Structural Outlier-Aware Post-Training Quantization for Monocular Depth Estimation

Anonymous Authors[1]

Abstract

ViT-based monocular depth estimation (MDE) models achieve strong accuracy but remain costly to deploy, motivating post-training quantization (PTQ). Existing PTQ methods are not well aligned with dense depth prediction, and recent depth-specific methods provide limited insight into which operators fail and why. We analyze 4-bit PTQ failure across operators using output sensitivity to clipping and cross-input range stability, showing that activation outliers play three distinct roles: range-dominating, signal-bearing, and input-dependent. Based on this analysis, we propose ORA-Q, an operator-role-aware PTQ framework that assigns each operator its grouping granularity, calibrator, and static or dynamic scaling mode. ORA-Q keeps most operators static and applies dynamic scaling only to input-dependent ranges. Experiments on Depth Anything variants and depth benchmarks show that ORA-Q consistently outperforms prior PTQ methods, improving δ_1 by 0.106 on average under 4-bit quantization.

1. Introduction

Monocular depth estimation (MDE) predicts a depth value for every pixel from a single RGB image (Birkl et al., 2023; Eigen et al., 2014), serving as a fundamental component of robotics, augmented reality, and 3D scene understanding (Kerbl et al., 2023; Mildenhall et al., 2021; Wofk et al., 2019). Foundation models such as Depth Anything v1 and v2 (Yang et al., 2024a;b) achieve strong zero-shot generalization, but their memory footprint and computational cost make deployment difficult. A practical solution is quantization, which replaces full-precision weights and activations with low-bit integers to reduce memory and arithmetic cost.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Quantization is commonly divided into quantization-aware training (QAT) and post-training quantization (PTQ). QAT can recover accuracy by retraining with quantization, but it requires training data and substantial retraining cost. PTQ instead determines quantization parameters from a small unlabeled calibration set, making it more practical, but low-bit PTQ often suffers from severe accuracy degradation, motivating various methods (Yuan et al., 2022; Li et al., 2023; Wu et al., 2024; Zhong et al., 2024; Fu et al., 2024).

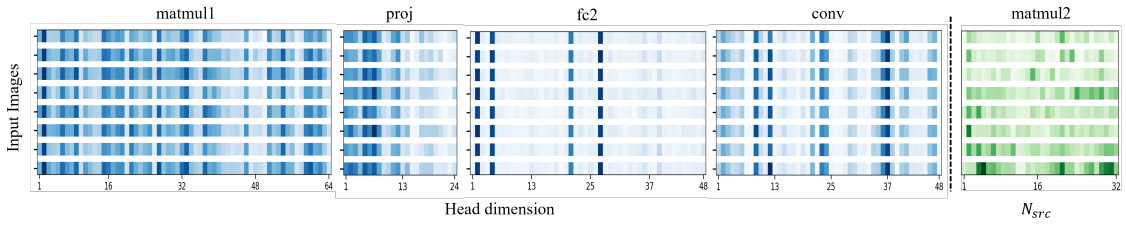
However, existing PTQ methods are often optimized for image classification and remain ill-suited to MDE, where dense predictions require fine-grained spatial structure. MDE-specific methods such as QSCA (Yang et al., 2025) and QuartDepth (Shen et al., 2025) improve accuracy through learned compensation, activation polishing, or weight compensation. Yet they mainly correct output errors, leaving unclear which operators fail and why different quantization strategies are needed.

We take a different perspective by analyzing where quantization errors arise, instead of addressing them afterward. This analysis shows that accuracy loss cannot be attributed to a single outlier type addressed by a unified strategy. Some outliers mainly increase the activation range and can be clipped with little impact, others contain important information and are sensitive to clipping, and another type exhibits input-dependent ranges that require dynamic scaling. To distinguish these roles, we use two analyses, output sensitivity to clipping and cross-input range stability. Together, they assign each operator to an outlier role that guides the quantization strategy.

