This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Quant Experts: Token-aware Adaptive Error Reconstruction with Mixture of Experts for Large Vision-Language Models Quantization

Chenwei Jia* Baoting Li* Xuchong Zhang† Mingzhuo Wei Bochen Lin Hongbin Sun
State Key Laboratory of Human-Machine Hybrid Augmented Intelligence
Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University
jiacw@stu.xjtu.edu.cn

Abstract

Post-Training Quantization (PTQ) has emerged as an effective technique for alleviating the substantial computational and memory overheads of Vision-Language Models (VLMs) by compressing both weights and activations without retraining the full model. Existing PTQ methods primarily rely on static identification and global compensation of sensitive or outlier channels, yet they often overlook the distributional differences of these important channels across inputs, leading to unsatisfactory quantization. In this work, we observe that the distributions and occurrence frequencies of important channels vary significantly both across modalities and among tokens, even within the same modality. Accordingly, we propose ***Quant Experts (QE)****, a token-aware adaptive error compensation with mixture-of-experts for VLMs quantization. QE divides the important channels into token-independent and token-dependent groups. For the former, a shared expert is designed for most tokens to compensate for global quantization error using a low-rank adapter. For the latter, routed experts including multiple routed low-rank adapters are elaborated to compensate for local quantization error related to specific tokens. Extensive experiments demonstrate that QE consistently enhances task accuracy across various quantization settings and model scales, ranging from 2B to 70B parameters, while maintaining performance comparable to full-precision models.*

1. Introduction

Model quantization has become a key technique for reducing the computational and memory costs of large-scale multimodal models [19]. By mapping weights and activations to low-bit representations, it achieves substantial compression and acceleration while preserving accuracy. Among existing approaches, Post-Training Quantization (PTQ) is widely adopted for its efficiency and compatibility without retraining the full model. However, low-bit quantization inevitably introduces numerical perturbations that degrade performance, a problem that is especially pronounced in multimodal scenarios [13].

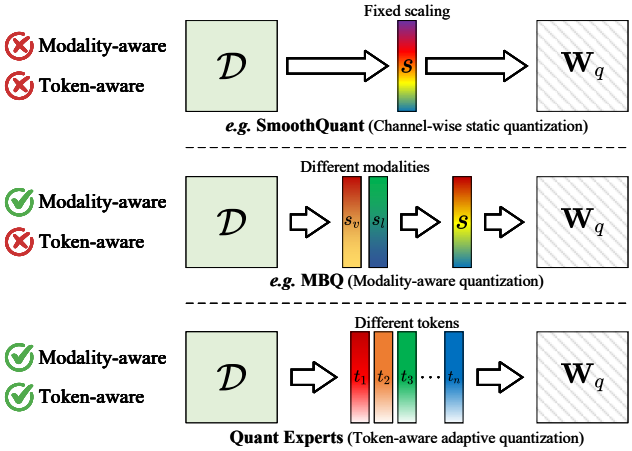


Figure 1. Illustration of Different Quantization Granularities. $\mathcal{D}$ is the calibration data, and $\mathbf{W}_q$ is the quantized weight matrix. SmoothQuant [44] employs channel-wise with fixed scaling coefficients to achieve global quantization. MBQ [23] presents a modality-aware strategy that focuses on sensitivities in input modalities. In contrast, we propose a token-aware adaptive quantization that considers both global and local error reconstruction.

Recent studies on large-scale model quantization reveal significant variation in channel value distributions and sensitivities, where a small set of outlier or highly sensitive channels plays a key role in preserving model expressiveness and accuracy. To mitigate quantization errors arising from these important channels, numerous methods have been proposed from different perspectives, including channel smoothing, mixed-precision strategy, Hessian-based op-

*Equal contribution.
†Corresponding author: zhangxc0329@xjtu.edu.cn.

timization, and low-rank reconstruction approaches. For instance, SmoothQuant [44] and AWQ [24] alleviate quantization errors by balancing activation and weight ranges via channel-wise scaling, with fixed coefficients for each channel estimated from the calibration dataset. SpQR [9] statically identifies highly sensitive channels and allocates higher precision to them, effectively reducing quantization errors. OBQ [11] and GPTQ [12] estimate a static Hessian matrix from inputs and perform channel- or block-wise quantization based on the computed sensitivity. LQER [48] and ASER [50] leverage singular value decomposition (SVD) to explicitly reconstruct quantization errors through a global low-rank adapter [16], *i.e.* all important channels are handled uniformly by a single adapter. Currently, in multimodal scenarios, MBQ [23] reveals substantial differences in channel sensitivities across modalities and introduces a modality-aware channel scaling strategy to improve the adaptability of channel-smoothing methods. Although these methods reduce quantization errors effectively, the fixed channel-importance estimation and global error compensation neglect the modeling of local feature variations, failing to adequately account for the dynamic relationships among important channel, error compensation, input modalities, and even tokens.

In this paper, we conduct a systematic analysis of large Vision-Language Models (VLMs) and observe that channel importance exhibits strong dynamics, varying significantly across modalities and tokens. In particular, only a small subset of important channels appears consistently among most tokens, while the majority exhibit strong input-dependent fluctuations, with their importance differing across modalities and tokens. To further minimize quantization errors, the dynamic nature of channel importance necessitates quantization methods that not only compensate for errors in globally activated important channels but also adaptively mitigate those arising in input-dependent ones.

Accordingly, we propose **Quant Experts (QE)**, a token-aware adaptive error reconstruction framework with the Mixture-of-Experts (MoE) [15, 36] for VLMs quantization, which considers both global and local variations in channel importance across different tokens. QE first estimates the occurrence frequency of important channels from calibration data and then divides them into *token-independent* and *token-dependent* channels. We then employ two types of experts, the shared expert and the routed experts, to model these two kinds of channel groups respectively. The shared expert employs a low-rank adapter to reconstruct global quantization errors that predominantly originate from token-independent important channels. Meanwhile, token-dependent channels are clustered into multiple sub-groups according to their co-occurrence relationships, and each sub-group is assigned a routed expert equipped with a routed low-rank adapter. During inference, QE fixedly employs the shared expert to compensate for global quantization errors and dynamically selects the most suitable routed expert according to the input tokens, thereby achieving adaptive compensation for local errors. Extensive experiments under various quantization settings demonstrate that QE effectively mitigates errors caused by dynamic channel importance and substantially restores model performance under low-bit quantization. This work makes the following contributions:

- We observe that the distributions and occurrence frequencies of important channels vary significantly both across modalities and among tokens, even within the same modality.
- We propose **QE**, a token-aware adaptive error compensation with MoE for VLMs quantization. It employs a fixed shared expert to compensate global quantization errors while dynamically selecting routed experts to correct token-dependent local errors.
- Extensive experiments across various quantization settings and multimodal benchmarks demonstrate that QE effectively mitigates low-bit performance degradation, achieving up to a 5.09% accuracy improvement under the challenging W4A6 quantization for the 72B model.

