INFOQUANT: Shaping Activation Distributions for Low-Bit LLM Quantization

Anonymous ACL submission

Abstract

Low-bit activation quantization remains a major bottleneck in efficient large language model (LLM) deployment. The difficulty is not only that activations contain outliers, but that their transformed distributions often remain poorly matched to a low-bit uniform quantizer. Existing post-training quantization (PTQ) methods suppress peaks, balance channels, or minimize reconstruction error, yet they rarely specify what activation distribution is actually easy to discretize. As a result, activations may appear numerically smoother while still incurring large quantization error because the quantization range remains wide or most values collapse into a few levels near the mean. We recast activation transformation as quantizer-facing distribution design and analyze quantization error from an information-theoretic perspective. Our analysis shows that quantization-friendly activations should jointly have a smaller numerical range and sufficient dispersion within that range. Guided by this analysis, we propose INFOQUANT, a train-free method that employs Peak Suppression Orthogonal Transformation (PSOT) to shape activations into more quantization-friendly distributions. We further introduce adaptive outlier-token selection to improve the robustness of PSOT during optimization. Across multiple LLM families, INFOQUANT consistently outperforms prior PTQ and end-to-end training baselines. Under W4A4KV4, it preserves 97% of floating-point accuracy on average, reduces the LLaMA-2 13B performance gap by 42% over the previous state of the art.

1 Introduction

Post-training quantization (PTQ) is one of the most practical ways to reduce the memory and compute cost of large language model (LLM) inference. Its main challenge, however, lies in low-bit activation quantization. Unlike weights, LLM activations often contain a small number of dominant coordinates that enlarge the quantization range and force round-to-nearest quantization to map many normal values to the same few levels. This mismatch becomes particularly severe in 4-bit settings, where limited quantization levels leave little room to preserve both rare extremes and dense central values.

Recent PTQ methods increasingly address this problem through activation transformations before quantization. SmoothQuant (Xiao et al., 2023) migrates activation difficulty into weights through diagonal scaling, while QuaRot (Ashkboos et al., 2024b) and SpinQuant (Liu et al., 2025b) use orthogonal rotations to redistribute activation energy; other methods further introduce more flexible affine transformations or reconstruction objectives (Ma et al., 2024; Sun et al., 2025). Although these methods differ in form, they share the same practical role: they change the activation distribution seen by the quantizer. Yet most of them are motivated by suppressing outliers, balancing channels, or reducing reconstruction error, rather than by defining what transformed distribution a low-bit quantizer can represent well. Consequently, as illustrated in Figure 1, they may reduce visible peaks without fully improving discretizability, or preserve small numerical error while still collapsing distributional resolution. The central question, therefore, is not only how to transform activations, but what transformed activation distribution is actually quantization-friendly.

We address this gap by recasting activation transformation as quantizer-facing distribution design. From an information-theoretic perspective, we analyze how quantization error depends on the activation distribution after transformation. Our theoretical and empirical study shows that lower quantization error is associated with two complementary properties: a smaller numerical range and greater dispersion within that range. This result reframes the role of activation transformation. Rather than merely suppressing outliers or minimizing heuristic

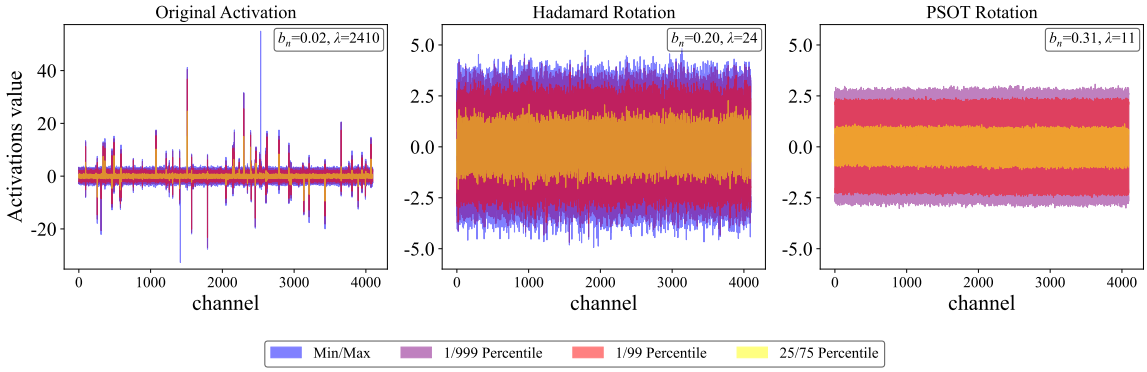

Figure 1: Activation distributions of the `LLaMA-2 7B, layer4 q/k/v_proj` input under three transformations: the original activations (left), Hadamard rotation (center), and the learned rotation from PSOT (right). Compared with the original and Hadamard-rotated activations, PSOT produces a more quantization-friendly distribution with a narrower numerical range and larger normalized dispersion. Here, b_n denotes the standard deviation after infinity-norm normalization, and a smaller $\lambda = \bar{s}/b_n$ indicates a lower normalized quantization-error bound, where $\bar{s}$ is the range-normalized quantization step size.

reconstruction losses, a PTQ method should explicitly shape activations into distributions that are easier for a low-bit quantizer to preserve. Since LLM activations are typically bell-shaped and often contain outliers (Liu et al., 2025a), they are naturally misaligned with this target, which explains why low-bit activation quantization remains difficult.

Guided by this principle, we propose INFOQUANT, a train-free PTQ method that learns orthogonal transformations to produce more quantization-friendly activation distributions. Its core component is Peak Suppression Orthogonal Transformation (PSOT), which applies an activation-wise peak suppression objective to reduce the numerical range while increasing normalized dispersion. We further introduce adaptive outlier-token selection to improve optimization robustness, and learn activation clipping parameters to refine the final quantization range after the distribution has been reshaped. Although INFOQUANT is designed around activation optimization, it remains compatible with standard weight quantization pipelines. Overall, our contributions can be summarized as follows:

- We introduce an information-theoretic framework for understanding activation quantization error and show, both theoretically and empirically, that quantization-friendly distributions should have a smaller numerical range and greater dispersion.
- We propose INFOQUANT, a hardware-efficient and train-free PTQ method centered on learned orthogonal activation shaping, together with adaptive outlier-token selection and learnable activation clipping for robust calibration.
- We demonstrate that activation-distribution optimization yields strong empirical and practical gains. In the W4A4KV4 setting, LLaMA-2 (7B, 13B, 70B) and LLaMA-3 (8B, 70B) retain an average of 97% of their original performance, and the 70B model can be quantized using only 24GB of GPU memory.

2 Related Work

Post Training Quantization for LLMs. PTQ is an efficient and widely used approach for compressing LLMs. Due to the flatness and uniform distribution of LLM weights, weight-only quantization typically results in minimal performance degradation. GPTQ (Frantar et al., 2023) uses Hessian-based error compensation to enable high compression with low accuracy loss. AWQ (Lin et al., 2024b) and OWO (Lee et al., 2024) further improve performance by mitigating the effects of activation outliers. QuIP (Chee et al., 2023) and QuIP# (Tseng et al., 2024) apply random Hadamard transforms for incoherent processing and employ vector quantization on weights, achieving better performance. In contrast, activation quantization remains more challenging due to the presence of rare but extreme outliers (Wei et al., 2023; Xiao et al., 2023), which can disproportionately affect accuracy.

Transformation-based Methods. These methods more effectively redistribute activation outliers across channels. Channel scaling (Xiao et al.,

2023) shifts part of this burden to weights, OmniQuant (Shao et al., 2024) and LRQuant (Zhao et al., 2024) optimize scaling parameters via MSE minimization. However, recent work (Yi et al., 2025) shows channel scaling alone fails under 4-bit settings, leading to notable degradation. AffineQuant (Ma et al., 2024) learns affine transformations to precondition activations. However, due to the significant overhead of full-size matrix multiplication, AffineQuant can only apply affine transformations to a small fraction of linear layers. FlatQuant (Sun et al., 2025) reduces this cost via Kronecker decomposition, applying affine transformation to every linear layer. Leveraging computational invariance (Ashkboos et al., 2024a), orthogonal transforms can be applied to weights and between-block activations without extra inference overhead. QuaRot (Ashkboos et al., 2024b) uses randomized Hadamard transforms to remove outliers. SpinQuant (Liu et al., 2025b) further optimizes learnable orthogonal matrices on the Stiefel manifold with task loss (e.g., cross-entropy) to find stable transformations. OSTQuant (Hu et al., 2025) combines channel scaling with orthogonal transforms and uses end-to-end distillation from original outputs to boost quantization. Kurtail (Akhondzadeh et al., 2025) facilitates quantization by controlling the kurtosis to make the distribution more uniform. BASE-Q (He et al., 2025) introduces an additional bias term to balance the mean values of different channels after rotation.