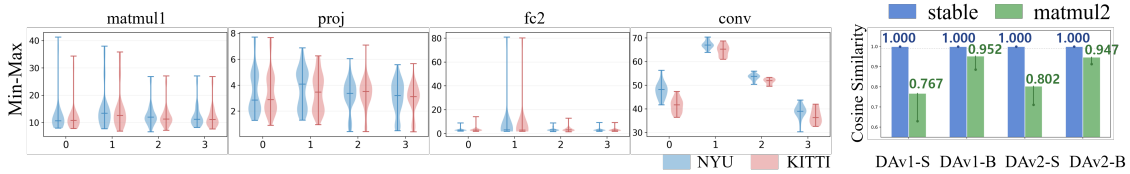
Building on this analysis, we propose ORA-Q, an outlier-role-aware framework for low-bit PTQ in Depth Anything (Yang et al., 2024a;b). ORA-Q diagnoses the functional role of outliers in each operator and assigns the corresponding grouping granularity, calibrator, and static or dynamic scaling mode. This replaces a single calibration strategy across the model and enables effective handling of both clip-insensitive and input-dependent behaviors within a unified framework.

Our main contributions are as follows:

- We analyze low-bit PTQ limitations in Depth Anything



(a) Cross-input stability



(b) Group/Dimension Index

(c) Cross-model consistency

Figure 1. Range stability across operators. (a) Per-group range heatmaps across calibration images. (b) Per-group range distributions on NYU and KITTI. (c) Cosine similarity of per-group range vectors across four Depth Anything variants.

Table 1. Operator selection under 4-bit quantization, averaged over four Depth Anything variants. (a) Individual restore starts from a fully 4-bit model and restores one operator to full precision. Others averages proj, fc2, qkv, and fc1. (b) Group-wise calibration effect after selecting matmul1, matmul2, and conv.

Config.	KITTI	NYU
All 4-bit	–	–
+ matmul2	+0.1230	+0.0456
+ conv	+0.0672	+0.0407
+ matmul1	+0.0132	+0.0158
+ others	-0.0036	-0.0009

(a) Individual restore

Config.	KITTI	NYU
base	–	–
+ fc2	+0.0112	+0.0057
+ proj	+0.0129	+0.0051
+ qkv	-0.0199	-0.0145
+ fc1	-0.0150	-0.0134

(b) Group-wise calibration

at the operator level and show that different operators contain outliers with functionally different roles, rather than a single shared degradation pattern.

- We propose ORA-Q, an outlier-role-aware framework that assigns each operator its grouping granularity, calibrator, and choice between static and dynamic quantization. Most operators become static once their outlier role is identified, while the attention-value product remains the primary case requiring dynamic scaling.
- We conduct experiments across diverse Depth Anything variants and depth benchmarks, demonstrating that ORA-Q consistently outperforms prior PTQ methods under 4-bit quantization.

2. Analyzing Outlier Roles in Depth Anything

In this section, we analyze how activation outliers behave across operators in Depth Anything (Yang et al., 2024a;b) to motivate our operator-specific quantization strategy. We first identify the operators that contribute most to quantization-induced degradation. For these operators, we then examine whether their quantization ranges are stable or input-dependent, and measure sensitivity to outlier clipping.

For clarity, we use short labels for the five operator classes analyzed throughout the paper: the attention score product $QK^\top$ (matmul1), the attention-value product AV (matmul2), the second FFN linear (fc2), the attention output projection (proj), and decoder convolutions (conv).

2.1. Operator Selection

To identify which operators contribute most to the accuracy drop under 4-bit quantization, we start from a fully 4-bit model and restore one operator at a time to full precision. As shown in Table 1(a), averaged over four Depth Anything variants, restoring matmul2 yields the largest recovery on both datasets, followed by conv and matmul1. The remaining operators, qkv, fc1, fc2, and proj, have negligible individual effects on average.

We then evaluate whether these remaining operators benefit from finer-grained calibration. Starting from a 4-bit base in which matmul1, matmul2, and conv are quantized, Table 1(b) shows that group-wise calibration improves δ_1 for fc2 and proj on both KITTI and NYU, but lowers δ_1 for qkv and fc1. We therefore focus on five operators, matmul1, matmul2, conv, fc2, and proj.

