SeGO: Sensitivity-Aware Golden Optimization for Large-Scale VLM Quantization

Tianqi Zhao[1] , **Xinrui Cheng**[1] , **Yang Su**[2] , **Weiyi Lu**[1] , **Zhaodong Zhang**[1] , **Zhongjie Wang**[1] and **Ruihan Hu**[1,*]

[1]Harbin Institute of Technology

[2]Taiyuan University of Technology

ztq@stu.hit.edu.cn, cxrxin0530@163.com, 2024521889@link.tyut.edu.cn, wylu651@qq.com, 25s003081@stu.hit.edu.cn, rainy@hit.edu.cn, rh.hu@hit.edu.cn

Abstract

The deployment of Vision-Language Models (VLMs) faces memory and computational bottlenecks because of the massive parameters and intensive computations. While Post-Training Quantization (PTQ) can reduce these costs, existing methods often overlook the heterogeneity of multimodal input when applied to VLMs, leading to quantization accuracy degradation. In this paper, we categorize the Transformer linear layers into Expansion Space (responsible for feature expansion) and Projection Space (responsible for feature aggregation) and reveal a cross-modal structural sensitivity asymmetry in VLMs: The Expansion Space is more sensitive to quantization accuracy than the Projection Space in both LLM and ViT encoders. Based on this, we propose SeGO, a unified structural sensitivity-aware sparse optimization framework. First, it constructs search interval for scaling factors using channel-wise statistics of LLM and ViT weights to exclude invalid solutions. Then it applies scaling protection only to the sensitive Expansion Space while directly quantizing the robust Projection Space. Moreover, SeGO uses a Golden-Section Solver to speedup the search of optimal scaling factors. Experiments show that SeGO achieves the balance among model parameter amount, quantization accuracy and scaling factors' search efficiency on InternVL2 and LLaVA series. Under the W3/W4A16 configuration, SeGO improves the accuracy of 7B-26B models by 3.8%-17.3% and enhances calibration efficiency by dozens of times.

1 Introduction

Recently, large Vision-Language Models (VLMs) have shown excellent performance across various tasks [Brown *et al.*, 2020; Yu *et al.*, 2025]. However, their massive parameters, such as up to 72B for InternVL2 [Chen *et al.*, 2024c], make deployment on resource-constrained devices extremely difficult. Post-Training Quantization (PTQ) has become a key method to reducing memory usage and accelerating inference without retraining [Shao *et al.*, 2023; Yuan *et al.*, 2023; Zhou *et al.*, 2025].

Although PTQ has succeeded in Large Language Models (LLMs) such as SmoothQuant [Xiao *et al.*, 2023], SpinQuant [Liu *et al.*, 2024d], FlatQuant [Sun *et al.*, 2024], Atom [Zhao *et al.*, 2024], existing researches indicate that the introduction of the visual modality leads to extreme heterogeneity with respect to input feature distributions [Luo *et al.*, 2024; Gong *et al.*, 2024]. Traditional PTQ methods often assume uniform sensitivity across layers and use global heuristic searches. They fail to handle complex cross-modal interactions in VLMs, leading to severe quantization accuracy drops at low bits. Furthermore, as the parameter amount of model increases, systematic outliers emerge in Transformer layers [Dettmers *et al.*, 2022], hence the performance degradation of traditional uniform quantization strategies becomes more pronounced, further exacerbating the difficulty of quantizing large-scale VLMs.

To address the problem of quantization accuracy degradation, some gradient-based PTQ methods such as OmniQuant [Shao *et al.*, 2023], SignRound [Li *et al.*, 2024c], QLLM [Liu *et al.*, 2024b], PB-LLM [Shang *et al.*, 2024b], NormTweaking [Li *et al.*, 2024b] recover quantization accuracy through backpropagation, but their high memory cost and hours of calibration time become prohibitive as model size grow. While GPU cluster optimizations [Li *et al.*, 2025] address hardware bottlenecks, algorithmic efficiency remains vital for VLM deployment. Moreover, from the perspective of model quantization coverage, most existing works [Zhou *et al.*, 2024; Lin *et al.*, 2024b; Li *et al.*, 2024d; Chen *et al.*, 2024b] focus on compressing the LLM while keeping the Vision Transformer (ViT) in FP16 format. This is mainly because, in early small-scale VLMs, the number of parameters in ViT (typically around 300M) is much smaller than that of LLMs. And visual inputs are more sensitive to quantization, the severe quantization accuracy degradation caused by quantizing the ViT is disproportionate to the limited memory reduction obtained. However, modern VLMs use much larger ViT. For instance, InternVL2-26B features a 6B-parameter ViT [Chen *et al.*, 2024a], which alone consumes more memory than many 7B-class LLMs. Under this trend, quantizing only the LLM will make the ViT the new resource bottleneck, hindering the model's deployment on edge devices. Therefore, achieving high-precision deployment of

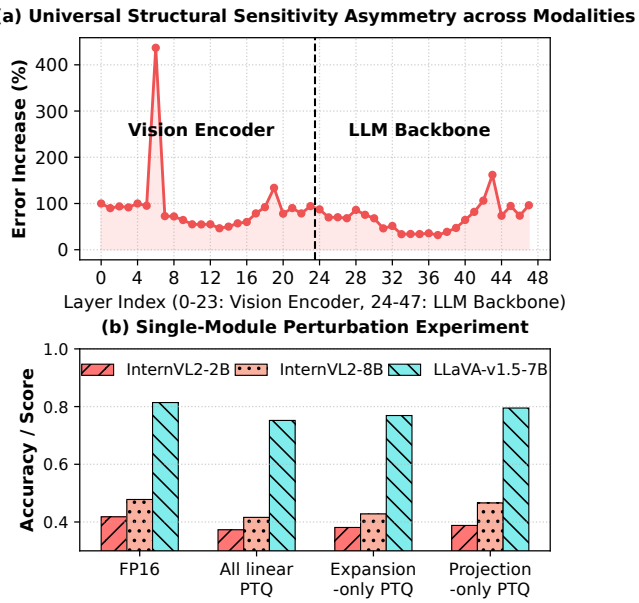

Figure 1: Motivation for Sparse Structural Intervention. (a) Quantization error is dominated by Expansion Space across the full VLM architecture. (b) Perturbation analysis reveals that Projection Space are robust to low-bit quantization, justifying a sparse, Expansion-focused optimization approach.

full-model VLMs remains an urgent challenge. This is critical for the realization of 6G endogenous AI networks, where edge-cloud collaboration is essential for distributing intelligence across the network [Wang *et al.*, 2025].

