SpecQuant: Spectral Decomposition and Adaptive Truncation for Ultra-Low-Bit LLMs Quantization

Zhixiong Zhao[1,2*†], **Fangxin Liu**[1,3*‡], **Junjie Wang**[1,3], **Chenyang Guan**[1], **Zongwu Wang**[1,3]
Li Jiang[1,3], **Haibing Guan**[1]

[1]Shanghai Jiao Tong University
[2]Nanyang Technological University
[3]Shanghai Qi Zhi Institute
zhixiong003@e.ntu.edu.sg, {liufangxin, ljiang_cs}@sjtu.edu.cn

Abstract

The emergence of accurate open large language models (LLMs) has sparked a push for advanced quantization techniques to enable efficient deployment on end-user devices. In this paper, we revisit the challenge of extreme LLM compression—targeting ultra-low-bit quantization for both activations and weights—from a Fourier frequency domain perspective. We propose SpecQuant , a two-stage framework that tackles activation outliers and cross-channel variance. In the first stage, activation outliers are smoothed and transferred into the weight matrix to simplify downstream quantization. In the second stage, we apply channel-wise low-frequency Fourier truncation to suppress high-frequency components while preserving essential signal energy, improving quantization robustness. Our method builds on the principle that most of the weight energy is concentrated in low-frequency components, which can be retained with minimal impact on model accuracy. To enable runtime adaptability, we introduce a lightweight truncation module during inference that adjusts truncation thresholds based on channel characteristics. On LLaMA-3 8B, SpecQuant achieves 4-bit quantization for both weights and activations, narrowing the zero-shot accuracy gap to only 1.5% compared to full precision, while delivering 2× faster inference and 3× lower memory usage.

Introduction

Large Language Models (LLMs) (Touvron et al. 2023; Guo et al. 2025; Achiam et al. 2023) have demonstrated impressive performance across a wide range of natural language processing tasks, including code generation and open-ended reasoning. These advances are largely driven by massive model scales and extensive pretraining data. However, the resulting memory footprint and computational cost pose significant challenges for deployment on edge devices (Lin et al. 2024a; Heo et al. 2023).

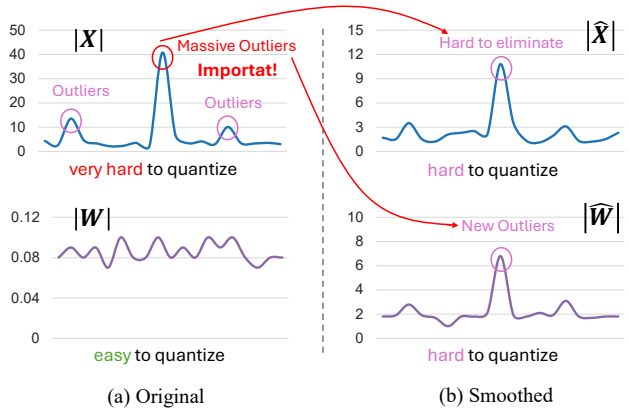

Figure 1: Activation and weight distributions before and after naive smoothing. While smoothing aims to mitigate activation outliers, it often transfers the quantization burden to weights, introducing new outliers and degrading the robustness of both activations and weights under quantization.

To reduce memory and accelerate inference, recent quantization methods (Liu et al. 2024; Hu et al. 2025a) aim to represent both weights and activations in low-bit formats (Lin et al. 2024b; Zhao et al. 2025), enabling faster matrix multiplications and smaller model sizes (Xiao et al. 2023). Despite their effectiveness, a core challenge remains: *activation outliers*, which expand the dynamic range and induce significant accuracy degradation when quantized (Wei et al. 2022).

Existing approaches attempt to mitigate outlier impact via distributional transformations. SmoothQuant (Xiao et al. 2023) shifts activation outliers into weights via layer-wise scaling; OSTQuant (Hu et al. 2025b), SpinQuant (Liu et al. 2024) and QuaRot (Ashkboos et al. 2024) employ lightweight rotation layers to reduce activation variance; SVDQuant (Li et al. 2024) applies global low-rank approximation to absorb outliers. However, these strategies face fundamental limitations. As shown in Figure 1, smoothing-based methods often transfer the burden of quantization from activations to weights without eliminating it. Rotation-based techniques introduce non-negligible runtime overhead, while SVD-based

*These authors contributed equally.
†This work was done when Zhixiong Zhao was an intern at Shanghai Jiao Tong University.
‡Fangxin Liu is the corresponding author.

approximations fail to preserve channel-wise outlier structures critical for contextual understanding.

Recent work (Jin et al. 2025) further reveals that extreme activation values often encode fine-grained contextual cues, essential for reasoning tasks. Thus, indiscriminate quantization of these values leads to substantial performance loss on long-context benchmarks. This motivates the need for a more principled and robust quantization strategy that preserves informative outliers without incurring high computational cost.

In this work, we propose SpecQuant, a novel compression framework based on **adaptive Fourier-domain decomposition**, which explicitly targets the spectral structure of weights induced by smoothed activations. SpecQuant consists of two stages: (1) activation smoothing to migrate outliers into the weight domain, and (2) channel-wise low-frequency truncation to suppress the transferred high-frequency noise while preserving signal fidelity. We observe that weights exhibit strong low-frequency bias in the Fourier domain. This property allows us to truncate high-frequency components with minimal impact on model accuracy, enabling more robust low-bit quantization: **Our contributions are summarized as follows:**

- **Frequency Domain Approximation:** We are the first to bridge a connection between frequency-domain compression and quantization robustness in LLMs. Our analysis leverages Fourier energy decay properties to provide theoretical guarantees for preserving accuracy under aggressive quantization.
- **Outlier-Resilient Spectral Quantization:** We propose SpecQuant , a novel two-stage quantization framework that first absorbs activation outliers via scaling-based smoothing and then performs adaptive, channel-wise spectral truncation in the Fourier domain, effectively mitigating quantization error caused by outlier redistribution.
- **Extensive Evaluation:** We evaluate SpecQuant on eight LLMs across ten datasets, showing up to 3× memory reduction and 1.7× speedup with only 1.5% accuracy drop, outperforming prior SOTA methods.