2.2. Structural Analysis

We examine whether each quantization axis is stable or varies across inputs. A stable axis can use static scales fixed during calibration, while an input-dependent axis cannot because its per-group range varies across inputs. For each group g and calibration image i, we compute

$$r_g^{(i)} = \max_j x_{g,j}^{(i)} - \min_j x_{g,j}^{(i)}, \tag{1}$$

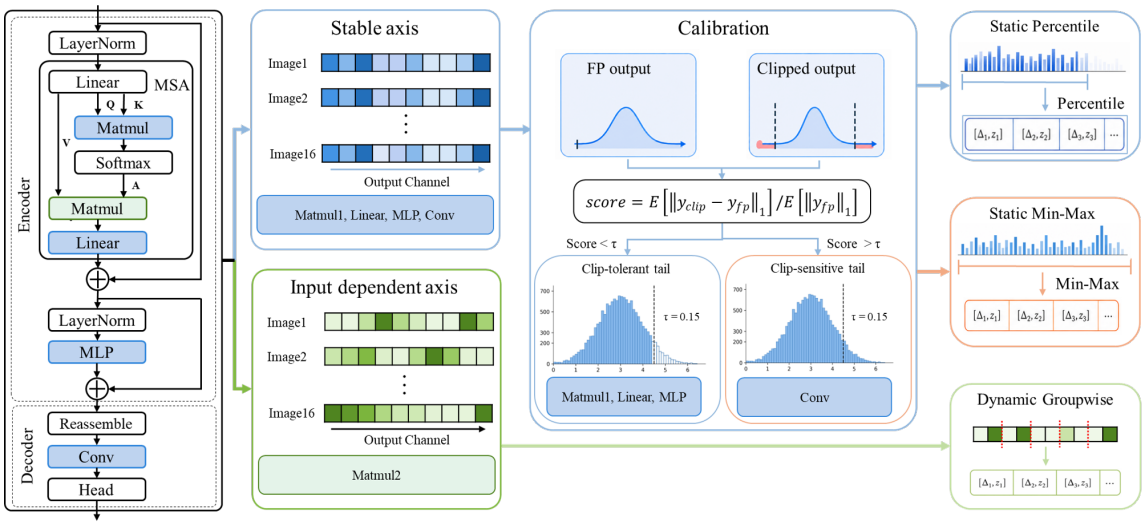


Figure 2. Pipeline of ORA-Q. Structural analysis separates stable and input-dependent axes, and distributional analysis further separates stable operators into clip-tolerant and clip-sensitive groups. Each group is assigned one quantization strategy, static percentile, static min-max, or dynamic groupwise.

where j indexes elements along the quantization axis. The stability score is the average per-group log-range standard deviation:

$$S = \frac{1}{G}\sum_{g=1}^{G} \mathrm{std}_i\left(\log r_g^{(i)}\right), \tag{2}$$

where G is the number of groups. A small S indicates stable axes, while a large S indicates input-dependent ones. The logarithm captures multiplicative range variation across different absolute scales.

Figure 1 analyzes the stability of per-group activation ranges. The four stable operators show nearly identical range patterns across KITTI images on DAv2-B and similar range distributions on NYU and KITTI, indicating that these patterns are not dataset-specific. The same trend holds across Depth Anything variants. As shown in Table 2 and Figure 1(c), the four stable operators achieve cross-model cosine similarities close to one, while matmul2 ranges from 0.767 to 0.952. We therefore use static scales fixed at calibration for the four stable operators and dynamic scales computed at runtime for matmul2.

2.3. Distributional analysis

For the four stable-axis operators, we further examine whether their outlier tails contain meaningful information. We clip the top 0.5% of activation values in the full-precision model and measure the resulting change in δ_1. An operator is classified as clip-tolerant if the change is negligible, and clip-sensitive otherwise. Table 2 reports representative values on DAv1-B and KITTI.

Among the four stable-axis operators, matmul1, fc2, and proj are clip-tolerant, since top-0.5% clipping causes negligible changes in δ_1. This suggests that their outlier tails primarily dominate the per-group range rather than carry depth-relevant information. In contrast, conv is clip-sensitive, with a substantial clipping-induced error. Its activations show a near-Gaussian distribution with no significant outlier tail, indicating that the conv operator performs depth-relevant computations and should be preserved. Table 2 summarizes the resulting per-operator quantization strategy.