To solve these problems, this paper first conducts an in-depth analysis of the sources of VLM quantization errors. We categorize the linear layers in VLMs into two types:

- Expansion Space: Map input hidden states to a high-dimensional space. Their output usually connects to nonlinear activation functions, leading to highly sparse distributions of activation values and significant outliers.
- Projection Space: Aggregate high-dimensional features back to the model's hidden dimension. It is more robust to quantization error.

We analyze the error propagation mechanism within the Transformer blocks of the full multimodal model. Under the low-bit quantization configuration, we perform a layer-by-layer sensitivity analysis on the 48-layer Transformer of the InternVL2-2B model (with the first 24 layers as the ViT encoder and the last 24 layers as the LLM encoder). In each layer, we test the layer-wise output errors generated by "quantizing only the Expansion Space" and "quantizing only the Projection Space". As shown in Fig. 1(a), the red line and its shaded area show the relative error increase percentage caused by quantizing the Expansion space alone, compared to quantizing the projection space alone at the current layer. Our analysis of the InternVL2-2B model shows that quantization errors are dominated by the Expansion Space across both the ViT and LLM. Since the InternVL2 series models use the most typical Transformer architecture, the structural sensitivity asymmetry observed in these models are also applicable to other VLMs based on the Transformer architecture. We extend the same experiments to other variants of InternVL2 and the LLaVA series, the findings are consistent with the results illustrated in Fig. 1(a). This indicates that the Expansion Space is the key bottleneck causing the degradation in VLM quantization accuracy.

To validate this finding, we further conduct single-module perturbation experiments on different models. We observe same structural sensitivity asymmetry across various quantization configurations. The results shown in the Fig. 1(b), which correspond to quantization at 4 bits, demonstrate that quantizing only the Expansion Space to 4 bits while keeping the Projection Space in FP16 results in accuracy that is very close to full quantization. Quantizing only the Projection Space to 4 bits while keeping the Expansion Space in FP16 almost restores the accuracy to the FP16 level. This phenomenon suggests that the quantization of the Expansion Space is the main factor contributing to the accuracy degradation, while the Projection Space exhibits strong robustness. This suggests that sparse intervention focusing on the Expansion Space is sufficient to restore accuracy, thereby saving the overhead of applying scaling protection to all linear layers. Considering that VLM activations change rapidly with text token and visual token and compared to the static distribution of weight features, forcibly quantizing the activations leads to a significant drop in the accuracy of VLMs [Jiang *et al.*, 2024; Shang *et al.*, 2024a], we focus on weight-only quantization and treat activation quantization as complementary future work.

Based on the above insights, we propose SeGO, a unified structural sensitivity-aware sparse optimization framework. Specifically, for the LLM encoder and ViT encoder, it efficiently constructs the search interval for scaling factors based on static weight statistics to exclude invalid solutions. Building upon this, we design a unified golden-section solver, leveraging its linear convergence property to speedup the search of optimal scaling factors for the Expansion Space. By only scaling the sensitive Expansion Space while directly quantizing the Projection Space, SeGO achieves efficient and consistent compression for the entire model. Our main contributions are:

- We reveal structural sensitivity asymmetry across modalities and systematically demonstrate for the first time that in the LLM encoder and ViT encoder of VLMs, the Expansion Space is a key bottleneck for quantization error.
- We propose the SeGO framework: we develope a unified parameter-solving strategy for the entire model. For the LLM encoder and ViT encoder, we innovatively use static weight statistics to construct the search interval for scaling factors.
- We introduce a Golden-Section Solver to speedup the search of optimal scaling factors without requiring gradients. It address the cross-modal optimization unification problem and significantly reduce computational overhead.