Related Works

LLM Compression and Outlier Mitigation

Recent progress in post-training quantization of LLMs below 4-bit precision—such as SmoothQuant (Xiao et al. 2023), OPTQ (Frantar et al. 2023), QuIP (Chee et al. 2023), SpinQuant (Liu et al. 2024), QuaRot (Ashkboos et al. 2024), SVDQuant (Li et al. 2024), and QuIP# (Tseng et al. 2024)—has focused on mitigating activation and weight outliers. These outliers, particularly extreme weight values, expand the quantization dynamic range, leading to significant accuracy degradation. SmoothQuant addresses this by shifting quantization difficulty from activations to weights via layer-wise scaling. SpinQuant and QuaRot introduce rotation-based strategies: SpinQuant learns orthogonal transforms to align weight-activation distributions, while QuaRot applies random rotations to suppress outliers. SVDQuant builds on SmoothQuant by decomposing weight matrices using singular value decomposition (SVD), isolating compressible residuals. Meanwhile, QuIP and QuIP# leverage randomized matrix transformations and E8 lattice structures for improved vector quantization. Both adopt error-compensated column-wise quantization from OPTQ, propagating residuals to enhance later quantization steps.

Despite these advances, challenges remain. Migration-based methods struggle with extreme or dense outlier distributions, rotation-based techniques introduce inference overhead, and global decomposition methods often fail to capture channel-specific outlier patterns. These limitations hinder robust ultra-low-bit quantization for real-world LLM deployment.

Frequency-Domain Optimization Strategies

Frequency-domain optimization has emerged as a powerful tool across machine learning, offering unique advantages in compression and efficiency. In computer vision, spectral representations help stabilize training and improve generalization. For example, FcaNet (Qin et al. 2021) applies discrete cosine transforms (DCT) to enhance channel attention, while F3D (Liu et al. 2021) replaces 3D convolutions with frequency-domain operations to improve hardware efficiency. In time-series forecasting, FreDF (Wang et al. 2024) uses spectral decomposition with consistency constraints to address temporal error propagation. Recently, frequency-domain principles have begun influencing LLM optimization. FourierFT (Gao et al. 2024) enables parameter-efficient fine-tuning by learning sparse frequency coefficients and reconstructing full model weights via inverse DFT. Notably, frequency projections align well with LLM channel structures and naturally absorb multi-scale activation outliers through bandwidth filtering.

This compatibility motivates the use of frequency-domain decomposition as an effective strategy for robust quantization. However, despite this promising synergy, systematic exploration of frequency-domain quantization for LLMs remains limited in current efforts.

Method

In this section, we first establish the theoretical foundation of Low-Frequency Fourier Projection for channel vectors, laying the groundwork for our subsequent methodology. We then introduce SpecQuant , a novel quantization paradigm specifically designed for LLMs. The core innovation of our approach is an auxiliary low-frequency spectral truncation branch that effectively mitigates channel-wise quantization challenges in both weight matrices and activation tensors, as illustrated in Figure 2.

Preliminaries

Quantization Multi-head Self-Attention (MSA) and Feed-Forward Network (FFN), the core components within each Transformer block of LLMs, fundamentally rely on standard linear transformations formulated as $Y = X \cdot W \in \mathbb{R}^{T \times C_{\text{out}}}$, where, $X \in \mathbb{R}^{T \times C_{\text{in}}}$ denotes the input activation matrix and $W \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$ is the corresponding weight matrix. In this work, we apply uniform integer quantization to both activations and weights to improve hardware efficiency. Specifi-

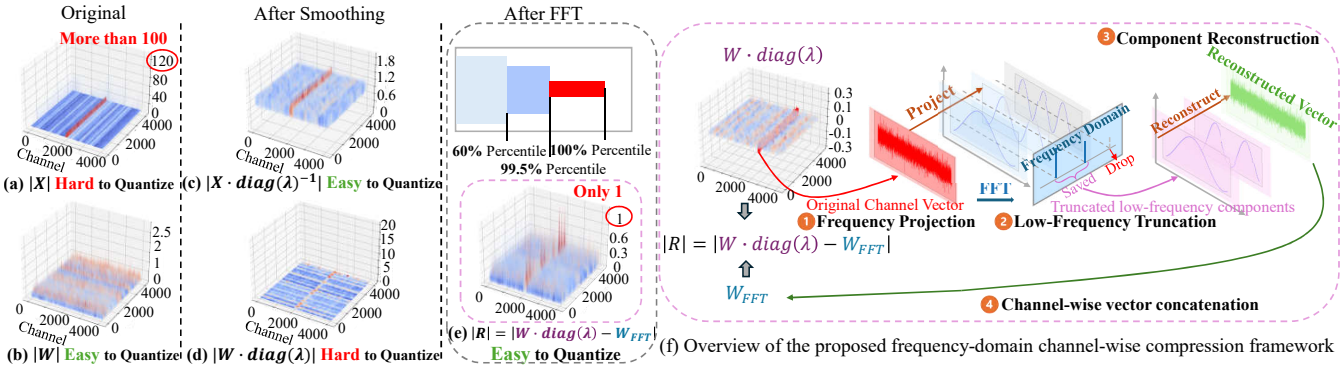

Figure 2: Overview of the proposed SpecQuant .

cally, a b-bit quantization maps a floating-point tensor X into a low-bit integer representation X_q as follows:

$$X_q = \left\lfloor \text{clamp}\left(\frac{X}{\Delta} + z,\, 0,\, 2^b - 1\right)\right\rceil \tag{1}$$

where $\Delta = \frac{\max(X)-\min(X)}{2^b-1}$ is the quantization step size, and $z = -\frac{\min(X)}{\Delta}$ is the zero-point offset. The operator $\lfloor\cdot\rceil$ denotes rounding to the nearest integer. Following prior works (Frantar et al. 2022; Lin et al. 2024a), we adopt per-token quantization for activations and per-channel quantization for weights.