3 Motivation

3.1 Quantization Preliminaries

Quantization maps high-precision values to a set of discrete levels. The process is detailed as follows:

$$\mathcal{Q}(\mathbf{X}) = \text{clamp}\left(\left\lfloor \frac{\mathbf{X}}{s} \right\rceil + z,\ 0,\ 2^N - 1\right) \quad (1)$$

Here, quantization step size is denoted by $s = \frac{\mathbf{X}_{\max} - \mathbf{X}_{\min}}{2^N - 1}$, and $z = -\left\lfloor \frac{\mathbf{X}_{\min}}{s} \right\rceil$ is the corresponding zero-point, $\lfloor \cdot \rceil$ denotes the rounding operation, and N represents the target bit-width. Given a floating-point tensor $\mathbf{X}$, the quantization function $\mathcal{Q}(\cdot)$ produces its integer-valued representation . Quantization error primarily arises from the rounding operation, which collapses all values within a single interval of size s into the same discrete level.

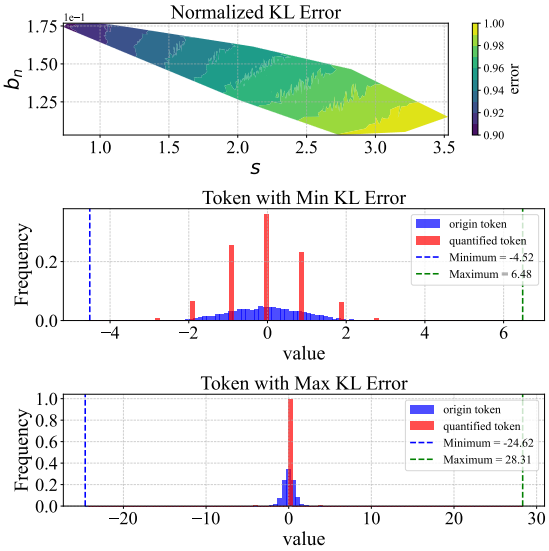

Figure 2: Quantization evaluation for the `LLaMA-2 7B,layer4 q/k/v_proj` input. Top: KL divergence between activation histograms before and after quantization, as a function of quantization step size s and the estimated dispersion parameter b_n, computed using a bin size of 15,000. Center and Bottom: Frequency histograms of token before and after quantization, corresponding to the smallest and largest quantization error.

3.2 Quantization Error of KL Divergence

Recent PTQ methods often optimize activation transformations with MSE-based objectives (Shao et al., 2024; Zhao et al., 2024; Sun et al., 2025). While MSE is a useful measure of numerical distortion, it does not fully capture the distributional mismatch introduced by low-bit quantization. This limitation is especially important for activation quantization under round-to-nearest (RTN), where many values may incur only small pointwise errors yet still be mapped to a small number of discrete levels. In such cases, the quantized activations can remain close in value to the original ones while losing substantial distributional resolution, which is not well reflected by MSE alone. Prior work (Liu et al., 2025a) has shown that activation distributions in LLMs are typically bell-shaped (e.g., Gaussian or Laplace). When such activations are quantized with a low-bit uniform RTN quantizer, a few large values can determine the quantization range, forcing most normal values to collapse toward levels near the mean (Figure 2, bottom). The resulting error is therefore not only a matter of local rounding distortion, but also of how poorly the available quantization levels match the underlying activation distribution. Although non-uniform quantizers can in principle better adapt to dense regions, they

usually introduce additional hardware complexity and are less attractive in practical low-bit deployment. These observations motivate an analytical metric that reflects both numerical deviation and quantization-induced distribution shift.

To address this, we use a smoothed KL divergence as an analytical metric that captures both numeric deviations and distributional shifts. Let $\mathbf{x}$ denote an activation token, and let each entry be denoted by a scalar $x \in \mathbf{x}$ with distribution $P(x)$. To connect the analysis to practical low-bit quantization, we consider a centered finite-bit clamped quantizer

$$\hat{x} = Q_{s,c}(x) = \mathrm{clip}\Big(s \left\lfloor \frac{x}{s} \right\rceil,\, -c,\, c\Big), \quad (2)$$

where s is the quantization step size and c is the clipping scale. Since quantization maps a continuous distribution to a discrete one, we adopt a continuous relaxation for tractability. Specifically, let q_i denote a quantization centroid, let $I_i = \{x \mid Q_{s,c}(x) = q_i\}$ be its quantization cell, and define the corresponding probability mass $p_i = \int_{I_i} P(x)\, dx$. We then approximate the quantized distribution as a smoothed density $Q_\theta^{(c)}(x) \approx \sum_i p_i \cdot \delta(x - q_i; \theta)$, where $\delta(x;\theta) = \frac{1}{2\theta} \exp\Big(-\frac{|x|}{\theta}\Big)$ is a Laplace kernel that converges to the Dirac delta function as $\theta \to 0$. Under the standard separation assumption $\theta \ll s$, the corresponding smoothed KL objective admits the approximation:

$$D_{\mathrm{KL}}\left(P \,\|\, Q_\theta^{(c)}\right) = \int_{-\infty}^{+\infty} P(x)\, \log\left(\frac{P(x)}{Q_\theta^{(c)}(x)}\right) dx$$
$$\approx -H(P) + H(\{p_i\}) + \log(2\theta) + \frac{1}{\theta}\, \mathcal{E}_{\mathrm{clip}}, \quad (3)$$

where $H(P)$ denotes the entropy of the original distribution $P(x)$, $H(\{p_i\})$ is the entropy of the quantized distribution, and $\mathcal{E}_{\mathrm{clip}} = \mathbb{E}|x - Q_{s,c}(x)|$ is the expected absolute error of the finite-bit clamped quantizer. Since activations typically have bounded yet varying ranges, we normalize this error by the standard deviation σ:

$$\mathcal{E}'_{\mathrm{clip}} = \frac{1}{\sigma}\, \mathbb{E}|x - Q_{s,c}(x)| \le \frac{s}{2\sigma} + \frac{1}{\sigma}\, \mathbb{E}[(|x| - c)_+]\,. \quad (4)$$

The first term is the in-range rounding error, while the second term is the clipping-tail error introduced by the finite quantization range. Let $\lambda = s/\sigma$ and $\kappa = c/\sigma$. Then

$$\mathcal{E}'_{\mathrm{clip}} \le \frac{\lambda}{2} + \tau_P(\kappa), \quad (5)$$

where $\tau_P(\kappa) = \mathbb{E}[(|Y| - \kappa)_+]$ for the normalized variable $Y = x/\sigma$. For common bell-shaped activation distributions such as Gaussian and Laplace, τ_P admits closed forms, and for a B-bit quantizer with $c = Ms$ and $M = 2^{B-1} - 1$, the resulting bound decreases as κ decreases in the tail-controlled regime relevant to calibrated PTQ.[1] Since $\kappa = c/\sigma = 1/b_n$ and $\lambda = s/\sigma = \bar{s}/b_n$ with $b_n = \sigma/c$, a smaller clipped range relative to the activation spread and a larger max-normalized dispersion jointly reduce the practical quantization error bound in this regime. Empirically, as shown in the top of Figure 2, KL divergence decreases consistently with smaller s and larger b_n. Together, these results provide strong evidence that enhancing activation dispersion while narrowing their numerical range yields a simple and effective principle for greatly mitigating quantization error.

4 Method

INFOQUANT is a train-free PTQ framework that optimizes the activation distribution seen by a low-bit quantizer. The analysis in Section 3.2 suggests that quantization error can be reduced by jointly narrowing the effective numerical range and increasing the dispersion within that range. Following this principle, INFOQUANT contains three components. First, Peak Suppression Orthogonal Transformation (PSOT) learns an orthogonal transformation that suppresses dominant activation coordinates while preserving rotation-based efficiency. Second, adaptive outlier-token selection (ASOT) identifies calibration tokens that expose reliable outlier behavior and assigns them larger optimization weights. Third, learnable activation clipping (LAC) refines the final quantization interval after the activation distribution has been reshaped.

4.1 Peak Suppression Orthogonal Transformation

PSOT learns a quantizer-facing orthogonal transformation by directly penalizing peak-dominated activation tokens. Let $\mathbf{x} \in \mathbb{R}^d$ denote one activation token, and let $\mathbf{R} \in \mathbb{R}^{d \times d}$ be a learnable orthogonal matrix with $\mathbf{R}^\top \mathbf{R} = \mathbf{I}$. We remove the coordinate-wise mean after rotation using the centering projector $\mathbf{P}_\perp = \mathbf{I} - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$, and define

[1] Proofs and the Gaussian/Laplace closed forms can be found in Appendix A.

