C-PTQ: UNIFIED CHANNEL-WISE SENSITIVITY AND SCALING FOR POST-TRAINING QUANTIZATION OF MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) require huge memory and computational costs, which limit their practical deployment. Post-training quantization (PTQ) techniques offer an efficient solution for model compression and inference acceleration. Yet, the quantized model faces performance degradation due to outlier channels, which are highly sensitive to quantization and substantially impair activation fidelity and task accuracy. To protect these salient channels during quantization, existing PTQ methods leverage modality- or token-level metrics to guide channel-wise scaling (CWS) of LLM decoders. However, these orthogonal measurements fail to capture channel-wise impacts on task-specific loss, and the misalignment between importance and scaling factors ultimately leads to suboptimal performance. To address this issue, we propose **C-PTQ**, a unified channel-wise PTQ method that harmonizes task-specific loss perturbation and quantization error. Motivated by second-order derivatives, we design a Fisher-weighted objective as a tractable Hessian approximation, seamlessly injecting task sensitivity into the scaling process. Notably, we achieve state-of-the-art performance without auxiliary modules like LoRA, thereby maintaining high efficiency. Experiments on Qwen2.5VL, InternVL2 and LLaVA-OV across 8 benchmarks demonstrate our effectiveness in both weight-only and weight-activation settings.

1 INTRODUCTION

The capacity of advanced MLLMs comes with huge computational costs, which limit their practical application. Much effort has been dedicated to reducing the computational overhead for developing efficient MLLMs, including distillation (Shu et al., 2025; Xu et al., 2024), pruning (Li et al., 2026; Fan et al., 2025) and quantization (Li et al., 2025; Xiang et al., 2026). Among them, *distillation* requires sufficient training data to achieve effective knowledge transfer (Lee et al., 2026), and *pruning* exhibits limited throughput gains relative to the pruning rate due to irregular memory access. In contrast, post-training *quantization* is hardware-friendly and data-efficient, requiring only a small calibration set to achieve rapid deployment on standard hardware. Since the majority of parameters and memory consumption in MLLMs reside in the LLM decoder, existing PTQ methods (Lin et al., 2024; Li et al., 2025; Xiang et al., 2026) typically focus on quantizing LLM decoders while keeping the vision encoder and multimodal projector in full precision. As illustrated in Fig. 1(a), outlier channels stretch the quantization range (dashed line), leading to large quantization error. To mitigate quantization error, channel-wise scaling (CWS) (Lin et al., 2024) searches for activation-aware scalers in every LLM block by minimizing Mean Squared Error (MSE) before and after quantization. However, preserving the task accuracy of MLLMs under low-bit quantization remains challenging, potentially due to inadequate optimization objectives.

Fig. 1(b) depicts the quantization error (MSE) and mean absolute gradients of the 14-th layer in InternVL2-8B (Chen et al., 2024b), where the red arrow highlights the misaligned optimization in the default CWS. To be concrete, the 1000-th and 2600-th channels exhibit small quantization error yet yield substantially large SFT gradients rather than near-zero values. As a result, the optimization towards minimizing quantization error may correspond to suboptimal task performance due to the large gradients. This motivates more fine-grained sensitivity or importance measures beyond CWS to align with SFT gradients. The block layouts in Fig. 1(c) represent the dimensional interest in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

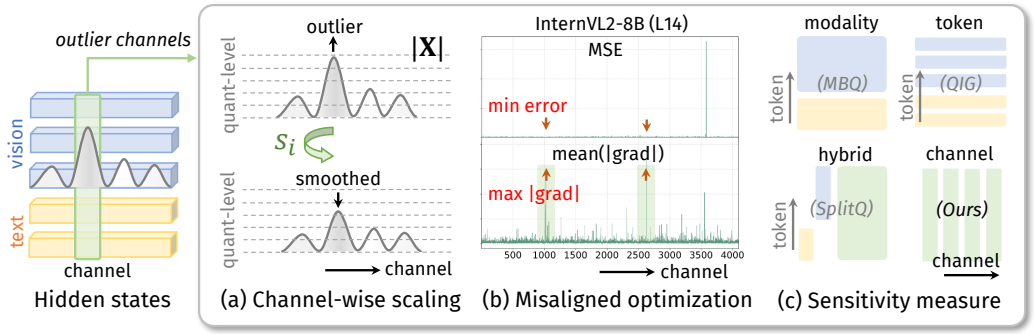

Figure 1: **Overview.** (a) Channel-wise scaling aims to smooth outlier channels with large magnitudes to reduce quantization error (*e.g.*, MSE). (b) However, channels with minor error can still exhibit substantial SFT gradients with degraded performance. (c) Our method aligns channel-wise sensitivity and scaling in objective for effective quantization.

sensitivity metrics, including *modality* (Li et al., 2025), *token* (Xiang et al., 2026; Zhou et al., 2026) and *hybrid* (Jia et al., 2026; Zhong et al., 2026) measures. Among them, MBQ (Li et al., 2025) and QIG (Xiang et al., 2026) extend the quantization objective to be modality-aware and token-aware, and SplitQ (Zhong et al., 2026)) adopts a hybrid mechanism by splitting modality-specific channels from the majority. While MBQ achieves the best efficiency compared to the following work (Xiang et al., 2026; Hu et al., 2026), its performance depends on the degree of modality imbalance in pretrained MLLMs. Recent work (Zhong et al., 2026; Jia et al., 2026) introduces low-rank matrices with hybrid importance measures to compensate for quantization error, which also brings more computational overhead. Interestingly, although the scaling factor is searched per channel, the quantization sensitivity was investigated from the orthogonal views. This leads to a natural question: *Can we directly inject channel-wise importance into CWS for aligned importance and scaling search?*

To support the synergy, we propose **C-PTQ**, a unified PTQ method that explicitly aligns channel-wise scaling with global task sensitivity. Our approach is built on the premise that minimizing quantization error alone is insufficient; **an effective strategy must reflect the channel-wise impact on task sensitivity**. Based on that, we leverage second-order curvature information as the sensitivity measure. Specifically, we introduce a Fisher-weighted quantization objective as a tractable approximation of the Hessian matrix. By applying a diagonal simplification (LeCun et al., 1989), we transform the complex loss perturbation into a decoupled objective of channel-wise weighted quantization error. Consequently, **C-PTQ** directly penalizes the reconstruction residual of each channel by its gradient variance. By integrating activation statistics and gradient sensitivity into the channel axis, our method achieves SOTA performance while avoiding computationally intensive operations.

Our key contributions are summarized as follows:

- We propose **C-PTQ**, a unified post-training quantization framework that explicitly aligns channel-wise importance with scaling factors, enabling fine-grained quantization.
- By employing the diagonal Fisher matrix as a tractable Hessian approximation, we derive an efficient Fisher-weighted objective that directly embeds task-specific sensitivity into the quantization process, thereby preserving downstream performance (Sec. 3).
- Extensive experiments on Qwen2.5VL, InternVL2 and LLaVA-OV show that our method achieves SOTA performance with favorable efficiency (Sec. 4).

2 Related Work

Post-training quantization. Despite the practical efficiency, quantization fundamentally struggles with large-magnitude outliers concentrated in specific channels, which stretch the quantization range and undermine low-bit representations. Considering the critical role of activations in determining quantization error, early **activation-aware** PTQ methods like AWQ (Lin et al., 2024) and SQ (Xiao et al., 2023) introduce channel-wise scaling by the activation distributions. Specifically, AWQ conducts grid search with MSE for the optimal channel-wise scales. Subsequent work adds sensitivity

measure into the objective (Li et al., 2025; Xiang et al., 2026). To capture structural importance beyond activation magnitudes, **curvature-based** methods like GPTQ (Frantar et al., 2023) incorporate second-order Hessian to apply closed-form error compensation during quantization. Inspired by fine-tuning techniques, **LoRA-based** methods (Hu et al., 2026; Jia et al., 2026) seek to push performance boundaries by applying external low-rank matrices for fine-grained compensation. Yet, LoRA itself incurs additional optimization and inference costs, motivating an efficient solution that directly aligns channel-wise scaling with task sensitivity.