Frequency Domain Projection Within each Transformer block, the linear transformation $Y = X \cdot W \in \mathbb{R}^{T\times C_{\text{out}}}$ remains the fundamental operation, where $X \in \mathbb{R}^{T\times C_{\text{in}}}$ and $W \in \mathbb{R}^{C_{\text{in}}\times C_{\text{out}}}$ are as defined above. To facilitate compression of weight matrices while maintaining hardware compatibility, we propose projecting weight matrices into the frequency domain using the Discrete Fourier Transform (DFT):

$$W_{\text{freq}}[k] = \sum_{n=0}^{N-1} W[n] \cdot e^{-i\frac{2\pi kn}{N}}, \quad k = 0, 1, \ldots, N-1 \tag{2}$$

where N denotes the size of the input dimension. Due to the $\mathcal{O}(N^2)$ complexity of naive DFT computation, we leverage the Fast Fourier Transform (FFT), which reduces this to $\mathcal{O}(N \log N)$, and is efficiently implemented in libraries such as CUDA FFT. The FFT can be recursively defined as:

$$W_{\text{freq}} = \text{FFT}(W) = \text{Combine}\left(\text{FFT}(W_{\text{even}}), \text{FFT}(W_{\text{odd}})\right) \tag{3}$$

By preserving low-frequency components, which concentrate the majority of the signal energy, and discarding less informative high-frequency components, this method achieves significant compression. Also, it maintains hardware-agnostic deployment flexibility, making it a practical and scalable solution for efficient LLM inference.

Low-Frequency Fourier Projection of Channel Vectors

Existing studies have confirmed that the outlier characteristics of LLM weights and activations exhibit significant cross-channel heterogeneity—dynamic range differences between channels can reach 3-4 orders of magnitude, and the proportion of high-frequency energy shows a strong negative correlation with task relevance (Lin et al. 2024a; Liu et al. 2024). To precisely capture such channel-specific properties, we propose a **channel-independent frequency-domain projection framework**, with the core assumption that each output channel's weight vector $\mathbf{W}[:, j] \in \mathbb{R}^{C_{\text{in}}}$ in LLMs can be treated as an independent stationary signal, whose energy is primarily concentrated in low-frequency components. This assumption is empirically supported: in the attention layers of LLaMA-2 7B, the average low-frequency (top 20% frequencies) energy proportion of 1000 randomly sampled channel vectors reaches 92.3%, with a standard deviation of only 3.7%.

For the weight vector $\mathbf{W}[:, j] = [w_0, w_1, \ldots, w_{C_{\text{in}}-1}]^\top \in \mathbb{R}^{C_{\text{in}}}$ of the j-th output channel (where $w_n \in \mathbb{R}$ are real-valued weights), its discrete Fourier transform is defined as:

$$\mathbf{W}_{\text{freq}}[k, j] = \sum_{n=0}^{C_{\text{in}}-1} w_n \cdot e^{-i\frac{2\pi kn}{C_{\text{in}}}}, \quad k = 0, 1, \ldots, C_{\text{in}} - 1 \tag{4}$$

where i is the imaginary unit, k is the frequency index, and $e^{-i\frac{2\pi kn}{C_{\text{in}}}} = \cos\left(\frac{2\pi kn}{C_{\text{in}}}\right) - i\sin\left(\frac{2\pi kn}{C_{\text{in}}}\right)$ is the complex exponential function. Since $\mathbf{W}[:, j]$ is a real-valued vector, its DFT coefficients satisfy **conjugate symmetry**: $\mathbf{W}_{\text{freq}}[k, j] = \overline{\mathbf{W}_{\text{freq}}[C_{\text{in}} - k, j]}$ (where $\bar{\cdot}$ denotes complex conjugation). This property allows us to store only the first $\lceil C_{\text{in}}/2 \rceil$ coefficients for complete reconstruction of the original vector, naturally reducing the frequency-domain storage overhead by 50%.

Specifically, each channel vector $\mathbf{W}[:, j] \in \mathbb{R}^{C_{\text{in}}}$ is independently transformed into the frequency domain via FFT:

$$\mathbf{W}_{\text{freq}}[:, j] = \texttt{FFT}(\mathbf{W}[:, j]) = \sum_{n=0}^{C_{\text{in}}-1} \mathbf{W}[n, j] \cdot e^{-i2\pi kn/C_{\text{in}}}, \quad k \in \{0, 1, \ldots, C_{\text{in}} - 1\} \tag{5}$$

This yields complex spectral coefficients $X_k \in \mathbb{C}$ with $O(C_{\text{in}} \log C_{\text{in}})$ complexity. By applying strategic spectral sparsification—contrasting with global SVD-based approximations—SpecQuant achieves fine-grained outlier suppression at channel level.

Similar to how SVD retains top-k singular values for approximation, Fourier analysis shows that smooth signals concentrate most energy in low-frequency components. This

provides a theoretical basis for our compression approach. Formally, given a real-valued discrete signal $x[n]$ of length N with FFT:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i2\pi kn/N}, \quad k = 0, 1, \ldots, N-1. \tag{6}$$

Parseval's theorem (Kelkar, Grigsby, and Langsner 1983) ensures energy preservation between time and frequency domains:

$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]|^2. \tag{7}$$

Here, the squared magnitude $|X[k]|^2$ represents the energy contribution at frequency k.