$\boldsymbol{\pi}_T(\mathbf{y}) = \mathrm{softmax}(|\mathbf{y}|/T)$:

$$\begin{aligned} \mathbf{y}(\mathbf{x};\mathbf{R}) &= \mathbf{x}\mathbf{R}\mathbf{P}_{\perp}, \\ \ell_{\mathrm{ps}}(\mathbf{x};\mathbf{R}) &= \left\|\boldsymbol{\pi}_T(\mathbf{y}) \odot \mathbf{y}\right\|_2, \end{aligned} \tag{6}$$

where T is the temperature and $\odot$ denotes element-wise multiplication. The softmax weights concentrate the objective on high-magnitude coordinates, so minimizing ℓ_{ps} suppresses the coordinates that dominate the quantization range.

This peak-suppression objective promotes the two distributional properties identified by the KL analysis. Since $\mathbf{R}$ is orthogonal, the rotation preserves token energy before centering. Reducing the largest centered coordinate therefore pushes the remaining energy to spread across more dimensions, which lowers the clipping scale and increases the max-normalized dispersion b_n. As illustrated in Figure 1, the learned rotation yields a narrower range and a more dispersed normalized distribution than both the original activations and a fixed Hadamard rotation. Appendix A gives the corresponding formal analysis.

4.2 Adaptive Outlier-Token Selection

PSOT obtains the strongest learning signal from tokens whose activations reveal quantization-sensitive outlier structure. Uniformly optimizing all calibration tokens can dilute this signal, because many tokens have weak or noisy peaks and contribute little to the desired range compression. We therefore introduce ASOT, which automatically selects reliable outlier tokens and reweights them during rotation optimization.

Let $\mathbf{X}^{(r)} \in \mathbb{R}^{n_r \times d}$ be the activation matrix of the r-th calibration sample, where $\mathbf{x}_i^{(r)}$ is its i-th token. For a threshold coefficient k, we define the selected token-index set as

$$\begin{aligned} o_i^{(r)} &= \left\| \frac{\mathbf{x}_i^{(r)} - \mu\mathbf{1}}{\sigma} \right\|_\infty, \\ \mathcal{T}^{(r)}(k) &= \left\{ i \mid o_i^{(r)} > k \right\}, \end{aligned} \tag{7}$$

where μ and σ are estimated from the corresponding calibration activations. This criterion selects tokens that contain at least one statistically extreme coordinate, rather than assuming that every token is equally informative for learning the rotation.

The threshold should select sparse outlier tokens while avoiding positions that are fixed across prompts. Prior work observes that activation outliers are more closely tied to token identity than absolute sequence position (Liu et al., 2024; Chen et al., 2025). We measure whether the selected positions vary across m calibration samples with the positional inconsistency score

$$\eta_m(k) = 1 - \frac{\frac{1}{m}\sum_{r=1}^{m} |\mathcal{T}^{(r)}(k)|}{\left|\bigcup_{r=1}^{m} \mathcal{T}^{(r)}(k)\right|}. \tag{8}$$

A larger $\eta_m(k)$ indicates that the selected outlier positions are less tied to fixed sequence locations and are therefore more sample-specific.

We choose the smallest threshold at which the inconsistency curve stabilizes while the selected tokens remain sparse:

$$k^\star = \min \left\{ k \in \mathcal{K} \;\middle|\; \begin{array}{l} |\nabla \eta_m(k)| < \delta, \\ \dfrac{1}{m}\displaystyle\sum_{r=1}^{m} |\mathcal{T}^{(r)}(k)| < \tau |\mathcal{I}| \end{array} \right\}, \tag{9}$$

where $\mathcal{K}$ is the ordered threshold grid, $\mathcal{I}$ is the token-index universe, δ controls the stabilization tolerance, and τ limits the selected-token ratio. With the selected sets $\mathcal{T}^{(r)}(k^\star)$, the final rotation objective is

$$\begin{aligned} &\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \sum_{r=1}^{m} \sum_{i \in \mathcal{I}} w_i^{(r)}\, \ell_{\mathrm{ps}}(\mathbf{x}_i^{(r)};\mathbf{R}), \\ &w_i^{(r)} = \begin{cases} \gamma, & i \in \mathcal{T}^{(r)}(k^\star), \\ 1, & \text{otherwise}. \end{cases} \end{aligned} \tag{10}$$

where $\gamma > 1$ emphasizes outlier tokens while retaining normal tokens as regularizing calibration samples. In all experiments, the ASOT hyperparameters are selected by grid search on the calibration set: we search over the threshold grid $\mathcal{K}$ to determine $k^\star$, set $\delta = 0.02$ and $\tau = 0.4$ for the stabilization and sparsity criteria, and choose γ based on validation perplexity. We study the effect of γ in Table 5.

4.3 Learnable Activation Clipping

After PSOT reshapes the activation distribution, the remaining error is governed by the final clipping interval used by the quantizer. We therefore apply learnable activation clipping to refine the lower and upper bounds. Inspired by (Zhao et al., 2024; Sun et al., 2025), we introduce two bounded parameters α and β and define the activation step size as

$$s(\alpha,\beta) = \frac{\alpha X_{\max} - \beta X_{\min}}{2^N - 1}, \tag{11}$$

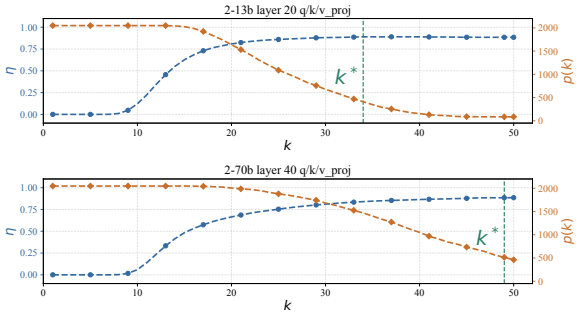

Figure 3: Adaptive threshold selection for ASOT. The positional inconsistency $\eta_{10}(k)$ increases with k and then plateaus, while the average number of selected outlier tokens $p(k) = \frac{1}{m}\sum_{r=1}^{m}|\mathcal{T}^{(r)}(k)|$ decreases. We choose $k^\star$ when $\eta_{10}(k)$ stabilizes and the selected token ratio satisfies the sparsity constraint in Eq. (9).

where N is the bit-width, and $X_{\max}$ and $X_{\min}$ are the observed maximum and minimum activation values. We parameterize α and β with a sigmoid function so that the learned clipping interval remains bounded during optimization.

For the l-th Transformer block, the clipping parameters are learned by matching the quantized block output to the full-precision block output:

$$\min_{\alpha,\beta}\left\|\mathcal{F}_l(\mathbf{X}_l) - \widehat{\mathcal{F}}_l(\mathbf{X}_l;\alpha,\beta)\right\|_2^2, \qquad (12)$$

where $\mathcal{F}_l$ denotes the full-precision Transformer block and $\widehat{\mathcal{F}}_l$ denotes the same block evaluated with activation quantization under the learnable clipping interval in Eq. (11). This final calibration step complements PSOT: the rotation first produces a more quantization-friendly distribution, and LAC then selects the interval that best preserves the block-level response.

5 Experiment

Models and Datasets. We evaluate whether activation-distribution optimization transfers across model families, scales, and evaluation metrics. The main comparison covers LLaMA-2 (7B–70B) (Touvron et al., 2023) and LLaMA-3 (8B–70B) (Grattafiori et al., 2024); additional Qwen2.5 (14B/32B) (Qwen et al., 2025) results are reported in Appendix E. We report WikiText2 perplexity (PPL) (Merity et al., 2016) as a sensitive language-modeling metric and use nine zero-shot tasks from `lm-evaluation-harness` (version 0.4.7) (Gao et al., 2024) to check whether lower quantization error translates to task-level behavior. The zero-shot suite includes BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), LAMBADA (OpenAI) (Radford et al., 2019), OpenBookQA (OBQA) (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC-Easy, and ARC-Challenge (Boratko et al., 2018).

Baselines. We compare with representative quantization methods that stress different parts of the design space: weight reconstruction methods GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024b), rotation-based methods QuaRot (Ashkboos et al., 2024b) and SpinQuant (Liu et al., 2025b), and recent low-bit LLM quantizers Kurtail (Akhondzadeh et al., 2025), BASE-Q (He et al., 2025), and OSTQuant (Hu et al., 2025). This comparison is useful because INFOQUANT changes the activation distribution before quantization, whereas several baselines mainly reconstruct weights, use fixed rotations, or rely on stronger supervision.

Implementation Details. The calibration process uses 128 samples from WikiText-2, each with a sequence length of 2048. Activations are quantized using per-token asymmetric quantization, while weights are quantized using asymmetric per-channel quantization with GPTQ (Frantar et al., 2023), applying a group size of 128 for key-value matrices. During the PSOT phase, we optimize block-diagonal orthogonal matrices initialized with Hadamard matrices via *Cayley SGD* (Li et al., 2020). The ASOT hyperparameters are selected by grid search on the calibration set, and the final search space and chosen values are reported in Appendix B and Appendix C. We report two variants to separate accuracy and deployment considerations. INFOQUANT uses block-diagonal rotations to reduce transformation overhead, while INFOQUANT* uses a full global orthogonal rotation similar to SpinQuant. More implementation details are provided in Appendix B.

