Stabilizing Ultra-Low-Bit Quantization of Multimodal LLMs via Global Bit Allocation

Guilin Li[1], Yuexiao Ma[1], Yue Zhang[1], Xinxiong Wu[1], Jiaqi Zhou[2], Qingheng Zhang[2], Yan Zhang[1], Fei Chao[1], Xiawu Zheng[1⋆], and Rongrong Ji[1]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, China
[2] Huawei Technologies Ltd., Nanjing, China

Abstract. Weight-only post-training quantization (PTQ) has proven highly effective for deploying large language models under memory-bound inference constraints. However, extending PTQ to Multimodal LLMs at ultra-low bit-widths (2–3 bits) presents a significant challenge. Existing uniform-precision methods suffer severe accuracy collapse due to their implicit assumption that all output channels within a layer are equally sensitive to quantization noise. This uniform bit-width assignment fails to account for the heterogeneous sensitivity of different channels. By lifting the analysis to the Transformer block level, we reveal a two-level anisotropy—inter-layer and intra-layer inter-channel. This anisotropy is further amplified by an outlier–sensitivity resonance mechanism, causing different output channels to contribute vastly different amounts of quantization error to the block output. Together, these findings provide the rigorous foundation for per-output-channel mixed-precision quantization. Building on the derived optimality condition, we propose GloBitQ, a framework that unfolds the anisotropic loss into principled bit allocation. It integrates three intermediate proxies: power-law balanced layer weighting, rank–value hybrid channel scoring, and global top-τ bit assignment. The resulting allocation provably approximates the optimum of the block-level quantization objective. On six multimodal benchmarks spanning Qwen2-VL-7B, Qwen2.5-VL-7B, and LLaVA-OneVision, GloBitQ consistently surpasses GPTQ, GPTAQ, and VLMQ: the 3-bit model performs within 2.1% of full precision(81.70 vs. 83.76), while the 2-bit model remains stable where prior methods collapse—all within a calibration-only, finetuning-free, hardware-compatible pipeline. Code is available in the supplementary material.

Keywords: Mixed-precision quantization · Differentiated calibration · Ultra-low-bit quantization

1 Introduction

Large vision–language models (VLMs) have recently demonstrated remarkable multimodal reasoning and perception capabilities, yet their immense computational and memory demands severely limit deployment on resource-constrained

⋆ Corresponding author

devices. Post-training quantization (PTQ) has emerged as a compelling compression strategy, significantly reducing memory footprint and inference latency by mapping high-precision floating-point values to low-bit formats without requiring retraining [4, 6, 7, 18–20, 24, 28, 31, 37]. Particularly suited for VLMs, weight-only quantization resolves the memory-bound challenges of inference by narrowing the gap between computational and memory bandwidth, thereby facilitating more efficient deployment. However, when the bit-width is pushed to ultra-low levels (e.g., 2–3 bits), conventional uniform-precision PTQ suffers severe accuracy collapse—a phenomenon that calls for deeper understanding of the quantization loss landscape itself [13, 22, 23, 27, 34].

We argue that the root cause of this collapse is the strongly anisotropic nature of the quantization loss. In pre-trained Transformers, a small, input-invariant subset of activation channels exhibits anomalously large magnitudes—the well-documented activation outlier phenomenon. These outliers cause the Hessian of the reconstruction loss to have a condition number far exceeding unity, so that a unit-norm weight perturbation along the most sensitive direction incurs orders of magnitude more loss than along the least sensitive direction. Uniform quantization, which assigns identical precision to all parameters, is oblivious to this anisotropy: it wastes bit budget on insensitive parameters while leaving critical ones under-protected. This motivates mixed-precision quantization, which selectively allocates higher bit-widths to the most sensitive parameters.

Existing mixed-precision methods, however, lack a principled basis for determining which parameters deserve higher precision and how much precision to assign [8, 11, 12, 16]. A key reason lies in the prevailing layer-wise reconstruction objective, which optimizes each linear layer independently. Under this objective, the quantization sensitivity of all output channels within the same layer is governed by a shared input covariance term, making them effectively indistinguishable. That is, the layer-wise objective itself offers no signal for telling apart important and unimportant output channels—the very granularity at which mixed-precision allocation must operate. This leaves existing approaches to rely on ad-hoc heuristics or secondary proxy metrics for bit allocation, often leading to suboptimal precision utilization in the ultra-low-bit regime.

Although many recent PTQ methods [15, 32, 36] have adopted block-level reconstruction objectives in practice for improved quantization quality, no prior work has formally analyzed why the block-level perspective is fundamentally different from the layer-wise one in terms of channel sensitivity differentiation. We provide this missing theoretical analysis. When quantization distortion is propagated through the full Transformer block—including attention mechanisms, layer normalization, and residual connections—the Kronecker symmetry present in the layer-wise formulation is broken, and different output channels acquire genuinely different sensitivities. Moreover, we show that activation outliers multiplicatively amplify these inter-channel sensitivity differences through a mechanism we term outlier–sensitivity resonance: the few outlier activation directions resonate with channel-dependent downstream propagation matrices, causing certain output channels to be disproportionately sensitive. This analysis reveals a

two-level anisotropy—both inter-layer (different layers have different average sensitivities) and intra-layer inter-channel (within each layer, output channels have non-uniform sensitivities)—which provides the first rigorous theoretical foundation for per-output-channel mixed-precision quantization.

Building on this analysis, we propose **GloBitQ**, a mixed-precision framework that translates the observed two-level anisotropy into a practical bit allocation strategy. Our analysis suggests that channels should be promoted to higher precision according to a composite importance score that combines block-level sensitivity with quantization range, and we further show that the mean absolute weight magnitude serves as an efficient proxy—reducing the cost from Hessian trace estimation to a simple $O(d)$ statistic while preserving the channel ranking. The two-level anisotropy is addressed by two corresponding components: a power-law balanced layer weighting that captures inter-layer sensitivity differences, and a rank–value hybrid channel scoring that captures intra-layer inter-channel differences, both governed by a single hyperparameter α that controls the fairness-versus-adaptivity trade-off across the entire model. Additionally, a learned affine smoothing is coupled with the mixed-precision quantizer during optimization: by jointly reducing the weight condition number and the activation covariance condition number, it narrows the gap between the assumptions underlying the bit allocation and the practical quantization setting, further improving the effectiveness of the derived precision map with negligible overhead.