Assuming $x[n]$ samples a continuous function $f(t)$ with r continuous derivatives ($f \in C^r$), classical Fourier theory states that smoother functions have faster decaying Fourier coefficients:

$$|X[k]| \leq \frac{C}{|k|^r}, \quad \text{for some constant } C, \tag{8}$$

As k increases, $|X[k]|$ rapidly decreases, making high-frequency energy negligible. Thus, most energy concentrates in low frequencies, justifying truncated Fourier representations to approximate the original signal with minimal loss.

Based on this, Figure 2 (f) illustrates the compression and reconstruction of a channel vector $\mathbf{W}[:, j] \in \mathbb{R}^{C_{\text{in}}}$ through four steps: **1) Frequency Projection:** Transform the channel vector to the frequency domain using FFT, yielding $N = C_{\text{in}}$ frequency components characterized by magnitude A_k, phase ϕ_k, and frequency f_k. **2) Low-Frequency Truncation:** Given a target compression ratio ρ, select $k = \lfloor \rho \cdot N/3 \rfloor$ low-frequency components, since each stores three parameters. **3) Component Reconstruction:** Reconstruct each retained component in time domain as $H_m[n] = A_m \cdot e^{i(2\pi f_m n + \phi_m)}$, for $n = 0, 1, \ldots, N-1$. **4) Channel-wise vector concatenation:** Concatenate reconstructed vectors channel-wise to form the compressed representation, preserving structure for downstream processing and quantization.

To further reduce memory, we exploit the relation $f_k = k/N$, allowing implicit frequency indexing without storing f_k. Thus, each component stores only (A_k, ϕ_k), reducing memory by 33%. The number of retained components updates to $k = \lfloor \rho \cdot N/2 \rfloor$, with reconstruction:

$$H_m[n] = A_m \cdot e^{i\left(\frac{2\pi m}{N} n + \phi_m\right)}, \quad n \in \{0, 1, \ldots, N-1\} \tag{9}$$

This implicit frequency indexing strategy ensures no additional approximation error, as spectral positions remain unchanged under channel-wise FFT.

Mathematical Guarantee: Let $\mathcal{F}^{-1}$ be the inverse FFT. The reconstruction error ϵ satisfies:

$$\epsilon = \left\| \mathbf{W}[:, j] - \mathcal{F}^{-1}\left(\sum_{m=0}^{k-1} H_m\right) \right\|_2 \leq \sqrt{\sum_{m=k}^{N-1} |A_m|^2} \tag{10}$$

As per the decay property above, this upper bound decreases rapidly for smooth vectors, allowing high compression with minimal fidelity loss.

SpecQuant Framework

Migrate outliers from activation to weight. To mitigate activation outliers, recent methods (Xiao et al. 2023; Lin et al. 2024b) propose jointly scaling activations and weights. Specifically, the input $\mathbf{X}$ is scaled by a per-channel factor $\boldsymbol{\lambda} \in \mathbb{R}^m$, yielding smoothed activations $\hat{\mathbf{X}} = \mathbf{X} \cdot \text{diag}(\boldsymbol{\lambda})^{-1}$, which suppresses extreme values and reduces activation quantization error (Figure 2 (a) and (c)). To preserve the original computation, the weight matrix is adjusted as $\hat{\mathbf{W}} = \mathbf{W} \cdot \text{diag}(\boldsymbol{\lambda})$. However, this increases the dynamic range and magnitude of the weight values (Figure 2 (b) and (d)), introducing new quantization challenges. This trade-off limits the net benefit of such smoothing-based approaches.

Absorbing Weight Outliers via Channel-wise Low-Frequency Spectral Approximation. To address the increased weight quantization difficulty after smoothing, we propose an activation-aware channel-wise spectral compression framework that absorbs outliers by retaining low-frequency weight components in the Fourier domain, while dynamically allocating frequency budget based on actual activation-weight interaction strength. Unlike traditional low-rank approximations such as SVD, our method preserves channel-wise structure, leverages intrinsic weight smoothness, and prioritizes components based on their real impact on model outputs.

Our key insight is twofold: (1) Weight outliers often manifest as sharp local variations corresponding to high-frequency components in the frequency domain; (2) Instead of relying solely on weight characteristics, channel importance should be determined by their interaction with input activations—since this interaction ultimately drives model outputs. By allocating more frequency budget to channels with more important channels, we adaptively preserve critical components while suppressing noise-like outliers.

For each channel $\hat{\mathbf{W}}[:, j] \in \mathbb{R}^{C_{\text{in}}}$, we first compute an activation-aware importance score that quantifies its contribution strength in activation-weight interactions:

$$\text{Score}(j) = \left| \bar{\mathbf{X}}_{:,j} \cdot \bar{\hat{\mathbf{W}}}_{:,j} \right| \tag{11}$$

where $\bar{\mathbf{X}}_{:,j} \in \mathbb{R}$ denotes the average value of the j-th channel in the input matrix, $\bar{\hat{\mathbf{W}}}_{:,j} \in \mathbb{R}$ represents the average value of the j-th channel in the weight matrix, and $\cdot$ indicates scalar multiplication. This score captures how significantly the channel influences outputs.

We then apply FFT to transform the channel vector into the frequency domain:

$$\hat{\mathbf{W}}_{\text{freq}}[:, j] = \mathcal{F}(\hat{\mathbf{W}}[:, j]) \tag{12}$$

A softmax-based normalization over the activation-aware scores determines the frequency budget allocation:

$$\rho_j = \frac{\exp(\alpha \cdot \text{Score}(j))}{\sum_{l=1}^{C_{\text{out}}} \exp(\alpha \cdot \text{Score}(l))}, \tag{13}$$

where α is a tunable temperature parameter. Each channel retains $k_j = \lfloor \rho_j \cdot C_{\text{in}} \rfloor$ lowest-frequency components, with channels having higher activation impact receiving more budget to preserve critical spectral information.