5.1 Main Results

Quantization Performance. Table 1 shows that the value of activation-distribution optimization becomes most visible when activations are actually quantized. In the weight-only 4-16-16 setting, INFOQUANT preserves 98.7%–99.5% of the floating-point zero-shot accuracy across the evaluated LLaMA models, but the gap among strong rotation-based methods is relatively small. This pattern is informative: when activations remain in high precision, reshaping them is not the dominant

#Bits W-A-KV	Method	LLaMA-3 8B 0-shot[9] Avg.(↑)	LLaMA-3 8B Wiki (↓)	LLaMA-3 70B 0-shot[9] Avg.(↑)	LLaMA-3 70B Wiki (↓)	LLaMA-2 7B 0-shot[9] Avg.(↑)	LLaMA-2 7B Wiki (↓)	LLaMA-2 13B 0-shot[9] Avg.(↑)	LLaMA-2 13B Wiki (↓)	LLaMA-2 70B 0-shot[9] Avg.(↑)	LLaMA-2 70B Wiki (↓)
16-16-16	FloatingPoint	68.09	6.14	73.81	2.86	65.21	5.47	67.61	4.88	71.59	3.32
4-16-16	GPTQ	61.03	7.43	31.45	9e3	60.86	9.84	64.71	5.79	70.96	3.94
	AWQ	67.03	7.36	68.92	5.92	63.89	5.83	66.25	5.07	70.88	4.03
	QuaRot	67.27	6.53	72.93	3.53	64.30	5.62	66.95	5.00	71.21	3.41
	SpinQuant	66.54	6.49	72.90	**3.49**	63.59	**5.58**	67.14	5.00	71.12	3.43
	OSTQuant	67.80	6.53	73.69	3.19	64.37	5.64	67.31	4.94	71.48	3.41
	INFOQUANT	**67.36**	**6.48**	**73.25**	3.50	**64.34**	5.60	**67.27**	**4.99**	**71.25**	**3.40**
4-4-16	QuaRot	61.69	8.02	65.56	6.35	61.87	6.05	65.13	5.35	69.96	3.78
	SpinQuant	64.11	7.28	66.99	6.10	57.37	6.78	63.23	5.24	70.58	3.68
	OSTQuant	65.14	7.24	72.21	3.97	63.90	5.60	66.24	5.14	70.92	3.57
	INFOQUANT	**65.74**	**7.07**	**70.71**	**5.24**	**62.84**	**5.86**	**66.71**	**5.15**	**70.82**	**3.62**
4-4-4	QuaRot	61.38	8.18	65.33	6.60	61.48	6.11	65.16	5.39	70.30	3.80
	SpinQuant	64.10	7.35	66.31	6.24	62.01	5.96	64.13	5.74	**70.57**	**3.61**
	Kurtail	-	7.20	-	7.20	-	5.90	-	5.20	-	
	OSTQuant	65.37	7.29	71.69	4.01	63.18	5.91	65.41	5.25	70.84	3.59
	OSTQuant†	65.13	**6.80**	-	-	62.45	**5.38**	-	-	-	
	BASE-Q	65.39	7.17	OOM	OOM	62.50	5.85	65.48	5.21	OOM	OOM
	INFOQUANT*	64.79	7.21	70.01	5.57	62.62	5.93	66.12	5.22	70.10	3.84
	INFOQUANT	**65.57**	7.16	**70.21**	**5.39**	**63.16**	5.89	**66.33**	**5.18**	70.35	3.64

Table 1: Comparison of perplexity on WikiText2 and averaged accuracy across nine diverse zero-shot tasks. Results for GPTQ, AWQ, QuaRot, SpinQuant, and OSTQuant are reported from the OSTQuant paper, while BASE-Q results are based on official code (Note: 'OOM' denotes Out of Memory in our device). INFOQUANT* denotes the application of a complete global orthogonal rotation. OSTQuant† refers to the OSTQuant without distillation.

bottleneck. The setting mainly verifies that the learned rotation does not damage the weight-only quantization path.

Once activations are quantized, the same table reveals a different behavior. Under 4-4-16, INFOQUANT improves the average zero-shot accuracy over SpinQuant by 2.91 points across the five LLaMA settings, with larger gains on LLaMA-2 7B (+5.47) and LLaMA-3 70B (+3.72). Under W4A4KV4, where weights, activations, and KV cache are all quantized to 4 bits, INFOQUANT preserves 96.9% of floating-point accuracy on average and improves over SpinQuant by 1.70 points. These results support the main design intuition of the paper: fixed or generic rotations are often sufficient to avoid catastrophic outliers, but low-bit activation quantization benefits from learning a distribution that uses the available quantization levels more evenly.

The comparison with OSTQuant also clarifies the boundary of the method. INFOQUANT outperforms the distillation-based OSTQuant baseline on LLaMA-3 8B and LLaMA-2 13B, while OSTQuant remains competitive on several 70B entries. This mixed pattern is useful rather than merely negative: it suggests that distribution shaping can recover a large part of the activation-quantization loss without full-precision supervision, but supervision and scale-specific calibration may still help in some large-model regimes. Full per-task results are reported in Appendix E.

Method	Prefill Speedup 2048	4096	8192	Memory Saving 2048	4096	8192
2-70B-INFOQUANT	2.46	2.11	1.97	2.91	2.59	2.22
2-70B-INFOQUANT*	2.61	2.26	2.10	3.25	2.84	2.36
3-8B-INFOQUANT	1.57	1.43	1.28	2.46	2.12	1.86
3-8B-INFOQUANT*	1.74	1.55	1.42	2.78	2.37	2.00

Table 2: Speedup and memory savings factors for LLaMA models of different sizes and sequence lengths, comparing 4-bit quantized implementations to FP16.

Speedup and Memory Savings. We evaluate inference efficiency using the W4A4 kernel from Ashkboos et al. (2024b). Table 2 reports prefill speedup and decoding memory savings relative to FP16 on a single Transformer block, using batch size 4 on an NVIDIA RTX 4090. On LLaMA-2 70B with sequence length 2048, INFOQUANT* achieves a 2.61× prefill speedup and 3.25× memory saving, while INFOQUANT achieves a 2.46× prefill speedup and 2.91× memory saving. The difference between INFOQUANT and INFOQUANT* exposes the main deployment trade-off: full rotations can be slightly faster in this implementation path, whereas block-diagonal rotations provide a more flexible per-layer optimization structure.

We further compare quantization-time memory overhead and inference-time transformation FLOPs

on LLaMA-3 70B. As shown in Figure 4, INFOQUANT* requires less than 24GB of GPU memory during quantization and introduces low additional transformation cost. The broader insight is that activation shaping should not be evaluated only by accuracy: a transformation that must be repeatedly applied online can erase part of the benefit of low-bit inference. FlatQuant, for example, performs multiple dynamic activation transformations during inference, which increases transform FLOPs and depends on specialized kernels. INFOQUANT instead keeps the rotation-based deployment path lightweight, making the method more compatible with existing low-bit kernels and consumer-grade quantization hardware.

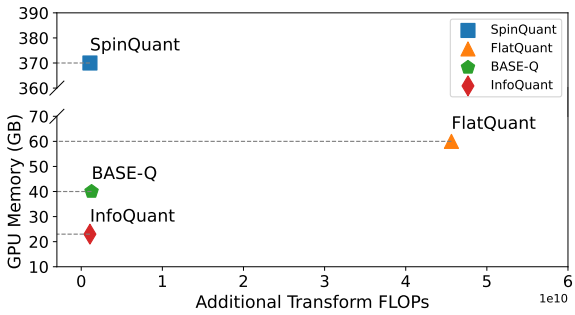

Figure 4: Comparison of memory overhead in quantization and transform FLOPs during inference across different methods. During inference, FlatQuant performs transformations on three activation values online and dynamically, while other methods apply the fast Hadamard transform only to the KV-cache efficiently.

5.2 Ablation Study

Module-wise Impact. Table 3 isolates how each component contributes to INFOQUANT under the W4A4KV4 setting. Replacing a fixed Hadamard rotation with the learned PSOT rotation reduces WikiText2 perplexity from 10.90 to 8.84 on LLaMA-3 8B and from 8.99 to 7.03 on LLaMA-2 7B. This is the largest single change in the ablation, which suggests that the core gain comes from learning where activation energy should be redistributed, not from simply adding more calibration stages.

The later components produce smaller but more diagnostic changes. Adding GPTQ further reduces perplexity to 7.53 and 6.01, indicating that activation-oriented rotation remains compatible with standard weight reconstruction. ASOT gives a modest improvement, which is consistent with its role as a signal reweighting mechanism rather than a separate quantizer. LAC then gives the best perplexity on both models (7.16 and 5.89). The ordering of these gains is important: first reshape the distribution, then reconstruct weights, then refine which calibration tokens and clipping ranges deserve attention.