3. Method

ORA-Q turns the per-operator analysis in Section 2 into a calibration pipeline that assigns each operator a grouping granularity, a calibration strategy, and a static or dynamic scaling mode. Figure 2 provides an overview of this decision flow.

3.1. Preliminary

Uniform quantization. We adopt the asymmetric uniform quantization scheme widely used in PTQ frameworks (Li et al., 2021; Wei et al., 2022; Yuan et al., 2022; Yang et al., 2025). Given a full-precision tensor X and bit-width b, quantization and dequantization are formulated as

$$\begin{aligned} \bar{X} &= \mathrm{clip}\big(\lfloor X/\Delta \rceil + z,\, 0,\, 2^b - 1\big), \\ \hat{X} &= \Delta\,(\bar{X} - z). \end{aligned} \tag{3}$$

Here, $\bar{X}$ and $\hat{X}$ denote the quantized integer tensor and the dequantized tensor. The scale Δ determines the step size, and the zero-point z compensates for asymmetric ranges. A min-max choice initializes them as

$$\Delta = \frac{\max(X) - \min(X)}{2^b - 1}, \qquad z = \left\lfloor -\frac{\min(X)}{\Delta} \right\rceil. \tag{4}$$

Table 2. Operator-level diagnosis of Depth Anything under 4-bit quantization and the resulting quantization strategy. Cosine similarities and $\Delta\delta_1$ are measured on DAv1-B (KITTI). Matmul2 diagnostics are omitted because its input dependence is architectural.

Operator	Role	Cosine		$\Delta\delta_1$ (top-0.5% clip)	Quantization strategy		
		Cross-dataset	Cross-model		Granularity	Calibrator	Status
matmul1	Clip-tolerant	0.9999	1.0000	−0.001	Per-dim (d=64)	Percentile	Static
fc2	Clip-tolerant	0.9989	0.9998	−0.001	Per-group (g=64)	Percentile	Static
proj	Clip-tolerant	0.9988	0.9996	−0.002	Per-group (g=32)	Percentile	Static
conv	Clip-sensitive	0.9989	0.9868	−0.116	Per-group (g=16)	Min-max	Static
matmul2	Input-dependent	–	–	–	Per-pair (g=2)	Min-max	Dynamic

Table 3. 4-bit quantization results on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012) for zero-shot relative depth estimation. Best results highlighted in bold.