2. Observation and Motivation

2.1. Observation One: The Positions of Important Channels Vary Across Tokens

As illustrated in Fig. 2, we visualize the value distributions of several tokens along with the positions of their corresponding important channels computed by Eqs. (1) and (2).

$$\mathbf{w} := \mathrm{Mean}_{\mathrm{row}}(|\mathbf{W}_f|), \tag{1}$$

$$\mathcal{C}_t = \mathrm{Top}\text{-}k(|x_t| \odot \mathbf{w}), \tag{2}$$

where $\mathbf{W}_f \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ is the weight matrix, d_{in} is the number of input channels, d_{out} is the number of output channels, $\mathrm{Mean}_{\mathrm{row}}$ computes the mean of absolute values along row, $\mathbf{w}$ is a d_{in}-dimensional vector, x_t is the t-th token, and $\mathcal{C}_t$ denotes its top-k important channels.

It is evident that, under the same layer weight, the positions of important channels vary not only across different modalities but also among tokens within the same modality. In the case of across modalities, the intrinsic differences in value distributions lead to cross-modal shifts in the positions of important channels. More importantly, even within the same modality, variations in token semantics and contextual information cause significant changes in activation distributions, resulting in dynamic migrations of important channels. These findings indicate that the positions of important channels are not static but dynamically adapt to both inter- and intra-modality variations.

This observation reveals a key limitation of those globally calibrated and compensated methods [9, 24, 50], just

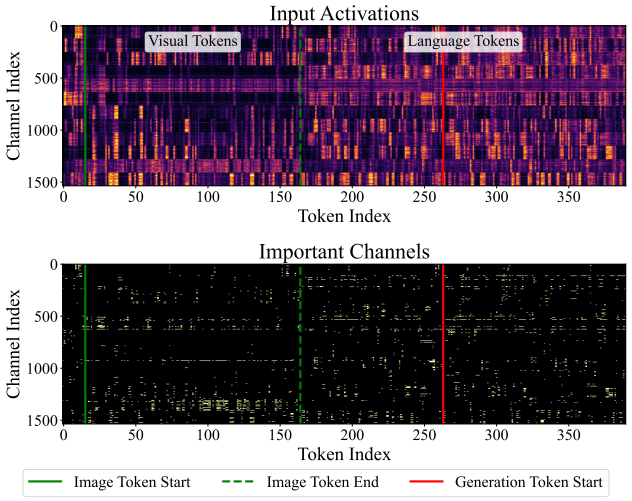


Figure 2. Visualization of a Transformer block in Qwen2VL-2B, illustrating token value distributions (top) and the positions of important channels (bottom). In the top panel, brightness reflects token magnitude, while in the bottom panel, highlighted positions denote important channels.

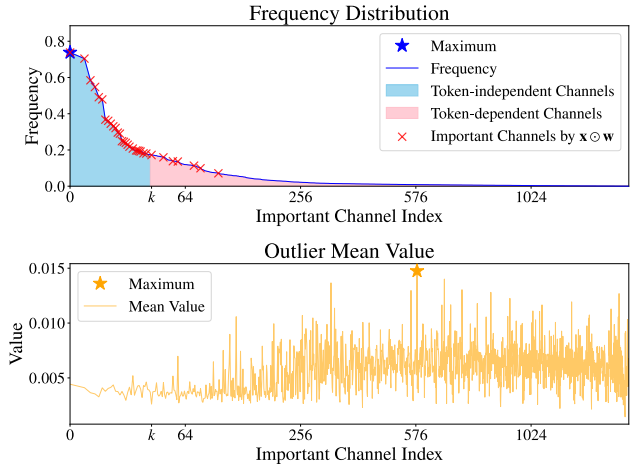


Figure 3. Visualization of a Transformer block in Qwen2VL-2B, illustrating important channel behavior. The top panel shows activation frequencies, while the bottom depicts average activation values. Red crosses mark important channel positions identified by the static global method on the calibration dataset.

identifying some fixed outliers and employing unified compensation, which fails to capture token-wise variations in channel importance.

2.2. Observation Two: Uneven Frequency Distribution of Important Channels

To gain deeper insight, we compute the occurrence frequency of each important channel using Eq. (3) and sort them in descending order of frequency.

$$f_c = k \times \frac{m_c}{\sum_{i=0}^{d_{\text{in}}} m_i}, \tag{3}$$

where m_c denotes the number of times channel c is identified as important, and f_c represents the proportion of tokens for which channel c is recognized as an important channel among all tokens.

As shown in Fig. 3, there is a small subset of important channels that occur across most tokens, while the majority are activated only for specific tokens. Meanwhile, those important channels with smaller frequency values may still exhibit larger outlier magnitudes, making their contribution to quantization error compensation equally non-negligible. Based on this observation, we define the channels, which consistently appear across diverse input tokens and play a key role in compensating global quantization errors, as *token-independent* important channels. Conversely, channels whose activation depends on specific tokens are regarded as *token-dependent* important channels, which reveal distinctive token properties of different tokens and the local nature of quantization errors.

This observation suggests that important channels in VLMs comprise both global and local components. Quantization error compensation that fails to properly address these two types inevitably results in performance degradation: insufficient compensation for token-independent important channels weakens the model's overall representational capacity, while neglecting token-dependent important ones compromises token-level semantic fidelity. Therefore, an effective quantization framework should adopt differentiated protection and compensation strategies across tokens to ensure both stability and adaptability in performance.