Sensitivity measure. Compared with LLMs, the quantization of MLLMs necessitates elaborate sensitivity measurements due to the intertwined multimodal nature. Motivated by the issue of **modality imbalance**, MBQ (Li et al., 2025) reweights the quantization error of AWQ (Lin et al., 2024) using the average vision and text gradients. Building upon this idea, MABA (Zhang et al., 2026) employs modality-specific gradients for mixed-precision bit allocation, while MASQuant (Hu et al., 2026) enhances the single factor in SQ (Xiao et al., 2023) with modality-aware smoothing factors. As an extension of GPTQ (Frantar et al., 2023), VLM-PTQ (Deng & Huang, 2026) handles vision and text Hessians separately before fusing them for quantization. Besides, QIG (Xiang et al., 2026) leverages integrated gradients to characterize **token-wise** sensitivity, enabling the quantization objective to place greater emphasis on critical tokens. The following work considers a **hybrid** heterogeneity, which compensates for a few token-dependent (Jia et al., 2026) or modality-specific (Zhong et al., 2026) channels with low-rank matrices. Despite the progress, we observe a fundamental mismatch: sensitivity is typically measured at the modality or token level, whereas scaling is optimized at the channel level. To address this discrepancy, we revisit **Fisher Information**, a classical sensitivity metric widely used for model compression (Cherilyn & Safaai, 2026) and parameter-efficient adaptation (Sung et al., 2021). Please refer to App. B for more illustration. By incorporating a Fisher-weighted objective into channel-wise scaling, we build a synergy between the activation statistics and the curvature information.

3 METHODS

3.1 PRELIMINARIES

Let $\mathbf{Y}{=}\mathbf{W}\mathbf{X}{=}[\mathbf{y}_1, \ldots, \mathbf{y}_{N_k}] \in \mathbb{R}^{N_c^{\text{out}} \times N_k}$ denotes the full-precision output of a linear layer, where N_c^{out} and N_k are the number of output channels and tokens, respectively. Accordingly, $\mathbf{y}_i \in \mathbb{R}^{N_c^{\text{out}}}$ denotes the i-th token in the output. Throughout this paper, we use WxAy to indicate x-bit weight and y-bit activation quantization. Taking weight-only quantization as an example, the quantization function Q maps the high-precision weights $\mathbf{W}$ to discrete levels (*e.g.*, from FP16 to INT4), yielding the quantized output $\hat{\mathbf{Y}}{=}Q(\mathbf{W})\mathbf{X}$. The resulting residual is defined as $\Delta\mathbf{Y} = \hat{\mathbf{Y}} - \mathbf{Y} = Q(\mathbf{W})\mathbf{X} - \mathbf{W}\mathbf{X}$. To protect important channels and mitigate the reconstruction error, channel-wise scaling approaches introduce a scaling vector $\mathbf{s}$, reformulating the residual as:

$$\Delta\mathbf{Y} = Q(\mathbf{W} * \mathbf{s})(\mathbf{s}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}. \tag{1}$$

To find the optimal scaling factor $\mathbf{s}$, AWQ (Lin et al., 2024) conducts a grid search for each LLM layer to minimize MSE of residuals:

$$\mathbf{s} = \arg\min_{\mathbf{s}} \|\Delta\mathbf{Y}\|_2^2. \tag{2}$$

Substituting Eqn. (1) into Eqn. (2), the full objective of AWQ is:

$$\mathbf{s} = \arg\min_{\mathbf{s}} \|Q(\mathbf{W} * \mathbf{s})(\mathbf{s}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}\|_2^2. \tag{3}$$

When extending channel-wise scaling to MLLMs, existing methods leverage modality- or token-level sensitivity for reweighted quantization objectives (Li et al., 2025; Xiang et al., 2026) or adaptive bit allocation (Zhang et al., 2026). For instance, MBQ (Li et al., 2025) splits the residual $\Delta\mathbf{Y}$ from Eqn. (1) into visual and textual components $\Delta\mathbf{Y}^v$ and $\Delta\mathbf{Y}^t$, to minimize a modality-balanced quantization error:

$$\mathbf{s} = \arg\min_{\mathbf{s}} \sum_{i}^{N_k^v} |\bar{\mathbf{g}}^v| \cdot \|\Delta\mathbf{y}_i^v\| + \sum_{i}^{N_k^t} |\bar{\mathbf{g}}^t| \cdot \|\Delta\mathbf{y}_i^t\|. \tag{4}$$

Here, $|\bar{\mathbf{g}}^{\cdot}| \in \mathbb{R}_{\geq 0}$ denotes the average absolute gradients from two modalities, and the residual is written as $\Delta\mathbf{Y}=[\Delta\mathbf{y}_1, \ldots, \Delta\mathbf{y}_{N_k}]$, where $\Delta\mathbf{y}_i=\hat{\mathbf{y}}_i-\mathbf{y}_i \in \mathbb{R}^{N_c^{\text{out}}}$ denotes the residual of the i-th token. Taking a step further into token-level granularity, QIG (Xiang et al., 2026) reweights the quantization objective by token-wise integrated gradients λ_i:

$$\mathbf{s} = \arg\min_{\mathbf{s}} \sum_{i=1}^{N_k} \lambda_i \cdot \|\Delta\mathbf{y}_i\|_2^2. \tag{5}$$

Recent work introduces a hybrid paradigm (Jia et al., 2026; Zhong et al., 2026), which splits a few outlier channels for separate compensation. A closer look at these approaches reveals a fundamental misalignment: although sensitivity is evaluated across the modality or token dimension, the optimization is executed per-channel, as shown in Fig. 1. This discrepancy leads to a natural question: *Can we directly align channel-wise importance with channel-wise scaling?*

3.2 Fisher-weighted Quantization

Loss perturbation. To explicitly align channel-wise quantization error with task-specific sensitivity, we bridge the output residual $\Delta\mathbf{Y}$ with the overall loss perturbation $\Delta\mathcal{L}$. Assuming that $\mathcal{L}(\mathbf{Y})$ is the SFT loss (*e.g.*, cross-entropy), we analyze its second-order Taylor expansion around the full-precision output $\mathbf{Y}$:

$$\mathcal{L}(\hat{\mathbf{Y}}) = \mathcal{L}(\mathbf{Y}) + \underbrace{\sum_{i=1}^{N_k} \mathbf{g}_i^\top \Delta\mathbf{y}_i}_{\text{first-order}} + \underbrace{\frac{1}{2}\sum_{i=1}^{N_k}\sum_{k=1}^{N_k} \Delta\mathbf{y}_i^\top \mathbf{H}_{ik} \Delta\mathbf{y}_k}_{\text{second-order}} + \mathcal{O}\left(\|\Delta\mathbf{Y}\|^3\right), \tag{6}$$

where $\mathbf{g}_i=\frac{\partial\mathcal{L}}{\partial\mathbf{y}_i} \in \mathbb{R}^{N_c^{\text{out}}}$ is the gradient with respect to the output of token i, and $\mathbf{H}_{ik}=\frac{\partial^2\mathcal{L}}{\partial\mathbf{y}_i\partial\mathbf{y}_k^\top} \in \mathbb{R}^{N_c^{\text{out}}\times N_c^{\text{out}}}$ is the corresponding Hessian block of token pairs. When $i\neq k$, $\mathbf{H}_{ik}$ captures the second-order interaction between different tokens. Dropping the third-order remainder (LeCun et al., 1989; Frantar et al., 2023), the loss perturbation caused by quantization is approximated as

$$\Delta\mathcal{L} \;=\; \mathcal{L}(\hat{\mathbf{Y}}) - \mathcal{L}(\mathbf{Y}) \;\approx \sum_{i=1}^{N_k} \mathbf{g}_i^\top \Delta\mathbf{y}_i + \frac{1}{2}\sum_{i=1}^{N_k}\sum_{k=1}^{N_k} \Delta\mathbf{y}_i^\top \mathbf{H}_{ik} \Delta\mathbf{y}_k. \tag{7}$$

Vanishing first-order terms. Following the Optimal Brain Damage (OBD) framework (LeCun et al., 1989), we assume that the pretrained model has already converged to a local optimum. Consequently, the expected gradient evaluated on the calibration set is close to zero, with a negligible first-order term. We provide a proof in App. C, and empirically verify in App. A.3 that retaining the first-order term leads to a relatively small contribution on our calibration set.

$$\Delta\mathcal{L} \approx \frac{1}{2}\sum_{i=1}^{N_k}\sum_{k=1}^{N_k} \Delta\mathbf{y}_i^\top \mathbf{H}_{ik} \Delta\mathbf{y}_k. \tag{8}$$

As shown in Eqn. (8), the residual $\Delta\mathbf{y}_i$ is modulated by Hessian blocks by a quadratic term, which accounts for the misplacement of MSE and grdients in Fig. 1(b).