	Method	W/A	Depth Anything v1 (Yang et al., 2024a)				Depth Anything v2 (Yang et al., 2024b)			
			E. ViT-S		*E.* ViT-B		*E.* ViT-S		*E.* ViT-B	
			$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓
NYUv2	FP	32/32	0.9720	0.0525	0.9791	0.0459	0.9736	0.0513	0.9770	0.0460
	MinMax (Gholami et al., 2022)	4/4	0.5024	0.2728	0.1972	1.5735	0.4873	0.2815	0.1102	2.3712
	Percentile (Li et al., 2019)	4/4	0.6542	0.2006	0.5430	0.2522	0.6675	0.1935	0.5379	0.2543
	BRECQ (Li et al., 2021)	4/4	0.5395	0.2535	0.4692	0.2886	0.5042	0.2714	0.4646	0.2910
	QDrop (Wei et al., 2022)	4/4	0.7166	0.1742	0.5785	0.2334	0.7115	0.1773	0.5794	0.2332
	PTQ4ViT (Yuan et al., 2022)	4/4	0.5693	0.2393	0.5895	0.2294	0.5034	0.2735	0.5183	0.2694
	RepQ-ViT (Li et al., 2023)	4/4	0.6639	0.1959	0.5410	0.2539	0.6464	0.2016	0.5465	0.2507
	ERQ (Zhong et al., 2024)	4/4	0.7062	0.1785	0.6126	0.2182	0.7140	0.1751	0.4705	0.2883
	QwT (Fu et al., 2024)	4/4	0.8007	0.1407	0.6486	0.2024	0.8050	0.1400	0.6589	0.1992
	QSCA (Yang et al., 2025)	4/4	0.8097	0.1377	0.6845	0.1875	0.8151	0.1361	0.6845	0.1875
	Ours	4/4	**0.8224**	**0.1331**	**0.8526**	**0.1218**	**0.8507**	**0.1200**	**0.8641**	**0.1192**
KITTI	FP	32/32	0.9369	0.0818	0.9396	0.0804	0.9340	0.0832	0.9389	0.0814
	MinMax (Gholami et al., 2022)	4/4	0.3441	0.3770	0.2058	1.9612	0.3423	0.3938	0.0832	4.4358
	Percentile (Li et al., 2019)	4/4	0.4099	0.3418	0.3327	0.3876	0.3780	0.3668	0.3275	0.3932
	BRECQ (Li et al., 2021)	4/4	0.3522	0.3719	0.3160	0.3989	0.3344	0.3906	0.3175	0.3990
	QDrop (Wei et al., 2022)	4/4	0.3234	0.3934	0.3338	0.3855	0.3748	0.3620	0.4082	0.3412
	PTQ4ViT (Yuan et al., 2022)	4/4	0.4106	0.3439	0.3251	0.3923	0.4187	0.3308	0.3200	0.3972
	RepQ-ViT (Li et al., 2023)	4/4	0.4159	0.3434	0.5410	0.2539	0.6464	0.2016	0.5465	0.2507
	ERQ (Zhong et al., 2024)	4/4	0.4847	0.3178	0.4241	0.3528	0.4616	0.3176	0.4066	0.3490
	QwT (Fu et al., 2024)	4/4	0.6862	0.1802	0.5867	0.2417	0.6941	0.1951	0.5346	0.2539
	QSCA (Yang et al., 2025)	4/4	0.7273	0.1874	0.6203	0.2365	0.6794	0.2067	0.6174	0.2296
	Ours	4/4	**0.7654**	**0.1641**	**0.6571**	**0.2100**	**0.8042**	**0.1472**	**0.7043**	**0.1883**

ORA-Q applies this quantizer per group. For stable operators, (Δ_g, z_g) is fixed at calibration, while for matmul2 it is computed at runtime.

3.2. Distributional classification

Section 2.3 classified stable-axis operators as clip-tolerant or clip-sensitive. To reproduce this classification during calibration, we compute a clipping score from the full-precision model. For each stable-axis operator and calibration image x_i, we run the model with the original operator and with the top 0.5% activations clipped to the 99.5th percentile, producing y_i^{fp} and y_i^{clip}. The score is

$$\text{score} = \frac{\mathbb{E}_i[\|y_i^{\text{clip}} - y_i^{\text{fp}}\|_1]}{\mathbb{E}_i[\|y_i^{\text{fp}}\|_1]}, \tag{5}$$

which measures the relative change in depth prediction. An operator is clip-tolerant if $\text{score} < \tau$ and clip-sensitive otherwise. We set $\tau = 0.15$, separating matmul1, fc2, and proj (0.03–0.10) from conv (0.24–0.40). The same classification holds across calibration sizes and all four Depth Anything variants.

3.3. Operator-specific quantization strategy

Calibration of stable operators. We calibrate (Δ_g, z_g) for each stable operator. The clip-tolerant operators matmul1, fc2, and proj use percentile calibration, while the clip-sensitive conv uses min-max calibration to preserve the full activation range. For percentile calibration, we search $p \in \{0.999, 0.9999, 0.99999\}$ and select the value that minimizes the squared error between original and dequantized activations. For conv, we set the range to $[\min(x), \max(x)]$

over calibration samples. Per-operator group sizes are chosen by the group-size sweep in Figure 4.

Dynamic per-pair quantization for matmul2. The structural analysis identifies matmul2 as the only operator with an input-dependent quantization axis, N_{src}. Since both operands $A \in \mathbb{R}^{N_q \times N_{\text{src}}}$ and $V \in \mathbb{R}^{N_{\text{src}} \times d}$ are indexed along this axis, their ranges are computed at runtime rather than fixed at calibration. Per-position quantization with g=1 gives unstable range estimates, while larger groups merge source positions with different attention patterns. We therefore use adjacent source-token pairs with g=2.