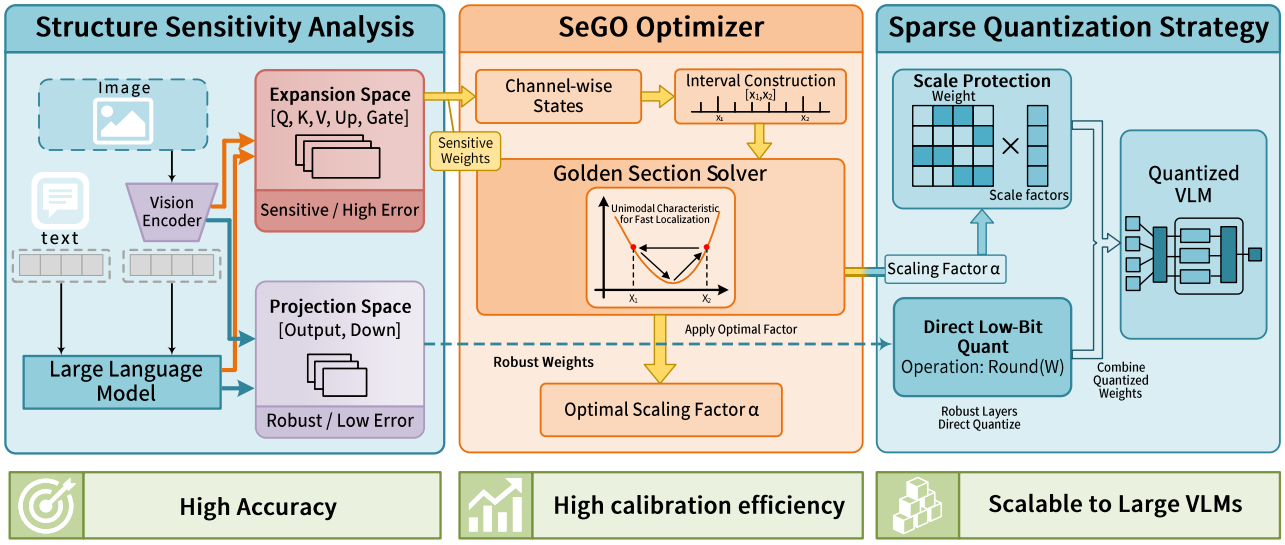

Figure 2: Overview of the SeGO framework. It consists of three phases: (1) Structure Sensitivity Analysis: Identifying the Expansion Space as the quantization bottleneck based on cross-modal structural sensitivity asymmetry. (2) SeGO Optimizer: Utilizing static weight statistics to construct search interval and employing a Golden-Section Solver to speedup the search of optimal scaling factors. (3) Sparse Quantization Strategy: Applying scaling protection to Expansion Space while directly quantizing robust Projection Space for high-accuracy VLM deployment.

2 Preliminaries and Analysis

2.1 Definition and Symbolic Representation

In the VLM Transformer architecture, we define the weight matrix set as $\mathbf{W}^{\mathrm{m}}_{\mathrm{type}}$, where $m \in \{\mathrm{LLM},\mathrm{ViT}\}$ and type $\in \{\mathrm{exp},\mathrm{proj}\}$.

Expansion Space($\mathbf{W}_{\mathrm{exp}}$): It is responsible for mapping the input features to a higher-dimensional space and the output is typically directly connected to a nonlinear activation function. $\mathbf{W}^{\mathrm{LLM}}_{\mathrm{exp}}$ includes the projections of Q, K, and V in the Self-Attention and the Gate and Up projections in the MLP module; $\mathbf{W}^{\mathrm{ViT}}_{\mathrm{exp}}$ includes the projections of q, k and v in the Multi-Head Attention and the first fully connected layer of the FFN. Due to the" dimensional Expansion + activation" operation, the weight distribution in the Expansion Space often has a large dynamic range and outliers.

Projection Space($\mathbf{W}_{\mathrm{proj}}$): It is responsible for aggregating high-dimensional features back to the backbone dimension. $\mathbf{W}^{\mathrm{LLM}}_{\mathrm{proj}}/\mathbf{W}^{\mathrm{ViT}}_{\mathrm{proj}}$ corresponds to the Output projection in each Attention and the Down projection in the MLP. Since the Projection Space performs dimensionality reduction and Down projection in the MLP. Since the projection space performs dimensionality reduction and weighted summation, it has a smoothing effect on the introduced quantization errors.

2.2 Quantization Basis

This paper aims to address the PTQ problem of the entire VLMs. Based on the definitions in Section 2.1, we focus on the quantization of the Expansion Space weights ($\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}$, m represents the LLM and ViT). First, we define the per-channel maximum absolute value vector of the weight matrix $\mathbf{v}^m \in \mathbb{R}^{C_{in}}$, where the c-th element is:

$$v^m_c = \max_j \left| \left(\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\right)_{j,c} \right|, \quad c \in \{1, ..., C_{in}\} \tag{1}$$

This statistic reflects the weight intensity distribution of the input channels in the Expansion Space across different modalities. We quantize the weights to b-bit integers, and the quantization function is represented as:

$$\widehat{\mathbf{W}}^{\mathrm{m}}_{\mathrm{exp}} = \Delta \cdot \mathrm{clip}\left(\lfloor \frac{\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}}{\Delta} \rceil + z, q_{\min}, q_{\max} \right) \tag{2}$$

Where Δ is the quantization step and z is the zero point. Our goal is to minimize the reconstruction error, as formulated in Eq.(3), under the given low-bit width. Since the structural sensitivity of $\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}$ is higher than $\mathbf{W}^{\mathrm{m}}_{\mathrm{proj}}$, our optimization focus will be entirely on the $\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}$.

$$\mathcal{L} = \left\| \widehat{\mathbf{W}}^{\mathrm{m}}_{\mathrm{exp}}\mathbf{X} - \mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\mathbf{X} \right\|^2_2 \tag{3}$$

2.3 Mitigating Errors through Channel-Wise Scaling Equivalent

To reduce above error, we introduce a channel-wise scaling factor α for the Expansion Space $\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}$, We reorganize the weight distribution through mathematical equivalence:

$$y = \mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\mathbf{X} = (\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\alpha)(\alpha^{-1}\mathbf{X}) \tag{4}$$

The quantized output error e after transformation satisfies is:

$$\begin{aligned} \|\mathbf{e}\|_2 &= \|(Q(\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\alpha) - \mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\alpha)\alpha^{-1}\mathbf{X}\|_2 \\ &\leq \|\mathbf{E}\|_2 \|\alpha^{-1}\|_2 \|\mathbf{X}\|_2 \end{aligned} \tag{5}$$

Where $\mathbf{E} = Q(\mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\alpha) - \mathbf{W}^{\mathrm{m}}_{\mathrm{exp}}\alpha$ represents the scaled weight error and Q means the quantization function. By carefully designing the scaling factor α, we can redistribute the weight's range, suppress truncation errors, and reduce rounding errors. From this derivation, we conclude that: by specifically optimizing the scaling factor for the Expansion Space, we can achieve precise protection of the structurally sensitive layers.

