Post-Training Quantization with Gradient-Projected Fisher Approximation for Vision Transformers

Jincheol Yang[1*] **Jaemin Choi**[1*] **Nahyun Lim**[1*] **Yun-Seong Jeong**[1*]
Matti Zinke[2*] **Hyunwoo Yu**[1*] **Bongjoon Hyun**[3] **Kyomin Sohn**[3] **Suk-Ju Kang**[1†]
[1]Department of Electronic Engineering, Sogang University
[2]Department of Computer Science, Sogang University [3]Samsung Electronics
{yjc3232, jam0225, julialim5, jus8077, mattizinke, hyunwoo137, sjkang}@sogang.ac.kr
{bongj.hyun, kyomin.sohn}@samsung.com

Abstract

Vision Transformers (ViTs) demonstrate strong performance, but their substantial memory footprint and computational overhead necessitate efficient compression techniques such as post-training quantization (PTQ). However, pushing ViTs to low-bit precision incurs significant accuracy degradation. Recent PTQ methods adopt block reconstruction-based optimization with curvature-aware objectives, typically implemented using Fisher-based approximations. These approaches rely on explicit curvature modeling based on diagonal or structured approximations, which still fail to capture cross-dimensional interactions. Moreover, they often involve matrix inversion, leading to numerical instability under ill-conditioned settings. To address these limitations, we propose Gradient-Projected Fisher Approximation for Quantization (GPFA-Q), a block reconstruction-based PTQ framework that avoids explicit curvature matrix construction while capturing off-diagonal interactions. First, we introduce Gradient-Projected Reconstruction (GPR), a reconstruction objective that captures cross-dimensional interactions without explicitly constructing curvature matrices. To further support GPR, we integrate Soft Grid Rounding (SGR), which reduces the mismatch between continuous reconstruction and discrete inference. Extensive experiments demonstrate that our GPFA-Q achieves the state-of-the-art performance in low-bit quantization across diverse vision tasks. Code is available at https://github.com/Nobody-96/GPFA-Q.

1 Introduction

Vision Transformers (ViTs) have emerged as a backbone in computer vision by achieving competitive performance across various tasks [1–3], powered by large-scale pretraining and the self-attention mechanism. However, the substantial number of parameters and the computational overhead of ViTs pose significant challenges for deployment on resource-constrained edge devices. To address this issue, various model compression techniques have been explored to reduce model complexity while maintaining performance. These compression techniques include model quantization [4–7], pruning [8–10], knowledge distillation [11–13], and lightweight architecture design [14–21].

Among these approaches, model quantization is one of the most effective approaches to significantly reduce the memory footprint and computational cost by mapping floating-point values to low-bit integers. Quantization is typically categorized into quantization-aware training (QAT) [22–24] and post-training quantization (PTQ) [25–34]. QAT retrains the model using the full training dataset, which achieves higher accuracy, but suffers from substantial training costs and requires the original

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

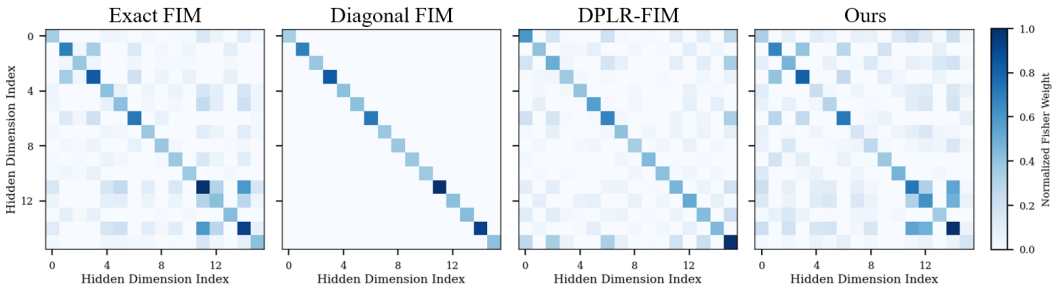


Figure 1: Visual comparison of FIM approximations on Block 11 of DeiT-Base under 3-bit quantization. The diagonal FIM discards off-diagonal correlations entirely, while the DPLR-FIM recovers part of this structure. Our method shows the highest similarity to the exact FIM, particularly in the off-diagonal elements. A 16×16 sub-matrix is shown for visual clarity. We show the full matrices in Section F of the appendix.

data. In contrast, PTQ seeks quantization parameters using a small unlabeled calibration set without additional training, making it widely preferred, as it is faster and less computationally demanding. To reduce the quantization error of PTQ methods, reconstruction-based methods [32, 33] have been proposed. These methods minimize block-wise output discrepancy, enabling low-bit quantization to achieve performance comparable to the full-precision models.

In particular, Hessian-guided metrics [27, 28, 32, 33] have been commonly used to measure the output discrepancy in block reconstruction-based PTQ. Since computing the exact Hessian matrix is computationally expensive, prior works rely on approximations based on the Fisher Information Matrix (FIM), which is commonly approximated using gradient outer products. A key design aspect in FIM-based PTQ methods is how to approximate the structure of the FIM. BRECQ [32] employs a diagonal approximation, which FIMA-Q [28] extends through a diagonal plus low-rank (DPLR) method to partially capture interactions. An accurate estimation of the FIM is important for reconstruction-based PTQ, as it captures the complex output correlations needed to correctly guide weight optimization across varying token and channel sensitivities. In low-bit settings, these cross-dimensional interactions become increasingly important. However, existing approximation methods introduce structural constraints for practical computation that severely limit correlation modeling. As shown in Figure 1, diagonal approximations completely ignore off-diagonal interactions. While the DPLR-FIM partially recovers them, its reliance on explicit matrix inversion may introduce numerical instability, which results in the approximation still deviating from the exact FIM.

To address these limitations, we propose Gradient-Projected Fisher Approximation for Quantization (GPFA-Q), a PTQ framework that models off-diagonal curvature interactions without explicit matrix construction. Instead of explicitly forming curvature matrices, we introduce Gradient-Projected Reconstruction (GPR), which projects block-wise quantization errors onto the task loss gradient subspace. This projection captures cross-dimensional, off-diagonal interactions by measuring reconstruction errors along gradient directions, where each projection represents interactions across multiple output dimensions and combines them without explicit curvature modeling. As a result, GPR provides a closer approximation to the underlying Fisher structure and yields a more effective reconstruction objective, as illustrated in Figure 1. To complement GPR, we incorporate Soft Grid Rounding (SGR), which mitigates the mismatch between continuous optimization and discrete inference. SGR evaluates reconstruction on the discrete quantization grid while preserving gradient flow via the Straight-Through Estimator (STE), ensuring that the gradient-projected objective remains consistent with the quantized model. Combined with the accurate curvature modeling of GPR, this leads to a highly effective reconstruction framework that significantly improves low-bit quantization performance for ViTs. Our main contributions are summarized as follows:

- We analyze block reconstruction-based PTQ from a curvature modeling perspective and show that explicit curvature approximations fail to capture cross-dimensional interactions across different feature dimensions, especially in low-bit settings.
- We propose GPFA-Q, a novel PTQ framework based on Gradient-Projected Fisher Approximation. For this, we introduce GPR, which reformulates the Fisher quadratic objective as gradient projections to implicitly model cross-dimensional interactions without explicit

matrix construction or inversion. To support GPR, we incorporate SGR to mitigate the mismatch between continuous optimization and discrete model inference.

- We conduct extensive experiments across diverse vision tasks and ViT-based architectures. Experimental results demonstrate that the proposed GPFA-Q achieves state-of-the-art performance in low-bit quantization.

2 Related Work

Post-Training Quantization (PTQ) has become the practical compression strategy for ViTs on edge devices, enabling efficient deployment without the need for full training data or extensive retraining costs. Early approaches focused on addressing the unique activation statistics of ViTs through heuristic metric adjustments and optimization strategies. FQ-ViT [31] introduced power-of-two scaling for LayerNorm and Log-Int-Softmax quantization. PTQ4ViT [26] utilized a twin-uniform quantization scheme with Hessian-guided metrics to handle the asymmetric distribution of activations. RepQ-ViT [30] further advanced this by applying scale reparameterization to decouple channel-wise variations in post-LayerNorm activations. ERQ [29] proposed a two-step error reduction framework, utilizing ridge regression and rounding refinement to minimize quantization errors sequentially. To further improve the representation capability for non-uniform distributions, AdaLog [35] developed an adaptive logarithm quantizer that dynamically selects the optimal logarithm base.