For each adjacent group $\mathcal{G}_k$, $A[:, \mathcal{G}_k]$ and $V[\mathcal{G}_k, :]$ use the same group boundary but are quantized independently:

$$\begin{aligned} \bar{X}_{\mathcal{G}_k} &= \text{clip}\big(\lfloor X_{\mathcal{G}_k}/\Delta_k \rceil + z_k,\, 0,\, 2^b - 1\big)\,, \\ \hat{X}_{\mathcal{G}_k} &= \Delta_k \left(\bar{X}_{\mathcal{G}_k} - z_k\right). \end{aligned} \tag{6}$$

All other operators use calibration-fixed static scales.

4. Experiments

Models, datasets, and metrics. We evaluate ORA-Q on Depth Anything (Yang et al., 2024a;b). NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012) serve as the main indoor and outdoor benchmarks, and we additionally report zero-shot results on Sintel (Butler et al., 2012), ETH3D (Schöps et al., 2017), and DIODE (Vasiljevic et al., 2019). Following standard practice (Yang et al., 2024a; 2025), we report threshold accuracy δ_1 with a threshold ratio of 1.25 and absolute relative error (AbsRel), where higher δ_1 and lower AbsRel indicate better depth estimation.

Implementation details. We quantize both weights and activations to 4 bits using an asymmetric uniform quantizer. ORA-Q uses 16 randomly sampled calibration images, fixes $\tau = 0.15$ across all settings, and requires no ground-truth depth. The pipeline runs on a single NVIDIA RTX 4090 GPU. The quantization granularity is per-dimension for matmul1 ($d = 64$) and group-wise for fc2 ($g = 64$), proj ($g = 32$), conv ($g = 8$), and matmul2 ($g = 2$).

4.1. Experimental results

Quantitative results. Table 3 reports 4-bit quantization results on NYUv2 and KITTI across Depth Anything variants. ORA-Q consistently outperforms prior PTQ methods, surpassing the previous state-of-the-art depth PTQ method, QSCA, by a clear margin across both datasets and model scales. The gain is largest on the Base models, consistent with prior observations that activation outliers in transformers become more prominent as model scale increases (Darcet et al., 2024; Dettmers et al., 2022).

Table 4. Ablation study on operator-wise calibrator assignment on DAv2-Base under 4-bit quantization for the NYU (Silberman et al., 2012) and KITTI (Geiger et al., 2012) datasets.

Configuration	KITTI δ_1	NYU δ_1
ORA-Q	**0.7043**	**0.8641**
conv: min-max → percentile	0.6512	0.8410
proj: percentile → min-max	0.6660	0.8388
fc2: percentile → min-max	0.4195	0.7710
All min-max	0.4156	0.7582
All percentile	0.6512	0.8410

Table 5. Progressive operator ablation on DAv2-B under 4-bit quantization. Base applies ORA-Q to matmul1, matmul2, and conv; remaining operators use per-tensor percentile calibration.

Config.	KITTI		NYUv2	
	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓
QSCA	0.6174	0.2296	0.6845	0.1875
Base	0.6751	0.1993	0.8291	0.1328
+ fc2	0.6829	0.1954	0.8446	0.1267
+ proj	0.6813	0.1966	0.8338	0.1313
Full	**0.6865**	**0.1930**	**0.8490**	**0.1253**

Qualitative results. Figure 3 compares depth predictions on NYUv2 and KITTI under the same 4-bit setting. ORA-Q better preserves scene structure, object boundaries, and depth discontinuities, while QSCA (Yang et al., 2025) produces visibly degraded depth maps with collapsed boundaries, blurred contours, and inconsistent depth in textured regions. Additional qualitative comparisons with 6-bit activations are provided in Appendix A.

4.2. Ablation studies

Calibrator assignment. Table 4 swaps the assigned calibrator of one stable operator at a time while leaving the rest of the pipeline unchanged. Switching the clip-sensitive conv from min-max to percentile reduces δ_1 by 0.035 on KITTI because percentile calibration clips the tail that carries depth-relevant information. The largest drop appears when fc2 is switched from percentile to min-max, where post-GELU outliers dominate the quantization range and the bulk of values lose precision. This confirms that the calibrator must be assigned per operator rather than shared across the model.