3. Method

As illustrated in Fig. 4, QE first estimates the distribution frequency of important channels from the calibration data $\mathcal{D}$ and partitions them into *token-independent* and *token-dependent* channels. It then employs the ***shared expert (SE)*** and the ***routed experts (REs)***, to handle these distinct channel groups, each implemented as a low-rank adapter. The shared expert focuses on global quantization error that is mainly caused by token-independent channels. Whereas token-dependent channels are clustered through co-occurrence analysis into multiple subgroups, and each is handled by a routed expert that specializes in local error compensation. During inference, a lightweight router dynamically selects the optimal routed expert for adaptive error compensation.

3.1. Calibration and Error Reconstruction

In this subsection, we illustrate how to minimize quantization error from token-aware adaptive optimization and establish the mapping between important channels and the mixture-of-experts. Let $\mathbf{W}_f^l$ denote the full-precision weight of the l-th layer, and $\mathbf{W}_q^l = Q(\mathbf{W}_f^l)$ its quantized

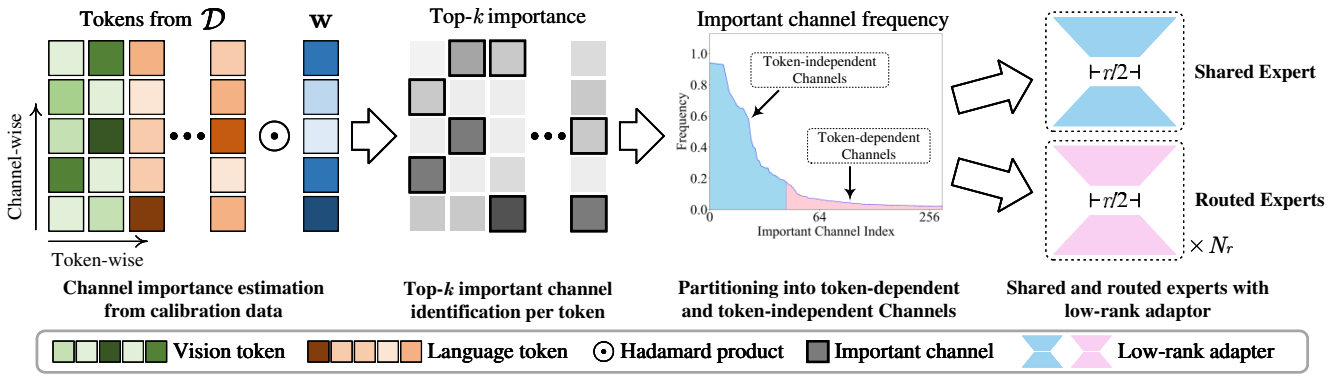


Figure 4. The framework of Quant Experts (QE). The token-independent channels are model by a shared expert, while token-dependent channels are captured by multiple routed experts. A lightweight low-rank adapter is implemented for each expert.

counterpart. As typical, the quantization error is defined as $\mathbf{E}^l = \mathbf{W}_f^l - \mathbf{W}_q^l$. Building on the analysis in Sec. 2, we further formulate quantization objective optimization as:

$$\arg\min_{\tilde{\mathbf{E}}^l} \|(\mathbf{E}^l - \underbrace{(\tilde{\mathbf{E}}_S^l + \tilde{\mathbf{E}}_R^l(x^l))}_{\tilde{\mathbf{E}}^l})x^l\|_F, \tag{4}$$

where x^l denotes the input tokens of the l-th layer, $\tilde{\mathbf{E}}_S^l$ represents the token-independent error approximation, and $\mathbf{E}_R^l(x^l)$ denotes the error approximation dependent on x^l.

Our approach leverages two complementary low-rank adapters to reconstruct the quantization error:

$$\tilde{\mathbf{E}}^l = \underbrace{\mathbf{L}_{SA}^l \mathbf{L}_{SB}^l}_{\tilde{\mathbf{E}}_S^l} + \underbrace{\mathbf{L}_{RA}^{l,i^*} \mathbf{L}_{RB}^{l,i^*}}_{\mathbf{E}_R^l(x^l)}, \tag{5}$$

$$i^* = \arg\min_i \left(\mathbf{R}^l |x^l|\right)_i, \tag{6}$$

where $\mathbf{L}_{SA}^l$ and $\mathbf{L}_{SB}^l$ form the shared expert that models token-independent error approximation, while $\mathbf{L}_{RA}^{l,i}$ and $\mathbf{L}_{RB}^{l,i}$ denote the routed experts adaptively modeling token-dependent error approximation. $\mathbf{R}^l$ is a lightweight router that predicts a score for each routed expert, and selects an expert with the minimal score. In this way, QE integrates global token-independent and local token-dependent compensation within a unified mixture-of-experts framework.

We quantify the distribution of channel importance using activation $X \in \mathbb{R}^{T \times d_{in}}$ collected from the calibration data $\mathcal{D}$. For the t-th input token x_t^l in layer l, we compute the per-token important channels $\mathcal{A}_t^l$ as Eq. (2). Then we aggregate all $\mathcal{A}_t^l$ over $\mathcal{D}$ to form a multiset $\mathcal{T}^l$, and compute the occurrence frequency f^l of each important channel according to Eq. (3). These channels are sorted by f and partitioned into two disjoint important channel sets: the first k channels $\mathcal{C}_s^l$, referred to as token-independent channels, and the subsequent $(N_r k)$ channels $\mathcal{C}_r^l$, referred to as token-dependent channels. To sum up, the pseudo-code of the dependence partitioning process is detailed in Algorithm 1.

Algorithm 1: Channel Dependence Partitioning

Input : Layer weight $\{\mathbf{W}_f^l\}_{l=1}^L$; calibration data $\mathcal{D}$; the number of important channels k; the number of routed experts N_r.
Output: Token-independent channels $\{\mathcal{C}_s^l\}$; token-dependent channels $\{\mathcal{C}_r^l\}$.