Comprehensive experiments on representative VLMs including Qwen2-VL-7B, Qwen2.5-VL-7B, and LLaVA-OneVision demonstrate that GloBitQ consistently outperforms existing PTQ baselines. Specifically, our 3-bit configuration restricts accuracy degradation to within 2.1% of the full-precision baseline, while the 2-bit setup maintains stability where prior methods fail. These results are achieved via a calibration-only, training-free, and hardware-compatible pipeline.

Our contributions can be summarized as follows:

1. We establish the first rigorous theoretical foundation for per-output-channel mixed-precision quantization by proving a two-level anisotropy (inter-layer and intra-layer inter-channel) in the block-level loss landscape and identifying an outlier–sensitivity resonance mechanism that multiplicatively amplifies inter-channel sensitivity gaps.
2. We derive GloBitQ, a mixed-precision framework that directly translates this theoretical insight into closed-form bit allocation rules unified by a single hyperparameter, and further integrate affine smoothing as a lightweight refinement to bridge the residual gap between theoretical assumptions and practice.
3. Extensive experiments on three state-of-the-art VLMs across six multimodal benchmarks show that GloBitQ consistently outperforms GPTQ, GPTAQ, and VLMQ under a calibration-only, hardware-friendly pipeline. At 3 bits, GloBitQ stays within 2.1% of full precision while surpassing the best baseline by up to 17%. At 2 bits, where per-channel GPTQ collapses entirely, GloBitQ remains stable and outperforms even group-128 baselines.

2 Related Work

2.1 Post-Training Quantization

Post-training quantization (PTQ) compresses large models without retraining. A central challenge is activation outliers: SmoothQuant [35] addresses this by migrating quantization difficulty from activations to weights, while OmniQuant [29] introduces learnable clipping and transformation modules to jointly reshape distributions. A complementary line of work reduces distortion via reparameterization: AffineQuant [26] learns invertible affine mappings, QuaRot [1] and SpinQuant [25] apply rotation-based transforms to mitigate outlier dominance, and FlatQuant [30] proposes lightweight affine smoothing to push PTQ toward sub-4-bit precision. Despite these advances, existing methods typically apply uniform bit allocation across all layers, limiting their effectiveness under ultra-low-bit or multimodal settings where per-component sensitivity varies dramatically.

2.2 Vision Language Model Quantization

Large vision–language models (VLMs) such as LLaVA [21] and Qwen-VL [2] pose additional quantization challenges due to heterogeneous activation distributions and cross-modal feature interactions. Recent efforts extend PTQ to multimodal settings: Q-VLM [32] groups layers by activation entropy, MBQ [15] applies sensitivity-driven granularity balancing, MQuant [38] adopts modality-specific scaling and rotation suppression, and VLMQ [36] incorporates token-level importance weighting into a Hessian-based framework. However, these methods generally rely on static precision assignments and do not exploit global sensitivity patterns spanning both visual and textual modules, limiting their effectiveness under tight bit budgets.

2.3 Mixed-Precision Quantization

Mixed-precision quantization assigns varying bit widths across a model to balance accuracy and efficiency. HAWQ [8] introduces Hessian-based sensitivity analysis for precision allocation; LLM-MQ [16] extends gradient-driven estimation to protect sparse outliers; SqueezeLLM [12] combines sensitivity-aware allocation with dense-and-sparse decomposition; and SliM-LLM [11] employs salience-driven bit weighting for hardware-friendly compression. These methods remain largely layer-wise or heuristic, lacking a global view of cross-layer and cross-channel importance, which can lead to imbalanced precision assignment in complex multimodal architectures.

3 Method

3.1 Preliminary

Definition 1 (Transformer Block). *Let the b-th Transformer block be a parameterized mapping* $\mathcal{B} : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times d}$*, comprising L linear projection lay-*

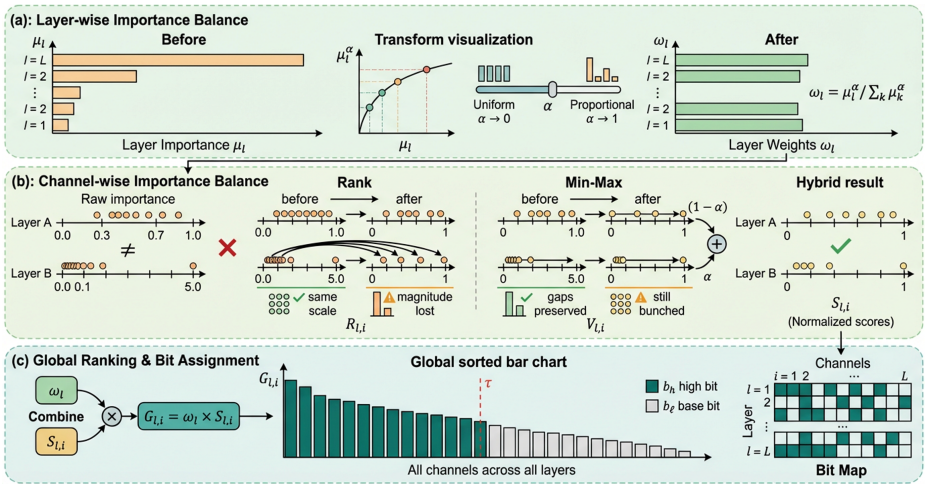

Fig. 1: Overview of the proposed GloBitQ. (a) Layer-wise balance: a concave power-law transform converts layer-average importance μ_l into normalized layer weights ω_l, with $\alpha \in (0, 1]$ interpolating between uniform and proportional allocation. (b) Channel-wise balance: a rank–value hybrid normalization produces cross-layer comparable scores $S_{l,i}$ from raw per-channel importance values. (c) Global ranking: the composite score $G_{l,i} = \omega_l \times S_{l,i}$ is sorted across all channels and thresholded at ratio τ to assign promoted (b_h) or base (b_ℓ) bit widths, yielding the final mixed-precision bit map.

ers (weights $\{W_l\}_{l=1}^{L}$*, e.g.* $W_Q, W_K, W_V, W_O, W_{\text{up}}, W_{\text{gate}}, W_{\text{down}}$*), layer normalization, attention mechanisms, nonlinear activations, and residual connections. The block receives input* $X \in \mathbb{R}^{T\times d}$ *and produces output* $Y = \mathcal{B}(X;\theta)$*, where* $\theta = \big(\text{vec}(W_1), \ldots, \text{vec}(W_L)\big) \in \mathbb{R}^P$ *with* $P = \sum_{l=1}^{L} m_l d_l$ *(*m_l*: output channels,* d_l*: input channels of layer* l*).*