Reconstruction-based methods, which optimize quantization parameters to minimize the difference between full-precision and quantized feature maps, have recently achieved state-of-the-art performance. This paradigm was pioneered by BRECQ [32] and QDrop [33] for Convolutional Neural Networks (CNNs). To bypass the non-differentiable nature of quantization, these methods typically rely on continuous surrogate variables for weight optimization. Inspired by their success, recent works have adapted these frameworks specifically for ViTs, focusing on metric robustness and distribution alignment. DopQ-ViT [36] explicitly tackles the distribution mismatch problem by introducing a distribution-friendly Tan quantizer and an outlier-aware scaling strategy. In terms of sensitivity metrics, APHQ-ViT [27] improved Hessian estimation using an average perturbation Hessian metric, enabling robust reconstruction against calibration data noise. More recently, FIMA-Q [28] reformulated block reconstruction with a Fisher-based objective and introduced a diagonal-plus-low-rank (DPLR) approximation to partially capture additional off-diagonal correlations. However, estimating such low-rank structures typically involves solving matrix systems, which may introduce additional numerical sensitivity in practice.

While these methods have advanced PTQ for ViTs, they struggle to accurately capture off-diagonal interactions in the Fisher structure. Our GPFA-Q addresses this limitation by enabling stable and accurate modeling of such interactions through gradient projections without explicit matrix inversion.

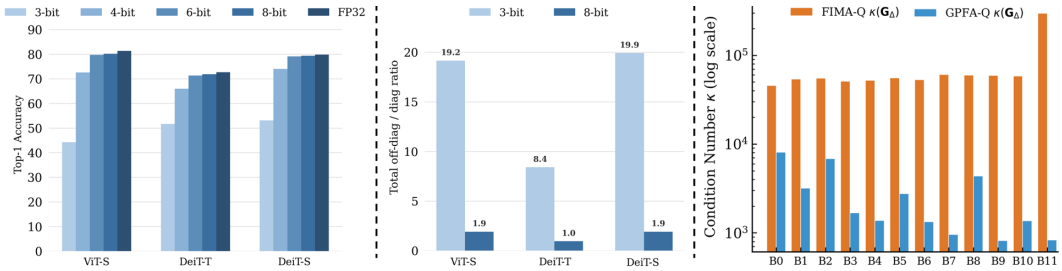
3 Method

In this section, we present our Gradient-Projected Fisher Approximation for Quantization (GPFA-Q) that models cross-dimensional interactions through gradient projections without explicit curvature construction. To motivate our approach, we first analyze the role of off-diagonal interactions in reconstruction-based PTQ in Section 3.1. Based on this analysis, Section 3.2 introduces GPFA-Q, which integrates Gradient-Projected Reconstruction (GPR) and Soft Grid Rounding (SGR) to enable stable low-bit reconstruction without explicit matrix inversion. For clarity, we provide the formulations of uniform quantization and block reconstruction used in our method in Appendix A.

3.1 Analysis of Off-Diagonal Modeling

Reconstruction-based PTQ minimizes the discrepancy between the full-precision block output and the quantized output. The effectiveness of reconstruction depends on the modeling of the curvature matrix, which determines the sensitivity to quantization perturbations. We first analyze when diagonal curvature is sufficient and when off-diagonal interactions become essential. Then, we show that existing explicit off-diagonal constructions can suffer from numerical instability as follows.

Effect of Bit-Width on Diagonal and Off-Diagonal Terms. We analyze the performance of diagonal curvature approximation across different bit-widths. Figure 2(a) shows that reconstruction-



(a) Performance of diagonal approximation (b) Aggregate off-diagonal term (c) Per-block κ at end of reconstruction

Figure 2: Analysis of diagonal and off-diagonal curvature modeling in block reconstruction. (a) Top-1 accuracy under diagonal-only reconstruction on ViT-Small (S), DeiT-Tiny (T), and DeiT-Small (S) across different bit-widths. (b) Ratio of off-diagonal to diagonal terms on ViT-S, DeiT-T, and DeiT-S at 3-bit and 8-bit. Off-diagonal terms have a substantially larger relative impact under low-bit quantization. (c) Per-block condition number $\kappa(\mathbf{G}_\Delta)$ at the end of 3-bit reconstruction on DeiT-S across transformer blocks. FIMA-Q shows substantially larger variation across blocks than GPFA-Q.

based diagonal curvature preserves most of the full-precision accuracy at 6-bit and 8-bit, but degrades substantially at 3-bit and 4-bit. This trend suggests that diagonal curvature provides a reasonable approximation when quantization perturbations are small, but becomes less effective in low-bit settings. To explain this observation, let $\mathbf{z} \in \mathbb{R}^D$ denote the full-precision block output, where D is the block-output dimension. The quantized output is denoted by $\hat{\mathbf{z}} = \mathbf{z} + \Delta\mathbf{z}$, where $\Delta\mathbf{z}$ is the quantization perturbation. Under a second-order approximation, the reconstruction objective can be approximated as:

$$\mathcal{L}(\mathbf{z} + \Delta\mathbf{z}) \approx \mathcal{L}(\mathbf{z}) + \frac{1}{2}\,\Delta\mathbf{z}^\top \mathbf{F}\,\Delta\mathbf{z}, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{D\times D}$ denotes a Fisher-based curvature approximation to the local curvature matrix. The quadratic form can be decomposed as:

$$\Delta\mathbf{z}^\top \mathbf{F}\,\Delta\mathbf{z} = \sum_i F_{ii}\,\Delta z_i^2 + \sum_{i\neq j} F_{ij}\,\Delta z_i\,\Delta z_j, \tag{2}$$

where the first term is the diagonal component and the second term captures cross-dimensional interactions. This decomposition indicates that the expected impact of the off-diagonal term depends not only on the curvature coefficients F_{ij}, but also on the correlation structure of the perturbations, as reflected in $\mathbb{E}[\Delta z_i\,\Delta z_j]$. When perturbations across different output dimensions are weakly correlated, which often occurs in high-bit quantization with small quantization steps, $\mathbb{E}[\Delta z_i\,\Delta z_j]$ tends to be small for $i \neq j$, and the diagonal term remains a useful approximation.

In contrast, under low-bit quantization, the larger quantization step size can lead to more structured perturbations. Because ViT activations exhibit dependencies across spatial and channel dimensions, partly induced by self-attention, low-bit quantization perturbations may inherit this structure, making $\mathbb{E}[\Delta z_i\,\Delta z_j]$ non-negligible across multiple pairs (i, j). Since the off-diagonal sum contains $D(D-1)$ cross-dimensional pairs, even moderate correlations can accumulate and have a noticeable impact on the quadratic form. Under approximately zero-mean perturbations, $\mathbb{E}[\Delta z_i \Delta z_j]$ can be approximated by $\mathrm{Cov}(\Delta z_i, \Delta z_j)$. Let $\mathrm{Cov}(\Delta\mathbf{z}) \in \mathbb{R}^{D\times D}$ denote the empirical covariance matrix of the block-output quantization error estimated over calibration samples, and define the aggregate off-diagonal ratio as:

$$\rho = \frac{\sum_{i\neq j} |[\mathrm{Cov}(\Delta\mathbf{z})]_{ij}|}{\sum_i [\mathrm{Cov}(\Delta\mathbf{z})]_{ii}}, \tag{3}$$

which compares the total magnitude of cross-dimensional error covariances with the total per-dimension variance, averaged over transformer blocks. Figure 2(b) supports this interpretation, showing that ρ is consistently and substantially larger at 3-bit than at 8-bit across evaluated models. These results suggest that aggressive quantization significantly increases such correlations, increasing the importance of off-diagonal terms. Together with the accuracy trend in Figure 2(a), these observations indicate that low-bit reconstruction benefits from objectives that account for cross-dimensional interactions. A formal analysis is provided in Section B of the appendix.