3 SeGO Framework

This section provides a detailed introduction to the SeGO framework and the overview is shown in Fig. 2. Based on the previously revealed structural sensitivity asymmetry, SeGO aims to achieve efficient quantization for the entire model through sparse intervention in the Expansion Space. We construct a unified optimization paradigm for the entire model. First, we use static weight statistics to construct the search interval for scaling factors. Next, within the determined search interval, we efficiently locate the optimal scaling factor using the golden-section solver. Finally, we perform unified quantization to both the Expansion Space and Projection Space.

3.1 Unified Scaling interval Initialization Based on Weight Magnitudes

Since the Expansion Space maps input features to a higher-dimensional space, its weight distribution often exhibits significant channel-wise variations. These variations are the cause of quantization errors. To construct an efficient and unified optimization framework, we use the statistical characteristics of multimodal model weights to initialize the scaling factors.

For both LLM and ViT Expansion Spaces weight $\mathbf{W}_{\mathrm{exp}}$, we calculate the channel-wise max absolute value v_c^m as formula(1). This metric directly reflects the weight intensity of the channel in the network transformation. Based on v_c^m, we define the initial scaling factor α_c^{init} for each channel. Our goal is to introduce the scaling factor to reduce the values of channels with large weights while equivalently amplifying the input features, thereby balancing the numerical range across channels. The α_c^{init} is calculated as follows:

$$\alpha_c^{init} = \mathrm{clip}\left(\frac{v_c^m}{\|v\|_2} \cdot \lambda, \alpha_{\min}, \alpha_{\max}\right) \tag{6}$$

Where $\|\mathrm{v}\|_2$ is the L2 norm of the channel magnitude, used to normalize the relative proportion between channels; λ is a hyperparameter controlling the overall balance strength; $\alpha_{\min}$ and $\alpha_{\max}$ are the preset lower and upper bounds for the scaling factor.

Although the scaling factors are initialized based on the statistical characteristics of the weights, the optimal solution may deviate from this static estimate. Therefore, we restrict the subsequent search to a local neighborhood centered $[\alpha_c^{init} - \delta, \alpha_c^{init} + \delta]$ around α_c^{init}, providing a high-quality starting point for fine-tuning the optimization for specific modal objectives.

3.2 Scaling Factor Solver Based on the Golden-Section

We need to find the optimal α^* to minimize the reconstruction error. This problem can be formalized as an optimization task to find the extremum of a univariate function within a continuous, bounded interval. Such problems typically have two key characteristics: First, the objective function often exhibits unimodality within a reasonably initialized interval; second, the optimization process needs to be completed with minimal computational overhead. To verify the critical assumption that"the loss function exhibits unimodal characteristics within the search interval", we analyze the Expansion Space at different depths in the InternVL2-8B model. For each channel's scaling factor optimization, we define a modality-weighted loss function. Previous studies have shown that the token of the text modality is typically more sensitive to quantization errors, and it often dominates the overall error [Luo *et al.*, 2024; Gong *et al.*, 2024]. Therefore, we split the output error into text and visual components and define the output reconstruction error as follows:

$$\mathcal{L}_{rec} = \min_{\alpha}\left(\beta_t\|\mathbf{y}_t^{fp} - \mathbf{y}_t^{q}\|_1 + \beta_v\|\mathbf{y}_v^{fp} - \mathbf{y}_v^{q}\|_1\right) \tag{7}$$

Where y^{fp} and y^q represent the full-precision and quantized output activation values. The subscripts t and v refer to the output corresponding to text tokens and visual tokens. β_t and β_v are the weighting coefficients used to balance the importance of different modalities, and $\|\cdot\|_1$ denotes the L1 norm, which is more robust to outlier errors.

For each channel, we perform grid sampling with a fine-grained step size of 0.05 within its determined search interval. As illustrated in Fig. 3, the reconstruction error curves exhibit a pronounced unimodal shape (convex-like) with unique minimum points. Moreover, the region around this minimum point shows smooth variation, indicating good local convexity.

For this typical scenario of optimizing a unimodal function within a limited interval, there exists a class of efficient and stable interval contraction algorithms in the field of numerical optimization. We adopt the golden-section method, one of the most representative algorithms, as the core solver. The advantage of this method is that it does not require derivative information and only uses function value comparisons to shrink the search interval, avoiding the memory and computational overhead caused by backpropagation. Each iteration reduces the interval by a fixed proportion, ensuring linear convergence and achieving a high-precision solution with very few evaluations.

Specifically, for each channel, we use the golden-section ratio $\varphi = \frac{\sqrt{5}-1}{2} \approx 0.618$ to iteratively shrink the search interval $[a, b]$ which is determined by Section 3.1:

1. **Probe Point Generation**: Compute two internal probe points:

$$x_1 = b - \varphi(b-a) \; and \; x_2 = a + \varphi(b-a) \tag{8}$$

2. **Interval Reduction**: Compare $\mathcal{L}_{rec}(x_1)$ and $\mathcal{L}_{rec}(x_2)$, whose calculation method is shown in Formula 6. If

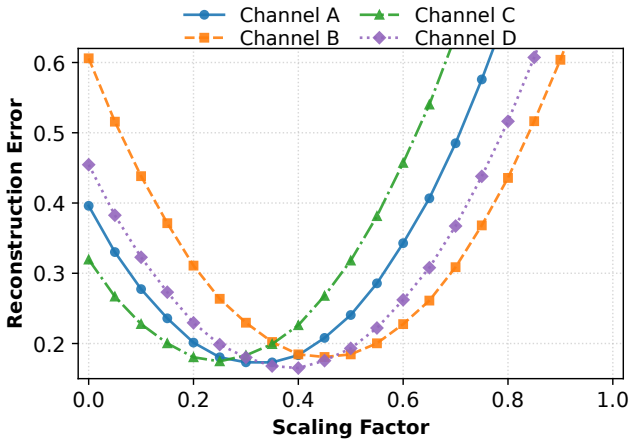

Figure 3: Visualization of the reconstruction error curve for different channels in the Expansion Space. The error curves exhibit distinct unimodal characteristics, validating the feasibility of our method. While grid search (dots, step=0.05) is limited by resolution, SeGO could locate the precise global minimum within the continuous domain, demonstrating superior precision.