Definition 2 (Block-Level Reconstruction Error). *For a joint parameter perturbation* $\Delta\theta$*, the block-level quantization error is*

$$\mathcal{E}_{\text{block}}(\Delta\theta) = \mathbb{E}_{X\sim\mathcal{D}}\Big[\big\|\mathcal{B}(X;\theta+\Delta\theta) - \mathcal{B}(X;\theta)\big\|_F^2\Big]. \tag{1}$$

Since $\mathcal{E}_{\text{block}}(\Delta\theta) \geq 0$ with equality at $\Delta\theta = 0$, the gradient $\nabla_{\Delta\theta}\mathcal{E}_{\text{block}}$ necessarily vanishes at the origin. Applying a first-order Taylor expansion to the block mapping itself, $\mathcal{B}(X;\theta+\Delta\theta) \approx \mathcal{B}(X;\theta) + J(X;\theta)\,\Delta\theta$, where $J \in \mathbb{R}^{Td\times P}$ is the Jacobian of $\text{vec}(\mathcal{B})$ with respect to θ, yields the Gauss–Newton approximation

$$\mathcal{E}_{\text{block}}(\Delta\theta) \;\approx\; \Delta\theta^\top \mathbf{H}_{\text{block}}\,\Delta\theta, \qquad \mathbf{H}_{\text{block}} \;\triangleq\; \mathbb{E}_X\big[J^\top J\big] \;\in\; \mathbb{R}^{P\times P}. \tag{2}$$

Within each layer l, the per-layer contribution to $\mathbf{H}_{\text{block}}$ is governed by the input second-moment matrix $H_l \triangleq \mathbb{E}[X_l^\top X_l] \in \mathbb{R}^{d_l\times d_l}$, which determines how weight perturbations in layer l propagate into output distortion.

Empirical evidence [6, 35] shows that LLM activations exhibit anomalously large magnitudes in a small, input-invariant subset of channels. We formalize this observation below.

Assumption 1 (Activation Outlier Channels)**.** *For each layer l, let $\mathbf{x}_l^{(j)} \in \mathbb{R}^T$ denote the j-th column of X_l. There exists an index set $\mathcal{S}_l \subset \{1, \ldots, d_l\}$ with $|\mathcal{S}_l| = k_l \ll d_l$, and positive constants $\alpha_l, \beta_l > 0$, such that*

$$\forall\, j \in \mathcal{S}_l : \quad [H_l]_{jj} = \|\mathbf{x}_l^{(j)}\|^2 \geq \alpha_l, \tag{3}$$

$$\forall\, j \in \mathcal{S}_l^c : \quad [H_l]_{jj} = \|\mathbf{x}_l^{(j)}\|^2 \leq \beta_l, \tag{4}$$

with outlier ratio $\gamma_l \triangleq \alpha_l/\beta_l \gg 1$.

From activation outliers to loss anisotropy. The outlier assumption directly implies severe anisotropy of H_l. By the Rayleigh quotient, $\lambda_{\min}(H_l) \leq [H_l]_{jj} \leq \lambda_{\max}(H_l)$; choosing outlier and normal channels respectively yields

$$\kappa(H_l) = \frac{\lambda_{\max}(H_l)}{\lambda_{\min}(H_l)} \geq \frac{\alpha_l}{\beta_l} = \gamma_l \gg 1. \tag{5}$$

Under uniform quantization with step size δ, the expected loss $\mathbb{E}[\mathcal{L}] \approx \frac{\delta^2}{12}\operatorname{tr}(H_l)$ is dominated by the outlier channels ($\sum_{j\in\mathcal{S}_l}[H_l]_{jj} \geq k_l\alpha_l \gg (d_l - k_l)\beta_l \geq \sum_{j\notin\mathcal{S}_l}[H_l]_{jj}$), showing the suboptimality of uniform precision.

This establishes input-dimension anisotropy within each layer but does not yet differentiate between output channels. We next show that block-level reconstruction $\mathcal{E}_{\text{block}}$ (Eq. 1) reveals a richer, practically actionable anisotropy across output channels.

3.2 Block-Level Output Channel Anisotropy

The input-dimension anisotropy in §3.1 explains why certain input channels are sensitive to quantization noise but does not distinguish output channels within the same layer. We now lift the analysis to the block level and show that a richer anisotropy emerges across output channels, founding per-output-channel mixed-precision quantization. Consider a Transformer block $\mathcal{B} : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times d}$ with L linear layers, where layer l has weight $W_l \in \mathbb{R}^{m_l \times d_l}$ and input $X_l \in \mathbb{R}^{T\times d_l}$.

Layer-wise metrics treat all output channels equally. The layer-wise reconstruction error $\mathcal{E}_l^{\text{layer}} = \mathbb{E}_X[\|X_l \Delta W_l^\top\|_F^2]$ assigns each output channel i the contribution $\mathbb{E}[\Delta w_{l,i}^\top H_l\, \Delta w_{l,i}]$ with the same Hessian $H_l = \mathbb{E}[X_l^\top X_l]$. Hence all output channels share identical quantization sensitivity, and no per-channel precision assignment is justified. This symmetry is inherent: layer-wise analysis cannot capture how downstream computation amplifies one channel's error more than another's.