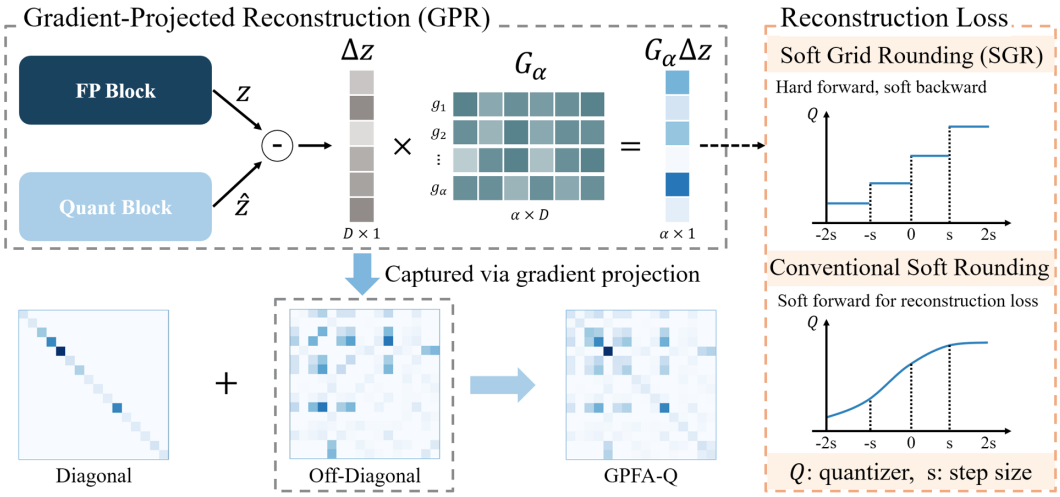


Figure 3: Overview of the GPFA-Q framework. Given full-precision and quantized block outputs, GPR projects the block-output quantization error $\Delta\mathbf{z}$ onto sampled task-loss gradients $\mathbf{G}_\alpha$, producing $\mathbf{G}_\alpha\Delta\mathbf{z}$. This projection implicitly models off-diagonal Fisher interactions without explicit curvature construction. It is combined with a diagonal term and SGR to form the GPFA-Q objective. SGR evaluates reconstruction using discretized outputs while propagating gradients through soft representations, reducing the mismatch between continuous optimization and discrete inference. For clarity, the projection is illustrated for a single error vector at the block output.

Numerical Instability of Explicit Curvature Construction. FIMA-Q [28] adopts DPLR FIM approximation, where the low-rank component is used to capture additional off-diagonal interactions. Let $\Delta\mathbf{Z} \in \mathbb{R}^{k\times D}$ denote the displacement matrix. Its rows consist of k displacement vectors collected during reconstruction. The low-rank formulation relies on the Gram matrix

$$\mathbf{G}_\Delta = \Delta\mathbf{Z}\Delta\mathbf{Z}^\top, \tag{4}$$

whose inversion is required for the low-rank correction. This inversion is sensitive to the conditioning of $\mathbf{G}_\Delta$. Each displacement vector captures the residual quantization error at a given stage of block reconstruction. As block reconstruction reduces this error, successive displacement vectors can become increasingly aligned while shrinking in magnitude. This alignment causes the rows of $\Delta\mathbf{Z}$ to become nearly collinear, thereby reducing its effective rank and decreasing its smallest singular value. As a result, the Gram matrix $\mathbf{G}_\Delta$ becomes increasingly ill-conditioned, with condition number

$$\kappa(\mathbf{G}_\Delta) = \kappa(\Delta\mathbf{Z})^2. \tag{5}$$

This makes its inverse highly sensitive to perturbations, leading to numerically unstable curvature estimation. Figure 2(c) empirically supports this analysis. The per-block condition numbers $\kappa(\mathbf{G}_\Delta)$ in FIMA-Q vary significantly across transformer blocks, indicating inconsistent numerical stability. Such variability leads to unreliable curvature modeling, especially in later reconstruction stages where the displacement matrix becomes nearly degenerate. Block reconstruction requires capturing off-diagonal interactions, but explicit low-rank construction based on matrix inversion can suffer from numerical instability. A detailed analysis is provided in Section C of the appendix.

3.2 Post-Training Quantization with Gradient-Projected Fisher Approximation

Based on the analysis in Section 3.1, we propose GPFA-Q, which addresses this challenge by capturing off-diagonal interactions through gradient projections without explicit curvature construction, as illustrated in Figure 3.

Gradient-Projected Reconstruction (GPR). The proposed GPR is motivated by the empirical FIM, approximated using sampled task-loss gradients:

$$\mathbf{F} = \frac{1}{N_s}\sum_{j=1}^{N_s}\mathbf{g}_j\mathbf{g}_j^\top, \tag{6}$$

where $\mathbf{g}_j = \nabla_{\mathbf{z}} \mathcal{L}_{\text{task}}$ denotes the gradient of the KL divergence between the final logits of the full-precision and quantized models, backpropagated to the block output. Under this formulation, the Fisher-weighted reconstruction objective can be written as:

$$\mathcal{L}_{\text{Fisher}} = \frac{1}{B} \sum_{i=1}^{B} \Delta \mathbf{z}_i^\top \mathbf{F} \Delta \mathbf{z}_i, \tag{7}$$

where B denotes the batch size used for block reconstruction. Expanding the quadratic form

$$\Delta \mathbf{z}_i^\top \mathbf{F} \Delta \mathbf{z}_i = \frac{1}{N_s} \sum_{j=1}^{N_s} \Delta \mathbf{z}_i^\top \mathbf{g}_j \mathbf{g}_j^\top \Delta \mathbf{z}_i, \tag{8}$$

which yields

$$\mathcal{L}_{\text{Fisher}} = \frac{1}{N_s B} \sum_{j=1}^{N_s} \sum_{i=1}^{B} (g_j^\top \Delta z_i)^2 = \frac{1}{N_s B} \|\mathbf{G}_s \Delta \mathbf{Z}_{\text{batch}}^\top\|_F^2, \tag{9}$$

where $\Delta \mathbf{Z}_{\text{batch}} \in \mathbb{R}^{B \times D}$ collects the quantization errors Δz_i as rows, and $\mathbf{G}_s \in \mathbb{R}^{N_s \times D}$ collects the gradient vectors. This formulation shows that the Fisher quadratic form can be computed through gradient projections without explicitly constructing the Fisher matrix. Each projection $\mathbf{g}_j^\top \Delta \mathbf{z}_i$ measures the reconstruction error along directions defined by the task-loss gradients, coupling errors across dimensions and thereby capturing off-diagonal interactions in the Fisher structure, which can be decomposed as:

$$(\mathbf{g}_j^\top \Delta \mathbf{z}_i)^2 = \sum_d g_{jd}^2 \Delta z_{id}^2 + \sum_{p \neq q} g_{jp} g_{jq} \Delta z_{ip} \Delta z_{iq}, \tag{10}$$

where the first term reflects per-dimension sensitivity, while the second term explicitly couples perturbations across different output dimensions ($p \neq q$). Therefore, GPR implicitly preserves cross-dimensional interaction terms without explicit construction of $\mathbf{F}$. When the quantization error is weakly aligned with the gradient direction, the projection magnitude is small, leading to a smaller penalty. In contrast, errors aligned with task-sensitive directions produce larger projections and incur a stronger penalty. GPR emphasizes coherent cross-dimensional error patterns while attenuating misaligned perturbations, thereby avoiding explicit curvature matrix construction and inversion.

To reduce reconstruction cost, we approximate the projection using a subset of gradients. Let $\mathbf{G}_\alpha \in \mathbb{R}^{\alpha \times D}$ denote a sampled gradient matrix, where α denotes the number of sampled gradients. The projection-based objective becomes:

$$\mathcal{L}_{\text{GPR}} = \frac{1}{\alpha B} \|\mathbf{G}_\alpha \Delta \mathbf{Z}_{\text{batch}}^\top\|_F^2, \tag{11}$$

Soft Grid Rounding (SGR). While our GPR optimizes reconstruction using soft-rounded outputs, inference is performed with discretely rounded weights, resulting in a mismatch between optimization and deployment. In conventional block reconstruction [32, 33, 27, 28], the reconstruction loss is evaluated on soft-rounded outputs while using STE only for gradient propagation. As a result, the optimized objective may not accurately reflect the output of the discretized model. To better align GPR with discrete inference, we introduce Soft Grid Rounding (SGR), which uses hard-rounded outputs in the forward pass and soft representations for gradient propagation:

$$\hat{\mathbf{Z}}^{\text{ste}} = \hat{\mathbf{Z}}^{\text{soft}} + \text{sg}\big(\hat{\mathbf{Z}}^{\text{hard}} - \hat{\mathbf{Z}}^{\text{soft}}\big), \tag{12}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, which blocks gradient flow in the backward pass. Under this formulation, the reconstruction error in GPR is evaluated as

$$\Delta \mathbf{Z}^{\text{ste}} = \hat{\mathbf{Z}}^{\text{ste}} - \mathbf{Z}, \tag{13}$$

which reflects the discrepancy between discretized outputs and full-precision outputs in the forward pass, while propagating gradients through soft representations. Let $\Delta \mathbf{Z}_{\text{batch}}^{\text{ste}}$ denote the matrix of these evaluated errors. The SGR term shares the same projection basis as GPR:

$$\mathcal{L}_{\text{SGR}} = \frac{1}{\alpha B} \|\mathbf{G}_\alpha \Delta \mathbf{Z}_{\text{batch}}^{\text{ste}}|_F^2. \tag{14}$$

which aligns the gradient-projected objective with the discrete inference process.