$\mathcal{L}_{rec}(x_1) < \mathcal{L}_{rec}(x_2)$, update the new interval as $[a, x_2]$, otherwise, update it as $[x_1, b]$.

3. **Convergence Check**: Repeat the above steps until the interval width is smaller than a pre-set convergence tolerance ϵ (for example $\epsilon = 10^{-2}$), which controls the precision of the optimal solution.

This solving process fully leverages the unimodal characteristic of the loss function within the interval. It efficiently locates the optimal scaling factor with logarithmic complexity $O(\log(1/\epsilon))$. And it perfectly meets the comprehensive requirements for precision and efficiency in VLM quantization.

3.3 Unified Quantization and Efficiency Analysis

After finding α^*, SeGO enters the final execution stage. We apply a unified mathematical transformation to reorganize the model weight. For each Expansion Space, we first adjust the weight distribution using the scaling factor and incorporate the inverse scaling factor into the preceding layer's parameters. This aims to maintain mathematical equivalence with the inference results. Subsequently, we quantize the weights of all linear layers (including the reorganized Expansion Space and the untreated Projection Space) to the target bit-width. This "structure alignment first, then unified quantization" paradigm ensures that the quantized model can be directly deployed on existing high-performance inference kernels without modifying any underlying operators.

Time Complexity: Thanks to the weight statistics-based initialization strategy proposed in Section 3.1, SeGO does not require data forward propagation during this phase. The search interval can be constructed using static weight statistics. Furthermore, the golden-section solver leverages its linear convergence property, requiring only a few iterations within the determined search interval to locate the optimal solution.

Space Complexity: The peak memory usage of SeGO is only slightly higher than that of model inference, allowing it to easily calibrate models with billions of parameters on consumer-grade GPUs.

4 Experiments

4.1 Experimental Setup

Datasets and Models. To activate both visual and linguistic capabilities during calibration, we construct a set of 128 image-text pairs randomly sampled from **MS COCO 2017** [Lin *et al.*, 2014] and **ShareGPT4V** [Chen *et al.*, 2023]. For a comprehensive evaluation, we select two VLM series across various scales: InternVL2 [Chen *et al.*, 2024c] (1B, 2B, 8B, 26B) and **LLaVA** (onevision-7B [Li *et al.*, 2024a] and v1.5-7B [Liu *et al.*, 2024a]). We evaluate performance via the **LMMs-Eval** [Zhang *et al.*, 2024] framework on diverse benchmarks, including **OCRBench** (text recognition) [Liu *et al.*, 2024c], **MMMU_val** (multi-discipline reasoning) [Yue *et al.*, 2024], **RealWorldQA** (real-world QA) [Fu *et al.*, 2024], and **AI2D** (diagram reasoning) [Kembhavi *et al.*, 2016].

Baselines and Configuration. We compare **SeGO** with three mainstream PTQ methods: (1) **RTN**, the basic round-to-nearest quantization; (2) **AWQ** [Lin *et al.*, 2024a], which utilizes channel-wise equalization to suppress outliers; and (3) **GPTQ** [Frantar *et al.*, 2023], a second-order approximation method based on the Hessian matrix. For a fair comparison, all methods adopt **weight-only, asymmetric group-wise quantization** (group size 128). We conduct evaluations under both **W3A16** and **W4A16** bit-width configurations.

4.2 Main Results

We evaluate SeGO against FP16 and mainstream PTQ baselines under the unified settings. Results for the InternVL2 series are summarized in Table 1–2, while results for LLaVA-onevision-7B and LLaVA-v1.5-7B are presented in Table 3–4.

Analysis of Results. As shown in Table 1-4, SeGO consistently outperforms baselines across different architectures and scales. In W4A16, all methods show some degradation on the small-scale model. The advantage of SeGO becomes more significant as model size increases (InternVL2-8B/26B). In the more challenging 3-bit scenario, the quantization error is significantly amplified. RTN almost fail on small models, while AWQ and GPTQ offer some relief but still exhibit significant accuracy degradation. SeGO maintains high accuracy. For InternVL2-8B on MMMU_val, SeGO reaches 0.4378, far exceeding RTN (0.2384) and GPTQ (0.2274). More importantly, SeGO ranks first for both LLaVA-onevision and LLaVA-v1.5, proving it effectively handles structural sensitivity asymmetry in different model architectures.