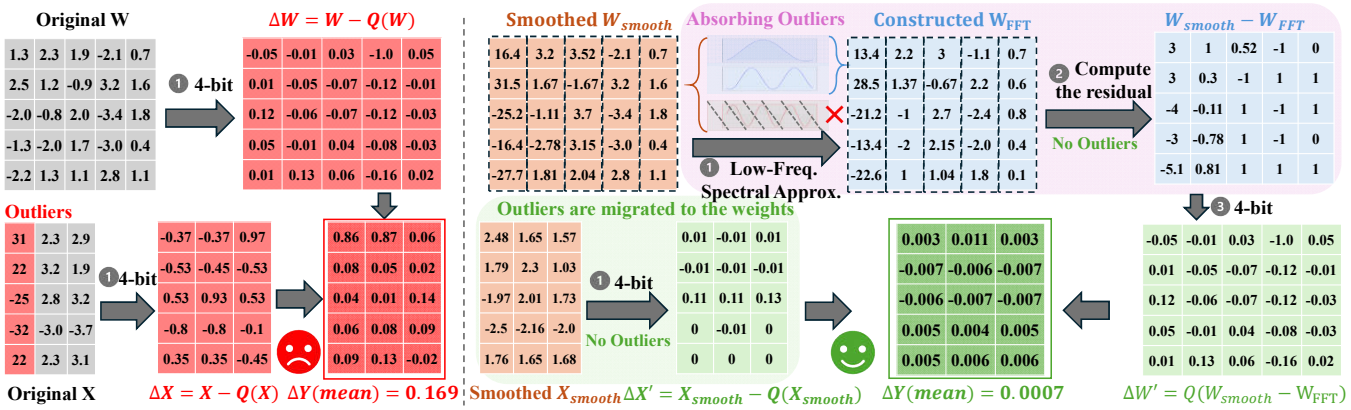

Figure 3: Comparison between conventional quantization and SpecQuant . Outlier channels in input activations are marked in red. SpecQuant adaptively absorbs these outliers using frequency-domain approximation, reducing overall quantization error.

The compressed weights are reconstructed via inverse FFT:

$$\mathbf{W}'[:, j] = \mathcal{F}^{-1}(\text{Truncate}(\hat{\mathbf{W}}_{\text{freq}}[:, j])) \tag{14}$$

This activation-aware strategy dynamically optimizes frequency retention at the channel level, ensuring that components most influential to actual model behavior are prioritized—resulting in a fine-grained compression scheme that balances efficiency and performance.

The residual $\mathbf{R} = \hat{\mathbf{W}} - \mathbf{W}'$ is quantized separately. The overall matrix product is approximated as:

$$\begin{aligned} \mathbf{XW} &= \hat{\mathbf{X}}\hat{\mathbf{W}} = \hat{\mathbf{X}}\mathbf{W}' + \hat{\mathbf{X}}\mathbf{R} \\ &\approx \underbrace{\hat{\mathbf{X}}\mathbf{W}'}_{\text{16-bit low-frequency branch}} + \underbrace{Q(\hat{\mathbf{X}})Q(\mathbf{R})}_{\text{4-bit residual}} \end{aligned} \tag{15}$$

This formulation retains the dominant low-frequency components in higher precision, while compressing the residual with low-bit quantization. Empirically, setting k to 16 or 32 per channel ensures both compression and accuracy. The overhead is negligible, adding only $2k/m$ to the overall cost, where m is the number of input channels.

From the spectral perspective, our method is inspired by Parseval's theorem, which ensures the energy of a signal is preserved in the frequency domain:

$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]|^2 \tag{16}$$

Moreover, for smooth functions $f \in C^r$, their Fourier coefficients decay polynomially as $|X[k]| \leq \frac{C}{|k|^r}$, indicating that most informative structure lies in low frequencies, justifying our truncation strategy for quantization robustness.

As illustrated in Figure 3(a), migrating activation outliers into weights without addressing the amplified weight magnitude leads to quantization artifacts, especially when activations exhibit large variance. This explains the limited effectiveness of conventional smoothing methods in compressing LLMs. In contrast, Figure 3(b) shows that SpecQuant absorbs migrated outliers through spectral approximation, yielding a quantization-friendly residual matrix $\Delta W'$. This design ensures that both smoothed activations and adapted weights are effectively quantized, resulting in improved end-to-end compression quality.

Experiments

Models and Datasets. We evaluate our method on the full LLaMA model family, including LLaMA-1 (7B–30B) (Touvron et al. 2023), LLaMA-2 (7B–70B) (Touvron et al. 2023), and LLaMA-3 (8B, 70B). Perplexity (PPL) is measured on the WikiText2 (Merity et al. 2016) test set. However, PPL alone does not fully reflect post-quantization performance, so we also report zero-shot accuracy on nine downstream tasks using the lm-evaluation-harness (v0.4.4) (Gao et al. 2023). These tasks include BoolQ (Clark et al. 2019), HellaSwag (Zellers et al. 2019), LAMBADA (OpenAI) (Radford et al. 2019), OpenBookQA (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 our method with standard RTN and several strong baselines, including SmoothQuant (Xiao et al. 2023), GPTQ (Frantar et al. 2022), Quarot (Ashkboos et al. 2024), and SpinQuant (Liu et al. 2024), covering both weight-only and weight-activation quantization. All activations are quantized using per-token asymmetric quantization. Residual weights are quantized using GPTQ (Frantar et al. 2022).

Implementation Details. To perform smoothing, we follow SmoothQuant and compute a per-channel smoothing factor. The optimal migration strength α for each layer is selected offline by minimizing the mean squared error (MSE) of the layer outputs after frequency-domain truncation on a calibration set. We apply different low-frequency group counts based on the target bit width. For example, in 4-bit quantization, each channel retains 16 low-frequency groups. To ensure fair comparison, we control the bit width of residual weights such that the total bit width after compression is aligned with baseline methods. A set of 256 randomly sampled examples from WikiText2 is used for calibration.