Block-level analysis breaks the symmetry. Practical per-channel quantizers process each output channel independently: for $i \neq j$, $\Delta w_{l,i}$ and $\Delta w_{l,j}$ are statistically independent with $\mathbb{E}[\Delta w_{l,i}] = \mathbf{0}$. Under this condition and the Gauss–Newton approximation (2), the expected block-level error of layer l decomposes as

$$\mathbb{E}\big[\mathcal{E}_l^{\text{block}}\big] = \sum_{i=1}^{m_l} \operatorname{tr}\Big(H_{l,i}^{\text{block}}\, \mathbb{E}\big[\Delta w_{l,i} \Delta w_{l,i}^\top\big]\Big), \tag{6}$$

where the channel-specific Hessian is $H_{l,i}^{\text{block}} \triangleq X_l^\top [\Phi_l]_{ii} X_l \in \mathbb{R}^{d_l \times d_l}$. Here $[\Phi_l]_{ii} \in \mathbb{R}^{T \times T}$ is the (i,i)-th diagonal block of $\Phi_l = M_l^\top M_l \in \mathbb{R}^{Tm_l \times Tm_l}$, with $M_l \in \mathbb{R}^{Td \times Tm_l}$ the Jacobian of $\operatorname{vec}(\Delta Z)$ with respect to $\operatorname{vec}(\Delta Y_l)$. The matrix $[\Phi_l]_{ii}$ captures how downstream modules propagate the error of output channel i to the block output. Because $[\Phi_l]_{ii}$ varies with i, each channel carries a distinct effective sensitivity

$$\sigma_{l,i}^{\text{block}} = \operatorname{tr}(H_{l,i}^{\text{block}}) = \operatorname{tr}\big([\Phi_l]_{ii} X_l X_l^\top\big). \tag{7}$$

Symmetry would require $[\Phi_l]_{ii}$ identical for all i, which is generically violated in Transformers because (a) head-dependent attention patterns induce distinct propagation paths for value-projection channels; (b) element-wise gating (e.g. SwiGLU) creates channel-dependent Jacobian entries; and (c) data-dependent layer normalization introduces channel-varying inter-channel coupling. We verify empirically in §4 that $\max_i \sigma_{l,i}^{\text{block}} / \min_i \sigma_{l,i}^{\text{block}}$ is consistently large.

Activation outliers amplify the anisotropy. The channel-dependent $[\Phi_l]_{ii}$ alone break symmetry; the outlier structure (Assumption 1) further amplifies the disparity. Under isotropic activations ($X_l X_l^\top = c\, I_T$),

$$R_l^{\text{iso}} = \frac{\max_i \ \sigma_{l,i}^{\text{iso}}}{\min_i \ \sigma_{l,i}^{\text{iso}}} = \frac{\max_i \ \operatorname{tr}([\Phi_l]_{ii})}{\min_i \ \operatorname{tr}([\Phi_l]_{ii})}. \tag{8}$$

Under realistic activations with outlier channels, the anisotropic $X_l X_l^\top$ interacts with the channel-varying $[\Phi_l]_{ii}$ to yield a compounded ratio

$$R_l^{\text{out}} = \kappa_\Phi \cdot R_l^{\text{iso}}, \qquad \kappa_\Phi \gg 1, \tag{9}$$

where κ_Φ reflects the alignment between dominant eigenvectors of $X_l X_l^\top$ and the per-channel propagation matrices. Thus the outlier-induced input anisotropy and the architecture-induced propagation anisotropy multiply, producing cross-channel sensitivity variation far exceeding either source alone.

Two-level anisotropy: the combined picture. At the inter-layer level, the layer-averaged sensitivities $\bar{\sigma}_l \triangleq \frac{1}{m_l} \sum_{i=1}^{m_l} \sigma_{l,i}^{\text{block}}$ are non-uniform across layers because Φ_l and activation statistics differ layer by layer. At the intra-layer level, individual $\sigma_{l,i}^{\text{block}}$ vary substantially within each layer, further amplified by activation outliers. This two-level anisotropy provides the theoretical foundation for per-output-channel mixed-precision quantization; a principled bit allocation must account for both levels simultaneously, as addressed in §3.3.

3.3 Global Importance-Balanced Bit Allocation

The analysis in §3.2 establishes why non-uniform bit allocation across output channels is necessary. We now turn to the question of how to allocate bits, deriving a concrete and computationally efficient scheme grounded in the block-level sensitivity framework.

Sensitivity-driven allocation criterion. Consider the practical setting where each output channel can be assigned either a base bit-width b_ℓ or a promoted width $b_h = b_\ell + 1$, and the per-channel promotion cost (in storage bits) is uniform. Under the block-level error decomposition (6) with independent per-channel quantization, promoting a channel reduces the block error by an amount proportional to the product of its sensitivity and the squared quantization range, $\sigma_{l,i}^{\text{block}} \cdot R_{l,i}^2$. Therefore, to minimize the total block error under a fixed bit budget, the channels promoted to b_h should simply be those with the largest values of $\sigma_{l,i}^{\text{block}} \cdot R_{l,i}^2$.

When layers have different input dimensions d_l, the storage cost of promoting output channel (l, i) becomes d_l bits rather than a constant; in that case the ranking criterion should be adjusted to $\sigma_{l,i}^{\text{block}} \cdot R_{l,i}^2 / d_l$. In standard Transformer architectures, however, most projection layers share $d_l = d$ (the model dimension), making the uniform-cost formulation adequate for our purposes.

Weight magnitude as an importance proxy. Computing the block-level Hessian trace for every output channel is expensive. We instead use the per-channel mean weight magnitude $I_{l,i} = \frac{1}{d_l} \sum_{j=1}^{d_l} |w_{l,i,j}|$ as a proxy. Empirically, the Spearman rank correlation between $\sigma_{l,i}^{\text{block}}$ and $I_{l,i}$ exceeds 0.8 across all layers and models tested (§4). Theoretically, this is consistent with the implicit-regularization view that gradient-based training distributes weight magnitude in proportion to functional importance, though a formal proof for general Transformers remains open. Under this sensitivity–magnitude correlation and mild sub-Gaussian assumptions on the weight distribution, $I_{l,i}$ serves as an approximately order-preserving proxy for $\sigma_{l,i}^{\text{block}} \cdot R_{l,i}^2$ within each layer, sufficient for the ranking-based bit assignment derived below. With this proxy in hand, we now construct the three components of our bit allocation scheme.

(a) Layer-wise importance balance. As shown in §3.2, different layers exhibit different average sensitivities $\bar{\sigma}_l$, which can be approximated by the layer-average importance $\mu_l = \frac{1}{m_l} \sum_{i=1}^{m_l} I_{l,i}$. Motivated by classical rate–distortion theory—which prescribes a sub-linear (logarithmic) dependence of optimal bit allocation on source variance for independent Gaussian sources—we adopt the power-law family as a flexible concave parameterization of the layer weights:

$$\omega_l = \frac{{\mu_l}^\alpha}{\sum_{k=1}^{L} {\mu_k}^\alpha}, \qquad \alpha \in (0, 1]. \tag{10}$$

When $\alpha \to 0$, all layers receive equal weight ($\omega_l \to 1/L$, ignoring sensitivity differences); when $\alpha = 1$, the allocation becomes strictly proportional to layer

sensitivity ($\omega_l \propto \mu_l$). The concavity of $t \mapsto t^\alpha$ for $\alpha < 1$ ensures diminishing returns from concentrating bits on high-sensitivity layers, in line with the qualitative prescription of rate–distortion theory.