Progressive operator analysis. Table 5 applies operator-specific decisions one at a time on top of a base configuration that applies ORA-Q strategies to matmul1, matmul2, and conv, with the remaining operators kept at per-tensor percentile calibration. The base alone already surpasses QSCA on both KITTI and NYU, confirming that these three operators are the main drivers of the improvement under 4-bit quantization. The full ORA-Q configuration further

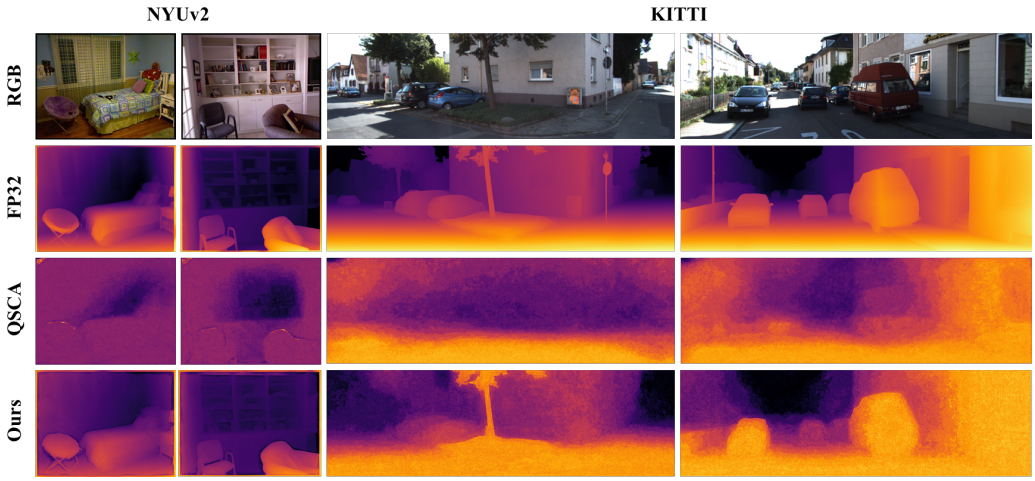


Figure 3. Qualitative results on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012) under 4-bit quantization. ORA-Q retains scene structure and object boundaries, while QSCA (Yang et al., 2025) produces visibly degraded depth maps.

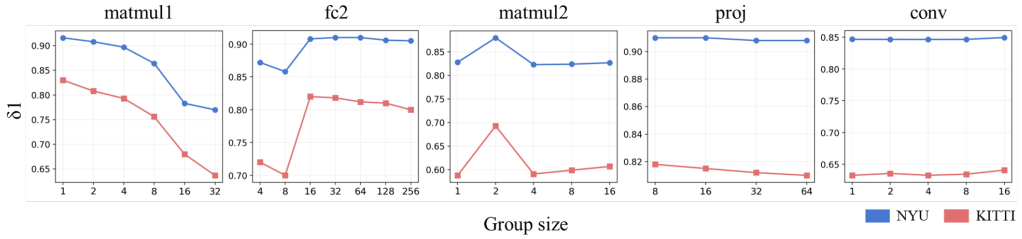


Figure 4. Group-size analysis on DAv2-B. matmul1 performs best with per-dimension quantization along the head dimension (d=64). fc2, proj, and conv plateau at g=64, g=32, and g=8, respectively, whereas matmul2 peaks sharply at g=2 and degrades with both smaller and larger groups.

adds complementary gains from fc2 and proj, achieving the highest δ_1 and the lowest AbsRel on both datasets.

4.3. Group size analysis

Figure 4 analyzes the group size of each calibrated operator while keeping the rest at their final ORA-Q settings. For matmul2, accuracy peaks sharply at $g = 2$, while $g = 1$ gives too few samples for reliable range estimation and $g \geq 4$ averages positions with different attention patterns. For matmul1, fc2, proj, and conv, accuracy plateaus after the selected group sizes. This motivates the group sizes reported in Table 2.