Diagonal approximation. To mitigate computational burden, we adopt the diagonal approximation as OBD (LeCun et al., 1989). Thereby, Eqn. (8) is simplified as:

$$\Delta\mathcal{L} \approx \frac{1}{2}\sum_{i=1}^{N_k} \Delta\mathbf{y}_i^\top \mathbf{H}_{ii} \Delta\mathbf{y}_i. \tag{9}$$

Then, we adopt Fisher Information to estimate diagonal elements of $\mathbf{H}_{ii}$.

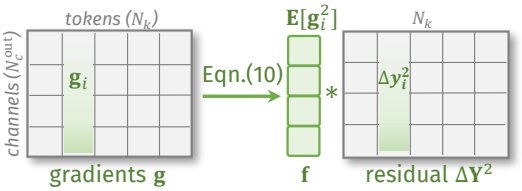

Figure 2: **Fisher-weighted objective.** Consistent with Fig. 1, the illustration here uses tokens N_k=5 and output channels N_c^{out}=4.

Algorithm 1 C-PTQ

Input: The gradient matrix $\mathbf{g}$ of output $\mathbf{Y}\in\mathbb{R}^{N_c^{\text{out}}\times N_k}$, weights $\mathbf{W}\in\mathbb{R}^{N_c^{\text{out}}\times N_c^{\text{in}}}$, input activations $\mathbf{X}\in\mathbb{R}^{N_c^{\text{in}}\times N_k}$, quantization function Q, calibration set $\mathcal{D}_{\text{cal}}$.
Objective: $\arg\min_{\mathbf{s}} L(\hat{\mathbf{Y}}) - L(\mathbf{Y})$
Output: Scaling factor $\mathbf{s} \in \mathbb{R}^{N_c^{\text{in}}}$.
// Step 1. Channel-wise sensitivity (diagonal Fisher)
$\mathbf{f} \in \mathbb{R}^{N_c^{\text{out}}} \leftarrow \text{mean}_{i,:}(\mathbf{g}_{ij}^2)$
// Step 2. Channel-wise scales (activation-aware)
$\bar{\mathbf{x}} \in \mathbb{R}^{N_c^{\text{in}}} \leftarrow \text{mean}_{i,:}(|\mathbf{X}_{ij}|)$, $\mathbf{s} \in \mathbb{R}^{N_c^{\text{in}}} \leftarrow \bar{\mathbf{x}}^{\alpha}$
// Step 3. Grid search with Fisher-weighted MSE over $\mathcal{D}_{\text{cal}}$ (: channel-wise multiplication)*
$\Delta\mathbf{Y} \leftarrow Q(\mathbf{W} * \mathbf{s})(\mathbf{s}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}$ or $\Delta\mathbf{Y} \leftarrow Q(\mathbf{W} * \mathbf{s})Q(\mathbf{s}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}$
$\text{Err}(\mathbf{s}) \leftarrow \|\sqrt{\mathbf{f}} * \Delta\mathbf{Y}\|_2^2$, $\mathbf{s} \leftarrow \arg\min_{\mathcal{D}_{\text{cal}}} \text{Err}(\mathbf{s})$
return $\mathbf{s}$.

From Hessian to Fisher Information. Since computing the exact Hessian matrix is less feasible for MLLMs, we approximate it using Fisher Information matrix ($\mathbb{E}\left[\mathbf{g}_i\mathbf{g}_i^\top\right]$). Adopting the diagonal approximation, we have:

$$\mathbf{H}_{ii} \approx \text{diag}(\mathbb{E}\left[\mathbf{g}_i^2\right]) := \text{diag}(\mathbf{f}). \tag{10}$$

Here, $\text{diag}(\mathbf{f})\in\mathbb{R}^{N_c^{\text{out}}\times N_c^{\text{out}}}$ denotes the diagonal Fisher matrix with $\mathbf{f}\in\mathbb{R}^{N_c^{\text{out}}}$ on its main diagonal. In practice, $\mathbf{f}$ is estimated from the calibration set by $\mathbf{f}=\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{g}_i^2$. Substituting Eqn. (10) into Eqn. (9), the loss perturbation is reformulated as:

$$\Delta\mathcal{L} \approx \frac{1}{2}\sum_{i=1}^{N_k}\Delta\mathbf{y}_i^\top \text{diag}(\mathbf{f})\Delta\mathbf{y}_i = \frac{1}{2}\sum_{i=1}^{N_k}\|\sqrt{\mathbf{f}}\odot\Delta\mathbf{y}_i\|_2^2, \tag{11}$$

where $\odot$ means elementwise multiplication. Eqn. (11) translates the loss perturbation into a simple, element-wise weighted residual, with which we successfully bridge the optimization gap between task-specific loss and quantization error. Unlike GPTQ (Frantar et al., 2023), which derives an activation-based Hessian ($\mathbf{H}=\mathbf{X}\mathbf{X}^\top$) to minimize the local reconstruction error, the Fisher matrix used here encapsulates the global task sensitivity. This explicit alignment guarantees that **C-PTQ** can efficiently execute channel-wise scaling while being strictly guided by task sensitivity. We provide empirical comparisons in Tab. 6 (Sec. 4.3).

3.3 Unified Channel-wise Scaling

Optimization objective. On the calibration set ($\mathcal{D}_{\text{cal}}$), the per-channel Fisher weights $\mathbf{f}_j\in\mathbb{R}_+$ are estimated by the average of squared gradients, and subsequently normalized by their mean across all channels to stabilize the magnitude. Omitting the constant factor in Eqn. (11), the objective is implemented as:

$$\text{Err}(\mathbf{s}) = \sum_{i=1}^{N_k}\sum_{j=1}^{N_c^{\text{out}}}\mathbf{f}_j\cdot\Delta\mathbf{y}_{ij}^2. \tag{12}$$

Substituting the residual in Eqn. (1) into Eqn. (12), our ultimate goal is to find the optimal scaling vector $\mathbf{s}$ that minimizes the loss perturbation:

$$\min_{\mathbf{s}}\|\sqrt{\mathbf{f}} * \left(Q(\mathbf{W} * \mathbf{s})\,(\mathbf{s}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}\right)\|_2^2, \tag{13}$$

where * denotes channel-wise multiplication. Recap the residual formulation in Eqn. (3); our objective can be interpreted as Fisher-weighted MSE. When extending our framework to weight-activation settings, the objective becomes:

$$\min_{\mathbf{s}}\|\sqrt{\mathbf{f}} * \left(Q(\mathbf{W} * \mathbf{s})\,Q(\mathbf{s}^{-1} * \mathbf{X}) - \mathbf{W}\mathbf{X}\right)\|_2^2. \tag{14}$$

Table 1: **Performance on Qwen2.5VL.** We report results on 7B and 32B models. The weight-only setting W3A16 quantizes weights to 3-bit, and weight-activation settings W4A8 and W4A6 quantize weights to 4-bit and activations to 8-bit and 6-bit, respectively.