Table 1: Comparison of the top-1 accuracy (%) across various ViTs on ImageNet [37]. Rec. indicates whether reconstruction is used. "W/A" indicates that the bit-width of the weight and activation are W and A bits, respectively. "*" indicates the results are reproduced by using the official code.

Method	Rec.	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-Precision	-	32/32	81.39	84.54	72.21	79.85	81.80	83.23	85.27
PTQ4ViT [26]	×	3/3	0.10	0.10	3.50	0.10	31.06	28.69	20.13
RepQ-ViT [30]	×	3/3	0.10	0.10	0.10	0.10	0.10	0.10	0.10
AdaLog [35]	×	3/3	13.88	37.91	31.56	24.47	57.47	64.41	69.75
I&S-ViT [38]	✓	3/3	45.16	63.77	41.52	55.78	73.30	74.20	69.30
DopQ-ViT [36]	✓	3/3	54.72	65.76	44.71	59.26	74.91	74.77	69.63
AIQViT [39]	✓	3/3	41.32	43.68	38.51	55.36	66.15	71.42	63.01
QDrop* [33]	✓	3/3	44.26	75.72	51.63	53.13	74.42	74.77	76.83
APHQ-ViT* [27]	✓	3/3	63.81	76.25	55.57	68.61	76.15	76.01	78.01
FIMA-Q* [28]	✓	3/3	63.17	77.18	55.81	69.34	76.38	76.68	78.75
Ours	✓	3/3	**64.68**	**77.79**	**56.92**	**70.26**	**76.97**	**77.74**	**79.77**
PTQ4ViT [26]	×	4/4	42.57	30.69	36.96	34.08	64.39	76.09	74.02
RepQ-ViT [30]	×	4/4	65.05	68.48	57.43	69.03	75.61	79.45	78.32
AdaLog [35]	×	4/4	72.75	79.68	63.52	72.06	78.03	80.77	82.47
I&S-ViT [38]	✓	4/4	74.87	80.07	65.21	75.81	79.97	81.17	82.60
DopQ-ViT [36]	✓	4/4	75.69	80.95	65.54	75.84	80.13	81.71	83.34
AIQViT [39]	✓	4/4	70.63	74.15	62.33	72.75	79.19	80.93	81.22
QDrop* [33]	✓	4/4	72.63	82.67	65.95	74.02	80.12	81.27	83.07
APHQ-ViT* [27]	✓	4/4	76.31	82.44	66.53	76.49	80.10	81.50	83.26
FIMA-Q* [28]	✓	4/4	76.27	82.81	66.90	76.76	80.29	81.68	83.47
Ours	✓	4/4	**76.67**	**83.16**	**67.34**	**76.93**	**80.36**	**82.07**	**83.81**

Overall GPFA-Q Framework. The GPFA-Q framework integrates GPR with SGR as shown in Figure 3. GPR captures cross-dimensional interactions through gradient projections, while SGR improves consistency between optimization and inference by aligning the forward computation with discrete rounding. To complement the projection term with explicit per-dimension sensitivity, we introduce a diagonal prior based on gradient second moments:

$$\hat{\mathbf{f}} = \mathbb{E}[\mathbf{g} \odot \mathbf{g}]. \tag{15}$$

Based on this statistic, the diagonal prior term is defined as:

$$\mathcal{L}_{\text{Diag}} = \frac{1}{B} \sum_{i=1}^{B} \Delta \mathbf{z}_i^{\top} \operatorname{diag}(\hat{\mathbf{f}}) \Delta \mathbf{z}_i. \tag{16}$$

The final GPFA-Q objective is then given by

$$\mathcal{L}_{\text{GPFA}} = \mathcal{L}_{\text{GPR}} + \mathcal{L}_{\text{Diag}} + \lambda(t)\mathcal{L}_{\text{SGR}}, \tag{17}$$

where $\lambda(t)$ increases linearly from 0 to $\lambda_{\max}$ after a warmup phase, gradually encouraging alignment with discrete inference behavior. The overall GPFA-Q framework pipeline for block-wise reconstruction is summarized in Algorithm 1 in the appendix.

4 Experiments

4.1 Experimental Setup

Datasets and Models. We evaluate our method on the ImageNet dataset [37] for image classification. We employ representative visual backbones, including ViT [1], DeiT [40], and Swin [41]. To further verify the generalization capability of GPFA-Q, we extend our evaluation to downstream tasks on the COCO dataset [42]. We evaluate object detection and instance segmentation performance using Mask R-CNN [43] and Cascade Mask R-CNN [44] with the Swin backbone.

Table 2: Ablation study of the proposed GPFA-Q across various ViTs on ImageNet [37], reporting the top-1 accuracy (%). Baseline indicates only a diagonal approximation of the Fisher matrix.

W/A	Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
3/3	Baseline	44.26	75.72	51.63	53.13	74.42	74.77	76.83
	+ GPR	63.99	77.38	56.53	69.80	76.46	77.57	79.54
	+ SGR (**GPFA-Q**)	**64.68**	**77.79**	**56.92**	**70.26**	**76.97**	**77.74**	**79.77**
4/4	Baseline	72.63	82.67	65.95	74.02	80.12	81.27	83.07
	+ GPR	76.48	82.97	67.18	76.79	80.30	81.86	83.75
	+ SGR (**GPFA-Q**)	**76.67**	**83.16**	**67.34**	**76.93**	**80.36**	**82.07**	**83.81**

Implementation Details. Our implementation is built upon the QDrop [33] framework, and for diagonal-based quantization, we follow prior methods such as BRECQ [32]. All pretrained ViTs are obtained from the timm library, while object detection models are obtained from MMDetection [45]. All experiments are conducted on a single RTX 4090 GPU. Following prior PTQ methods [32, 33, 27, 28], we randomly sample 1,024 and 256 unlabeled images from the ImageNet and COCO datasets, respectively, as calibration sets. We use channel-wise uniform quantization for weights and layer-wise uniform quantization for activations, including post-Softmax and post-GELU activations. We set $\alpha = 32$ and $\lambda_{\max} = 0.5$. We strictly follow the original configurations and hyperparameters provided by the official implementations of each baseline method to ensure reproducibility and fair comparison.

4.2 Experimental Results

We evaluate top-1 accuracy across various ViTs on ImageNet [37] under 3-bit and 4-bit quantization in Table 1. We compare with recent PTQ methods including APHQ-ViT [27] and FIMA-Q [28]. Overall, our method consistently achieves the best performance across architectures and bit-widths. In 4-bit quantization, it maintains strong accuracy and outperforms all compared methods. In the more challenging 3-bit quantization, the performance gap becomes larger. This trend is consistently observed across different architectures, indicating the robustness of our method. Notably, our method improves over FIMA-Q and APHQ-ViT on DeiT-T, achieving 56.92 compared to 55.81 and 55.57, respectively, demonstrating its effectiveness in low-bit quantization. These results highlight that capturing cross-dimensional interactions while maintaining numerical stability is essential for reconstruction-based PTQ, and that the proposed method effectively addresses these challenges.

4.3 Ablation Study

Effect of the Number of Sampled Gradients α. We analyze the impact of the number of sampled gradients α on quantization performance and reconstruction time under 3-bit quantization, as shown in Figure 4, for DeiT [40] and ViT [1] models. Increasing α does not yield consistent accuracy improvements across models. This suggests that a small number of sampled gradients is already sufficient to capture the dominant gradient directions for block reconstruction. In contrast, the reconstruction time increases steadily with α, reflecting the higher computational cost observed when using more sampled gradients. Based on this observation, we choose $\alpha = 32$ as a practical default, which achieves a favorable balance between performance and computational cost.

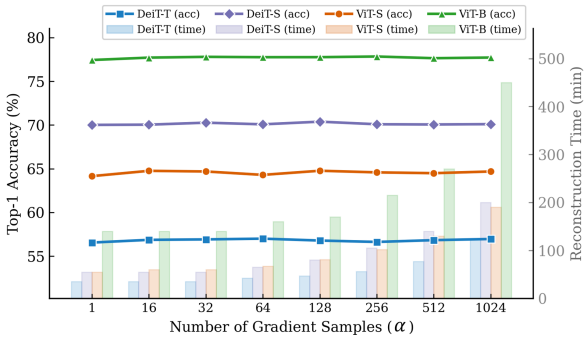


Figure 4: Effect of the number of sampled gradients α on top-1 accuracy and reconstruction time.