4.3 Calibration Efficiency and Cost Comparison Experiment

In addition to its accuracy advantage, another contribution is its exceptional calibration efficiency. To show this advantage,

Model	Bitwidth	Method	MMMU	RWQA	OCR	AI2D
Intern VL2-1B	FP16	–	0.3422	0.5203	0.7400	0.6231
	W4A16	RTN	0.3187	0.4464	0.6966	0.5584
		AWQ	**0.3222**	0.4444	**0.7021**	0.5613
		GPTQ	0.2972	0.4324	0.6571	0.5343
		SeGO	0.3089	**0.4876**	0.6911	**0.5774**
Intern VL2-2B	FP16	–	0.3422	0.4183	0.7510	0.7247
	W4A16	RTN	0.3181	0.3733	0.6627	0.6951
		AWQ	0.3317	0.3725	0.6732	0.6972
		GPTQ	0.2837	0.3655	0.6452	0.6932
		SeGO	**0.3322**	**0.3778**	**0.7288**	**0.7124**
Intern VL2-8B	FP16	–	0.4778	0.6562	0.7655	0.8235
	W4A16	RTN	0.4161	0.6036	0.7293	0.7561
		AWQ	0.4111	0.6026	0.7400	0.7591
		GPTQ	0.3741	0.5896	0.7195	0.7584
		SeGO	**0.4756**	**0.6327**	**0.7611**	**0.8164**
Intern VL2-26B	FP16	–	0.4711	0.6771	0.7794	0.8293
	W4A16	RTN	0.3672	0.5786	0.7123	0.6695
		AWQ	0.3322	0.5373	0.7237	0.6522
		GPTQ	0.3523	0.5528	0.6871	0.6402
		SeGO	**0.4722**	**0.6667**	**0.7555**	**0.8277**

Table 1: Performance comparison of various quantization methods on the InternVL2 series under W4A16 configuration.

Model	Bitwidth	Method	MMMU	RWQA	OCR	AI2D
Intern VL2-1B	FP16	–	0.3422	0.5203	0.7400	0.6231
	W3A16	RTN	0.2742	0.3326	**0.3960**	0.4031
		AWQ	**0.2822**	**0.3386**	0.3820	0.4011
		GPTQ	0.2572	0.3266	0.3710	**0.4035**
		SeGO	0.2789	0.3273	0.3890	0.3999
Intern VL2-2B	FP16	–	0.3422	0.4183	0.7510	0.7247
	W3A16	RTN	0.2900	0.3412	0.6380	0.6637
		AWQ	**0.3200**	0.3477	0.6460	0.6658
		GPTQ	0.2800	0.3431	0.6040	0.6615
		SeGO	0.2989	**0.3634**	**0.6990**	**0.6667**
Intern VL2-8B	FP16	–	0.4778	0.6562	0.7660	0.8235
	W3A16	RTN	0.2384	0.5464	0.7140	0.7379
		AWQ	0.2444	0.5539	0.7210	0.7484
		GPTQ	0.2274	0.5494	0.7180	0.7459
		SeGO	**0.4378**	**0.6327**	**0.7430**	**0.8028**
Intern VL2-26B	FP16	–	0.4711	0.6771	0.7790	0.8293
	W3A16	RTN	0.3327	0.6119	0.7340	0.6879
		AWQ	0.3667	0.6256	0.7380	0.7032
		GPTQ	0.3357	0.6237	0.7050	0.6952
		SeGO	**0.4644**	**0.6627**	**0.7570**	**0.8196**

Table 2: Performance comparison of various quantization methods on the InternVL2 series under W3A16 configuration.

we compare the calibration time overhead of different methods.

Model	Bitwidth	Method	MMMU	RWQA	OCR	AI2D
LLaVA -onevision -7B	FP16	–	0.4956	0.6654	0.6230	0.8138
	W4A16	RTN	0.4857	0.6361	0.6110	0.7520
		AWQ	0.4826	0.6526	0.6150	**0.8093**
		GPTQ	0.4791	0.6437	0.6140	0.7630
		SeGO	**0.4922**	**0.6601**	**0.6170**	0.8070
	W3A16	RTN	0.4180	0.5654	0.4380	0.7329
		AWQ	0.4336	0.5716	0.5650	0.7735
		GPTQ	0.4235	0.5723	0.5320	0.7576
		SeGO	**0.4556**	**0.6196**	**0.5590**	**0.7895**

Table 3: Quantization performance on the LLaVA-onevision-7B model across W4A16 and W3A16 configurations.

Model	Bitwidth	Method	MMMU	RWQA	OCR	AI2D
LLaVA -v1.5 -7B	FP16	–	0.3622	0.5608	0.3150	0.5515
	W4A16	RTN	0.3498	0.5552	0.3090	0.5340
		AWQ	0.3493	0.5502	**0.3100**	0.5300
		GPTQ	0.3503	0.5525	0.3040	0.5355
		SeGO	**0.3556**	**0.5556**	0.3070	**0.5398**
	W3A16	RTN	0.2490	0.3515	0.2150	0.3461
		AWQ	0.2670	0.5066	0.2630	0.4887
		GPTQ	0.3030	0.4916	0.2710	0.4697
		SeGO	**0.3060**	**0.5085**	**0.2980**	**0.5194**

Table 4: Quantization performance on the LLaVA-v1.5-7B model across W4A16 and W3A16 configurations.

Weight-only Quantization vs Weight-Activation Joint Quantization

To validate the efficiency of SeGO as a weight-only quantization strategy, we conduct a systematic comparison between SeGO (W4A16) and the current mainstream weight-activation joint quantization method, SmoothQuant (W4A8). We record the total calibration time (including data loading, statistical computation, and parameter search) and the final average quantization accuracy for VLM models of different sizes. The comparison results are shown in the Fig. 4.

SmoothQuant requires online statistics of the activation value distribution and scaling factor search for each layer of the entire network. In contrast, SeGO only needs to optimize the parameters for the Expansion Space and avoids the dynamic quantization of activation values. Results show that SeGO achieves significant acceleration across all models and its accuracy performance under W4A16 significantly outperforms the W4A8 SmoothQuant. Specifically, SeGO achieve average accuracy improvements of +0.65%, +2.23%, and +2.00% on three models. This is primarily due to the structural sensitivity protection strategy we propose: by precisely optimizing the weight distribution of the key Expansion Space, SeGO reduces quantization errors caused by extreme outliers in VLM activation values.