#Bits	Method	LLaMA-3 8B		LLaMA-3 70B		LLaMA-2 7B		LLaMA-2 13B		LLaMA-2 70B		LLaMA 7B		LLaMA 13B		LLaMA 30B	
		0-shot[9]	Wiki	0-shot[9]	Wiki	0-shot[9]	Wiki	0-shot[9]	Wiki	0-shot[9]	Wiki	0-shot[9]	Wiki	0-shot[9]	Wiki	0-shot[9]	Wiki
W-A-KV		Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)	Avg.(↑)	PPL(↓)
16-16-16	FP16	68.09	6.14	73.81	2.86	65.21	5.47	67.61	4.88	71.59	3.32	64.48	5.68	66.67	5.09	70.00	4.10
4-16-16	RTN	63.70	8.13	31.15	1e5	61.27	7.02	60.24	6.39	69.62	3.87	62.67	7.94	63.45	8.60	65.69	6.13
	SmoothQuant	62.79	8.12	67.94	6.70	58.88	8.03	62.03	5.86	65.93	5.50	62.24	7.46	62.69	18.75	65.69	5.80
	GPTQ	61.03	7.43	31.45	9e3	60.86	9.84	64.71	5.79	70.96	3.94	60.15	7.93	64.36	6.58	66.95	5.26
	Omniquant	65.66	7.19	-	-	63.19	5.74	66.38	5.02	71.04	3.47	63.42	5.86	66.22	5.21	69.07	4.25
	AWQ	67.03	7.36	68.92	5.92	63.89	5.83	66.25	5.07	70.88	4.03	63.30	5.97	65.58	5.28	69.44	4.28
	QuaRot	67.27	6.53	72.93	3.53	64.30	5.62	66.95	5.00	71.21	3.41	63.40	5.83	65.91	5.20	69.73	4.27
	SpinQuant	66.54	6.49	72.90	**3.49**	63.59	5.58	**67.14**	5.00	71.12	3.43	63.94	**5.76**	66.32	**5.16**	69.62	4.21
	SpecQuant	**66.88**	**6.48**	**72.98**	3.53	**63.99**	**5.56**	67.10	**4.98**	**71.12**	**3.40**	**63.99**	5.85	**66.35**	5.21	**69.77**	**4.21**
4-4-16	RTN	33.42	6e2	31.21	8e3	32.44	nan	30.86	8e3	30.90	7e4	32.51	7e3	31.63	3e4	31.57	2e3
	SmoothQuant	33.04	1e3	34.67	2e2	32.13	nan	34.26	1e3	35.86	3e2	34.42	3e2	33.29	6e2	34.64	1e3
	GPTQ	32.98	5e2	31.47	4e4	32.72	nan	30.11	4e3	30.86	nan	32.12	1e3	31.51	3e3	30.88	2e3
	QuaRot	61.69	8.02	65.56	6.35	61.87	6.05	**65.13**	5.35	69.96	3.78	61.76	6.22	64.46	5.50	68.14	4.57
	SpinQuant	64.11	7.28	66.99	6.10	57.37	6.78	63.23	5.24	70.58	3.68	61.82	6.08	64.59	**5.36**	68.08	4.53
	SpecQuant	**64.75**	**7.25**	**69.75**	**5.12**	**62.88**	**5.88**	65.12	**5.18**	**70.77**	**3.63**	**61.85**	**6.05**	**64.77**	5.45	**68.25**	**4.50**
4-4-4	RTN	33.18	7e2	30.82	8e3	32.67	nan	30.93	7e3	31.73	7e4	32.87	1e4	31.33	3e4	31.64	2e3
	SmoothQuant	32.96	1e3	33.76	3e2	32.12	nan	33.36	1e3	35.54	3e2	34.42	3e2	33.28	5e2	34.65	1e3
	GPTQ	33.71	6e2	31.20	4e4	33.52	nan	27.85	5e3	31.09	nan	31.80	2e3	30.63	3e3	31.07	2e3
	Omniquant	32.33	4e2	-	-	48.40	14.26	50.35	12.30	-	-	48.46	11.26	45.63	10.87	45.04	12.35
	QuaRot	61.38	8.18	65.33	6.6	61.48	6.11	65.16	5.39	70.30	3.80	61.22	6.26	64.59	5.53	68.08	4.60
	SpinQuant	64.10	7.35	66.31	6.24	62.01	5.96	64.13	5.74	70.57	3.61	61.32	6.12	64.95	5.39	**68.14**	4.55
	SpecQuant	**64.75**	**7.33**	**69.43**	**5.77**	**62.12**	**5.95**	**65.33**	**5.35**	**70.77**	**3.60**	**61.59**	**6.12**	**64.99**	**5.39**	68.12	**4.53**

Table 1: Comparison of perplexity on WikiText2 and averaged accuracy on nine Zero-Shot tasks.

Overall Results

Quantization Performance. As shown in Table 1, SpecQuant outperforms SOTA across diverse models and quantization settings. In the 4-16-16 setting, it retains over 99% FP16 zero-shot accuracy, outperforming activation-aware and weight-only baselines. Versus weight-only methods (e.g., GPTQ, AWQ), it narrows the FP performance gap, especially on LLaMA-3-8B, with only a 1.21% accuracy drop, significantly lower than competitors' $>$1.55% degradation.

In the stricter 4-4-16 setting (constraining both weights and activations), SpecQuant outperforms SpinQuant by over 1 percentage point across benchmark models. Even in extreme 4-4-4 quantization, it delivers meaningful accuracy gains, showing robustness to aggressive compression. These results confirm our frequency-domain approach's effectiveness: the low-frequency truncation branch outperforms rotation/smoothing-only methods by suppressing activation outliers and resolving distributional imbalances, boosting quantization stability.