(b) Channel-wise importance balance. Within layer l, the channel sensitivities $\sigma_{l,i}^{\text{block}}$ vary with i and are approximated by $I_{l,i}$. Because the distributional shapes of $\{I_{l,i}\}$ can differ drastically across layers—some near-uniform, some heavy-tailed—directly using raw importance values for cross-layer ranking would introduce systematic bias. We resolve this via a rank–value hybrid normalization:

$$S_{l,i} = (1-\alpha)\,\frac{\text{rank}(I_{l,i})}{n_l} + \alpha\,\frac{I_{l,i} - I_l^{\min}}{I_l^{\max} - I_l^{\min}}, \tag{11}$$

where n_l is the number of channels in layer l, and $I_l^{\min}, I_l^{\max}$ are the layer-wise extremes. The rank component $\text{rank}(I_{l,i})/n_l$ is equivalent to the empirical CDF transform, providing a distribution-free normalization robust to magnitude outliers. The value component preserves absolute magnitude proportionality, allowing highly sensitive channels to receive commensurately higher scores.

The mixture coefficient α embodies a bias–variance trade-off: the rank component is a low-variance, high-bias estimator (discards magnitude information but is robust to proxy noise), while the value component is a low-bias, high-variance estimator (retains magnitude information but is more noise-sensitive). We reuse the same α as in Eq. (10) as a pragmatic simplification that reduces the hyperparameter space to a single scalar; the theoretical framework does not require this coupling, and independent tuning could yield marginal improvements at the cost of a two-dimensional search.

(c) Global ranking and bit assignment. Following the sensitivity-driven criterion derived above, optimal bit allocation ranks channels by $\sigma_{l,i}^{\text{block}} \cdot R_{l,i}^2$. We approximate this target quantity through a multiplicative decomposition into a layer-level factor and a within-layer factor:

$$\sigma_{l,i}^{\text{block}} \cdot R_{l,i}^2 \;\approx\; \underbrace{f_{\text{inter}}(\mu_l)}_{\text{approximated by } \omega_l(\alpha)} \cdot \underbrace{g_{\text{intra}}(I_{l,i})}_{\text{approximated by } S_{l,i}(\alpha)}, \tag{12}$$

where the approximations rely on the sensitivity–magnitude correlation, the power-law layer weighting (10), and the rank–value normalization (11). Combining both factors yields the unified global importance score

$$G_{l,i}(\alpha) = S_{l,i}(\alpha) \cdot \omega_l(\alpha), \tag{13}$$

and bit widths are assigned via a top-τ rule:

$$b_{l,i} = \begin{cases} b_h, & \text{if } G_{l,i} \in \text{Top}_\tau(G), \\ b_\ell, & \text{otherwise}, \end{cases} \tag{14}$$

where τ is the fraction of channels promoted to the higher bit-width. While each step in the approximation chain (12) introduces some error, only the rank

order of the channel scores needs to be preserved for the top-τ assignment to match the theoretically optimal ranking. We empirically validate in §4 that the resulting bit maps closely track the rankings obtained from exact Hessian-trace estimation. The bit map is computed once and fixed before the subsequent affine optimization stage.

3.4 Affine Smoothing as Approximation Refinement

The bit allocation in §3.3 relies on two approximations: i.i.d. quantization noise across input dimensions, and weight magnitude as a sensitivity proxy (Assumption 3.3). A learned per-layer invertible affine transformation can reduce the effective errors from both sources, serving as a lightweight refinement for the derived bit allocation rather than an independent module.

In practice, we adopt a two-stage procedure: first compute the bit map $\{b_{l,i}\}$ from the original weights and fix it, then optimize the affine transforms with the mixed-precision quantizer Q_{mix} inside the optimization loop. This couples the affine alignment with the mixed-precision structure while avoiding the complexity of iterative alternation between bit-map computation and affine optimization.

4 Experiments

4.1 Experimental Setup

Models and Benchmarks. We evaluate the proposed GloBitQ framework on three representative large-scale vision–language models (VLMs): Qwen2-VL-7B-Instruct [33], Qwen2.5-VL-7B-Instruct [3], and LLaVA-OneVision [14]. Six widely used multimodal benchmarks are adopted to cover diverse visual and textual reasoning abilities, including *ChartQA*, *DocVQA*, *OCRBench*, *ScienceQA*, *Seed-Bench2Plus*, and *TextVQA*. All results are obtained following the official evaluation protocols of each dataset and computed using the open-source VLMEvalKit [9], which provides a unified and reproducible benchmarking interface for multimodal models.

Baselines and Implementation Details. We compare GloBitQ with a set of PTQ quantization baselines: GPTQ [10], an efficient layer-wise quantization solver; GPTAQ [17], a Hessian-aware extension of GPTQ; and VLMQ [36], a recent ultra-low-bit PTQ framework tailored to multimodal models. All baseline results are reported under the same evaluation settings and quantization granularity (per-channel or group-wise) for a fair comparison, following the configurations provided in VLMQ. Unless otherwise specified, GloBitQ performs mixed-precision quantization with bit widths drawn from $\{2, 3, 4\}$ and optimized via the solver described in Eq. (14). Affine Smoothing optimization is conducted for 15 epochs with a learning rate of 5×10^{-3}, using instruction-tuning data sampled from the COCO-based subset of ShareGPT4V [5]. The corresponding images are loaded from the COCO dataset folder, where 512 samples are randomly selected for calibration. For the balance factor α, we set 0.7 for both

Table 1: Quantitative comparison of different 3-bit quantization methods on six benchmarks. "Full Precision" denotes the original model without quantization. "AVG" is the average score across all datasets.