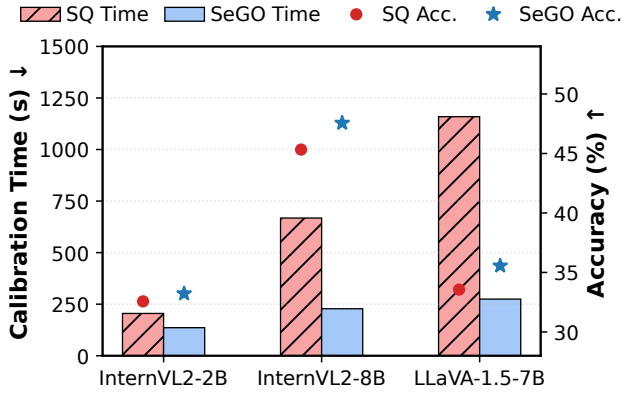

Figure 4: Efficiency and accuracy comparison between SeGO (W4A16) and SmoothQuant (W4A8) across different models.

Optimization Strategy	Gradient Required?	Calibration Time	Speedup vs. Gradient	W3A16 Accuracy (MMMU)
Grid Search	No	227.9 s	50×	41.3%
Gradient-Based	Yes	11,400 s (∼3.2 h)	1× (Baseline)	43.9%
Ours	No	135.0 s	∼85×	43.8%

Table 5: Performance and efficiency comparison of different optimization solvers for scaling factor search on InternVL2-8B.

Optimality Analysis of the Golden-Section Solver

To verify the superiority of the core solver, we design a set of rigorous controlled experiments. We compare the calibration time and quantization accuracy of three different optimization strategies on the InternVL2-8B model. The three strategies are: (1) Grid Search, (2) Gradient-Based Search and (3) our Golden-Section Solver. The comparison results are shown in the Table 5.

While grid search doesn't require gradients, its solution quality is limited by the predefined step size, often only finding suboptimal solutions at discrete points. In contrast, the golden-section solver iteratively approximates the optimal point in the continuous interval, achieving high numerical precision. Our golden-section solver reach nearly the same accuracy as the gradient-based search. It confirm that quantization error function exhibits good unimodal characteristics within a constrained local interval, allowing precise solutions without expensive gradient optimization.

Moreover, our solver's calibration time is just 1/85 that of the gradient-based method (135 seconds vs 3.2 hours). This means SeGO can perform dozens of model iterations within the same computational budget, significantly lowering the deployment threshold for VLMs on edge devices.

Datasets	Method	Expansion Space Types	W3A16 acc	W4A16 acc
OCR	Baseline PTQ	None	0.7140	0.7290
	SeGO–MLP (A)	Gate/Up	0.7310	0.7540
	SeGO–Attn (B)	Q/K/V	0.7350	0.7550
	SeGO–Full (A+B)	Q/K/V + Gate/Up	0.7430	0.7610
AI2D	Baseline PTQ	None	0.7379	0.7561
	SeGO–MLP (A)	Gate/Up	0.8015	0.8151
	SeGO–Attn (B)	Q/K/V	0.7966	0.8177
	SeGO–Full (A+B)	Q/K/V + Gate/Up	0.8028	0.8174
RWQA	Baseline PTQ	None	0.5464	0.6036
	SeGO–MLP (A)	Gate/Up	0.6145	0.6258
	SeGO–Attn (B)	Q/K/V	0.6131	0.6301
	SeGO–Full (A+B)	Q/K/V + Gate/Up	0.6327	0.6353

Table 6: Ablation study of scaling intervention on different Expansion Space for InternVL2-8B across various benchmarks.

4.4 Ablation Study

To verify whether SeGO's introduction of scaling only in the Expansion Space is truly "effective and necessary", we conduct a structural-level ablation experiment on the InternVL2-8B model. We compare scaling only MLP (SeGO-MLP), only Attention (SeGO-Attn), both (SeGO-Full) and Baseline PTQ. Results are shown in Table 6.

The results show that scaling only the Expansion Space of MLP and Attention brings improvements, confirming that the Expansion Space is more sensitive observed in earlier figure. Quantitatively, the precision recovery achieved by optimizing the Expansion Space is significantly higher than that of the Projection Space, which validates our hypothesis that the feature expansion process in Transformers is the primary source of quantization error. SeGO-Full consistently outperforms the case where scaling is applied to only a single Expansion Space across three benchmarks and two bit-widths. It indicates that the MLP and Attention Expansion Space complement each other in error propagation. Correcting one type of Expansion Space only mitigates part of the structural asymmetry, while joint processing maximizes error suppression. By concurrently stabilizing expansion layers in both MLP and Attention, SeGO preserves the consistency of semantic manifold representations.

5 Conclusion

This paper proposes SeGO, a unified structural sensitivity-aware sparse optimization framework for VLMs quantization. We reveal that the Expansion Space is the key bottleneck for quantization error. By applying scaling protection only to Expansion Space and using a Golden-Section Solver, SeGO establishes a new benchmark for VLM parameter amount-quantization accuracy-calibration efficiency trade-offs. It outperforms existing methods significantly in W3A16 configurations and provides an 85x speedup over gradient-based methods. SeGO enables the efficient deployment of large-scale VLMs at low calibration costs. Future work will explore its generalization to larger models (70B+) and other modalities.

Contribution Statement

Tianqi Zhao and Ruihan Hu contributed equally to this work. *Ruihan Hu is the corresponding author.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported in part by the National Key Research and Development Program under Grant 2023YFF0905500, in part by the National Key Research and Development Program of Heilongjiang under Grant 2022ZX01A11, in part by the Natural Science Foundation of Heilongjiang Province under Grant LH2023F001, in part by the ZTE Industry-University-Institute Cooperation Funds under Grant No. 2022H2ZTE02-01-P4, and in part by the Spring Goose Talent Team Program of Heilongjiang.