Speedup and Memory Savings. SpecQuant achieves 4-bit quantization with negligible accuracy loss, enabling practical low-bit inference. To assess its efficiency, we follow the measurement setup used in prior work (Hu et al. 2025a), and evaluate the LLaMA series on an NVIDIA 3090 GPU. As shown in Table 2, SpecQuant delivers over 2× speedup across models, reaching nearly 2.5× on the challenging LLaMA-30B, due to efficient low-precision computation and reduced memory access overhead. Additionally, SpecQuant achieves over 3× memory savings on average. The added cost from the low-frequency truncation branch is minimal, introducing no significant computational overhead.

Ablation Study

To validate the contribution of each component in our method, we conducted detailed ablation experiments. As shown in Table 3, direct W4A4 quantization leads to a notable drop in performance compared to the full-precision model. Introducing the Smooth operation prior to quantization yields modest gains, but the improvement is limited. Adding a low-frequency truncation branch alone also fails to deliver satisfactory results. In contrast, our full method significantly improves quantization performance by first migrating activation outliers into the weight parameters via smoothing, and then quantizing the residual weights after decomposition. The low-frequency truncation branch effectively absorbs the migrated outliers, resulting in a robust low-bit model.

Trade-off of Increasing Truncation Groups

Table 4 summarizes the impact of varying the number of truncation groups in SpecQuant 's low-frequency branch. As the number of groups increases, quantization performance improves consistently. However, this gain comes with additional parameter storage and higher inference latency. To strike a balance between accuracy and efficiency, we adopt a configuration with 16 truncation groups. Furthermore, by applying mixed-precision quantization to the residual weights, SpecQuant maintains the same parameter count as baseline methods while delivering superior performance.

Model	Seqlen	Prefill Time		Prefill Speedup	Memory		Memory Saving
		FP16	INT4		FP16	INT4	
LLaMA3-8B	256	8.035ms	3.510ms	2.289x	0.430GB	0.126GB	3.413x
	512	15.545ms	6.663ms	2.333x	0.442GB	0.134GB	3.299x
	1024	29.169ms	13.086ms	2.229x	0.466GB	0.151GB	3.086x
	2048	57.470ms	26.312ms	2.188x	0.513GB	0.188GB	2.729x
	4096	117.523ms	53.082ms	2.214x	0.608GB	0.278GB	2.187x
	8192	256.394ms	119.198ms	2.151x	0.795GB	0.400GB	1.988x
LLaMA2-13B	256	11.449ms	4.667ms	2.453x	0.634GB	0.178GB	3.562x
	512	21.195ms	8.602ms	2.464x	0.663GB	0.192GB	3.453x
	1024	41.752ms	17.182ms	2.430x	0.723GB	0.220GB	3.286x
	2048	81.965ms	34.864ms	2.351x	0.841GB	0.283GB	2.972x
	4096	199.046ms	81.710ms	2.436x	1.079GB	0.404GB	2.671x
	8192	359.409ms	162.335ms	2.214x	1.551GB	0.642GB	2.416x
LLaMA-30B	256	18.682ms	6.485ms	2.881x	1.047GB	0.285GB	3.674x
	512	34.393ms	12.743ms	2.699x	1.085GB	0.305GB	3.557x
	1024	66.880ms	24.835ms	2.693x	1.162GB	0.343GB	3.388x
	2048	157.500ms	59.886ms	2.630x	1.315GB	0.422GB	3.116x
	4096	272.355ms	105.523ms	2.581x	1.625GB	0.577GB	2.816x
	8192	576.555ms	234.086ms	2.463x	2.242GB	0.889GB	2.522x

Table 2: Prefill time and Memory usage of LLaMA models with different parameter sizes and sequence lengths, compared between our 4-bit implementation and FP16. All tests were conducted on a Transformer block with batch size 4 on a 3090 GPU.

Method			LLaMA-7B		LLaMA2-7B		LLaMA3-8B	
Quant	Smooth	Trunc.	Wiki ↓	0-shot[9] ↑	Wiki ↓	0-shot[9] ↑	Wiki ↓	0-shot[9] ↑
✓			9e3	25.34	nan	26.44	8e3	24.42
✓	✓		3e2	34.42	nan	32.13	1e3	33.04
✓		✓	24.57	54.72	26.79	52.88	27.75	55.08
✓	✓	✓	**6.05**	**61.85**	**5.88**	**62.88**	**7.25**	**64.75**

Table 3: Ablation study on the impact of different methods on WikiText2 perplexity (PPL) and zero-shot[9] accuracy for LLaMA-7B, LLaMA2-7B, and LLaMA3-8B.

Trunc. Groups	Model Size Overhead	Latency Overhead	LLaMA-7B		LLaMA2-7B		LLaMA3-8B	
			Wiki↓	0-shot[9] ↑	Wiki↓	0-shot[9] ↑	Wiki ↓	0-shot[9] ↑
16	2.7%	5.2%	6.04	61.89	5.87	62.91	7.24	64.78
32	5.5%	7.3%	6.03	62.01	5.85	63.11	7.21	65.12
64	11.2%	12.1%	5.99	62.88	5.81	64.09	7.08	66.03

Table 4: Impact of the number of truncation groups in the low-frequency truncation branch on WikiText perplexity (PPL) and zero-shot[9] accuracy for LLaMA models.

Comparison of Importance Metrics for Truncation

We compare four importance metrics for allocating compression budgets under a fixed 20% compression ratio: Mean Absolute Value (Abs Mean), Maximum Absolute Value (Abs Max), L2 Norm, and our proposed spectral entropy. These metrics, commonly used in quantization and pruning, capture different aspects of channel salience. As shown in Table 5, spectral entropy consistently yields the lowest perplexity across all benchmarks and models, validating its robustness and effectiveness. Unlike magnitude-based metrics, it captures the energy distribution across frequency components, enabling better identification of structurally important channels. These findings empirically support our frequency-aware design and highlight the advantage of spectral-domain importance estimation in compression-aware allocation.