Hadamard	PSOT	GPTQ	ASOT	LAC	WikiText2 3-8B	WikiText2 2-7B
✓					10.9	8.99
	✓				8.84	7.03
	✓	✓			7.53	6.01
	✓	✓	✓		7.42	5.97
	✓	✓	✓	✓	**7.16**	**5.89**

Table 3: Component ablation of INFOQUANT under the W4A4KV4 configuration. Lower WikiText2 perplexity indicates better quantization quality.

More Ablations. Additional ablations in Appendix C study the weighting factor γ, the temperature T, initialization robustness, clipping-ratio sensitivity, and the block size of block-diagonal orthogonal matrices. These experiments provide practical guidance for using INFOQUANT: objective reweighting and temperature control the stability of peak suppression, while block size and clipping range determine how much distribution-shaping flexibility can be traded for lower inference cost and easier calibration.

6 Conclusion

This work studies low-bit LLM quantization from the perspective of quantizer-facing activation distribution design. Instead of treating activation transformation only as outlier suppression or reconstruction-error minimization, we analyze the distributional error introduced by quantization and show that quantization-friendly activations should jointly have a smaller numerical range and sufficient dispersion within that range. Guided by this principle, INFOQUANT reshapes activations with Peak Suppression Orthogonal Transformation, emphasizes informative calibration tokens with ASOT, and refines the final quantization interval with learnable activation clipping. Our results suggest that activation quantization should be evaluated not only by whether outliers are reduced, but also by whether the transformed distribution uses the available low-bit levels effectively. A natural next step is to extend this analysis beyond bell-shaped activation assumptions and develop transformations that adapt to broader distributional regimes.

Limitations

The findings of this paper should be interpreted within the scope of the evaluated settings. We validate INFOQUANT on a limited set of model families, tasks, and quantization configurations, and do not claim that the same gains will automatically transfer to substantially different architectures, activation regimes, or deployment scenarios without additional study. Our method is motivated by activation patterns that have been widely observed in prior studies and are also present in the models evaluated in this paper. While these observations are sufficient to support the improvements reported here, broader validation would still be useful to determine how consistently the same behavior holds across other model families and quantization settings. From a practical perspective, the method still relies on calibration data and implementation choices such as transformation and clipping settings. Although the approach is intended for practical post-training quantization, its effectiveness may therefore vary when calibration conditions differ substantially from those used in evaluation.

References

Mohammad Sadegh Akhondzadeh, Aleksandar Bojchevski, Evangelos Eleftheriou, and Martino Dazzi. 2025. KURTAIL : KURTOSIS-BASED LLM QUANTIZATION. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024a. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. 2024b. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240.

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.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, and 1 others. 2018. A systematic classification of knowledge, reasoning, and context within the arc dataset. *arXiv preprint arXiv:1806.00358*.

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.

Mengzhao Chen, Yi Liu, Jiahao Wang, Yi Bin, Wenqi Shao, and Ping Luo. 2025. Prefixquant: Eliminating outliers by prefixed tokens for large language models quantization. *Preprint*, arXiv:2410.05265.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan-Adrian Alistarh. 2023. Optq: Accurate post-training quantization for generative pre-trained transformers. In *11th International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liulu He, Shenli Zheng, Karwei Sun, Yijiang Liu, Yufei Zhao, Chongkang Tan, Huanrui Yang, Yuan Du, and Li Du. 2025. Base-q: Bias and asymmetric scaling enhanced rotational quantization for large language models. *arXiv preprint arXiv:2506.15689*.

Xing Hu, Yuan Cheng, Dawei Yang, Zhixuan Chen, Zukang Xu, JiangyongYu, XUCHEN, Zhihang Yuan, Zhe jiang, and Sifan Zhou. 2025. OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. In *The Thirteenth International Conference on Learning Representations*.

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.

Jun Li, Fuxin Li, and Sinisa Todorovic. 2020. Efficient riemannian optimization on the stiefel manifold via the cayley transform. In *International Conference on Learning Representations*.

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.

James Liu, Pragaash Ponnusamy, Tianle Cai, Han Guo, Yoon Kim, and Ben Athiwaratkun. 2025a. Training-free activation sparsity in large language models. In *The Thirteenth International Conference on Learning Representations*.

Ruikang Liu, Haoli Bai, Haokun Lin, Yuening Li, Han Gao, Zhengzhuo Xu, Lu Hou, Jun Yao, and Chun Yuan. 2024. Intactkv: Improving large language model quantization by keeping pivot tokens intact. In *ACL (Findings)*.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025b. Spinquant: LLM quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations*.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. 2024. Affinequant: Affine transformation quantization for large language models. In *ICLR*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. 2025. Flatquant: Flatness matters for llm quantization. *Preprint*, arXiv:2410.09426.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. 2024. QuIP$\#$: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Forty-first International Conference on Machine Learning*.

Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. 2023. Outlier suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1648–1665.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.

Ke Yi, Zengke Liu, jianwei zhang, Chengyuan Li, Tong Zhang, Junyang Lin, and Jingren Zhou. 2025. Rotated runtime smooth: Training-free activation smoother for accurate INT4 inference. In *The Thirteenth International Conference on Learning Representations*.

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*.

Jiaqi Zhao, Miao Zhang, Chao Zeng, Ming Wang, Xuebo Liu, and Liqiang Nie. 2024. Lrquant: Learnable and robust post-training quantization for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2240–2255.

Appendix Overview

Appendix Overview

- Section A: Theory proofs.
- Section B: Additional implementation details.
- Section C: More ablations.
- Section D: Limitations.
- Section E: Full results.
- Section F: Visualization results.

A Theory Proofs

Theorem 1 *Let $X \sim P$ be a centered symmetric bell-shaped activation variable with standard deviation σ, and let*

$$Q_{s,c}(x) = \mathrm{clip}\Big(s\left\lfloor \frac{x}{s} \right\rceil,\ -c,\ c\Big)$$

be a finite-bit uniform round-to-nearest quantizer with step size s, clipping scale $c = Ms$, and $M = 2^{B-1} - 1$. Define the quantization centroids and cells by

$$\begin{aligned} q_i &= is, \qquad i = -M, \dots, M, \\ I_i &= \{x \mid Q_{s,c}(x) = q_i\}, \end{aligned}$$

and the smoothed quantized density

$$\begin{aligned} Q_\theta^{(c)}(x) &= \sum_{i=-M}^{M} p_i\, \delta(x - q_i; \theta), \\ p_i &= \int_{I_i} P(x)\, dx, \end{aligned}$$

where $\delta(x;\theta) = \frac{1}{2\theta}\exp(-|x|/\theta)$ and $\theta \ll s$. Then the smoothed KL objective admits the approximation

$$\begin{aligned} D_{\mathrm{KL}}(P \| Q_\theta^{(c)}) \approx{}& -H(P) + H(\{p_i\}) + \log(2\theta) \\ &+ \frac{1}{\theta}\mathcal{E}_{\mathrm{clip}}, \end{aligned}$$

where $\mathcal{E}_{\mathrm{clip}} = \mathbb{E}|X - Q_{s,c}(X)|$ is the expected absolute error of the clamped quantizer. Moreover, the normalized clipped error

$$\mathcal{E}'_{\mathrm{clip}} = \frac{1}{\sigma}\mathbb{E}|X - Q_{s,c}(X)|$$

satisfies

$$\begin{aligned} \mathcal{E}'_{\mathrm{clip}} &\le \frac{\lambda}{2} + \tau_P(\kappa), \\ \lambda &= \frac{s}{\sigma}, \qquad \kappa = \frac{c}{\sigma}, \end{aligned}$$

where $\tau_P(\kappa) = \mathbb{E}[(|Y| - \kappa)_+]$ for $Y = X/\sigma$. With fixed bit-width B, this bound can be written as

$$\begin{aligned} F_{P,B}(\kappa) &= \frac{\kappa}{2M} + \tau_P(\kappa), \\ \kappa &= \frac{1}{b_n}. \end{aligned}$$

For Gaussian and Laplace activations, τ_P has the closed forms given below, and $F_{P,B}(\kappa)$ is increasing in κ once the clipping tail is sufficiently controlled. Equivalently, in this tail-controlled regime, a smaller clipped range relative to the activation spread and a larger max-normalized dispersion $b_n = \sigma/c$ tighten the bound.