Model	Granularity	Bit	ChartQA	DocVQA	OCRBench	ScienceQA	Seed2Plus	TextVQA	AVG
Qwen2-VL-7B-Instruct-Int3									
Full Precision	-	-	83.16	93.8	86.5	85.7	69.1	84.3	83.76
GPTQ	Per-Channel	3	58.84	74.55	61.50	65.13	50.37	73.49	63.98
GPTAQ	Per-Channel	3	58.76	74.03	61.10	64.25	52.17	73.18	63.91
VLMQ	Per-Channel	3	59.44	76.59	61.90	63.05	52.92	74.54	64.74
GloBitQ	Per-Channel	3.01	79.28	92.82	82.70	84.33	67.32	83.76	**81.70**
Qwen2.5-VL-7B-Instruct-Int3									
Full Precision	-	-	86.32	94.91	88.4	72.68	69.43	85.31	82.84
GPTQ	Per-Channel	3	62.44	89.33	77.70	79.39	66.58	78.38	75.63
GPTAQ	Per-Channel	3	63.68	90.41	79.60	80.52	66.40	77.94	76.42
VLMQ	Per-Channel	3	65.32	91.36	80.00	81.70	67.32	79.28	77.49
GloBitQ	Per-Channel	3.01	82.0	93.77	86.3	59.34	64.82	84.08	**78.3**
LLaVA-OneVision-Int3									
Full Precision	-	-	80.12	87.05	62.3	95.33	65.43	75.93	77.69
GPTQ	Per-Channel	3	74.24	77.82	56.60	85.26	60.65	71.90	71.07
GPTAQ	Per-Channel	3	75.72	78.07	58.30	85.97	61.84	72.69	72.09
VLMQ	Per-Channel	3	74.56	78.21	59.20	86.37	61.05	73.66	72.17
GloBitQ	Per-Channel	3.01	78.52	85.17	62.10	94.69	64.86	74.65	**76.66**

Qwen2-VL-7B and *Qwen2.5-VL-7B*, and 0.5 for *LLaVA-OneVision*, which balances global and local smoothness. All experiments are executed on NVIDIA A800 GPUs (80 GB). For mixed-bit configurations, "3.01-bit" corresponds to a composition of 99% 3-bit and 1% 4-bit channels, while "2.01-bit" denotes 99% 2-bit and 1% 3-bit precision, maintaining overall compactness under the global bit-budget constraint.

3-bit Quantization. Tab. 1 reports the quantitative results of various 3-bit post-training quantization methods across six benchmarks. Overall, GloBitQ consistently achieves the highest performance on all tested models, clearly surpassing conventional per-channel quantization baselines such as GPTQ, GPTAQ, and VLMQ. On *Qwen2-VL-7B-Instruct-Int3*, our method attains an average score of **81.70**, which corresponds to an improvement of +**17.7 %** over GPTQ and +**16.96 %** over VLMQ, while reducing the performance drop relative to the full-precision model to only **2.1 %** (81.70 vs. 83.76). A similar trend is observed on Qwen2.5-VL-7B-Instruct-Int3, where GloBitQ reaches an average of **78.3**, outperforming all baselines by a wide margin and demonstrating stable generalization across both document-centric (DocVQA, OCRBench) and reasoning-intensive (ScienceQA, Seed2Plus) tasks. Finally, on *LLaVA-OneVision-Int3*, GloBitQ yields an average score of **76.66**, improving upon GPTQ and GPTAQ by roughly +**5.6 %** and +**4.5 %** , respectively, and recovering more than **70%** of the accuracy loss induced by standard 3-bit quantization. This performance margin confirms that globally balanced precision allocation effectively redistributes bit widths to the most informative channels,

Table 2: Quantitative comparison of different 2-bit quantization methods on six benchmarks. "Full Precision" denotes the original model without quantization. "NAN" denotes model collapse with meaningless outputs in this configuration. "AVG" is the average score across all datasets.

Model	Granularity	Bit	ChartQA	DocVQA	OCRBench	ScienceQA	Seed2Plus	TextVQA	AVG
Qwen2-VL-7B-Instruct-Int2									
Full Precision	-	-	83.16	93.8	86.5	85.7	69.1	84.3	83.7
GPTQ	Group-128	2	56.44	74.90	62.60	50.37	51.78	71.93	61.3
GPTAQ	Group-128	2	56.08	72.57	61.30	59.70	51.25	67.98	61.4
VLMQ	Group-128	2	55.32	75.76	62.50	62.32	53.80	74.80	64.0
GPTQ	Per-Channel	2.01	-	-	-	-	-	-	NAN
GloBitQ	Per-Channel	2.01	59.44	79.15	66.6	68.22	53.22	73.35	**66.6**
Qwen2.5-VL-7B-Instruct-Int2									
Full Precision	-	-	86.32	94.91	88.4	72.68	69.43	85.31	82.8
GPTQ	Group-128	2	57.72	78.79	70.80	55.77	48.40	74.36	64.3
GPTAQ	Group-128	2	53.48	74.82	67.70	2.95	17.83	73.79	48.4
VLMQ	Group-128	2	59.68	73.20	69.30	62.72	57.27	65.88	64.6
GPTQ	Per-Channel	2.01	-	-	-	-	-	-	NAN
GloBitQ	Per-Channel	2.01	64.8	81.17	74.1	48.78	54.89	74.69	**66.4**
LLaVA-OneVision-Int2									
Full Precision	-	-	80.12	87.05	62.3	95.33	65.43	75.93	77.6
GPTQ	Group-128	2	61.08	62.62	48.60	67.88	51.25	67.29	59.7
GPTAQ	Group-128	2	62.12	62.81	49.90	66.12	50.94	67.35	59.8
VLMQ	Group-128	2	62.76	64.82	51.40	69.04	53.32	68.20	61.5
GPTQ	Per-Channel	2.01	-	-	-	-	-	-	NAN
GloBitQ	Per-Channel	2.01	59.48	64.06	52.1	73.72	53.53	61.94	60.8
GloBitQ	Per-Channel	2.05	62.63	68.03	52.7	77.29	55.59	62.83	**63.1**

yielding more expressive mixed-precision configurations under the same bit budget.

2-bit Quantization. Pushing quantization into the 2-bit regime exposes the fundamental limitations of existing post-training methods. At such extreme precision, quantization noise dominates the representational space, leading most frameworks to diverge during calibration or inference. This phenomenon is evident in Tab. 2, where conventional approaches such as GPTQ and GPTAQ frequently produce invalid outputs—shown as "NAN" entries—once the bit width falls below three. Group-wise variants can remain numerically stable but suffer severe accuracy degradation across all benchmarks. In contrast, GloBitQ sustains coherent behavior and delivers competitive results under every tested setting. For *Qwen2-VL-7B-Instruct-Int2*, our method attains an average score of **66.6 %**, outperforming VLMQ by +**2.6 %** and surpassing GPTQ by over +**5 %**. The robustness persists on *Qwen2.5-VL-7B-Instruct-Int2*, where GloBitQ reaches **66.4 %**, markedly higher than group-level baselines (+**2.1**–**18 %** gain) even when competing methods partially collapse on reasoning-focused datasets such as ScienceQA and Seed2Plus. Such cross-model stability reveals that globally balanced precision reallocation mitigates the sensitivity spikes that typically cripple low-bit quantizers.