Setting	Method	MMMU	SQA	VizWiz	TextVQA	OCRBench	ChartQA	AI2D	DocVQA	Avg	%
FP16	Qwen2.5VL-7B	50.8	87.2	70.5	83.1	83.9	83.1	82.9	95.7	79.7	100
W3A16	$\text{AWQ}_{\text{MLSys'24}}$	46.7	84.9	65.0	80.9	81.0	80.4	79.9	94.3	76.6	-3.9
	$\text{MBQ}_{\text{CVPR'25}}$	46.8	85.2	66.2	**81.0**	81.6	**82.0**	79.3	94.0	77.0	-3.4
	$\text{QIG}_{\text{CVPR'26}}$	44.2	84.0	**66.9**	80.8	81.7	81.4	**80.2**	94.3	76.7	-3.8
	$\text{MASQuant}_{\text{CVPR'26}}$	41.9	81.0	65.4	74.3	57.9	77.1	75.6	92.0	70.7	-11.3
	C-PTQ	**48.4**	**85.2**	64.6	80.2	**82.6**	80.9	79.9	**94.9**	**77.1**	**-3.3**
W4A8	$\text{SQ}_{\text{ICML'23}}$	41.6	73.0	65.5	**77.3**	77.0	79.9	73.4	91.1	72.4	-9.2
	$\text{MBQ}_{\text{CVPR'25}}$	**45.0**	84.1	59.2	71.8	72.3	73.6	79.4	86.2	71.5	-10.3
	$\text{QIG}_{\text{CVPR'26}}$	41.9	84.4	63.6	76.0	**78.2**	76.7	79.4	**90.7**	73.9	-7.3
	$\text{MASQuant}_{\text{CVPR'26}}$	43.3	**85.7**	**66.4**	77.0	72.8	**82.8**	**80.6**	87.4	74.5	-6.5
	C-PTQ	44.0	85.0	64.8	77.1	77.2	81.6	78.9	90.0	**74.8**	**-6.1**
W4A6	$\text{SQ}_{\text{ICML'23}}$	29.0	75.0	63.9	74.6	76.1	73.2	70.3	92.4	69.3	-13.0
	$\text{MBQ}_{\text{CVPR'25}}$	38.4	73.0	48.2	65.2	65.9	58.8	70.6	84.5	63.1	-20.8
	$\text{QIG}_{\text{CVPR'26}}$	39.1	71.5	56.9	70.9	74.3	67.4	71.5	89.8	67.7	-15.1
	$\text{MASQuant}_{\text{CVPR'26}}$	29.7	79.7	62.6	72.9	70.3	**73.3**	**78.3**	85.1	69.0	-13.4
	C-PTQ	**41.8**	**81.6**	**62.8**	**76.2**	**78.3**	70.6	75.7	**92.4**	**72.4**	**-9.2**
FP16	Qwen2.5VL-32B	60.2	91.6	64.7	77.3	79.8	69.2	84.5	94.6	77.7	100
W3A16	$\text{AWQ}_{\text{MLSys'24}}$	55.7	89.2	63.0	77.8	77.6	**69.4**	82.8	95.4	76.4	-1.7
	$\text{MBQ}_{\text{CVPR'25}}$	**58.8**	87.4	62.4	78.0	77.0	73.4	82.5	95.0	76.6	-1.4
	$\text{QIG}_{\text{CVPR'26}}$	55.6	88.2	62.3	77.1	**78.5**	65.6	82.8	95.2	75.7	-2.6
	C-PTQ	57.3	**89.8**	**64.6**	**78.4**	78.2	67.8	**83.9**	**96.3**	**77.0**	**-0.9**
W4A8	$\text{SQ}_{\text{ICML'23}}$	56.2	87.4	**66.3**	75.9	75.6	**69.6**	81.9	93.5	75.8	-2.4
	$\text{MBQ}_{\text{CVPR'25}}$	54.4	87.5	57.3	73.7	75.4	67.5	81.1	92.0	73.6	-5.3
	$\text{QIG}_{\text{CVPR'26}}$	54.9	**87.7**	64.2	76.1	78.1	**69.6**	82.2	94.5	75.9	-2.3
	C-PTQ	**58.0**	87.5	65.1	**76.3**	**79.1**	66.2	**82.4**	**94.6**	**76.2**	**-1.9**
W4A6	$\text{SQ}_{\text{ICML'23}}$	51.1	85.1	58.2	**73.4**	75.2	**61.5**	79.5	92.9	72.1	-7.2
	$\text{MBQ}_{\text{CVPR'25}}$	49.7	82.1	54.0	68.5	72.0	49.6	80.0	86.8	67.8	-12.7
	$\text{QIG}_{\text{CVPR'26}}$	49.1	84.3	**61.4**	72.4	75.2	54.0	78.7	**93.0**	71.0	-8.6
	C-PTQ	**55.8**	**86.0**	59.1	71.5	**77.6**	57.4	**81.2**	92.7	**72.7**	**-6.4**

As illustrated in Fig. 2, this objective selectively penalizes channels that are highly critical to the task loss $\mathcal{L}$, overcoming the limitation of the vanilla MSE in AWQ (Lin et al., 2024) while remaining activation-aware. Unlike MBQ (Li et al., 2025) and QIG (Xiang et al., 2026), our reweighting and scaling directions are aligned along the channel axis. The overall pipeline of **C-PTQ** is presented in Alg. 1, with i, j denoting the channel and token axes, respectively. Specifically, the scaling factor $\mathbf{s}$ is parameterized as a power function of the average input activation, *i.e.*, $\mathbf{s}{=}\bar{\mathbf{x}}^{\alpha}$, where $\bar{\mathbf{x}}$ denotes the channel-wise averaged input activation. Consequently, the optimization reduces to a one-dimensional grid search over $\alpha{\in}[0, 1)$. Although a dense Fisher matrix theoretically accounts for cross-channel dependencies, our empirical results indicate that our diagonal simplification performs on par with the dense one (see details in Sec. 4.3). Given the high computational cost of dense matrices, the proposed diagonalized approximation provides a practical and efficient alternative.

4 Experiments

4.1 Setup

Models and baselines. To evaluate our approach, we apply quantization to the LLM decoders of three representative pretrained MLLMs: Qwen2.5VL (Team, 2025), InternVL2 (Chen et al., 2024b), and LLaVA-OV (Li et al., 2024). These models are selected to cover varying degrees of modality bias and activation characteristics. For a fair comparison, all channel-wise scaling methods, including AWQ (Lin et al., 2024), MBQ (Li et al., 2025), QIG (Xiang et al., 2026) and our **C-PTQ** conduct a 40-step grid search to identify the optimal scaling factors. For LoRA-based MASQuant (Hu et al., 2026), we train it for 2 epochs following the default configuration. All baselines are implemented using their official repositories, and the final evaluations are conducted on the LMMs-Eval framework (Zhang et al., 2025). Please refer to App. A.1 for implementation details.

Evaluation. For the quantization process, we randomly sample 128 image-caption pairs from COCO caption dataset (Lin et al., 2014) proposed by ShareGPT4V (Chen et al., 2024a), following previous work (Li et al., 2025; Xiang et al., 2026). To ensure a fair comparison, all baselines are calibrated on this same subset. Once calibrated, we evaluate the models on a comprehensive suite of benchmarks: MMMU (Yue et al., 2024) for visual reasoning, VizWiz (Gurari et al., 2018)

Table 2: **Performance on InternVL2 and LLaVA-OV.** We report results on InternVL2-8B and LLaVA-OV-7B, where **C-PTQ** demonstrates SOTA performance consistently.

Setting	Method	MMMU	SQA	VizWiz	TextVQA	OCRBench	ChartQA	AI2D	DocVQA	Avg	%
FP16	InternVL2-8B	47.9	97.1	61.1	77.1	76.9	82.4	82.4	91.8	77.1	100
W3A16	$AWQ_{MLSys'24}$	**47.1**	96.3	**60.3**	74.5	74.7	79.2	80.2	90.4	75.3	-2.3
	$MBQ_{CVPR'25}$	45.2	96.3	58.5	74.9	75.3	79.9	79.5	90.0	75.0	-2.7
	$QIG_{CVPR'26}$	46.2	96.3	59.9	74.3	74.1	**80.5**	80.0	**90.6**	75.2	-2.5
	C-PTQ	45.5	**96.5**	59.9	**75.0**	**75.9**	80.4	**80.4**	90.5	**75.5**	**-2.1**
W4A8	$SQ_{ICML'23}$	42.2	96.0	52.1	72.3	72.0	78.3	77.1	87.8	72.2	-6.4
	$MBQ_{CVPR'25}$	43.9	96.1	55.9	71.9	72.9	78.2	79.4	88.8	73.4	-4.8
	$QIG_{CVPR'26}$	43.6	95.1	56.6	72.4	72.0	77.5	77.3	88.1	72.8	-5.6
	C-PTQ	**45.3**	**96.8**	**57.7**	**72.4**	**73.6**	**78.4**	**80.3**	**89.1**	**74.2**	**-3.8**
W4A6	$SQ_{ICML'23}$	**43.8**	**95.0**	51.1	69.9	70.6	**75.3**	74.8	86.7	70.9	-8.0
	$MBQ_{CVPR'25}$	41.4	94.7	53.3	68.9	69.4	74.6	**77.0**	87.0	70.8	-8.2
	$QIG_{CVPR'26}$	39.9	91.9	34.3	69.6	67.7	72.6	74.7	86.5	67.2	-12.8
	C-PTQ	42.1	94.9	**55.3**	**70.3**	**71.1**	75.0	76.9	**87.3**	**71.6**	**-7.1**
FP16	LLaVA-OV-7B	46.0	95.8	60.4	76.1	62.2	80.2	81.3	90.2	74.0	100
W3A16	$AWQ_{MLSys'24}$	43.1	94.5	60.0	73.2	59.8	76.8	78.3	86.2	71.5	-3.4
	$MBQ_{CVPR'25}$	**45.0**	**94.9**	60.0	72.8	58.9	**77.7**	79.1	86.3	71.8	-3.0
	$QIG_{CVPR'26}$	44.7	94.6	**61.2**	**73.8**	59.8	77.0	79.1	86.9	72.1	-3.6
	C-PTQ	44.5	94.8	60.6	73.4	**60.3**	77.4	**79.5**	**86.9**	**72.2**	**-2.4**
W4A6	$SQ_{ICML'23}$	39.8	90.9	54.3	55.9	36.9	65.0	74.1	64.6	60.2	-18.6
	$MBQ_{CVPR'25}$	40.0	90.1	55.9	64.0	49.3	72.3	74.1	79.5	65.6	-11.4
	$QIG_{CVPR'26}$	**42.3**	91.2	56.4	65.1	**50.1**	**73.1**	75.3	80.0	66.7	-9.9
	C-PTQ	40.6	**91.6**	**61.6**	**66.0**	49.7	72.7	**75.7**	**81.4**	**67.4**	**-8.9**