A.1 Proof of Theorem 1

For the finite-bit clamped quantizer, the interior cells are

$$\begin{aligned} I_i &= \left[\left(i - \frac{1}{2}\right)s, \left(i + \frac{1}{2}\right)s\right), \\ &|i| < M, \end{aligned}$$

while the boundary cells absorb the clipped tails:

$$\begin{aligned} I_M &= \left[c - \frac{s}{2}, \infty\right), \\ I_{-M} &= \left(-\infty, -c + \frac{s}{2}\right]. \end{aligned}$$

For $x \in I_i$, the quantizer outputs the centroid $q_i = is$. Under the kernel-separation assumption $\theta \ll s$, neighboring kernels contribute negligibly around each cell, and we may approximate

$$\begin{aligned} Q_\theta^{(c)}(x) &\approx p_i\, \delta(x - q_i; \theta) \\ &= \frac{p_i}{2\theta}\exp\left(-\frac{|x - q_i|}{\theta}\right), \qquad x \in I_i. \end{aligned}$$

Substituting this expression into the KL divergence gives

$$\begin{aligned} D_{\mathrm{KL}}(P \| Q_\theta^{(c)}) &= \sum_{i=-M}^{M} \int_{I_i} P(x) \log\left(\frac{P(x)}{Q_\theta^{(c)}(x)}\right) dx \\ &\approx \sum_{i=-M}^{M} \int_{I_i} P(x) \log P(x)\, dx \\ &\quad - \sum_{i=-M}^{M} \log(p_i) \int_{I_i} P(x)\, dx \\ &\quad + \frac{1}{\theta}\sum_{i=-M}^{M} \int_{I_i} |x - q_i| P(x)\, dx \\ &\quad + \log(2\theta). \end{aligned} \tag{13}$$

Using

$$\sum_i \int_{I_i} P(x)\log P(x)\,dx = -H(P),$$
$$\sum_i \log(p_i)\int_{I_i} P(x)\,dx = \sum_i p_i \log p_i = -H(\{p_i\}),$$

we obtain

$$D_{\mathrm{KL}}(P\|Q_\theta^{(c)}) \approx -H(P) + H(\{p_i\}) + \log(2\theta) + \frac{1}{\theta}\mathcal{E}_{\mathrm{clip}}, \tag{14}$$

where

$$\mathcal{E}_{\mathrm{clip}} = \sum_{i=-M}^{M}\int_{I_i} |x-q_i|P(x)\,dx = \mathbb{E}|X - Q_{s,c}(X)|.$$

To upper bound the practical clamped-quantization error, define the clipping projection $\Pi_c(x) = \mathrm{clip}(x,-c,c)$. Then

$$|x - Q_{s,c}(x)| \le |x - \Pi_c(x)| + |\Pi_c(x) - Q_{s,c}(x)| \le (|x|-c)_+ + \frac{s}{2}. \tag{15}$$

The second term is the standard in-range round-to-nearest error, while the first term is the clipping-tail error. Taking expectations and dividing by σ gives

$$\mathcal{E}'_{\mathrm{clip}} := \frac{1}{\sigma}\mathbb{E}|X - Q_{s,c}(X)| \le \frac{s}{2\sigma} + \frac{1}{\sigma}\mathbb{E}[(|X|-c)_+]\,. \tag{16}$$

With the normalized variable $Y = X/\sigma$, $\lambda = s/\sigma$, and $\kappa = c/\sigma$, this becomes

$$\mathcal{E}'_{\mathrm{clip}} \le \frac{\lambda}{2} + \tau_P(\kappa),$$
$$\tau_P(\kappa) := \mathbb{E}[(|Y|-\kappa)_+]\,. \tag{17}$$

Because $c = Ms$ for a fixed B-bit quantizer, we have $\kappa = M\lambda$ and therefore

$$F_{P,B}(\kappa) = \frac{\kappa}{2M} + \tau_P(\kappa).$$

Gaussian activations. If $X \sim \mathcal{N}(0,\sigma^2)$, then $Y \sim \mathcal{N}(0,1)$ with density $\phi(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$ and survival function $\bar{\Phi}(y) = 1 - \Phi(y)$. The clipping-tail term is

$$\tau_G(\kappa) = 2\int_\kappa^\infty (y-\kappa)\phi(y)\,dy = 2\big(\phi(\kappa) - \kappa\bar{\Phi}(\kappa)\big). \tag{18}$$

Hence

$$F_{G,B}(\kappa) = \frac{\kappa}{2M} + 2\big(\phi(\kappa) - \kappa\bar{\Phi}(\kappa)\big),$$

and

$$F'_{G,B}(\kappa) = \frac{1}{2M} - 2\bar{\Phi}(\kappa).$$

Therefore $F_{G,B}$ is increasing whenever

$$\kappa \ge \kappa_G^\star := \Phi^{-1}\left(1 - \frac{1}{4M}\right).$$

In this tail-controlled regime, decreasing $\kappa = c/\sigma$ reduces the bound.

Laplace activations. If $X \sim \mathrm{Laplace}(0,b)$, then $\sigma = \sqrt{2}\,b$. We keep the same normalized error definition as above,

$$\mathcal{E}'_{\mathrm{clip}} = \frac{1}{\sigma}\mathbb{E}|X - Q_{s,c}(X)|,$$

so the Laplace case is parameterized by the scale b only through the identity $\sigma = \sqrt{2}b$. With the normalized variable $Y = X/\sigma$, the density becomes

$$f_Y(y) = \frac{1}{\sqrt{2}}e^{-\sqrt{2}|y|}.$$

The clipping-tail term becomes

$$\tau_L(\kappa) = 2\int_\kappa^\infty (y-\kappa)\frac{1}{\sqrt{2}}e^{-\sqrt{2}y}\,dy = \frac{1}{\sqrt{2}}e^{-\sqrt{2}\kappa}. \tag{19}$$

Hence

$$F_{L,B}(\kappa) = \frac{\kappa}{2M} + \frac{1}{\sqrt{2}}e^{-\sqrt{2}\kappa},$$
$$F'_{L,B}(\kappa) = \frac{1}{2M} - e^{-\sqrt{2}\kappa}.$$

Therefore $F_{L,B}$ is increasing whenever

$$\kappa \ge \kappa_L^\star := \frac{\log(2M)}{\sqrt{2}}.$$

Again, in this regime, decreasing κ tightens the bound. Equivalently, one may express the same condition in terms of b via $\kappa = c/(\sqrt{2}b)$, but the

normalized error itself remains defined with respect to σ.

Finally, the normalized dispersion metric used in the main text is

$$b_n = \frac{\sigma}{c} = \frac{1}{\kappa}, \qquad \lambda = \frac{s}{\sigma} = \frac{s/c}{\sigma/c} = \frac{\bar{s}}{b_n}.$$

Therefore, for a fixed bit-width and a controlled clipping tail, a smaller clipped range relative to the activation spread and a larger max-normalized dispersion b_n both correspond to a smaller κ and a tighter error bound.

A.2 Why b_n Tends to Increase Under PSOT

Based on the analysis in the motivation section, the normalized surrogate error is controlled by the ratio between the quantization step size and the distributional spread. Since practical activation quantization uses a clipped numerical range, we normalize both quantities by the clipping scale. This yields the dispersion metric b_n used in the main text.

Suppose we apply a rotation to an activation token $\mathbf{t}$ with d dimensions using an orthogonal matrix $\mathbf{A}$, and then center it as in Eq. (6), yielding $\mathbf{t}' = \mathbf{tA} - \mathbb{E}[\mathbf{tA}]$. The optimization objective aims to reduce $\|\mathbf{t}'\|_\infty$, which lowers the clipping scale that determines the quantization step size. Since $\mathbf{A}$ is orthogonal, the Euclidean norm remains invariant under rotation, and centering only removes the mean component:

$$\|\mathbf{t}'\|_2 \leq \|\mathbf{tA}\|_2 = \|\mathbf{t}\|_2 \tag{20}$$

For a centered token, the range-normalized dispersion satisfies

$$b_n(\mathbf{t}') = \sqrt{\frac{1}{d}} \frac{\|\mathbf{t}'\|_2}{\|\mathbf{t}'\|_\infty}. \tag{21}$$

Comparing the transformed and original tokens gives

$$\frac{b_n(\mathbf{t}')}{b_n(\mathbf{t})} = \frac{\|\mathbf{t}'\|_2}{\|\mathbf{t}\|_2} \cdot \frac{\|\mathbf{t}\|_\infty}{\|\mathbf{t}'\|_\infty}. \tag{22}$$

Therefore, when PSOT suppresses the peak value while preserving most of the centered token energy, the decrease in $\|\mathbf{t}'\|_\infty$ can dominate the mild change in $\|\mathbf{t}'\|_2$, causing $b_n(\mathbf{t}')$ to increase. This is a conditional mechanism rather than a universal guarantee, but it explains why peak suppression empirically tends to produce both a smaller effective clipping range and a more dispersed normalized distribution.

B Additional Implementation Details

Additional Setup. To improve efficiency, block diagonal matrices are used, partitioned into two blocks for 7B–13B models and four blocks for the 70B model. The PSOT temperature is set to $T = 2$, with a batch size of 4 and an initial learning rate of 2, linearly decayed over 15 epochs. For ASOT, we use $m = 10$ calibration samples and select the hyperparameters by grid search on the calibration set; the final setting uses $\delta = 0.02$, $\tau = 0.4$, and $\gamma = 30$. In the subsequent **LAC** phase, the initial clipping ratio is set to 0.95, constrained within the interval $[0.5, 1]$. We use the AdamW optimizer with an initial learning rate of 0.05, applying a cosine annealing decay schedule over 5 epochs with a batch size of 4.