Table 3: Ablation study on 2-bit quantization of **Qwen2-VL-7B-Instruct**. GloBitQ achieves a +15.5% gain over GPTQ and FlatQuant with less than 0.1% extra memory, verifying the effectiveness of global importance–balanced mixed-precision allocation.

Method	Granularity	Bit	ChartQA	SeedBench 2 Plus	DocVQA	OCRBench	ScienceQA	TextVQA	AVG
GPTQ	Group-128	2	52.64	58.04	56.9	54.73	40.79	68.31	53.5
FlatQuant	Per-Channel	2	56.48	58.39	62.9	52.40	46.20	59.47	51.1
GloBitQ	Per-Channel	2.01	59.44	79.15	66.6	68.22	53.22	73.35	66.6

The same trend is observed on *LLaVA-OneVision-Int2*. While group-wise methods plateau around 60%, GloBitQ already achieves **60.8 %** under a strict 2.01-bit configuration. When slightly extending the fractional precision to 2.05 bits, performance further climbs to **63.1 %**, showing significant recovery—particularly on ScienceQA (+7.6 %). This moderate bit interpolation demonstrates that continuous precision adjustment can provide a smooth transition between compactness and fidelity. Such stability at 2-bit precision illustrates the effectiveness of mix-precision with affine smoothing in flattening the quantization landscape and improving numerical conditioning.

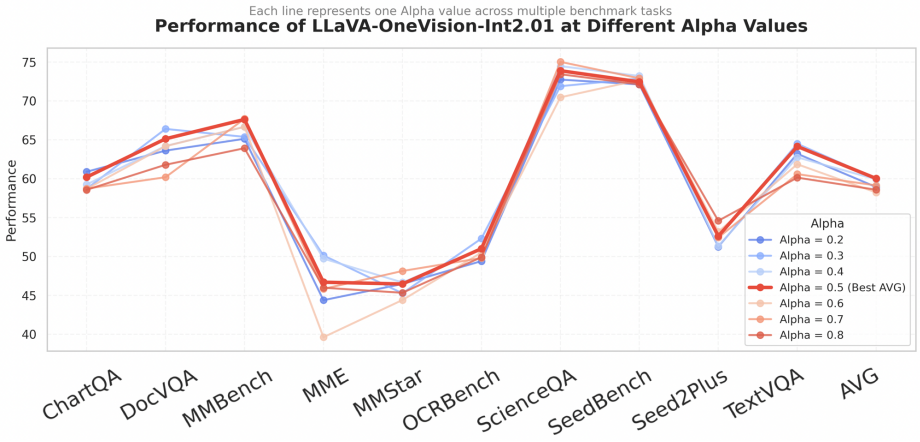

Fig. 2: Ablation on the balance factor α in LLaVA-OneVision-Int2.01. A moderate value ($\alpha = 0.5$) achieves the best trade-off between global and local precision allocation, confirming the importance of balanced sensitivity weighting for stable ultra-low-bit quantization.

Ablation on the Balance Factor α To further investigate the effect of the global–local precision trade-off in GloBitQ, we conduct an ablation study on the balance factor α using *LLaVA-OneVision-Int2.01* as a representative case. As shown in Fig. 2, α controls the extent to which the global mixed-precision allocator emphasizes channel-wise sensitivity when distributing bits across linear layers

within each transformer block. Smaller values of α (*e.g.*, 0.2–0.3) encourage nearly uniform bit allocation, resulting in stable but sub-optimal precision utilization, whereas higher values (*e.g.*, 0.7–0.8) over-emphasize sensitive channels and may harm generalization on certain tasks such as *MME* and *TextVQA*. A moderate setting of $\alpha = 0.5$ achieves the best overall trade-off, yielding the highest average performance across nine benchmarks. This tendency highlights that the proposed global–local balance mechanism is key to maintaining adaptive yet stable quantization behavior in the ultra-low-bit regime.

Ablation Study on the Effect of Global Mixed Precision. To investigate the contribution of each component in our proposed framework, we conduct an ablation study focusing on the role of global importance–balanced mixed-precision allocation under ultra-low-bit quantization. Tab. 3 summarizes the results on Qwen2-VL-7B across six multimodal benchmarks. For fair comparison, all GPTQ results are reproduced without using the act-reordering option. As seen in the table, directly applying FlatQuant at 2-bit precision without any mixed-precision extension leads to substantial performance degradation, particularly in complex reasoning- and OCR-related tasks (ScienceQA, SeedBench 2 Plus, and DocVQA). Although it performs slightly better than GPTQ on average, its uniform bit assignment cannot effectively preserve information within sensitive channels. When the proposed GloBitQ mechanism is introduced—even with a marginal bit increase to 2.01 bits—the model achieves a remarkable +15.5 % gain in overall accuracy, while incurring less than 0.1% additional memory cost. These results clearly demonstrate that the global importance–balanced precision allocation in GloBitQ enables targeted protection of critical weight channels and layers, significantly improving quantization robustness and representational fidelity in the few-bit regime.

5 Conclusion

We presented GloBitQ, a mixed-precision framework grounded in a two-level anisotropy analysis—inter-layer and intra-layer inter-channel—of the block-level quantization loss. We find that layer-wise reconstruction cannot distinguish output-channel sensitivities, and that block-level error propagation coupled with outlier–sensitivity resonance breaks this symmetry, providing the first rigorous basis for per-output-channel mixed-precision quantization. GloBitQ translates the resulting optimality condition into three practical intermediate proxies—power-law layer weighting, rank–value hybrid channel scoring, and global top-τ bit assignment—yielding an allocation scheme that provably approximates the optimum of the block-level quantization objective at negligible computational cost. Experiments on multiple VLMs show that GloBitQ consistently surpasses existing baselines, preserving near-full-precision accuracy at 3 bits and remaining stable at 2 bits where prior methods collapse.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2025YFE0113500), the National Natural Science Foundation of China (No. 62576299), the Fundamental Research Funds for the Central Universities and Xiaomi Young Talents Program.