Effect of GPR and SGR. We examine the effect of GPR and SGR under 3-bit and 4-bit quantization in Table 2. We use a baseline that relies on a diagonal approximation of the Fisher matrix. Incorporating GPR results in substantial performance improvements, highlighting the importance of modeling cross-dimensional interactions for accurate reconstruction. Adding SGR to GPR yields

Table 3: Comparison of off-diagonal curvature modeling methods across various ViTs on ImageNet [37], reporting the top-1 accuracy (%). "LR-FIM" denotes the low-rank Fisher formulation used in FIMA-Q [28].

W/A	Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
3/3	LR-FIM (FIMA-Q)	63.21	77.06	54.99	69.06	76.17	76.35	78.83
	GPR (Ours)	64.08	77.43	56.52	69.68	76.61	77.50	79.50
	GPFA-Q (Ours)	**64.68**	**77.79**	**56.92**	**70.26**	**76.97**	**77.74**	**79.77**
4/4	LR-FIM (FIMA-Q)	76.23	82.63	66.86	76.50	80.12	81.54	83.36
	GPR (Ours)	76.53	83.03	67.18	76.83	80.32	81.93	83.55
	GPFA-Q (Ours)	**76.67**	**83.16**	**67.34**	**76.93**	**80.36**	**82.07**	**83.81**

Table 4: Comparison of block reconstruction time (in minutes) across different PTQ methods on various ViTs. All measurements are conducted on a single RTX 4090 GPU. Baseline uses only a diagonal approximation of the Fisher matrix.

Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
APHQ-ViT [27]	65	150	40	60	150	160	215
FIMA-Q [28]	55	130	35	55	130	160	200
Baseline	40	100	30	40	100	100	135
+ GPR	45	105	30	45	105	105	140
+ SGR (GPFA-Q)	50	125	35	50	125	125	180

consistent further improvements. While the gains are smaller compared to those from GPR, they demonstrate the benefit of improving optimization consistency with respect to discrete quantization. These results demonstrate that GPR provides the primary performance gains, while SGR offers complementary improvements, with the combined framework achieving the highest accuracy.

Effectiveness of Off-Diagonal Curvature Modeling. Table 3 compares different strategies for modeling off-diagonal curvature terms during block reconstruction. LR-FIM corresponds to the low-rank Fisher formulation adopted in FIMA-Q, while GPR denotes the proposed projection-based formulation. GPR consistently outperforms LR-FIM across all evaluated architectures, with gains observed under both low-bit settings. These results indicate that the proposed projection-based formulation captures cross-dimensional interactions more effectively than the low-rank approximation. By avoiding explicit matrix inversion while preserving informative off-diagonal structure, GPR provides a stronger reconstruction objective for low-bit PTQ.

Reconstruction Time Analysis. We compare block reconstruction time to evaluate the computational overhead of GPFA-Q against recent PTQ methods. Table 4 reports the reconstruction time of APHQ-ViT [27], FIMA-Q [28], and our variants on a single RTX 4090 GPU. Compared with the baseline diagonal formulation, introducing GPR increases runtime due to the additional gradient projection computations, while adding SGR incurs further overhead. Nevertheless, the GPFA-Q framework remains competitive with prior state-of-the-art methods. Across most architectures, GPFA-Q achieves reconstruction times comparable to FIMA-Q and consistently lower than or similar to APHQ-ViT. Combined with the superior accuracy reported in Table 1, these results demonstrate a favorable trade-off between reconstruction cost and quantization performance.

5 Conclusion

In this paper, we investigated reconstruction-based post-training quantization (PTQ) from a curvature modeling perspective and showed that accurate modeling of off-diagonal interactions is critical in low-bit quantization. We proposed GPFA-Q, which captures informative cross-dimensional interactions through gradient projections without explicit curvature matrix construction or inversion. The proposed Gradient-Projected Reconstruction (GPR) provides an effective surrogate objective for modeling off-diagonal interactions, while Soft Grid Rounding (SGR) improves consistency between reconstruction

optimization and discrete inference. Extensive experiments across diverse Vision Transformers (ViTs) demonstrate that GPFA-Q outperforms state-of-the-art PTQ methods under low-bit settings.

Limitations & Future Works. Despite the effectiveness of GPFA-Q, some limitations still remain. First, like most reconstruction-based PTQ methods, the current framework relies on fake quantization, where quantization is simulated rather than executed with integer arithmetic. As a result, a residual gap may remain between fake quantization and integer inference. Second, our current study focuses primarily on ViT-based architectures across standard vision tasks. Although the proposed method can be applied beyond ViTs, its effectiveness on other architectures and larger multimodal models remains to be further explored. In addition, applying the proposed projection-based curvature approximation to large-scale vision-language models and other efficient low-bit systems is a promising direction.

Acknowledgements

This work was supported by Samsung Electronics Co., Ltd(IO251218-14799-01) and the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2025-02263706, Development of an analog-digital mixed ultra-low power neuromorphic edge SoC, 25%) and the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (IITP-2026-RS-2023-00260091, 25%) and Institute of Information communications Technology Planning Evaluation (IITP) under the Artificial Intelligence Innovation Human Resources Development (IITP-RS-2026-25547954, 25%) grant funded by the Korea government(MSIT) and the K-On-Device AI Semiconductor RD Program (RS-2026-25620710, Development of On-Device AI Semiconductor-Based AI Algorithms, Intelligence Interconnection Framework, and Software Platforms for AI Operations, 25%) funded By the Ministry of Trade, Industry and Energy(MOTIE, Korea).

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[4] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

[5] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[6] Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602*, 2020.

[7] Minghao Fu, Hao Yu, Jie Shao, Junjie Zhou, Ke Zhu, and Jianxin Wu. Quantization without tears. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4462–4472, 2025.

[8] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[9] Fang Yu, Kun Huang, Meng Wang, Yuan Cheng, Wei Chu, and Li Cui. Width & depth pruning for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3143–3151, 2022.

[10] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Guo-Hua Wang, Yifan Ge, and Jianxin Wu. Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8183–8195, 2021.

[13] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

[14] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5785–5795, 2023.

[15] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[16] Beoungwoo Kang, Seunghun Moon, Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. Metaseg: Metaformer-based global contexts-aware network for efficient semantic segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 434–443, 2024.

[17] Hyunwoo Yu, Yubin Cho, Beoungwoo Kang, Seunghun Moon, Kyeongbo Kong, and Suk-Ju Kang. Embedding-free transformer with inference spatial reduction for efficient semantic segmentation. In *European Conference on Computer Vision*, pages 92–110. Springer, 2024.

[18] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2263–2271, 2023.

[19] Jongbeom Lee, Hyunwoo Yu, Jincheol Yang, Jaemin Choi, and Suk-Ju Kang. Sparstar: Sparse attention for spacetime autoregressive video synthesis. *arXiv preprint arXiv:2608.10519*, 2026.

[20] Seunghun Moon, Hyunwoo Yu, Haeuk Lee, and Suk-Ju Kang. Wimfris: Window mamba fusion and parameter efficient tuning for referring image segmentation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[21] Hyunwoo Yu, Jongbeom Lee, Hyunbin Kim, Sungjun Park, Jaemin Choi, Jincheol Yang, Seunghun Moon, Bongjoon Hyun, Kyomin Sohn, Kyeongbo Kong, et al. Openattention: One percent is enough with tail proxies. *arXiv preprint arXiv:2610.30951*, 2026.

[22] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

[23] Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 696–697, 2020.

[24] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

[25] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pages 7197–7206. PMLR, 2020.

[26] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European conference on computer vision*, pages 191–207. Springer, 2022.

[27] Zhuguanyu Wu, Jiayi Zhang, Jiaxin Chen, Jinyang Guo, Di Huang, and Yunhong Wang. Aphq-vit: Post-training quantization with average perturbation hessian based reconstruction for vision transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9686–9695, 2025.

[28] Zhuguanyu Wu, Shihe Wang, Jiayi Zhang, Jiaxin Chen, and Yunhong Wang. Fima-q: Post-training quantization for vision transformers by fisher information matrix approximation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14891–14900, 2025.

[29] Yunshan Zhong, Jiawei Hu, You Huang, Yuxin Zhang, and Rongrong Ji. Erq: Error reduction for post-training quantization of vision transformers. In *Forty-first International Conference on Machine Learning*, 2024.

[30] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17227–17236, 2023.

[31] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1173–1179, 2022.