Computational Graph. To better accommodate the varying degrees of discrepancy across activation at different layers, our proposed INFOQUANT method employs PSOT to optimize the distribution of each layer's activation individually. This per-layer optimization yields improved quantization performance. However, to maintain computational consistency in the presence of residual connections, an additional matrix multiplication is required specifically at these connection points, as illustrated in Figure 6. To mitigate the overhead introduced by this operation, we adopt block-diagonal orthogonal matrices as the rotation matrices, which significantly reduce inference cost and enhance the efficiency of the PSOT process. For fair comparison, the INFOQUANT* employs a global rotation matrix, as shown in Figure 5. This approach eliminates the need for additional computation at residual connections and results in a computation graph equivalent to that of SpinQuant (Liu et al., 2025b).

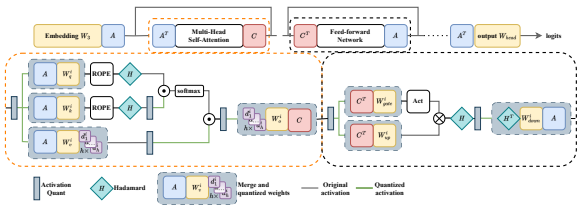

Figure 5: overall diagram of rotation of INFOQUANT*.

Runtime and Memory Overhead. Table 4 presents the optimization runtime and memory consumption for a single Transformer layer of the LLaMA-2 models (7B, 13B, and 70B), measured

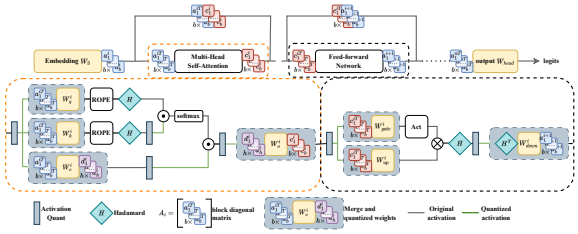

Figure 6: overall diagram of rotation of INFOQUANT.

Model	PSOT		LAC	
	Time	Memory	Time	Memory
LLaMA-2 7B	∼6.0	∼6	∼3.0	∼10
LLaMA-2 13B	∼7.5	∼7	∼2.0	∼12
LLaMA-2 70B	∼13.0	∼11	∼3.5	∼23

Table 4: Per-layer optimal time (min) and GPU memory usage (GB) on NVIDIA RTX 4090.

on an NVIDIA RTX 4090 GPU. Remarkably, our method enables quantization of the 70B model using just 24GB of GPU memory, demonstrating compatibility with consumer-grade hardware. Note that GPUs with larger memory capacities can reduce optimization time by enabling greater parallelism, potentially achieving multi-fold speedups.

C More Ablations

Weighting factor γ	Wikitext2 PPL
1	7.03
30	**6.89**
50	6.97
80	6.93
$+\infty$	7.02

Table 5: Ablation study of weighting factor γ on WikiText2 PPL using RTN quantization under W4A4KV4 configuration for LLaMA-2 7B.

Influence of Weighting Factor γ. The weighting factor γ controls how strongly ASOT emphasizes outlier tokens in Eq. (10). Table 5 shows that a moderate emphasis works best: $\gamma = 30$ achieves the lowest perplexity, while both weak emphasis ($\gamma = 1$) and exclusive focus on outlier tokens ($\gamma = +\infty$) are worse. This trend suggests that outlier tokens provide high-value learning signals for peak suppression, but normal tokens are still necessary to preserve the overall activation distribution. In other words, ASOT should rebalance the optimization objective rather than collapse it into an outlier-only objective.

Influence of Temperature T in PSOT. The temperature T controls how selectively PSOT suppresses high-magnitude activation coordinates. As shown in Table 6, $T = 2$ achieves the best perplexity, while both sharper weighting ($T = 0.3$) and flatter weighting ($T = 8$) degrade performance. This pattern reveals a useful design trade-off. If T is too small, the objective concentrates on only a few extreme coordinates and may overfit the calibration activations. If T is too large, the softmax weights become nearly uniform and the objective loses its ability to target peaks. A moderate temperature therefore provides the best balance between peak suppression and distribution-level stability.

Temperature T	Wikitext2 PPL
0.3	7.20
1	7.11
2	**6.89**
4	7.04
8	7.22

Table 6: Effect of the PSOT temperature T on WikiText2 perplexity using RTN quantization under the W4A4KV4 configuration for LLaMA-2 7B.

Block Sizes of Block-diagonal Orthogonal Matrices. To enhance computational and memory efficiency during activation rotation and inference, we adopt block-diagonal orthogonal matrices. However, this structure inherently limits interaction to within individual blocks, thereby restricting global information aggregation. As shown in Table 7, increasing the number of blocks, which corresponds to a reduction in block size, results in a decline in quantization performance. This observation aligns with findings from (Lin et al., 2024a), which attributes the degradation to inter-block shifts in activation means that hinder quantization efficiency. We further hypothesize that full orthogonal matrices enable more effective redistribution of activation values due to the cumulative contribution of their unit vectors across higher dimensions. In contrast, block-diagonal matrices reduce this capability, thereby diminishing the range compression essential for effective quantization. To balance these competing considerations, we select an intermediate block size, as detailed in the Implementation details.

Robustness of Initialization A. To investigate the robustness of the proposed PSOT algorithm with respect to orthogonal initialization strategies, we conduct an ablation study comparing Hadamard and randomly generated orthogonal matrices. For

Table 7: Impact of block number in block-diagonal orthogonal matrices on quantization performance under W4A4KV4 configuration with GPTQ.

Block number	LLaMA-3 8B	LLaMA-2 13B
1	**7.13**	**5.16**
2	7.18	5.18
4	7.17	5.32
8	7.39	5.37

the latter, we start from uniformly sampled random matrices and apply QR decomposition to ensure orthogonality. As shown in Table 8, we evaluate both initialization methods on LLaMA-2 13B and LLaMA-3 8B models. While previous work (Ashkboos et al., 2024b) reported a notable performance gap in favor of Hadamard-based initialization, our findings demonstrate that, after optimization via PSOT, both initialization schemes yield similarly stable and effective results. This suggests that PSOT is robust to the choice of orthogonal basis at initialization.

Table 8: Impact of initialization of PSOT under W4A4KV4 configuration with GPTQ.

Model	Initialization	WikiText2 PPL
LLaMA-3 8B	Hadamard	7.18
	Random	7.18
LLaMA-2 13B	Hadamard	5.18
	Random	5.19

Clipping Ratio Ablation. The orthogonal matrix obtained by PSOT reduces quantization error through rotation, but this benefit comes at the cost of increased sensitivity to clipping thresholds. As shown in Table 9. Small changes in the clipping thresholds (α, β) can lead to significant variations in the model's performance, highlighting a trade-off between reduced quantization error and the difficulty of fine-tuning the clipping parameters for optimal results.

Table 9: WikiText perplexity of LLAMA 2-7B after PSOT, evaluated with different clipping ratios. To assess the sensitivity to various clipping ratios, all results were obtained using RTN quantization with W4A4KV4.

clip radio α, β	Wikitext2 PPL
1	7.21
0.95	7.06
0.9	6.98
0.85	7.00
0.8	8.02

D Limitations

Our work proposes a post-training quantization method that achieves state-of-the-art performance in quantizing large language models (LLMs) under strict hardware constraints. However, it also has several limitations. First, we computes the Kullback-Leibler divergence before and after quantization to capture quantization error, and this final approach has shown promising results. Nevertheless, the precise mechanisms by which activation distribution shapes affect LLM performance are not yet fully understood. Second, our method PSOT is designed based on the observation that neural network activations generally follow a ball-shaped distribution. However, our approach has not been rigorously tested on distributions that deviate from this assumption. Future research should investigate more flexible transformation strategies that make a broader range of activation distributions amenable to quantization.

E More Results

Results for Qwen family For a more comprehensive evaluation, we further test our method on the Qwen-2.5 models in Table 10. Experimental results show that under 4-4-4 quantization, our method maintains superior performance on both the 14B and 32B models, highlighting its robustness and scalability.

#Bits W-A-KV	Method	Qwen-2.5 14B 0-shot[9] Avg.(↑)	Qwen-2.5 14B Wiki (↓)	Qwen-2.5 32B 0-shot[9] Avg.(↑)	Qwen-2.5 32B Wiki (↓)
16-16-16	FloatingPoint	70.95	5.29	71.11	5.02
4-4-4	QuaRot	67.23	6.77	68.14	6.04
	SpinQuant	67.29	6.55	68.51	5.88
	OSTQuant	67.81	6.37	OOM	OOM
	INFOQUANT	**67.65**	**6.30**	**69.90**	**5.69**

Table 10: Evaluation results on Qwen2.5 models. The results for QuaRot and SpinQuant are reproduced using their respective official open-source implementations. Due to the memory limitations, the Qwen-2.5 32B models were not evaluated using the OSTQuant codebase.