References

1. Ashkboos, S., Mohtashami, A., Croci, M.L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., Hensman, J.: Quarot: Outlier-free 4-bit inference in rotated llms. Advances in Neural Information Processing Systems **37**, 100213–100240 (2024)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chee, J., Cai, Y., Kuleshov, V., De Sa, C.M.: Quip: 2-bit quantization of large language models with guarantees. Advances in Neural Information Processing Systems **36**, 4396–4429 (2023)
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
6. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. Advances in neural information processing systems **35**, 30318–30332 (2022)
7. Dettmers, T., Svirschevski, R., Egiazarian, V., Kuznedelev, D., Frantar, E., Ashkboos, S., Borzunov, A., Hoefler, T., Alistarh, D.: Spqr: A sparse-quantized representation for near-lossless llm weight compression. arXiv preprint arXiv:2306.03078 (2023)
8. Dong, Z., Yao, Z., Gholami, A., Mahoney, M.W., Keutzer, K.: Hawq: Hessian aware quantization of neural networks with mixed-precision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 293–302 (2019)
9. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11198–11201 (2024)
10. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022)
11. Huang, W., Qin, H., Liu, Y., Li, Y., Liu, Q., Liu, X., Benini, L., Magno, M., Zhang, S., Qi, X.: Slim-llm: Salience-driven mixed-precision quantization for large language models. arXiv preprint arXiv:2405.14917 (2024)
12. Kim, S., Hooper, C., Gholami, A., Dong, Z., Li, X., Shen, S., Mahoney, M.W., Keutzer, K.: Squeezellm: Dense-and-sparse quantization. arXiv preprint arXiv:2306.07629 (2023)
13. Lee, J.H., Kim, J., Kwon, S.J., Lee, D.: Flexround: Learnable rounding based on element-wise division for post-training quantization. In: International Conference on Machine Learning. pp. 18913–18939. PMLR (2023)

14. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
15. Li, S., Hu, Y., Ning, X., Liu, X., Hong, K., Jia, X., Li, X., Yan, Y., Ran, P., Dai, G., et al.: Mbq: Modality-balanced quantization for large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4167–4177 (2025)
16. Li, S., Ning, X., Hong, K., Liu, T., Wang, L., Li, X., Zhong, K., Dai, G., Yang, H., Wang, Y.: Llm-mq: Mixed-precision quantization for efficient llm deployment. In: The Efficient Natural Language and Speech Processing Workshop with NeurIPS. vol. 9, p. 3 (2023)
17. Li, Y., Yin, R., Lee, D., Xiao, S., Panda, P.: Gptaq: Efficient finetuning-free quantization for asymmetric calibration. arXiv preprint arXiv:2504.02692 (2025)
18. Li, Z., Yang, T., Wang, P., Cheng, J.: Q-vit: Fully differentiable quantization for vision transformer. arXiv preprint arXiv:2201.07703 (2022)
19. Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.M., Wang, W.C., Xiao, G., Dang, X., Gan, C., Han, S.: Awq: Activation-aware weight quantization for on-device llm compression and acceleration. Proceedings of machine learning and systems **6**, 87–100 (2024)
20. Lin, Y., Zhang, T., Sun, P., Li, Z., Zhou, S.: Fq-vit: Post-training quantization for fully quantized vision transformer. arXiv preprint arXiv:2111.13824 (2021)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
22. Liu, J., Gong, R., Wei, X., Dong, Z., Cai, J., Zhuang, B.: Qllm: Accurate and efficient low-bitwidth quantization for large language models. arXiv preprint arXiv:2310.08041 (2023)
23. Liu, S.Y., Liu, Z., Cheng, K.T.: Oscillation-free quantization for low-bit vision transformers. In: International conference on machine learning. pp. 21813–21824. PMLR (2023)
24. Liu, Y., Yang, H., Dong, Z., Keutzer, K., Du, L., Zhang, S.: Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20321–20330 (2023)
25. Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., Blankevoort, T.: Spinquant: Llm quantization with learned rotations. arXiv preprint arXiv:2405.16406 (2024)
26. Ma, Y., Li, H., Zheng, X., Ling, F., Xiao, X., Wang, R., Wen, S., Chao, F., Ji, R.: Affinequant: Affine transformation quantization for large language models. arXiv preprint arXiv:2403.12544 (2024)
27. Moon, J., Kim, D., Cheon, J., Ham, B.: Instance-aware group quantization for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16132–16141 (2024)
28. Ramachandran, A., Kundu, S., Krishna, T.: Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In: European Conference on Computer Vision. pp. 307–325. Springer (2024)
29. Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., Luo, P.: Omniquant: Omnidirectionally calibrated quantization for large language models. arXiv preprint arXiv:2308.13137 (2023)
30. Sun, Y., Liu, R., Bai, H., Bao, H., Zhao, K., Li, Y., Hu, J., Yu, X., Hou, L., Yuan, C., et al.: Flatquant: Flatness matters for llm quantization. arXiv preprint arXiv:2410.09426 (2024)

31. Tseng, A., Chee, J., Sun, Q., Kuleshov, V., De Sa, C.: Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. arXiv preprint arXiv:2402.04396 (2024)
32. Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., Lu, J.: Q-vlm: Post-training quantization for large vision-language models. Advances in Neural Information Processing Systems **37**, 114553–114573 (2024)
33. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
34. Wu, Z., Chen, J., Zhong, H., Huang, D., Wang, Y.: Adalog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In: European Conference on Computer Vision. pp. 411–427. Springer (2024)
35. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: Smoothquant: Accurate and efficient post-training quantization for large language models. In: International conference on machine learning. pp. 38087–38099. PMLR (2023)
36. Xue, Y., Huang, Y., Shao, J., Zhang, J.: Vlmq: Efficient post-training quantization for large vision-language models via hessian augmentation. arXiv preprint arXiv:2508.03351 (2025)
37. Yao, Z., Yazdani Aminabadi, R., Zhang, M., Wu, X., Li, C., He, Y.: Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. Advances in neural information processing systems **35**, 27168–27183 (2022)
38. Yu, J., Zhou, S., Yang, D., Wang, S., Li, S., Hu, X., Xu, C., Xu, Z., Shu, C., Yuan, Z.: Mquant: Unleashing the inference potential of multimodal large language models via full static quantization. arXiv preprint arXiv:2502.00425 (2025)