[32] Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

[33] Wei, X., Gong, R., Li, Y., Liu, X., and Yu, F. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022.

[34] Jincheol Yang, Jaemin Choi, Matti Alexander Zinke, and Suk-Ju Kang. Qsca: Quantization with self-compensating auxiliary for monocular depth estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[35] Zhuguanyu Wu, Jiaxin Chen, Hanwen Zhong, Di Huang, and Yunhong Wang. Adalog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *European Conference on Computer Vision*, pages 411–427. Springer, 2024.

[36] Lianwei Yang, Haisong Gong, Haokun Lin, Yichen Wu, Zhenan Sun, and Qingyi Gu. Dopq-vit: Towards distribution-friendly and outlier-aware post-training quantization for vision transformers. *arXiv preprint arXiv:2408.03291*, 2024.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[38] Yunshan Zhong, Jiawei Hu, Mingbao Lin, Mengzhao Chen, and Rongrong Ji. I&s-vit: An inclusive & stable method for pushing the limit of post-training vits quantization. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 2025. doi: 10.1109/TPAMI.2025.3610466.

[39] Runqing Jiang, Ye Zhang, Longguang Wang, Pengpeng Yu, and Yulan Guo. Aiqvit: Architecture-informed post-training quantization for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17635–17643, 2025.

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021.

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[43] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[44] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[45] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arxiv 2019. *arXiv preprint arXiv:1906.07155*, 5, 1906.

Appendix

A Preliminaries

Uniform Quantization. We adopt the asymmetric uniform quantization scheme widely used in PTQ frameworks [27, 28, 32, 33]. Given a full-precision tensor $\mathbf{X}$ and a target bit-width b, the quantization and dequantization processes are formulated as:

$$\bar{\mathbf{X}} = \text{clip}\left(\left\lfloor \frac{\mathbf{X}}{s} \right\rceil + z, 0, 2^b - 1\right), \quad \hat{\mathbf{X}} = s\left(\bar{\mathbf{X}} - z\right), \tag{18}$$

where $\bar{\mathbf{X}}$ and $\hat{\mathbf{X}}$ denote the quantized integer tensor and the dequantized tensor, respectively. $\lfloor \cdot \rceil$ denotes rounding-to-nearest, and $\text{clip}(\cdot)$ restricts the quantized values to $[0, 2^b - 1]$. The scale s determines the quantization step size, and the zero-point z compensates for asymmetric ranges. These parameters are initialized using the min-max method:

$$s = \frac{\max(\mathbf{X}) - \min(\mathbf{X})}{2^b - 1}, \quad z = \left\lfloor \frac{-\min(\mathbf{X})}{s} \right\rceil. \tag{19}$$

Block Reconstruction. To mitigate quantization errors accumulated across layers, block reconstruction is widely adopted in PTQ methods [27, 28, 32, 33, 36, 38, 39]. Prior works motivate this approach by analyzing the loss increase caused by quantization using a second-order Taylor expansion. Let $\Delta\mathbf{W} = \hat{\mathbf{W}} - \mathbf{W}$ denote the weight perturbation induced by quantization, and define its vectorized form as $\Delta\mathbf{w} = \text{vec}(\Delta\mathbf{W})$. The change in the task loss can be approximated as:

$$\Delta\mathcal{L} \approx \Delta\mathbf{w}^\top \nabla_{\mathbf{w}}\mathcal{L} + \frac{1}{2}\Delta\mathbf{w}^\top \mathbf{H} \Delta\mathbf{w}, \tag{20}$$

where $\mathbf{H}$ denotes the Hessian matrix with respect to $\mathbf{w}$. For model weights, since a pretrained model is assumed to be close to a local optimum, $\nabla_{\mathbf{W}}\mathcal{L} \approx \mathbf{0}$. As in standard block reconstruction, we also drop the first-order term in activation space under the assumption that the quantization error has zero mean and is uncorrelated with the task-loss gradient. This makes the second-order curvature term the dominant factor. Optimizing this objective over the entire model is computationally expensive. Block reconstruction therefore reformulates the optimization at the block level by matching the outputs of each full-precision block and its quantized block. Consider the l-th transformer block $\mathcal{B}_l(\cdot; \mathbf{W}_l)$ and its quantized block $\hat{\mathcal{B}}_l(\cdot; \hat{\mathbf{W}}_l)$. Given calibration inputs $\mathbf{A} \sim \mathcal{D}$, we denote the full-precision and quantized block outputs as

$$\mathbf{z}_l = \mathcal{B}_l(\mathbf{A}; \mathbf{W}_l), \quad \hat{\mathbf{z}}_l = \hat{\mathcal{B}}_l(\mathbf{A}; \hat{\mathbf{W}}_l), \tag{21}$$

and define the block-output quantization perturbation as

$$\Delta\mathbf{z}_l = \hat{\mathbf{z}}_l - \mathbf{z}_l. \tag{22}$$

The baseline block reconstruction objective is then written as:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{A}\sim\mathcal{D}}\left[\|\Delta\mathbf{z}_l\|_2^2\right], \tag{23}$$

where $\boldsymbol{\theta}$ denotes the optimized quantization parameters, including the weight rounding variables and the activation quantization parameters. This objective defines the standard block reconstruction setting used throughout this work.

B Analysis of Expected Off-Diagonal Terms

Let $\mathbf{F} \in \mathbb{R}^{D\times D}$ denote a symmetric Fisher-based curvature matrix, and let $\Delta\mathbf{z} \in \mathbb{R}^D$ denote the quantization perturbation. Consider the quadratic form

$$Q(\Delta\mathbf{z}) = \Delta\mathbf{z}^\top \mathbf{F} \Delta\mathbf{z} = Q_{\text{diag}} + Q_{\text{off}}, \tag{24}$$

where

$$Q_{\text{diag}} = \sum_i F_{ii}\Delta z_i^2, \qquad Q_{\text{off}} = \sum_{i\neq j} F_{ij}\Delta z_i \Delta z_j.$$

The expected off-diagonal term over the calibration data is given by:

$$\mathbb{E}[Q_{\text{off}}] = \sum_{i \neq j} F_{ij}\, \mathbb{E}[\Delta z_i \Delta z_j]. \tag{25}$$

This quantity can be decomposed as:

$$\mathbb{E}[\Delta z_i \Delta z_j] = \text{Cov}(\Delta z_i, \Delta z_j) + \mathbb{E}[\Delta z_i]\mathbb{E}[\Delta z_j]. \tag{26}$$

When the perturbation mean is small, i.e., $\mathbb{E}[\Delta z_i] \approx 0$, we have the approximation

$$\mathbb{E}[\Delta z_i \Delta z_j] \approx \text{Cov}(\Delta z_i, \Delta z_j). \tag{27}$$

Therefore, this quantity is primarily determined by cross-dimensional perturbation covariance. In particular, assume that the perturbations across different dimensions satisfy

$$|\text{Cov}(\Delta z_i, \Delta z_j)| \leq \epsilon_{ij}, \qquad (i \neq j), \tag{28}$$

and $\mathbb{E}[\Delta z_i] \approx 0$. Using (25) and (27), we obtain the approximate bound

$$|\mathbb{E}[Q_{\text{off}}]| \;\lesssim\; \sum_{i \neq j} |F_{ij}|\, \epsilon_{ij}. \tag{29}$$

If additionally $\epsilon_{ij} \leq \epsilon$ uniformly for all $i \neq j$, then

$$|\mathbb{E}[Q_{\text{off}}]| \leq \epsilon \sum_{i \neq j} |F_{ij}|. \tag{30}$$

Equation (30) establishes an upper bound that inherently scales with $O(D^2)$ due to the summation over $D(D-1)$ off-diagonal pairs. Therefore, for the off-diagonal term to remain strictly bounded against dimensional scaling, the magnitude of the cross-dimensional Fisher elements must asymptotically shrink, satisfying $|F_{ij}| = O(1/D^2)$. Under high-bit quantization, the perturbation covariance bound ϵ is sufficiently small, which, combined with the asymptotic decay of $|F_{ij}|$, suppresses the expected off-diagonal impact. Conversely, when low-bit quantization induces larger and more structured perturbations, the covariance magnitudes $|\text{Cov}(\Delta z_i, \Delta z_j)|$ may increase across many pairs (i, j). In that case, the accumulated off-diagonal term can become significant, making diagonal-only curvature approximations less reliable. Therefore, diagonal approximations are better justified when perturbations are weakly correlated, whereas low-bit block reconstruction can benefit from objectives that explicitly account for cross-dimensional interactions.