for visual perception, SQA-IMG (Lu et al., 2022) (abbreviated as SQA in tables), OCRBench (Liu et al., 2024), and TextVQA (Singh et al., 2019) for text recognition, ChartQA (Masry et al., 2022) and AI2D (Kembhavi et al., 2016) for structural understanding, and DocVQA (Mathew et al., 2021) for document comprehension. Dataset descriptions and our application on LLMs can be found in App. A.2 and A.3.

4.2 Main Results

Weight-only quantization. We first evaluate weight-only quantization under the W3A16 setting. As reported in Tab. 1, the fine-grained sensitivity metrics of QIG (Xiang et al., 2026) and MASQuant (Hu et al., 2026) do not translate into noticeable average gains, while AWQ (Lin et al., 2024) remains a highly competitive baseline under multimodal calibration data. Against these competitive baselines, **C-PTQ** still achieves consistent improvements. Notably, it is the only method that surpasses AWQ on InternVL2-8B (Tab. 2), while also maintaining the top performance on LLaVA-OV-7B. In general, the performance gap between weight-only and the full-precision model is relatively small, with all PTQ methods averaging merely a 2-3% drop on InternVL2-8B. We subsequently explore weight-activation quantization, where the performance degradation becomes more severe.

Weight-activation quantization. As observed in Tab. 1, MBQ (Li et al., 2025) exhibits the worst performance on W4A8 and W4A6 scenarios of Qwen2.5VL. The token-wise QIG (Xiang et al., 2026) and hybrid MASQuant (Hu et al., 2026) work better for weight-activation settings, while lagging behind ours by average scores. We attribute their unsatisfactory results to the misaligned importance and quantization, where our unified solution offers more benefits. Meanwhile, we observe notable performance fluctuations of SQ (Xiao et al., 2023) and MBQ (Li et al., 2025) across different MLLMs in Tab. 1 and Tab. 2. For example, SQ outperforms MBQ on Qwen2.5VL-7B and -32B, whereas MBQ achieves a substantial advantage over SQ on LLaVA-OV-7B. One possible explanation is that the cascaded paradigm in early models like LLaVA-OV introduces stronger modality bias, which suits well with the motivation behind MBQ. In contrast, more advanced models such as Qwen2.5VL suffer less from these challenges. As a result, the gradient-based modality bias of MBQ becomes less reliable and even introduces additional noise, causing worse performance than the simpler SQ. In addition, the variance among datasets appears in low bitwidth settings. Considering the W4A6 on Qwen2.5VL-7B, SQ (Xiao et al., 2023) gets merely 29.0 in MMMU, but yields the same accuracy score 92.4 as ours on DocVQA. In comparison, **C-PTQ** benefits from the unified channel-wise sensitivity and quantization to maintain SOTA performance across different datasets and MLLM models.

Table 3: **Quantization costs including gradient collection and (denoted as "+") CWS.** The standard (7B, 8B) and larger (26B, 32B) models are tested on one and four A100 cards, respectively.

Setting	Method	Qwen2.5VL-7B	InternVL2-8B	Qwen2.5VL-32B	InternVL2-26B
W3A16	$\text{MBQ}_{\text{CVPR'25}}$	40s+10m42s	40s+8m33s	2m37s+43m21s	1m34s+20m34s
	C-PTQ	43s+10m42s	41s+8m47s	2m50s+44m01s	1m39s+21m01s
	$\text{QIG}_{\text{CVPR'26}}$	41m15s	27m0s	89m38s	59m35s

Efficiency analysis. We analyze the efficiency from two aspects. First, as a channel-wise scaling approach, **C-PTQ** inherently offers superior quantization and inference efficiency compared to LoRA-based methods (Hu et al., 2026; Zhong et al., 2026), since a grid search is sufficient for the optimal channel-wise scales. Second, within the CWS family, the speed bottleneck is primarily determined by the reweighting strategy. As reported in Tab. 3, **C-PTQ** achieves comparable quantization runtime with MBQ, while being substantially faster than QIG. To improve the binary importance metrics (from two modalities), QIG introduces a 32-step integration for sensitivity estimation. Despite the fine-grained measure, it requires substantially longer time than MBQ. In comparison, we still maintain acceptable quantization costs.

4.3 Ablation Study

How significant is our channel-wise metric? A natural hypothesis is that channel-wise sensitivity correlates with either the magnitude of activations or the magnitude of gradients. To examine this hypothesis and justify the use of Fisher Information, we replace the weighting factor $\mathbf{f}_j$ in Eqn. (12) with two alternatives: the mean absolute activation values (*i.e.*, $\text{mean}_{i,:}(|\mathbf{y}_{ij}|)$) and the mean absolute gradient values (*i.e.*, $\text{mean}_{i,:}(|\mathbf{g}_{ij}|)$). The results in Fig. 3 validate the effectiveness of our Fisher-weighted objective over these heuristic designs across many settings.

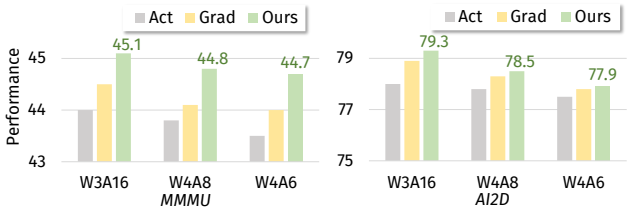

Figure 3: **Ablation study on channel-wise importance (LLaVA-OV-7B).** We compare our diagonal Fisher against the mean absolute activation and gradient-based channel-wise importance metrics to highlight our significance.

Table 4: **Ablation study on the distribution of calibration set (Qwen2.5VL-7B).** We report results on different COCO subsets, sample sizes and Pile dataset (text-only). The default 128 COCO samples used in the main results are annotated in gray.

Setting	Method	MMMU	SQA	VizWiz	TextVQA	OCRBench	ChartQA	AI2D	DocVQA	Avg	%
W3A16	COCO_{64}	48.2	85.0	64.4	79.5	82.4	80.7	79.7	94.7	76.8	-3.6
	COCO_{128a}	48.4	85.2	64.6	80.2	82.6	80.9	79.9	94.9	77.1	-3.3
	COCO_{128b}	48.2	85.6	64.0	79.9	82.8	80.7	80.1	94.7	77.0	-3.4
	COCO_{256}	48.9	85.3	65.0	79.5	82.2	80.4	79.9	94.9	77.0	-3.4
	Pile_{128}	47.8	85.5	66.0	80.6	81.0	79.1	80.7	94.7	76.9	-3.5
W4A8	COCO_{64}	45.4	84.7	65.6	77.0	77.2	81.3	79.2	90.1	75.1	-5.8
	COCO_{128}	44.0	85.0	64.8	77.1	77.2	81.6	78.9	90.0	74.8	-6.1
	COCO_{256}	44.6	84.9	64.5	77.1	77.6	81.4	79.0	90.2	74.9	-6.0
	Pile_{128}	44.3	82.2	63.5	75.6	81.0	79.1	80.7	88.0	74.3	-6.8

How robust is C-PTQ to the calibration set? To evaluate the robustness of our method, we conduct ablation study on $\mathcal{D}_{\text{cal}}$. As shown in Tab. 4, two randomly sampled COCO subsets ($\text{COCO}_{128a,128b}$) yield nearly identical results. This demonstrates that **C-PTQ** is insensitive to the choice of specific calibration samples. Furthermore, when the calibration size is adjusted to 64 or 256 samples, the average performance fluctuates within $\pm 0.3\%$, confirming the method's robustness to sample size variations. Finally, while calibrating on text-only data from Pile causes a minor performance drop of 0.2$\sim$0.5, **C-PTQ** flexibly supports both multimodal and text-only calibration. In contrast, modality-aware approaches like MBQ are strictly limited to multimodal data.