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[Chen *et al.*, 2023] 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. *arXiv preprint arXiv:2311.12793*, 2023.

[Chen *et al.*, 2024a] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26564–26574, 2024.

[Chen *et al.*, 2024b] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Igel: Interpretable geometric evaluation for low-bit quantized vision-language models. *arXiv preprint arXiv:2412.04423*, 2024.

[Chen *et al.*, 2024c] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[Dettmers *et al.*, 2022] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332, 2022.

[Frantar *et al.*, 2023] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

[Fu *et al.*, 2024] Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhu, Haozhe Zhao, Yan-Bin Luk, Yunhang Shen, Mengdan Zhang, Yu Qiao, and Peng Gao. A survey on evaluation of large multimodal models. *arXiv preprint arXiv:2307.03126*, 2024.

[Gong *et al.*, 2024] Zhuocheng Gong, Junteng Ma, Cheng Chen, Shiyi Zhu, Lingpeng Kong, and Chuan Wu. Mquant: Unleashing the inference potential of multimodal large language models via static quantization. *arXiv preprint arXiv:2405.17644*, 2024.

[Jiang *et al.*, 2024] Xiaoyan Jiang, Hang Yang, Kaiying Zhu, Xihe Qiu, Shibo Zhao, and Sifan Zhou. Ptq4ris: Post-training quantization for referring image segmentation. *arXiv preprint arXiv:2409.17020*, 2024.

[Kembhavi *et al.*, 2016] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*, pages 235–251. Springer, 2016.

[Li *et al.*, 2024a] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[Li *et al.*, 2024b] Liang Li, Qingyuan Li, Bo Zhang, and Xiangxiang Chu. Norm tweaking: High-performance low-bit quantization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18536–18544, 2024.

[Li *et al.*, 2024c] Xiaocheng Li, Yukun Yue, Zukang Xu, et al. Signround: Interpretable sign-based optimization for post-training quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5797–5806, 2024.

[Li *et al.*, 2024d] Xiaocheng Li, Yukun Yue, Zukang Xu, Yuzhang Shang, and Cheng-Lin Liu. Bi-llava: Hierarchical binary vision-language models. *arXiv preprint arXiv:2404.14444*, 2024.

[Li *et al.*, 2025] Baojia Li, Chunzhi He, Yinben Xia, et al. Star pulse network: Collaborative optimization for GPU cluster collective communication and centralized routing. *ZTE Technology*, 2025. (in Chinese).

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2024a] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *Proceedings of the 1st Conference on Machine Learning and Systems (MLSys)*, 2024.

[Lin *et al.*, 2024b] Ji Lin, Hongxu Yin, Wei Ping, Jiaqi Ma, Mohammed Shoeybi, Kai Wang, Jue Wang, Guangxuan Xiao, Haotian Liu, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] Jing Liu, Ruihao Gong, Xiuying Wei, Ning Liu, Haoyuan Huang, Jinyang Guo, Budan Wan, Xianglong Liu, and Dachao Qiao. Qllm: Accurate and efficient low-bitwidth quantization for large language models. In *International Conference on Learning Representations*, 2024.

[Liu *et al.*, 2024c] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2024.

[Liu *et al.*, 2024d] Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024.

[Luo *et al.*, 2024] Jiajun Luo, Tanghui Jia, Wenqi Shao, Mingze Gao, Lu Hou, and Ping Luo. Mbq: Modality-balanced quantization for large vision-language models. *arXiv preprint arXiv:2410.15065*, 2024.

[Shang *et al.*, 2024a] Yuzhang Shang, Ze-Quan Shang, Sifan Zhou, Siyuan Huang, and Yan Yan. Evaluating the robustness of quantized large vision-language models. *arXiv preprint arXiv:2406.11306*, 2024.

[Shang *et al.*, 2024b] Ze-Quan Shang, Yuxuan Hui, Yan Yan, Liang Liu, and Cheng-Lin Liu. Pb-llm: Partially binarized large language models. In *International Conference on Learning Representations*, 2024.

[Shao *et al.*, 2023] 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. *CoRR*, abs/2308.13137, 2023.

[Sun *et al.*, 2024] Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, et al. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*, 2024.

[Wang *et al.*, 2025] Zhiqin Wang, Jizhe Zhou, and Kaifeng Han. End-to-edge collaborative 6G endogenous AI networks. *ZTE Technology*, 31(4):29–33, 2025. (in Chinese).

[Xiao *et al.*, 2023] 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.

[Yu *et al.*, 2025] Li Yu, Jianhua Zhang, and Yichen Cai. Three enabling technologies for 6G digital twin channels: Multi-modal sensing, environment knowledge, and large models. *ZTE Technology*, 31(4):19–28, 2025. (in Chinese).

[Yuan *et al.*, 2023] Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. Rptq: Reorderbased post-training quantization for large language models. *arXiv preprint arXiv:2304.01089*, 2023.

[Yue *et al.*, 2024] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[Zhang *et al.*, 2024] 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. *arXiv preprint arXiv:2407.12772*, 2024.

[Zhao *et al.*, 2024] 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 improved quantization for large language models. In *Proceedings of Machine Learning and Systems*, volume 6, pages 196–209, 2024.

[Zhou *et al.*, 2024] Qing-Yuan Zhou, Ziyang Luo, Taolin Zhang, et al. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

[Zhou *et al.*, 2025] Sifan Zhou, Shuo Wang, Zhihang Yuan, Mingjia Shi, Yuzhang Shang, and Dawei Yang. Gsq-tuning: Group-shared exponents integer in fully quantized training for llms on-device fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22971–22988, 2025.