C Numerical Instability of the Displacement Gram Matrix

During iterative block reconstruction, dominant quantization error directions are often gradually reduced. As reconstruction proceeds, the residual displacement vectors can become increasingly concentrated in a lower-dimensional subspace. Let

$$\Delta \mathbf{Z} = \begin{bmatrix} \delta_1^\top \\ \delta_2^\top \\ \vdots \\ \delta_k^\top \end{bmatrix} \in \mathbb{R}^{k \times D} \tag{31}$$

denote the displacement matrix, where D is the block-output dimension, k is the number of collected reconstruction steps, and $\delta_t \in \mathbb{R}^D$ denotes the residual block-output quantization error at step t. Suppose that these displacement vectors approximately lie in an r-dimensional subspace with $r \ll D$, i.e.,

$$\delta_t \approx \sum_{m=1}^{r} a_{t,m} \mathbf{u}_m, \qquad t = 1, \ldots, k, \tag{32}$$

where $\{\mathbf{u}_m\}_{m=1}^r$ are basis directions and $a_{t,m}$ are scalar coefficients. Then $\Delta \mathbf{Z}$ is approximately rank-r, and its trailing singular values $\sigma_{r+1}, \ldots, \sigma_{\min(k,D)}$ become small.

In the extreme case where all displacement vectors are nearly aligned,

$$\delta_t \approx \alpha_t \mathbf{u}, \qquad \mathbf{u} \in \mathbb{R}^D, \tag{33}$$

the displacement matrix becomes

$$\Delta\mathbf{Z} \approx \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} \mathbf{u}^\top, \tag{34}$$

which is rank-one. Therefore, as reconstruction proceeds, $\Delta\mathbf{Z}$ can become nearly rank-deficient. This degeneracy directly affects the conditioning of the Gram matrix used in low-rank curvature construction:

$$\mathbf{G}_\Delta = \Delta\mathbf{Z}\Delta\mathbf{Z}^\top \in \mathbb{R}^{k\times k}. \tag{35}$$

Let the singular value decomposition of $\Delta\mathbf{Z}$ be

$$\Delta\mathbf{Z} = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min(k,D)}), \tag{36}$$

where U and V are orthogonal matrices. Then

$$\mathbf{G}_\Delta = \Delta\mathbf{Z}\Delta\mathbf{Z}^\top = U\Sigma\Sigma^\top U^\top. \tag{37}$$

Hence, the spectral-norm condition number satisfies

$$\kappa_2(\mathbf{G}_\Delta) = \frac{\sigma_1^2}{\sigma_s^2} = \kappa_2(\Delta\mathbf{Z})^2, \quad s = \min(k, D), \tag{38}$$

Therefore, even a moderate increase in $\kappa_2(\Delta\mathbf{Z})$ is amplified quadratically in the Gram matrix. This is problematic because low-rank correction requires matrix inversion. For an invertible matrix A and perturbation matrix E,

$$\begin{gathered} (A+E)^{-1} - A^{-1} = -(A+E)^{-1}EA^{-1}, \\ \|(A+E)^{-1} - A^{-1}\| \le \|(A+E)^{-1}\|\,\|E\|\,\|A^{-1}\|, \\ \frac{\|(A+E)^{-1} - A^{-1}\|}{\|A^{-1}\|} \lesssim \kappa(A)\frac{\|E\|}{\|A\|}. \end{gathered} \tag{39}$$

Applying this result to $A = \mathbf{G}_\Delta$, we obtain that

$$(\Delta\mathbf{Z}\Delta\mathbf{Z}^\top)^{-1} \tag{40}$$

becomes highly sensitive to perturbations when G_Δ is ill-conditioned. Consequently, explicit low-rank curvature constructions that require inversion of G_Δ can become numerically unreliable during later stages of block reconstruction, when the displacement matrix is close to degenerate.

D Pseudo Algorithm

Algorithm 1 GPFA-Q Block Reconstruction

Input: Calibration data $\mathcal{D}_{\text{calib}}$, target block $\mathcal{B}$, quantized block $\hat{\mathcal{B}}$, number of sampled gradients α, warmup ratio τ, iterations T, and SGR schedule $\lambda(t)$.
Output: Optimized quantized block $\hat{\mathcal{B}}$.
Collect gradient projection basis
Compute full-precision block outputs $\mathbf{Z}$ using $\mathcal{B}$ and $\mathcal{D}_{\text{calib}}$.
Compute gradients of the KL divergence with respect to block outputs and sample α gradients to form $\mathbf{G}_\alpha$.
Compute the diagonal gradient statistic $\hat{\mathbf{f}}$.
# Optimize quantized block
for $t = 1, \ldots, T$ **do**
    Sample mini-batch $(\mathbf{X}_t, \mathbf{Z}_t)$.
    Compute soft-rounded output $\hat{\mathbf{Z}}_t^{\text{soft}}$ from $\hat{\mathcal{B}}$.
    Compute $\Delta\mathbf{Z}_t = \hat{\mathbf{Z}}_t^{\text{soft}} - \mathbf{Z}_t$.
    Calculate $\mathcal{L}_{\text{GPR}}$ using Eq. 11.
    Calculate $\mathcal{L}_{\text{Diag}}$ using Eq. 16.
    if $t > \tau T$ **then**
        Compute hard-rounded output $\hat{\mathbf{Z}}_t^{\text{hard}}$.
        Construct $\hat{\mathbf{Z}}_t^{\text{ste}}$ using Eq. 12.
    else
        Set $\hat{\mathbf{Z}}_t^{\text{ste}} = \hat{\mathbf{Z}}_t^{\text{soft}}$.
    end if
    Compute $\Delta\mathbf{Z}_t^{\text{ste}} = \hat{\mathbf{Z}}_t^{\text{ste}} - \mathbf{Z}_t$.
    Calculate $\mathcal{L}_{\text{SGR}}$ using Eq. 14.
    Calculate $\mathcal{L}_{\text{GPFA}}$ using Eq. 17.
    Update AdaRound [25] rounding variables and activation scaling factors in $\hat{\mathcal{B}}$.
end for
return $\hat{\mathcal{B}}$.

E Additional Experimental Results

Table 5: Comparison results on COCO for the object detection and instance segmentation tasks. AP^b and AP^m indicate AP^{box} and AP^{mask}, respectively. "W/A" indicates that the bit-width of the weight and activation are W and A bits, respectively. "*" indicates the results are reproduced by using the official code."†" indicates the results reported in the original papers.

Method	W/A	Mask R-CNN				Cascade Mask R-CNN			
		Swin-T		Swin-S		Swin-T		Swin-S	
		AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AP^b	AP^m
Full-Precision	32/32	46.0	41.6	48.5	43.3	50.4	43.7	51.9	45.0
I&S-ViT [38]	4/4	37.5	36.6	43.4	40.3	48.2	42.0	50.3	43.6
DopQ-ViT [36]	4/4	37.5	36.5	43.5	40.4	48.2	42.1	50.3	43.7
AIQViT [39]	4/4	38.2	36.7	44.1	40.4	47.1	41.4	49.8	43.4
QDrop [33]	4/4	36.2	35.4	41.6	39.2	47.0	41.3	49.0	42.5
APHQ-ViT* [27]	4/4	38.8	**38.0**	$43.4_{44.1\dagger}$	**41.0**	**48.7**	42.5	50.3	43.7
FIMA-Q* [28]	4/4	38.7	37.7	$43.3_{\mathbf{44.2}\dagger}$	40.7	48.5	42.4	50.3	43.7
Ours	4/4	**39.0**	**38.0**	43.4	**41.0**	**48.7**	**42.6**	**50.5**	**43.9**

Additional Results on Object Detection and Instance Segmentation tasks. We further evaluate the proposed method on object detection and instance segmentation tasks using the COCO [42] benchmark. Table 5 reports the results under 4-bit quantization for Mask R-CNN [43] and Cascade

Table 6: Comparison of structural fidelity to the exact FIM using cosine similarity. Higher values indicate better alignment.

W/A	Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B
3/3	FIMA-Q	51.5	62.5	54.1	35.9	38.4
	GPFA-Q	62.1	82.5	73.0	57.8	62.8
4/4	FIMA-Q	49.0	67.9	56.8	45.7	44.8
	GPFA-Q	62.1	83.9	74.1	63.2	63.3

Mask R-CNN [44] with Swin [41] backbones. Overall, GPFA-Q achieves consistently competitive performance compared to prior reconstruction-based PTQ methods across both detection (AP^b) and segmentation (AP^m) metrics. In particular, it shows clear improvements on Cascade Mask R-CNN with Swin small, achieving 50.5 AP^b and 43.9 AP^m. These results demonstrate that the proposed method generalizes beyond image classification and remains effective for more complex downstream tasks, where accurate modeling of cross-dimensional interactions is also important. For Mask R-CNN with the Swin-S backbone, we observe discrepancies between reproduced results and those reported in the original papers, and we mark such cases with †.