Table 5: **Ablation study on dense Fisher and first-order derivatives (Qwen2.5VL-7B).** We report average performance over four benchmarks and the quantization time in the last three columns.

Setting	Method	MMMU	SQA	OCRBench	ChartQA	Avg	%	Time
W3A16	Dense	47.7	84.2	82.2	79.9	73.5	100%	50 min
	+1st order	48.3	85.3	82.4	80.5	74.1	+0.8%	12 min
	C-PTQ	48.4	85.2	82.6	80.9	74.3	+1.2%	**11 min**
W4A8	Dense	44.8	84.7	77.3	81.7	72.1	100%	62 min
	+1st order	44.4	85.0	77.0	81.5	72.0	-0.1%	19 min
	C-PTQ	44.0	85.0	77.2	81.6	72.0	-0.1%	**18 min**

Does the diagonal approximation remain faithful to the original dense Fisher matrix? We validate the effectiveness of our diagonal approximation by comparing it directly against the full dense Fisher matrix. Tab. 5 demonstrates that the diagonal simplification successfully captures important channels with comparable or even higher performance. Meanwhile, our diagonal Fisher achieves substantial savings in computational resources, as shown in the quantization time.

Can the first-order term be safely ignored? In line with OBD (LeCun et al., 1989), we omit the first-order term in the loss perturbation expansion to minimize computational costs. As demonstrated in Tab. 5, incorporating this term introduces extra overhead with a marginal difference, thereby validating our objective in Eqn. (11). Further details are provided in App. A.3.

Comparison with RTN and GPTQ. As a Hessian-motivated method, we compare **C-PTQ** with GPTQ (Frantar et al., 2023) and round-to-nearest (RTN) in Tab. 6. Consistent with previous observations (Li et al., 2025; Deng & Huang, 2026), GPTQ performs even worse than the naive RTN on the latest MLLMs. This degradation suggests that solely minimizing layer-wise quantization error does not guarantee optimal performance, as the activation-derived Hessian in GPTQ is not explicitly aligned with the SFT objective. Ultimately, this highlights the effectiveness of our framework, which directly aligns the quantization error with the task-specific loss landscape rather than merely relying on reconstruction error.

Table 6: **Comparison with GPTQ and RTN (W3A16).** GPTQ shows even worse performance than the naive round-to-nearest (RTN), reflecting the significance of alignment between quantization error and loss perturbation.

Model	Method	MMMU	SQA	VizWiz	TextVQA	OCRBench	ChartQA	AI2D	DocVQA	Avg	%
Qwen2.5VL-7B	RTN	48.0	83.6	67.4	79.1	78.8	75.0	79.0	92.5	75.4	-5.4
	$\text{GPTQ}_{\text{ICLR'23}}$	48.2	73.4	65.5	78.7	76.4	62.5	79.2	87.8	71.5	-10.3
	C-PTQ	48.4	85.2	64.6	80.2	82.6	80.9	79.9	94.9	**77.1**	**-3.3**
InternVL2-8B	RTN	43.8	96.2	56.1	74.6	73.7	79.4	80.5	90.2	74.3	-3.6
	$\text{GPTQ}_{\text{ICLR'23}}$	43.1	94.3	59.8	74.5	74.1	76.4	76.6	90.2	73.6	-4.5
	C-PTQ	45.5	96.5	59.9	75.0	75.9	80.4	80.4	90.5	**75.5**	**-2.1**

5 Conclusion

In this paper, we propose **C-PTQ**, a unified channel-wise post-training quantization method that explicitly connects task-specific loss perturbation with quantization error. To align the sensitivity metric and quantization error at the channel level, we employ a diagonal Fisher matrix from SFT loss as a tractable Hessian approximation to measure the channel-wise sensitivity. Based on that, we propose a Fisher-weighted objective, selectively penalizing error in loss-critical channels while significantly reducing the computational overhead compared to dense matrices. Operating as a lightweight framework, **C-PTQ** achieves SOTA performance with exceptional efficiency. Extensive experiments on Qwen2.5VL, InternVL2, and LLaVA-OV across eight benchmarks demonstrate the effectiveness of our approach in both weight-only and weight-activation quantization regimes.

Limitations and future work. Despite the effectiveness of our unified framework, our current quantization scope is limited to LLM decoders, leaving the vision encoder and projection layers in full precision. Furthermore, similar to existing CWS approaches, **C-PTQ** experiences severe performance degradation under extreme compression settings (*e.g.*, W4A4). Extra compensation is necessary for extremely low-bit scenarios, which is considered as our future work.

AI USE STATEMENT

In this work, we used generative AI tools for polishing to improve readability. We have not used generative AI tools for idea and data generation, and the remaining required disclosure tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We confirm that it does not involve human subjects, personal data, or applications with significant societal impact that require separate ethical discussion.

REPRODUCIBILITY STATEMENT

We describe the quantization objective in Sec. 3.2 and 3.3, and discuss the experimental details in Sec. 4.1. We'll release the code upon acceptance.

REFERENCES

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, 2020.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Audrey Cherilyn and Houman Safaai. Supernodes and halos: Loss-critical hubs in llm feed-forward layers. *arXiv preprint arXiv:2604.23475*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

Juncan Deng and Kejie Huang. Vlm-ptq: Efficient post-training quantization for large vision-language models. In *CVPR*, 2026.

Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In *EMNLP*, pp. 18896–18913, 2025.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training compression for generative pretrained transformers. In *ICLR*, 2023.

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

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

Lulu Hu, Wenhu Xiao, Xin Chen, Xinhua Xu, Bowen Xu, Kun Li, and Yongliang Tao. Masquant: Modality-aware smoothing quantization for multimodal large language models. In *CVPR*, 2026.

Chenwei Jia, Baoting Li, Xuchong Zhang, Mingzhuo Wei, Bochen Lin, and Hongbin Sun. Quant experts: Token-aware adaptive error reconstruction with mixture of experts for large vision-language models quantization. In *CVPR*, 2026.

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

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *NeurIPS*, 1989.

Byung-Kwan Lee, Yu-Chiang Frank Wang, and Ryo Hachiuma. Masking teacher and reinforcing student for distilling vision-language models. In *CVPR*, 2026.

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

Jiameng Li, Aleksei Tiulpin, and Matthew B. Blaschko. MI-pruner: Crossmodal mutual information-guided token pruner for efficient mllms. *arXiv preprint arXiv:2604.03072*, 2026.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. Mbq: Modality-balanced quantization for large vision-language models. In *CVPR*, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*, 2024.

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 *ECCV*, 2014.

Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group fisher pruning for practical network compression. In *ICML*, 2021.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL Findings*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

Cuong Pham, Anh Dung Hoang, Cuong C Nguyen, Trung Le, Gustavo Carneiro, and Thanh-Toan Do. Layer-wise high-impact parameter ratio optimization in post-training quantization for large language models. In *ACL*, 2026.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Weilong Dai, Zhelun Yu, Wanggui He, et al. Llava-mod: Making llava tiny via moe-knowledge distillation. In *ICLR*, 2025.

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

Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *NeurIPS*, 2021.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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

Ziwei Xiang, Fanhu Zeng, Hongjian Fang, Rui-Qi Wang, Renxing Chen, Yanan Zhu, Yi Chen, Peipei Yang, and Xu-Yao Zhang. Fine-grained post-training quantization for large vision language models with quantization-aware integrated gradients. In *CVPR*, 2026.

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 *ICML*, 2023.

Chi Xu, Gefei Zhang, Yantong Zhu, Luca Benini, Guosheng Hu, Yawei Li, and Zhihong Zhang. Towards extreme pruning of llms with plug-and-play mixed sparsity. *arXiv preprint arXiv:2503.11164*, 2025.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

Xi Zhang, Hanwei Zhu, Jiamang Wang, Xiaolin Wu, and Weisi Lin. Modality-aware bit allocation for mixed-precision quantization of vision-language models. In *CVPR Findings*, 2026.