for *sample* $d \in \mathcal{D}$ **do**
  Inference d and store input activations in $\mathbf{X}$
for $l \leftarrow 1$ **to** L **do**
  Initialize $\mathcal{T}^l \leftarrow \emptyset$
  Compute $\mathbf{w}^l$ from $\mathbf{W}_f^l$ according to Eq. (1)
  for *token* x_t^l *in* $\mathbf{X}^l$ **do**
    $\mathcal{A}_t^l \leftarrow \text{Top-}k(|x_t^l| \cdot \mathbf{w}^l)$ `// indices`
    $\mathcal{T}^l \leftarrow \mathcal{T}^l \uplus \mathcal{A}_t^l$
  $(\mathcal{C}_{\text{unsorted}}^l, f^l) \leftarrow \texttt{unique}(\mathcal{I}^l)$
  $\pi \leftarrow \texttt{argsort}(f^l, \text{descending})$
  Reorder $\mathcal{C}^l \leftarrow \mathcal{C}_{\text{unsorted}}^l[\pi]$
  Token-independent channels $\mathcal{C}_s^l \leftarrow \mathcal{C}_{[1:k]}^l$
  Token-dependent channels $\mathcal{C}_r^l \leftarrow \mathcal{C}_{[k+1:(N_r+1)k]}^l$
return $\{\mathcal{C}_s^l\}_{l=1}^L$, $\{\mathcal{C}_r^l\}_{l=1}^L$

3.2. *SE* for Token-Independent Channels

We introduce the shared expert to reconstruct the quantization error that primarily arises from token-independent channels. As in prior work [12, 24, 40], the performance degradation and quantization error stem mainly from these outlier channels, underscoring the need to mitigate their impact on both weight and activation quantization.

Building upon the above motivation, the shared expert is designed following [50] to focus on the token-independent channels $\mathcal{C}_s^l$. Specifically, to alleviate weight quantization errors and suppress the interference of their outlier magnitudes on other channels, token-independent channels are

Algorithm 2: Building Routed Experts

Input : Per-layer data $\{\mathbf{X}^l, \mathcal{O}^l, \mathbf{E}_S^l, \mathcal{C}_r^l\}_{l=1}^L$; the number of routed experts N_r; rank r.

Output: REs $\{(\mathbf{L}_{RA}^{l,i}, \mathbf{L}_{RB}^{l,i})\}$ and Router $\{\mathbf{R}^l\}$.

for $l \leftarrow 1$ **to** L **do**
- Initialize router parameter $\mathbf{R}^l \in \mathbb{R}^{d_{\text{in}} \times N_r}$
- Compute $\mathbf{x}^l \leftarrow \text{Mean}_{\text{row}}(|\mathbf{X}^l|)$
- $\mathbf{S}^l \leftarrow \text{NPMI_Similarity}(\mathcal{O}^l)$
- $\mathcal{N}^l \leftarrow \text{Normed_Laplacian}(\mathbf{S}^l)$
- $\mathbf{U}^l \leftarrow \text{Eigenvectors}(\mathcal{N}^l)$
- Cluster $\mathbf{\Gamma}^l \leftarrow \text{KMeans}(\mathbf{U}^l_{[:,2:N_r+1]},\ N_r)$
- **for** *cluster labels* $\gamma \in \mathbf{\Gamma}^l$ **do**
  - Initialize weight $\omega \leftarrow \mathbf{1} \in \mathbb{R}^{d_{\text{in}}}$
  - Compute $\omega \leftarrow (\mathbf{x}_\gamma^l / \min(\mathbf{x}_\gamma^l)) \odot \mathbf{x}^l$
  - Normalize $\omega \leftarrow \omega / \sqrt{\min(\omega)\max(\omega)}$
  - Perform SVD: $U\Sigma V^\top = \mathbf{E}_S^l \text{diag}(\omega)$
  - $\mathbf{L}_{RA}^{l,i} \leftarrow U_r \Sigma_r$
  - $\mathbf{L}_{RB}^{l,i} \leftarrow V_r^\top S^{-1} \text{diag}(1/\omega)$
  - $R_{[:,i]}^l \leftarrow \text{Mean}_{\text{row}}(|\mathbf{E}_S^l - \mathbf{L}_{RA}^{l,i}\mathbf{L}_{RB}^{l,i}|)$

return $\{(\mathbf{L}_{RA}^{l,i}, \mathbf{L}_{RB}^{l,i})\}_{l=1,i=1}^{L,N_r}$, $\{\mathbf{R}^l\}_{l=1}^L$

exempted from direct quantization and reconstructed by a low-rank adapter $\{(\mathbf{L}_{SA}^l, \mathbf{L}_{SB}^l)\}_l^L$ using whitening SVD. To mitigate token-wise activation quantization errors, the shared expert employs channel-wise scaling, which reduces activation magnitudes while proportionally amplifying the corresponding weights. Through the initial reconstruction for $\mathbf{E}^l$ performed by the shared expert, QE achieves precise recovery of token-independent channels, while the remaining error $\mathbf{E}_S^l = \mathbf{E}^l - \mathbf{L}_{SA}^l \mathbf{L}_{SB}^l$ is subsequently refined by the routed experts associated with token-dependent channels. The algorithm for the shared expert is detailed in the supplementary Sec. 7.1.

3.3. *REs* for Token-Dependent Channels

In this subsection, we focus on the quantization error compensation for token-dependent important channels. Ideally, one would tailor an individual compensation strategy to each token, but the virtually unbounded value combinations make such token-specific designs computationally infeasible. Therefore, it becomes essential to design a constrained yet effective compensation strategy that approximates the optimal solution. We observe that token-dependent channels exhibit correlated occurrence patterns across different tokens. Based on this insight, we empirically compute their co-occurrence statistics and cluster channels with strong mutual association. We then employ routed experts, each employed with a low-rank adapter dedicated to modeling its corresponding channel cluster. During inference, a lightweight router estimates the final error of each expert and activates the one predicted to yield the lowest error.

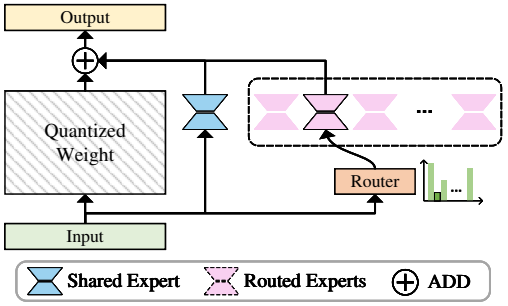


Figure 5. Illustration of the Inference Computation Process of QE.

Specifically, we construct the co-occurrence matrix following Eq. (7) to capture the token-level correlation patterns among token-dependent channels from all tokens $\mathbf{X}^l$.

$$\mathcal{O}_{t,i}^l = \mathbf{1}\big(c_i \in \mathcal{C}_r^l \cap \mathcal{A}_t^l\big), \quad \mathcal{O}^l \in \{0,1\}^{T \times (N_r k)}, \tag{7}$$

where $\mathcal{O}^l$ denotes the co-occurrence matrix over T tokens in l-th layer, and the indicator function $\mathbf{1}(\cdot)$ returns 1 if the i-th channel c_i is important for the t-th token, and 0 otherwise.