Structural Fidelity Analysis via Cosine Similarity. We measure the cosine similarity between the approximated Fisher matrix $\hat{\mathbf{F}}$ and the exact empirical FIM $\mathbf{F}_{\text{exact}}$. The exact matrix $\mathbf{F}_{\text{exact}}$ is computed using per-sample gradients from the `[CLS]` token representation. Cosine similarity is evaluated for each transformer block and averaged across all blocks. As shown in Table 6, GPFA-Q consistently achieves higher similarity than FIMA-Q across models and bit-widths. This indicates that GPFA-Q better preserves the curvature structure by maintaining gradient directions without destructive absolute value operations.

F Additional Visual Comparison of FIM Approximations

Building on the structural fidelity analysis in Appendix E, this result is consistent with the qualitative observations in Figure 1, where GPFA-Q preserves the off-diagonal patterns of the exact FIM, while FIMA-Q fails to retain them. To further validate this observation, we visualize the full FIM across all dimensions in Figure 5, extending the partial visualization shown earlier. While the earlier figure presents a subset for clarity, this full visualization confirms that the rich off-diagonal curvature structure persists across the entire dimensional space. Consistent with the quantitative results, GPFA-Q better preserves these structures than FIMA-Q.

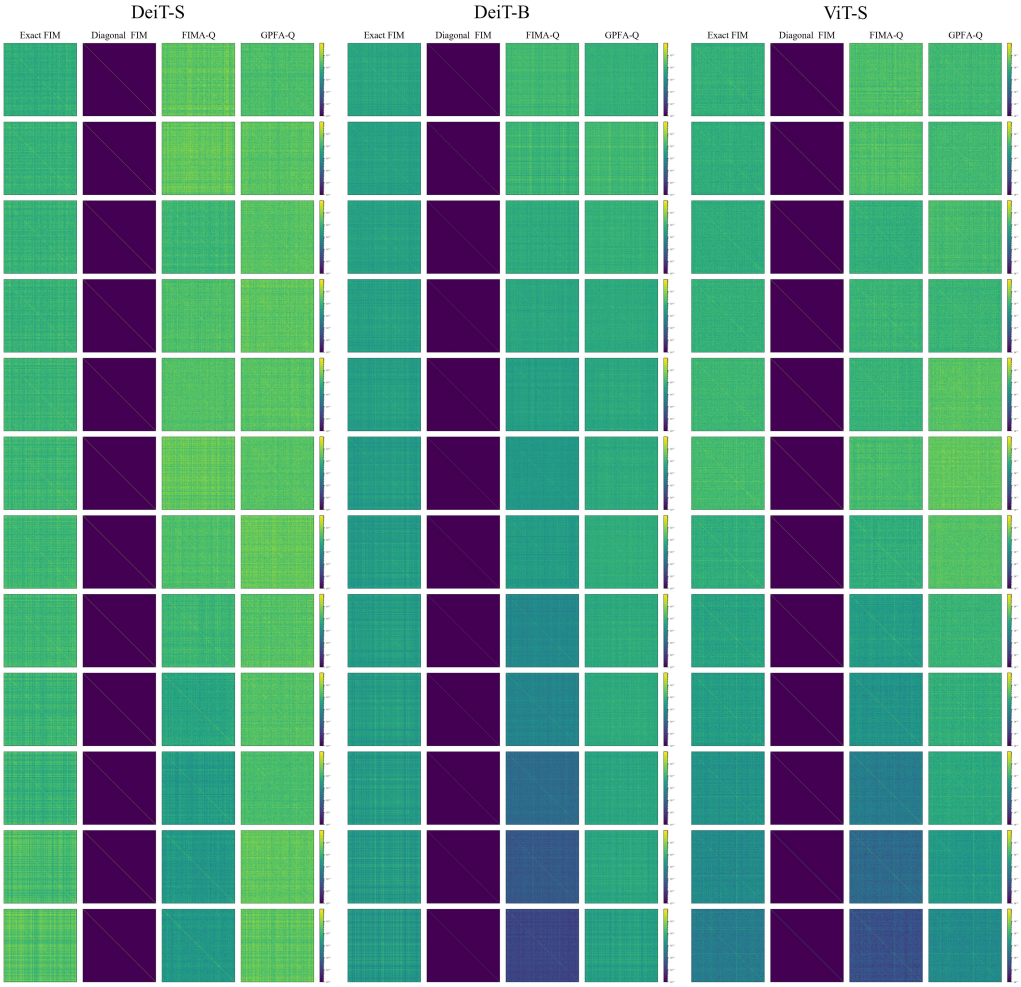


Figure 5: Full-dimensional visualization of FIM structures across models, showing comparisons between Exact FIM, Diagonal FIM, FIMA-Q, and GPFA-Q.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the contributions of GPFA-Q, including gradient-projected reconstruction and its ability to model off-diagonal interactions without explicit curvature construction. The claims are consistent with both theoretical motivation and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses practical limitations of the proposed method. In particular, it highlights the gap between fake quantization and integer inference, which may introduce residual discrepancies during deployment. It also acknowledges that the current evaluation focuses primarily on ViT-based architectures, and that further validation on other architectures and large-scale multimodal models remains for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results such as theorems or lemmas requiring complete proofs. Instead, it provides analytical derivations and reformulations of the Fisher-based objective (e.g., Eq. (6)–(11)) to motivate the proposed method. Additional supporting analyses are included in the appendix (e.g., expected off-diagonal behavior and conditioning of the Gram matrix), but these are intended to provide intuition and empirical justification rather than formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details to reproduce the main experimental results. It clearly specifies the datasets (ImageNet and COCO), model architectures (ViT, DeiT, Swin), and evaluation protocols. The quantization setting, including bit-width configurations, calibration dataset size, and quantization schemes (channel-wise for weights and layer-wise for activations), is fully described. Key hyperparameters such as the number of sampled gradients ($\alpha = 32$) and the SGR scheduling are also provided. In addition, the reconstruction procedure is explicitly described in Algorithm 1, including the optimization steps and loss components, enabling reproduction of the proposed method without requiring access to the original code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not release code or additional data at submission time. However, all experiments are conducted using publicly available datasets (ImageNet and COCO) and existing frameworks (e.g., timm and MMDetection). Detailed descriptions of the experimental setup, quantization configurations, and implementation procedures are provided to facilitate reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient experimental details to understand and interpret the results. It specifies the datasets (ImageNet and COCO), model architectures (ViT, DeiT, Swin), and evaluation metrics (Top-1 accuracy). The quantization settings, including bit-width configurations, calibration dataset size, and quantization schemes, are clearly described. Key hyperparameters such as the number of sampled gradients and reconstruction settings are also provided. Additional implementation details are included in the appendix to further support clarity.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports deterministic evaluation metrics (Top-1 accuracy) without error bars or statistical significance analysis. The experiments follow standard evaluation protocols in post-training quantization, where results are obtained from fixed pretrained models without stochastic training procedures. Therefore, variability across runs is minimal, and results are directly comparable under identical settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies that all experiments are conducted on a single NVIDIA RTX 4090 GPU, providing a clear description of the compute hardware used. While detailed runtime or total compute estimates are not explicitly reported, the experimental setup follows standard post-training quantization protocols with moderate computational requirements. The provided information is sufficient to reproduce the experiments under comparable hardware settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It does not involve human subjects, sensitive data, or high-risk applications. All experiments are conducted on publicly available datasets using standard evaluation protocols, and no ethical concerns specific to this work are identified.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a dedicated discussion of broader societal impacts. The work focuses on improving the efficiency of Vision Transformers through post-training quantization, and its potential impacts are generally aligned with those of existing research on model compression and efficient deep learning.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not introduce or release new datasets, pretrained models, or systems that pose a high risk for misuse. The work focuses on post-training quantization methods applied to existing publicly available models and datasets, and therefore does not require specific safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available datasets and frameworks, including ImageNet, COCO, timm, and MMDetection, all of which are properly cited. The usage of these assets follows their respective terms of use, and appropriate credit is given to the original sources.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release any new datasets, models, or code assets. Therefore, documentation of new assets is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, and therefore IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are not used as part of the core methodology or experimental pipeline in this research. Any potential use of LLMs is limited to writing or editing assistance and does not affect the scientific content or conclusions of the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.