Importance Metrics	LLaMA-7B		LLaMA2-7B	
	WikiText2 ↓	PTB ↓	WikiText2 ↓	PTB ↓
Original	6.60	66.00	6.33	33.63
Abs Mean	6.58	56.60	6.32	33.22
Abs Max	6.72	53.67	6.45	34.17
L2 Norm	6.59	55.19	6.31	33.08
Spectral Entropy	**6.55**	**47.52**	**6.30**	**32.64**

Table 5: Comparison of Importance Metrics for Compression Allocation in SpecQuant under a 20% compression ratio on LLaMA-7B and LLaMA2-7B.

Conclusion

In this paper, we propose SpecQuant, a novel frequency-domain quantization method for ultra-low-bit LLMs, which effectively addresses the long-standing challenge of outlier management through spectral-domain processing. SpecQuant combines activation smoothing with channel-adaptive Fourier decomposition, ensuring that amplified weights in the frequency domain do not impact the low-frequency components, which retain most of the signal energy. Experiments across multiple zero-shot tasks validate the effectiveness of our method across various network architectures.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ashkboos, S.; Mohtashami, A.; Croci, M.; Li, B.; Cameron, P.; Jaggi, M.; Alistarh, D.; Hoefler, T.; and Hensman, J. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37: 100213–100240.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.

Boratko, M.; Padigela, H.; Mikkilineni, D.; Yuvraj, P.; Das, R.; McCallum, A.; Chang, M.; Fokoue-Nkoutche, A.; Kapanipathi, P.; Mattei, N.; et al. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. *arXiv preprint arXiv:1806.00358*.

Chee, J.; Cai, Y.; Kuleshov, V.; and De Sa, C. M. 2023. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36: 4396–4429.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D.-A. 2023. OPTQ: Accurate post-training quantization for generative pre-trained transformers. In *11th International Conference on Learning Representations*.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Noac'h, A. L.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2023. A Framework for Few-Shot Language Model Evaluation. https://zenodo.org/records/10256836.

Gao, Z.; Wang, Q.; Chen, A.; Liu, Z.; Wu, B.; Chen, L.; and Li, J. 2024. Parameter-efficient fine-tuning with discrete fourier transform. *arXiv preprint arXiv:2405.03003*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Heo, J. H.; Kim, J.; Kwon, B.; Kim, B.; Kwon, S. J.; and Lee, D. 2023. Rethinking channel dimensions to isolate outliers for low-bit weight quantization of large language models. *arXiv preprint arXiv:2309.15531*.

Hu, X.; Cheng, Y.; Yang, D.; Xu, Z.; Yuan, Z.; Yu, J.; Xu, C.; Jiang, Z.; and Zhou, S. 2025a. OstQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting. *arXiv preprint arXiv:2501.13987*.

Hu, X.; Cheng, Y.; Yang, D.; Xu, Z.; Yuan, Z.; Yu, J.; Xu, C.; Jiang, Z.; and Zhou, S. 2025b. OstQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting. arXiv:2501.13987.

Jin, M.; Mei, K.; Xu, W.; Sun, M.; Tang, R.; Du, M.; Liu, Z.; and Zhang, Y. 2025. Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding. arXiv:2502.01563.

Kelkar, S. S.; Grigsby, L. L.; and Langsner, J. 1983. An Extension of Parseval's Theorem and Its Use in Calculating Transient Energy in the Frequency Domain. *IEEE Transactions on Industrial Electronics*, IE-30(1): 42–45.

Li, M.; Lin, Y.; Zhang, Z.; Cai, T.; Li, X.; Guo, J.; Xie, E.; Meng, C.; Zhu, J.-Y.; and Han, S. 2024. Svdqunat: Absorbing outliers by low-rank components for 4-bit diffusion models. *arXiv preprint arXiv:2411.05007*.

Lin, H.; Xu, H.; Wu, Y.; Cui, J.; Zhang, Y.; Mou, L.; Song, L.; Sun, Z.; and Wei, Y. 2024a. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. *Advances in Neural Information Processing Systems*, 37: 87766–87800.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024b. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6: 87–100.

Liu, B.; Jiang, Z.; Wu, J.; Chen, X.; Han, Y.; and Liu, P. 2021. F3D: Accelerating 3D convolutional neural networks in frequency space using ReRAM. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, 571–576. IEEE.

Liu, Z.; Zhao, C.; Fedorov, I.; Soran, B.; Choudhary, D.; Krishnamoorthi, R.; Chandra, V.; Tian, Y.; and Blankevoort, T. 2024. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Qin, Z.; Zhang, P.; Wu, F.; and Li, X. 2021. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 783–792.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tseng, A.; Chee, J.; Sun, Q.; Kuleshov, V.; and De Sa, C. 2024. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *arXiv preprint arXiv:2402.04396*.

Wang, H.; Pan, L.; Chen, Z.; Yang, D.; Zhang, S.; Yang, Y.; Liu, X.; Li, H.; and Tao, D. 2024. Fredf: Learning to forecast in frequency domain. *arXiv preprint arXiv:2402.02399*.

Wei, X.; Zhang, Y.; Zhang, X.; Gong, R.; Zhang, S.; Zhang, Q.; Yu, F.; and Liu, X. 2022. Outlier suppression: Pushing the limit of low-bit transformer language models. *Advances in Neural Information Processing Systems*, 35: 17402–17414.

Xiao, G.; Lin, J.; Seznec, M.; Wu, H.; Demouth, J.; and Han, S. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 38087–38099. PMLR.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhao, Z.; Li, H.; Liu, F.; Lu, Y.; Wang, Z.; Yang, T.; Jiang, L.; and Guan, H. 2025. QUARK: Quantization-Enabled Circuit Sharing for Transformer Acceleration by Exploiting Common Patterns in Nonlinear Operations. arXiv:2511.06767.