Full Results In Table 11, we show the complete results of INFOQUANT in experiment Section. We compare the perplexity score on WikiText2 and the accuracy on nine zero-shot tasks using the `lm-evaluation-harness` (version 0.4.7)(Gao et al., 2024), including BoolQ(Clark et al., 2019), HellaSwag (Zellers et al., 2019), LAMBADA (OpenAI)(Radford et al., 2019), OpenBookQA (OBQA)(Mihaylov et al., 2018), PIQA (Bisk

et al., 2020), SIQA (Sap et al., 2019), Wino-Grande (Sakaguchi et al., 2021), ARC-Easy, and ARC-Challenge (Boratko et al., 2018).

F Visualization results

Figure 7 and Figure 8 shows the activation distribution of different layers in LLaMA-2-7B and LLaMA-3-8B.

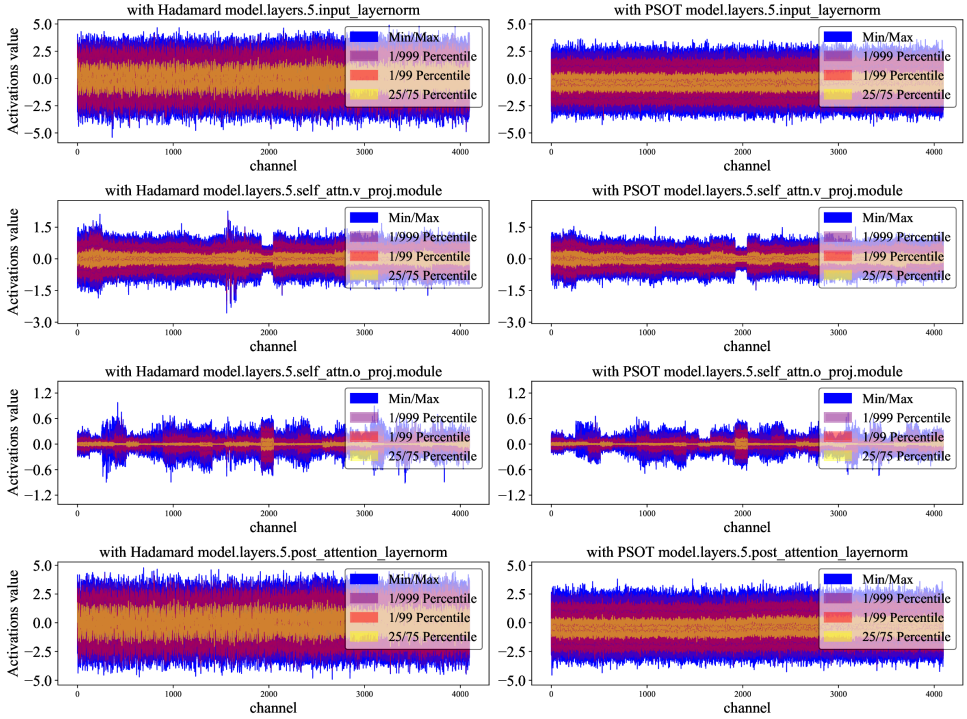

Figure 7: The rotated activation distribution of different layers in LLaMA-2 7B with Hadamard and PSOT.

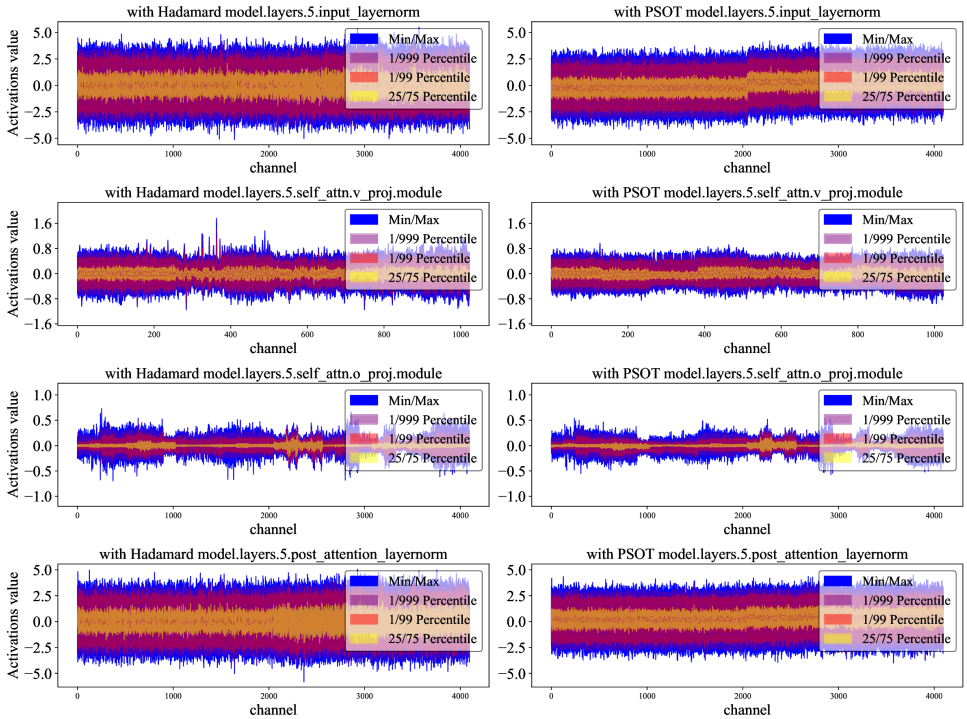

Figure 8: The rotated activation distribution of different layers in LLaMA-3 8B with Hadamard and PSOT.

Table 11: Full INFOQUANT's results of the perplexity score on WikiText2 and averaged accuracy on all task on **LLaMA-2 & 3**.

Model	#Bits W-A-KV	ARC-c (↑)	ARC-e (↑)	BoolQ (↑)	HellaS. (↑)	Lam. (↑)	OBQA (↑)	PIQA (↑)	SIQA (↑)	WinoG. (↑)	Avg. (↑)	Wiki2 (↓)
2-7B	16-16-16	46.42	74.33	77.71	75.94	73.69	44.20	79.16	45.91	69.53	65.21	5.47
	4-16-16	44.62	73.86	76.67	75.22	72.87	43.60	78.02	45.65	68.43	64.34	5.60
	4-4-16	43.34	71.25	75.44	74.10	71.90	40.80	77.09	44.98	66.61	62.84	5.86
	4-4-4	43.17	71.59	75.60	73.92	72.39	42.20	77.69	45.14	66.77	63.16	5.89
2-13B	16-16-16	49.15	77.53	80.58	79.39	76.62	45.20	80.63	47.49	71.90	67.61	4.88
	4-16-16	48.89	77.31	79.82	78.81	76.40	45.20	79.92	46.72	72.38	67.27	4.99
	4-4-16	48.98	75.72	80.06	78.40	75.39	44.80	79.43	46.32	71.19	66.71	5.15
	4-4-4	48.04	75.34	79.36	77.97	75.47	44.40	79.54	45.60	71.27	66.33	5.18
2-70B	16-16-16	57.42	81.02	83.79	83.81	79.60	48.80	82.70	49.18	77.98	71.59	3.32
	4-16-16	57.25	80.86	82.96	83.37	79.66	48.20	82.92	48.82	77.27	71.25	3.40
	4-4-16	55.72	80.01	82.35	82.86	79.64	49.00	82.05	48.52	77.27	70.82	3.62
	4-4-4	55.89	80.05	81.99	82.53	78.92	47.60	82.15	48.67	75.37	70.35	3.64
3-8B	16-16-16	53.50	77.74	81.10	79.18	75.74	44.80	80.63	47.08	73.01	68.09	6.14
	4-16-16	53.07	77.44	78.44	78.18	74.77	44.00	80.30	46.62	73.40	67.36	6.48
	4-4-16	48.81	77.02	77.77	76.67	73.30	43.80	78.45	44.52	71.35	65.74	7.07
	4-4-4	50.00	75.98	75.65	76.28	72.46	44.60	78.67	45.50	71.03	65.57	7.16
3-70B	16-16-16	64.42	85.98	85.14	84.95	79.47	48.46	84.39	50.82	80.66	73.81	2.86
	4-16-16	62.63	85.19	86.21	84.35	78.27	47.00	84.49	50.73	80.43	73.25	3.50
	4-4-16	57.34	81.14	84.92	82.88	77.94	45.80	81.66	48.41	76.32	70.71	5.24
	4-4-4	56.66	81.31	83.61	82.24	76.23	45.60	82.05	47.65	76.56	70.21	5.39