Next, we employ spectral clustering [34, 37] to partition the token-dependent important channels based on their most likely co-occurrence patterns. Firstly, the co-occurrence matrix $\mathcal{O}^l$ is transformed into a similarity matrix $\mathbf{S}^l \in \mathbb{R}^{(N_r k)\times(N_r k)}$ by using the normalized pointwise mutual information (NPMI) [2, 21], which quantifies the association strength between co-occurring channels as follows:

$$p(i) = \tfrac{1}{T} \sum_t^T \mathcal{O}_{t,i}^l, \tag{8}$$

$$p(i,j) = \tfrac{1}{T} \sum_t^T (\mathcal{O}_{t,i}^l \mathcal{O}_{t,j}^l), \tag{9}$$

$$\mathbf{S}_{i,j} = (\log \tfrac{p(i,j)}{p(i)p(j)}) \big/ -\log p(i,j), \tag{10}$$

where i, j represent the channel indices, and $\mathbf{S}_{i,j}$ represents the NPMI value between channels i and j. A higher NPMI value indicates a stronger likelihood that these channels will occur as important channels within a single token.

For the similarity matrix $\mathbf{S}^l$, we perform eigen decomposition on its normalized Laplacian $\mathcal{N}^l$ and take the next N_r eigenvectors $\mathbf{U}^l = [u_2, \ldots, u_{N_r+1}]$ as spectral embeddings. K-Means then partitions $\mathcal{C}_r^l$ into N_r clusters of token-dependent channels using spectral embeddings $\mathbf{U}^l$. For each cluster, we apply a weighting vector ω on $\mathbf{E}_S^l$ to enhance the reconstruction accuracy for its token-dependent channels γ. Next, we perform SVD to reconstruct each weighted $\mathbf{E}_S^l$ and truncate the rank to r. The N_r low-rank adapters together constitute the routed experts. Finally, we use the absolute mean of the remaining error

$\mathbf{E}_R^{l,i} = \mathbf{E}_S^l - \mathbf{L}_{RA}^{l,i}\mathbf{L}_{RB}^{l,i}$ for i-th routed expert as the parameters of router $\mathbf{R}_i^l$, which estimates the error of any input token under i-th routed expert. The above procedure is detailed in Algorithm 2. The inference process is illustrated in Fig. 5, where the shared expert provides stable global compensation and the router adaptively activates the optimal routed expert for dynamic local compensation.

To further alleviate the performance degradation introduced by post-training quantization, we design an optional lightweight refinement strategy. Specifically, only routed experts $(\mathbf{L}_{RA}^l, \mathbf{L}_{RB}^l)$ and the router $\mathbf{R}^l$ are trainable, while all other parameters remain frozen. Moreover, this refinement is performed layer-wise without end-to-end training of all parameters. The refinement is detailed and formulated in the supplementary material Sec. 7.2.

4. Experiments

4.1. Experimental Setup

Models. We conduct comprehensive PTQ experiments on several representative open-source VLMs, covering the Qwen2VL [43] series (2B, 7B, 72B) and the InternVL2 [8] series (2B, 8B). All model weights are obtained from the official repositories.

Baselines. We perform systematic comparisons with popular open-source PTQ methods. For weight-activation quantization, QE is evaluated under W4A6 and W4A8 settings against round-to-nearest (RTN), channel-scaling SmoothQuant (SQ) [44], modality-balanced MBQ [23], and low-rank reconstruction LQER [48]. Activations and weights are quantized using per-token and per-output-channel symmetric schemes, respectively. For weight-only quantization, we adopt AWQ [24] as the channel-scaling baseline under the W3A16 configuration, applying group-wise asymmetric quantization with 128 group size. Throughout, "WxAy" denotes weight and activation bitwidths of x and y, while #W and #A represent their bitwidths respectively.

Evaluation Metrics. To comprehensively assess the performance of our method, we experiment across diverse multimodal tasks. Text recognition and understanding are tested on OCRBench [25] and TextVQA [38]; document and infographic comprehension on DocVQA [32] and InfoVQA [33]; chart reasoning on ChartQA [31]; and general visual perception on VizWiz-VQA [14]. ScienceQA [29] and MMMU [46] evaluate scientific and general reasoning, while MMStar [6] and MuirBench [42] assess overall multimodal and multi-image understanding. AI2D [18] measures diagram comprehension, ensuring coverage across all major aspects of multimodal reasoning. We evaluate on the open source evaluation framework LMMs-Eval [49].

Experimental Details. We follow MBQ [23] and use the enhanced COCO Caption dataset [7] from ShareGPT4V [5], randomly sampling 128 image-caption pairs as the calibration set. In both LQER and QE, the total SVD rank r is set to 64. Since QE employs both shared and routed experts, the total rank of 64 is split into $\frac{64}{2}$ for each type, ensuring that the overall rank matches LQER. The k is fixed to 32, and the N_r is set to 8. For refinement, we set epochs for 16 with 100 iterations per epoch using the AdamW [28] optimizer (learning rate 1×10^{-4}, no weight decay) and a cosine annealing schedule [27]. The refinement coefficients are $\tau = 0.5$, $\alpha = 1.0$, and $\beta = 0.05$. Smaller models are evaluated on $4\times$ RTX 4090 24G GPUs, and the 72B model is on $4\times$ A800 80G GPUs.

4.2. Main Results

Smaller VLMs are generally more sensitive to quantization, making them more challenging benchmarks [22, 23]. We report results on Qwen2VL-2B and InternVL2-8B in Tab. 1 and Tab. 2, and on the larger 72B model in Tab. 3.

Weight-Activation Quantization. QE consistently surpasses both the modality-aware baseline MBQ [23] and the static low-rank method LQER [48]. In the challenging W4A6 setting, it improves Qwen2VL-2B accuracy by 4.01% over MBQ, with only a 4.23% drop from full precision, and gains 3.13% on InternVL2-2B. At W4A8, the performance gap to full precision narrows to within 2%, demonstrating strong robustness. On the larger Qwen2VL-72B, QE achieves a remarkable average accuracy improvement of 5.09% under the W4A6 quantization setting, nearly matching full-precision performance.