Yi Zhong, Haotong Qin, Xindong Zhang, Lei Zhang, and Guolei Sun. Breaking modality heterogeneity in low-bit quantization for large vision-language models. *arXiv preprint arXiv:2605.19929*, 2026.

Jundong Zhou, Tianao Cai, Yujie Huang, Xinbing Wang, Guang-Zhong Yang, and Nanyang Ye. Impquant: Fine-grained importance-aware quantization for large vision-language models. In *ICML*, 2026.

Appendix

A Extended Experiments

A.1 Implementation Details

The smaller models ($\leq$8B) are evaluated on a single A100 GPU, and larger models (26B$\sim$32B) are deployed on 4 A100 GPUs. We summarize the LLM backbone and number of layers for MLLMs used in the main paper in Tab. 7. To ensure a fair comparison, all PTQ baselines are evaluated using the same subset of the COCO dataset, sampled with a fixed seed of 42. We implement FlashAttention (Dao, 2024) for QwenVL- and InternVL-series, and adopt eager attention for the LLaVA-OV model. Due to the vast numerical discrepancy between gradients and MSE, we normalize the Fisher factors by their mean value, written as $\mathbf{f}_j = \mathbf{f}_j / \bar{\mathbf{f}}$. Experiments validate that the normalization choice doesn't affect the performance, *e.g.*, mean, MinMax and so on.

Table 7: **Overview of MLLMs.**

MLLM	LLM	# Layers
Qwen2VL-2B	Qwen2-2B	28
Qwen2VL-7B	Qwen2-7B	28
Qwen2.5VL-7B	Qwen2.5-7B	28
Qwen2.5VL-32B	Qwen2.5-32B	64
InternVL2-8B	InternLM2.5-7B	32
InternVL2-26B	InternLM2-20B	48
LLaVA-OV-7B	Qwen2-7B	28

A.2 Dataset Description

MMMU (Yue et al., 2024). MMMU was designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. The questions in MMMU span 30 subjects and 183 subfields. Meanwhile, it comprises 32 image types, such as charts, diagrams, maps, tables, music sheets and chemical structures. MMMU highlights advanced perception and reasoning with domain-specific knowledge, which demands that MLLMs tackle tasks requiring expert-level proficiency. The validation set used in our experiments has **900** multiple-choice question-answer pairs.

SQA-IMG (Lu et al., 2022). We denote SQA for ScienceQA benchmark, which employs multiple-choice questions to assess a model's performance in the scientific domains. The dataset spans three primary subjects, *i.e.*, natural science, language science and social science. It features a hierarchical structure organized by topic, category and skill. This hierarchy comprises 26 topics, 127 categories and 379 skills. While the questions are accompanied by relevant illustrations, a portion of the dataset is text-only. We evaluate on the SQA-IMG subset, which consists of **2,017** multimodal pairs where both images and questions are present.

VizWiz (Gurari et al., 2018). VizWiz is a real-world visual question answering dataset designed to assist people who are blind or visually impaired. Unlike traditional VQA datasets that are collected from carefully curated images, VizWiz consists of images captured by visually impaired users using their mobile devices and the corresponding spoken questions they asked. As a result, the dataset contains many real-world challenges, including poor image quality, occlusions, blur, improper framing and unanswerable questions. These characteristics make it a more realistic benchmark for evaluating the robustness and practical applicability of VQA systems. In this paper, we test on the validation set with **4,319** questions.

OCRBench (Liu et al., 2024). OCRBench is a comprehensive benchmark designed to evaluate the Optical Character Recognition (OCR) capabilities of MLLMs. It contains question-answer pairs collected from 29 publicly available datasets and covers five representative OCR-related tasks: text recognition, scene text-centric VQA, document-oriented VQA, key information extraction and handwritten mathematical expression recognition. While traditional OCR benchmarks focus solely on

text transcription accuracy, OCRBench provides a unified evaluation framework for assessing both text perception and reasoning abilities in multimodal models. OCRBench contains **1,000** questions.

TextVQA (Singh et al., 2019). The TextVQA benchmark aims to assess a model's proficiency in reading and reasoning over visual text. It emphasizes the integration of Optical Character Recognition (OCR) with natural language understanding. The images, primarily sourced from OpenImages-v3 (Krasin et al., 2017), feature diverse real-world scenarios—such as street signs, billboards and product packaging—that are rich in textual content. Alongside the raw images, ground-truth OCR tokens are provided as auxiliary input. To arrive at the correct answer, models must either extract text directly from the image or perform contextual reasoning based on the identified characters. We report evaluation results on the validation set, which comprises **5,000** image-question pairs.

ChartQA (Masry et al., 2022). ChartQA is a dataset for chart question answering that evaluates a model's ability to understand and reason over information presented in charts. ChartQA emphasizes both visual understanding and complex logical reasoning, requiring models to perform operations such as comparison, aggregation, arithmetic calculation and trend analysis based on chart content. The benchmark covers various chart types, including bar charts, line charts and pie charts, and often requires integrating visual cues with underlying numerical information. Overall, it provides **2,500** questions for evaluation.

AI2D (Kembhavi et al., 2016). AI2D focuses on diagrams, which are common tools for representing complex concepts, relationships and events. The challenge lies in identifying the diagram structure, the constituent semantics and their relationships. Subsequently, it provides detailed ground truth for diagram parsing, including the segmentation of visual elements like arrows, text boxes and blobs, as well as their syntactic relations. AI2D consists of **3,088** questions.

DocVQA (Mathew et al., 2021). DocVQA is a large-scale dataset specifically designed for VQA tasks on document images, including thousands of pages from diverse real-world documents. It challenges AI models to understand both text and layout structure by requiring them to answer natural language questions based on invoices, reports, forms and scanned scripts. The validation set has **5,349** questions.

A.3 More Results

Here, we extend the discussion in Sec. 4.3. First, we demonstrate that **C-PTQ** is robust across various model scales ranging from 2B to 72B. Next, we provide the derivation incorporating the first-order term reported in Tab. 5, along with its application to LLM quantization. Then, we present additional visualizations illustrating the misalignment between the quantization error and SFT gradients on Qwen2.5VL and InternVL2. Finally, we provide a comparison with VLM-PTQ (Deng & Huang, 2026).

Does C-PTQ maintain SOTA performance across varying model scales? Beyond the 7B to 32B models evaluated in Sec. 4.2, we further quantize the 2B and 72B versions of Qwen2VL. Fig. 4 compares our approach with MBQ and QIG under the W3A16 setting across three datasets. The consistent improvements indicate that our unified channel-wise scaling framework generalizes highly well to both small (2B) and large (72B) models.

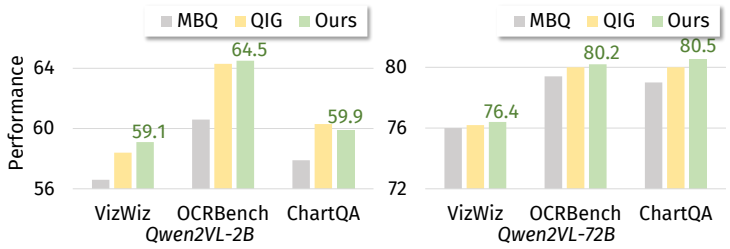

Figure 4: **Ablation study on more model scales (W3A16).** Our method maintains SOTA performance across scales.

