each Transformer layer. Storing all of these activations of a layer on the GPU at once is prohibitively expensive for ordinary users due to the large memory footprint. An intuitive alternative is to move them to the CPU and send them back to the GPU on demand, but this leads to a substantial increase in runtime because of frequent CPU–GPU data transfers. Each Transformer layer consists of seven linear layers, and due to their structural properties, several of these linear layers share identical input activations. As shown in Table 11, this fact motivates us to group the seven linear layers into four distinct groups, where each group shares the same full-precision activations for calibration. This grouping reduces the activation-storage overhead to some extent.

E.2 Sliding-Window Prefetching

Additionally, since current quantization algorithms process groups sequentially, we further observe two key properties: 1) **Temporal Separation of Access Patterns**: different groups require activation access at clearly different time points during quantization.

2) **A Deterministic Gap Between Computation and Transfer**: the time required for quantization computation is significantly longer than the corresponding CPU–GPU transfer time. These observations motivate the design of a pipeline-coordinated sliding-window mechanism. The mechanism keeps only a limited number of activation groups on the GPU and asynchronously prefetches the next group during the quantization of the current one. Because the transfer time is fully hidden behind computation, this approach achieves nearly negligible time overhead while greatly reducing GPU memory usage. Specifically, during the quantization phase, we maintain a sliding window of size two, ensuring that at most two activation groups reside on the GPU at any moment. The window simultaneously covers the group currently being quantized and the next group in the processing order. When quantizing the current group, the algorithm initiates an asynchronous prefetch of the next group using an independent CUDA stream, enabling full overlap between data transfer and computation. Since quantization time far exceeds transfer time, the prefetched group is always ready by the time the window advances, ensuring seamless switching. As the window slides forward, processed groups are immediately evicted from GPU memory and replaced by newly prefetched ones, forming a pipeline-style data flow. Table 12 ablates this activation-management strategy. Compared with keeping all activation groups on GPU, GFD w/ SWP substantially reduces GPU memory. Compared with naive CPU offloading, it avoids most transfer overhead by overlapping CPU–GPU transfer with quantization computation. This verifies that sliding-window prefetching tightly constrains the activation memory cost while maintaining comparable quantization time.

Linear layer	self_attn.q_proj	self_attn.k_proj	self_attn.v_proj	self_attn.o_proj	mlp.up_proj	mlp.gate_proj	mlp.down_proj
Group domain	g1	g1	g1	g2	g3	g3	g4

Table 11: Grouping scheme for activation storage in GFD.

Method	LLaMA-3-8B Time	LLaMA-3-8B Memory	LLaMA-3-70B Time	LLaMA-3-70B Memory
GFD w/o SWP (ALL_ON_GPU)	45.9s	25.86GB	140.2s	63.13GB
GFD w/o SWP (ALL_TO_CPU)	55.58s	12.38GB	170.4s	35.28GB
GFD w/ SWP	46.5s	10.54GB	141.4s	30.85GB

Table 12: Ablation of sliding-window prefetching (SWP) on GPU memory and per-layer quantization time for LLaMA-3-8B and LLaMA-3-70B.