Weight-Only Quantization. QE again outperforms MBQ and LQER across all models. MBQ's distribution reshaping offers limited benefit over AWQ due to capacity limits, while LQER's static compensation yields only minor recovery. These results highlight the necessity of dynamic compensation to handle modality- and token-level distribution shifts that static methods cannot capture.

Additional results on Qwen2VL-7B and InternVL2-2B (Supplementary Tab. 9 and Tab. 10) show consistent gains, further confirming the generalizability of our method across diverse vision-language models.

4.3. Ablation Studies

Effect of Each Component. Our method integrates a shared expert (SE) and routed experts (REs). Main results show that their combination yields notable gains across quantization settings. To further verify each expert type, we conduct ablation studies in Tab. 4. Results show that removing either expert consistently degrades performance. The *random routing* experiment demonstrates that the proposed routing method can adaptively select the suitable routed expert to recover model accuracy. Similarly, the *random clustering* experiment confirms that the proposed co-occurrence-based clustering substantially enhances quanti-

Method	#W	#A	MMMU	OCRBench	ScienceQA	TextVQA	VizWiz	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	MuriBench	Avg. (↑)
Qwen2VL-2B	16	16	39.89	74.90	76.96	77.72	65.73	70.01	72.04	87.28	58.49	43.46	26.19	62.97
RTN	4	6	34.00	59.80	64.70	67.58	55.62	59.26	56.08	75.78	45.12	40.67	31.19	53.62
SQ (ICML'23)	4	6	30.44	59.60	65.25	65.88	53.90	59.16	40.44	70.73	39.36	38.20	30.00	50.27
LQER (ICML'24)	4	6	33.00	65.80	68.32	69.37	55.91	62.56	**62.68**	81.02	48.92	37.82	29.77	55.92
MBQ (CVPR'25)	4	6	**34.44**	61.10	67.08	69.45	57.19	60.91	60.08	76.24	43.13	**42.61**	29.77	54.73
QE	4	6	33.78	**68.20**	**71.84**	**73.18**	**59.62**	**65.45**	64.60	**82.75**	**51.84**	42.04	**32.88**	**58.74**
RTN	4	8	35.00	65.20	70.85	72.46	56.96	65.45	65.24	80.10	49.03	40.75	**34.23**	57.75
SQ (ICML'23)	4	8	32.11	65.80	68.02	68.61	57.43	61.95	44.08	75.07	42.06	38.93	29.58	53.06
LQER (ICML'24)	4	8	35.67	69.80	71.99	73.08	56.81	67.55	66.68	83.68	52.73	39.88	31.50	59.03
MBQ (CVPR'25)	4	8	34.33	62.30	70.85	72.36	58.32	65.19	62.32	78.19	47.93	**42.57**	32.62	57.00
QE	4	8	**37.33**	**72.10**	**74.67**	**75.34**	**61.59**	**68.36**	**69.28**	**84.46**	**53.97**	41.29	34.12	**61.14**
RTN	3	16	32.44	65.80	67.63	70.43	55.81	61.59	62.80	78.36	46.32	35.24	38.54	55.91
AWQ (MLSys'24)	3	16	33.22	63.40	68.32	70.60	56.88	61.72	62.08	78.70	45.02	36.48	35.62	55.64
LQER (ICML'24)	3	16	**34.56**	67.50	69.86	70.04	58.68	62.76	65.36	81.19	47.54	36.35	**38.58**	57.49
MBQ (CVPR'25)	3	16	33.44	63.90	69.16	70.75	52.59	63.28	63.56	79.27	45.87	34.66	34.50	55.54
QE	3	16	33.89	**70.10**	**72.09**	**74.46**	**60.15**	**64.96**	**68.52**	**82.97**	**51.31**	**39.66**	34.12	**59.29**

Table 1. Main results on the model of Qwen2VL-2B.

Method	#W	#A	MMMU	OCRBench	ScienceQA	TextVQA	VizWiz	AI2D	ChartQA	DocVQA	InfoVQA	MMStar	MuriBench	Avg. (↑)
InternVL2-8B	16	16	48.00	76.90	97.12	76.91	60.61	82.09	82.60	89.97	66.92	59.36	36.12	70.60
RTN	4	6	37.00	69.20	93.16	69.96	55.34	73.77	74.16	83.81	55.82	50.03	30.12	62.94
SQ (ICML'23)	4	6	40.44	69.50	94.84	69.83	50.91	75.13	74.92	84.28	56.77	49.59	32.00	63.47
LQER (ICML'24)	4	6	40.22	72.20	94.74	71.83	58.03	76.49	77.88	85.80	60.47	49.69	30.85	65.29
MBQ (CVPR'25)	4	6	43.67	71.00	95.49	70.26	52.90	77.59	75.64	84.21	58.09	53.89	32.27	65.00
QE	4	6	**44.89**	**74.30**	**96.23**	**74.77**	**59.00**	**79.40**	**80.48**	**87.77**	**63.12**	**55.27**	**34.15**	**68.13**
RTN	4	8	43.33	72.80	95.93	73.11	56.96	79.40	79.36	86.55	61.47	55.23	**35.38**	67.23
SQ (ICML'23)	4	8	42.22	72.20	95.54	72.58	51.68	77.04	77.48	85.48	59.07	51.98	32.58	65.26
LQER (ICML'24)	4	8	44.44	75.10	96.68	75.06	57.49	80.99	80.44	88.11	63.85	54.77	33.15	68.19
MBQ (CVPR'25)	4	8	44.44	73.50	**96.78**	72.00	56.67	79.21	77.72	86.42	61.48	55.32	32.77	66.94
QE	4	8	**45.56**	**75.60**	96.73	**76.08**	**59.77**	**81.06**	**81.60**	**88.31**	**64.83**	**56.09**	34.31	**69.09**
RTN	3	16	44.22	74.20	96.18	74.64	55.99	80.47	79.32	87.96	62.61	55.37	32.88	67.62
AWQ (MLSys'24)	3	16	45.67	74.60	96.33	74.97	59.18	80.47	80.08	88.01	63.55	54.55	34.85	68.39
LQER (ICML'24)	3	16	45.33	74.90	**96.38**	74.77	57.12	80.60	80.04	88.10	63.61	55.65	32.73	68.11
MBQ (CVPR'25)	3	16	**46.11**	75.20	96.18	74.97	58.43	79.70	79.72	88.00	63.16	54.96	**35.00**	68.31
QE	3	16	45.78	**75.90**	96.28	**75.50**	**59.46**	**80.99**	**80.76**	**88.79**	**64.35**	**57.14**	33.38	**68.94**

Table 2. Main results on the model of InternVL2-8B.