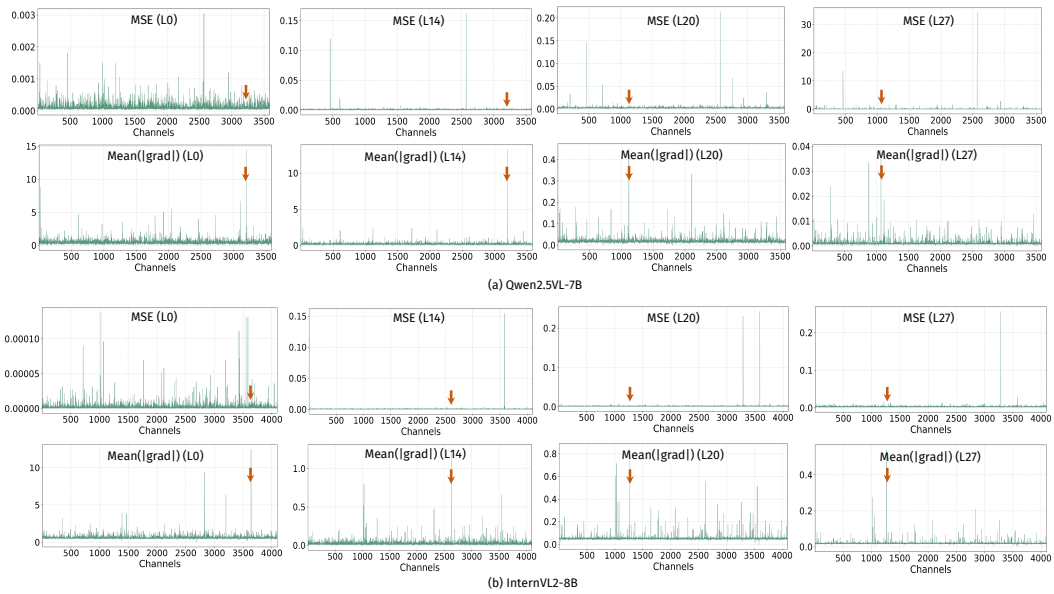

Figure 5: **Visualization of channel-wise quantization error and gradients (Qwen2.5VL-7B, InternVL2-8B).** We recommend viewing the layer-wise subfigure vertically. The red arrows indicate misaligned channels, where small error corresponds to disproportionately large gradients. Under such misalignment, optimizing towards minimal quantization error can induce excessively large gradients and poor downstream performance, which motivates our channel-wise aligned method.

Can the first-order term be safely ignored? If we consider both first- and second-order terms in Eqn. (7), then the loss perturbation $\Delta\mathcal{L}$ in Eqn. (11) is formulated as:

$$\Delta\mathcal{L} \approx \sum_{i=1}^{N_k} (\mathbf{g}^\top \Delta\mathbf{y}_i + \frac{1}{2}\|\sqrt{\mathbf{f}} \odot \Delta\mathbf{y}_i\|^2). \tag{15}$$

As reported in Tab. 5, retaining the first-order term yields negligible benefits yet introducing substantial computational overhead, which justifies our elegant formulation in Eqn. (11). Please refer to App. C for theoretical proof.

Table 8: **Performance on LLM quantization (W4A16).**

Method	PPL$_\downarrow$	PIQA$_\uparrow$	ARC-easy$_\uparrow$	ARC-chal.$_\uparrow$	Avg$_\uparrow$
Llama2-7B					
AWQ	5.59	78.3	68.9	41.1	62.8
C-PTQ	5.60	78.5	69.6	41.5	63.2
Llama3-8B					
AWQ	6.53	78.9	79.6	49.9	69.5
C-PTQ	6.49	79.4	80.1	50.3	69.9

Applications in LLMs. Our method is also readily applicable to LLMs such as Llama2-7B (Touvron et al., 2023) and Llama3-8B (Grattafiori et al., 2024). Following AWQ (Lin et al., 2024), we quantize LLMs using 128 calibration samples from the Pile dataset under the W4A16 setting. Tab. 8 reports the evaluation across four standard benchmarks: perplexity (PPL), PIQA (Bisk et al., 2020) and the easy and challenge subsets of ARC (Clark et al., 2018). Specifically, we utilize the validation set of PIQA (3,676 samples) and the test sets of ARC-easy (9,496 samples) and ARC-challenge (4,687 samples). The average scores across PIQA, ARC-easy, and ARC-challenge indicate that **C-PTQ** consistently outperforms AWQ. Ultimately, this successful extension demonstrates the broad applicability and effectiveness of our framework.

Visualization of the misaligned quantization error and SFT gradients. As shown in Fig. 5, this misalignment is evident across multiple layers in both Qwen2.5VL and InternVL2. Specifically, the red arrows in the top panel of Fig. 5(a) and (b) highlight channels with small quantization

error, which vanilla AWQ (Lin et al., 2024) treats as an indicator of low sensitivity. However, we observe that these corresponding channels actually exhibit large SFT gradients. This discrepancy demonstrates that the optimization towards minimizing local quantization error does not necessarily preserve downstream performance. Ultimately, this visualization corroborates the necessity of the quadratic formulation in Eqn. (8).

Comparison with VLM-PTQ. As a follow-up to GPTQ (Frantar et al., 2023), VLM-PTQ (Deng & Huang, 2026) derives a closed-form correction term that combines the propagated quantization residual with the corresponding inverse-Hessian column to relocate the quantization target beyond standard round-to-nearest (RTN). As in MBQ (Li et al., 2025), it introduces a modality-aware quantization scheme that differentiates channel importance between vision and language tokens through a lightweight fusion-coefficient search. Since VLM-PTQ hasn't released the code, we copy the results from their paper in Tab. 9, where **C-PTQ** remains superior.

Table 9: **Comparison with VLM-PTQ (Qwen2.5VL-7B)**.

Setting	Method	TextVQA	SQA	OCRBench	ChartQA	DocVQA
W3A16	VLM-PTQ	81.4	84.1	80.2	69.8	92.3
	C-PTQ	80.2	85.2	82.6	80.9	94.9

B Extended Related Work

Fisher Information for channel-wise sensitivity. While Fisher Information is widely adopted in model compression (Liu et al., 2021; Xu et al., 2025), its application varies significantly. For instance, Supernode (Cherilyn & Safaai, 2026) leverages a Fisher-based metric to assess the loss perturbation of masking a channel, and defines a "supernode" from the Fisher-style loss proxy (LP) for channel pruning. Since the top 1% channels dominate the overall LP mass, they protect these "supernodes" from being pruned under a given sparsity budget. In contrast, we employ Fisher as a weighting factor for post-training quantization. Meanwhile, Supernode tackles the intertwined expectation of the joint second moments of activation and gradient in a computationally expensive manner. As described in Sec. 3.2, we adopt a diagonal approximation for efficient measurement of channel-wise sensitivity. Another relevant work, High-Impact-Ratio (Pham et al., 2026) proposes a mixed-precision weight quantization for LLMs, optimizing the layer-wise ratio of high-impact parameters instead of a fixed ratio across layers. The average Fisher Information serves as an impact score to determine the optimal ratio of high-impact parameters per layer under a fixed resource budget. Later, the selected high-impact parameters are quantized with a higher bit-width and a stronger quantization method, and vice versa. Unlike such bit-width scheduling methods (Pham et al., 2026), **C-PTQ** is a principled PTQ framework designed to fix the objective misalignment inherent in channel-wise scaling. Furthermore, mixed-precision quantization often brings challenges for hardware deployment due to heterogeneous bit-width execution, whereas **C-PTQ** offers a hardware-friendly quantization scheme to avoid such issues.

C Extended Proof

Vanishing first-order terms. We take the weight-only quantization as an example. The output residual of token i can be written as $\Delta\mathbf{y}_i{=}\Delta\mathbf{W}\mathbf{x}_i$, where $\mathbf{W}$ and $\mathbf{x}_i$ denote the weight matrix and corresponding input. By the chain rule, the gradient with respect to the weight matrix $\mathbf{W}$ is $\frac{\partial\mathcal{L}}{\partial\mathbf{W}}{=}\sum_{i=1}^{N_k}\mathbf{g}_i\mathbf{x}_i^\top$, where $\mathbf{g}_i{=}\frac{\partial\mathcal{L}}{\partial\mathbf{y}_i}$ is the gradient with respect to output $\mathbf{y}_i$. Subsequently, the first-order term in Eqn. (7) can be formulated as:

$$\sum_{i=1}^{N_k}\mathbf{g}_i^\top\Delta\mathbf{y}_i=\sum_{i=1}^{N_k}\mathbf{g}_i^\top\Delta\mathbf{W}\mathbf{x}_i=\left\langle\frac{\partial\mathcal{L}}{\partial\mathbf{W}},\Delta\mathbf{W}\right\rangle, \tag{16}$$

where $\langle\cdot,\cdot\rangle$ means Frobenius inner product. In the main paper, we assume that the pretrained model is close to a local optimum, such that $\frac{\partial\mathcal{L}}{\partial\mathbf{W}}$ is small. Therefore, the first-order loss perturbation induced by weight quantization (Eqn. (16)) is negligible compared with the second-order curvature term, aligning with prior studies (Frantar et al., 2023; LeCun et al., 1989; Cherilyn & Safaai, 2026).