Setting	Method	MMMU	OCRBench	ScienceQA	TextVQA	VizWiz
FP16	-	61.44	78.70	91.22	82.26	76.27
W4A6	RTN	51.33	59.10	85.52	72.12	64.29
	SQ	53.33	61.20	86.96	74.42	63.45
	LQER	52.33	59.60	86.71	73.84	66.64
	MBQ	52.67	69.70	86.32	76.08	67.99
	QE	**58.11**	**76.60**	**90.33**	**79.27**	**73.91**
W4A8	RTN	57.11	66.50	90.08	75.96	71.14
	SQ	56.44	65.10	90.03	76.37	66.70
	LQER	57.44	70.60	91.32	78.15	74.28
	MBQ	58.33	73.90	89.24	79.19	72.24
	QE	**58.89**	**78.30**	**91.47**	**81.47**	**75.83**

Table 3. Main results of Qwen2VL-72B model (higher is better).

Setting	Component	MMMU (↑)	ScienceQA (↑)
FP16	-	39.89	76.95
W4A6	routed experts (REs)	34.56	68.72
	shared expert (SE)	35.22	69.61
	SE + random routing	35.89	70.00
	SE + random clustering	35.33	69.71
	QE (SE+REs)	**36.89**	**70.85**
W4A8	routed experts (REs)	36.00	71.94
	shared expert (SE)	36.78	73.13
	SE + random routing	37.89	73.67
	SE + random clustering	37.22	73.82
	QE (SE+REs)	**38.00**	**74.37**

Table 4. Ablation study results on Qwen2VL-2B model.

zation performance. Furthermore, we visualize the clustering results in Fig. 6 according to Sec. 3.3. The clustered similarity matrix and t-SNE [30] projection indicate that our method effectively identifies and partitions these co-occurring clusters.

Effect of Refinement. Tab. 5 presents the ablation results of the proposed refinement for routed experts under the Qwen2VL W4A6 quantization. As observed on both 2B and 7B models, applying refinement consistently leads to notable accuracy improvements across multiple tasks compared to the non-refinement counterparts.

Effect of the Number of Routed Experts. We analyze in Tab. 6 the impact of the number of routed experts on model accuracy. Here, N_r denotes the number of routed experts.

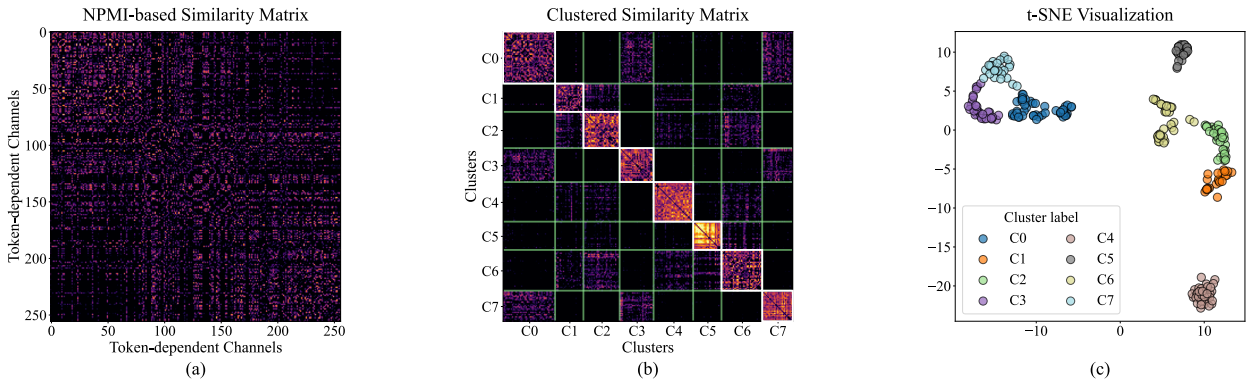


Figure 6. Illustration of the Co-Occurrence-Based Clustering in a Transformer Block of Qwen2VL-2B. (a) Similarity matrix $\mathbf{S}^l$ showing mutual co-occurrence among token-dependent channels, with brightness indicating similarity. (b) Channels with strong co-occurrence are grouped into the same cluster. (c) t-SNE [30] projection demonstrates that the clustering effectively captures their co-occurrence relations.

Model	Ref.	MMMU	OCRBench	ScienceQA	TextVQA	VizWiz
2B	✗	33.78	68.20	**71.84**	73.18	59.62
	✓	**36.89**	**69.60**	70.85	**73.30**	**60.58**
7B	✗	**45.44**	73.00	79.87	71.63	64.18
	✓	44.00	**74.60**	**80.61**	**77.58**	**65.11**

Table 5. Ablation study of Refinement on Routed Experts. "✓" indicates applied refinement; "✗" indicates none.

It can be observed that as N_r increases, the model performance gradually improves; however, a larger N_r also implies higher memory overhead for routed experts.

N_r	OCRBench	TextVQA	VizWiz	Avg. (↑)
2	68.40	73.14	59.70	67.08
4	68.50	73.13	60.41	67.35
8	69.60	73.30	60.58	67.83
16	69.90	73.52	60.75	68.06

Table 6. Impact of the number of routed experts on the performance of Qwen2VL-2B under the W4A6 quantization setting.

4.4. Overheads Analysis and Kernel Performance

The low-rank adapter introduces additional computation and memory overhead from the lightweight auxiliary matrices $\mathbf{L}_{:A}$ and $\mathbf{L}_{:B}$. Let s denote the sequence length, d hidden size, $r(\ll d)$ the rank, and $N_r(\ll d)$ the routed experts count. The complexity analysis of a single layer is provided in Tab. 7, with computation quantified in terms of floating-point operations (FLOPs) and memory overhead evaluated by the total number of parameters. Notably, QE incurs only minimal additional computational and memory costs compared to the original linear, yet enables the quantized model to recover accuracy comparable to full precision.

To assess hardware efficiency, we develop an analytical performance model following the FlightLLM [47] accelerator architecture. Prefill-stage kernel speedups of linear layers are measured with sequence length 128 using Qwen2VL-7B weight shapes under various quantization settings. As shown in Tab. 8, QE achieves 3.5×-4.5× acceleration, highlighting its strong potential for hardware-level efficiency gains.

Complexity	Origin	QE
Computation	sd^2	$sd^2 + sd(2r + N_r)$
Memory	d^2	$d^2 + rd(1 + N_r)$

Table 7. Complexity analysis of the linear layer in QE method.

Shape (IC × OC)	W4A6	W4A8	W3A16
3584 × 3584	3.56×	3.50×	4.10×
3584 × 18944	3.60×	3.59×	4.50×
18944 × 3584	3.84×	3.77×	4.50×

Table 8. NPU speedup ratios of QE for Qwen2VL-7B linear layers compared with the fp16 model, measured during the prefill stage with a sequence length of s=128. "IC" and "OC" denote the input and output channel dimensions, respectively.

5. Conclusion

In this work, we reveal a key observation that the distributions and occurrence frequencies of important channels vary significantly both across modalities and among tokens, even within the same modality. Building on this insight, we propose Quant Experts (QE), a token-aware adaptive error compensation for VLMs quantization, that dynamically adapts to such variations. Specifically, QE employs a shared expert to robustly reconstruct token-independent channels and routed experts to adaptively compensate token-dependent ones, with each expert implemented as a low-rank adapter. Comprehensive evaluations on diverse VLMs show that QE consistently outperforms globally static PTQ baselines across different quantization configurations.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2025ZD0122404), the National Natural Science Foundation of China (Grant No. 92464302), the China Postdoctoral Science Foundation (Grant No. 2025M770535), and the Sanqin Talent Special Support Plan (Grant No. 2024STZZK09).

References

[1] Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated llms. In *NeurIPS*, pages 100213–100240, 2024. 1

[2] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of GSCL*, pages 31–40, 2009. 5

[3] Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. QBB: Quantization with binary bases for llms. In *NeurIPS*, pages 3209–3228, 2024. 1

[4] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *NeurIPS*, 2023. 1

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pages 370–387, 2024. 6

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024. 6

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 6

[9] Tim Dettmers, Ruslan A. Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024. 2, 1

[10] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020. 1

[11] Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal Brain Compression: a framework for accurate post-training quantization and pruning. In *NeurIPS*, 2022. 2, 1

[12] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. 2, 4, 1

[13] Zhuocheng Gong, Jiahao Liu, Jingang Wang, Xunliang Cai, Dongyan Zhao, and Rui Yan. What makes quantization for large language model hard? an empirical study from the lens of perturbation. In *AAAI*, pages 18082–18089, 2024. 1

[14] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 6

[15] Ethan He, Abhinav Khattar, Ryan Prenger, Vijay Korthikanti, Zijie Yan, Tong Liu, Shiqing Fan, Ashwath Aithal, Mohammad Shoeybi, and Bryan Catanzaro. Upcycling large language models into mixture of experts. *arXiv preprint*, 2024. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[17] Wenjing Ke, Zhe Li, Dong Li, Lu Tian, and Emad Barsoum. DL-QAT: Weight-decomposed low-rank quantization-aware training for large language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 113–119, 2024. 1

[18] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 6

[19] Gun Il Kim, Sunga Hwang, and Beakcheol Jang. Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Computing Surveys*, 57 (10):1–39, 2025. 1

[20] Junhan Kim, Chungman Lee, Eulrang Cho, Kyungphil Park, Ho young Kim, Joonyoung Kim, and Yongkweon Jeon. Towards next-level post-training quantization of hyper-scale transformers. In *NeurIPS*, 2024. 1

[21] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Conference on Computational Natural Language Learning*, pages 171–180, 2014. 5

[22] Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Evaluating quantized large language models. *arXiv preprint arXiv:2402.18158*, 2024. 6

[23] Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. MBQ: Modality-balanced quantization for large vision-language models. In *CVPR*, pages 4167–4177, 2025. 1, 2, 6

[24] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024. 2, 4, 6, 1

[25] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: on the hidden mystery of ocr in large multimodal models. In *Science China Information Sciences*, page 220102, 2024. 6

[26] Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics*, pages 467–484, 2024. 1

[27] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 6

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 6

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 7, 8

[31] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Association for Computational Linguistics*, 2022. 6

[32] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 6

[33] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 6

[34] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2001. 5

[35] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. 1

[36] Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. 2

[37] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000. 5

[38] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 6

[39] Siqing Song, Chuang Wang, Rui-Qi Wang, Yi Yang, and Xu-Yao Zhang. Achieving binary weight and activation for LLMs using post-training quantization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8782–8795. Association for Computational Linguistics, 2025. 1

[40] Hanlin Tang, Yifu Sun, Decheng Wu, Kai Liu, Jianchen Zhu, and Zhanhui Kang. EasyQuant: An efficient data-free quantization algorithm for LLMs. In *Empirical Methods in Natural Language Processing*, pages 9119–9128, 2023. 4

[41] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better llm quantization with hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning*, 2024. 1

[42] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *ICLR*, 2025. 6

[43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint*, 2024. 6

[44] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099, 2023. 1, 2, 6

[45] Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. Exploring post-training quantization in llms from comprehensive study to low rank compensation. In *AAAI*, 2024. 1

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 6

[47] Shulin Zeng, Jun Liu, Guohao Dai, Xinhao Yang, Tianyu Fu, Hongyi Wang, Wenheng Ma, Hanbo Sun, Shiyao Li, Zixiao Huang, et al. Flightllm: Efficient large language model inference with a complete mapping flow on fpgas. In *ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, pages 223–234, 2024. 8

[48] Cheng Zhang, Jianyi Cheng, George A. Constantinides, and Yiren Zhao. LQER: low-rank quantization error reconstruction for llms. In *International Conference on Machine Learning*, 2024. 2, 6, 1

[49] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-eval: Reality check on the evaluation of large multimodal models. In *Association for Computational Linguistics*, pages 881–916, 2025. 6

[50] Weibo Zhao, Yubin Shi, Xinyu Lyu, Wanchen Sui, Shen Li, and Yong Li. ASER: activation smoothing and error recon-

struction for large language model quantization. In *AAAI*, 2025. 2, 4, 1

[51] Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. Atom: Low-bit quantization for efficient and accurate llm serving. In *Machine Learning and Systems*, 2024. 1