5. Conclusion

We presented Operator-Role-Aware Quantization (ORA-Q), a 4-bit post-training quantization framework for monocular depth estimation. By analyzing how quantization affects individual operators in Depth Anything models, we identified three distinct outlier roles and converted them into operator-specific choices of grouping granularity, calibration strategy, and scaling mode while maintaining a uniform quantizer. Across all examined Depth Anything models and five depth benchmarks, ORA-Q consistently outperforms prior PTQ methods.

Limitations and future work. Extending ORA-Q to larger backbones and Depth Anything 3 (Lin et al., 2025) is a natural next step. Depth Anything 3 adopts a broader any-view geometry formulation, and its monocular-specific variant is available only at the Large scale, so including it would mix changes in model family, formulation, and scale, making operator-wise analysis less controlled. ORA-Q is evaluated with fake quantization rather than hardware integer kernels, and measuring end-to-end latency on deployment hardware remains future work. Applying the same per-operator analysis to other depth-related tasks such as multi-view stereo, video depth estimation, and 3D reconstruction is another promising direction.

Impact Statement

This work aims to improve the practical deployment of monocular depth estimation models by reducing their computational and memory cost through post-training quantization. More efficient depth estimation can make foundation depth models more accessible in resource-constrained environments, including mobile, embedded, and robotic systems. However, monocular depth estimation is often used in safety-

relevant settings such as robotics and autonomous driving, where unreliable depth predictions may affect downstream decisions. Although ORA-Q improves the accuracy of 4-bit quantized models, a gap to full precision remains. Quantized depth models should therefore be validated carefully before deployment, especially in safety-critical applications. We expect the main impact of this work to be positive, as it advances efficient deployment of low-bit depth estimation models while highlighting the need for reliability-aware evaluation.

References

Birkl, R., Wofk, D., and Müller, M. MiDaS v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2022.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2014.

Fu, M., Yu, H., Shao, J., Zhou, J., Zhu, K., and Wu, J. Quantization without tears. *arXiv preprint arXiv:2411.13918*, 2024.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. A survey of quantization methods for efficient neural network inference. In *Low-power Computer Vision*, pp. 291–326. Chapman and Hall/CRC, 2022.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.

Li, R., Wang, Y., Liang, F., Qin, H., Yan, J., and Fan, R. Fully quantized network for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Li, Z., Xiao, J., Yang, L., and Gu, Q. RepQ-ViT: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Lin, H., Chen, S., Liew, J., Chen, D. Y., Li, Z., Shi, G., Feng, J., and Kang, B. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., and Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Shen, X., Ma, W., Liu, J., Yang, C., Ding, R., Wang, Q., Ding, H., Niu, W., Wang, Y., Zhao, P., Lin, J., and Gu, J. QuartDepth: Post-training quantization for real-time depth estimation on the edge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F. Z., Daniele, A. F., Mostajabi, M., Basart, S., Walter, M. R., et al. DIODE: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

Wei, X., Gong, R., Li, Y., Liu, X., and Yu, F. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Wofk, D., Ma, F., Yang, T.-J., Karaman, S., and Sze, V. FastDepth: Fast monocular depth estimation on embedded systems. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Wu, Z., Chen, J., Zhong, H., Huang, D., and Wang, Y. AdaLog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Yang, J., Choi, J., Zinke, M., and Kang, S.-J. QSCA: Quantization with self-compensating auxiliary for monocular depth estimation. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2025.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. Depth anything v2. *Advances in Neural Information Processing Systems*, 2024b.

Yuan, Z., Xue, C., Chen, Y., Wu, Q., and Sun, G. PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Zhong, Y., Hu, J., Huang, Y., Zhang, Y., and Ji, R. ERQ: Error reduction for post-training quantization of vision transformers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

A. Appendix

A. Additional Experimental Results

Additional quantitative results under W4A6. Table 7 and Table 6 report the W4A6 results on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012), and the 4-bit zero-shot results on Sintel (Butler et al., 2012), ETH3D (Schöps et al., 2017), and DIODE (Vasiljevic et al., 2019), respectively. Under W4A6, ORA-Q achieves the best δ_1 on all Base models across NYUv2 and KITTI, while remaining within a small margin of the strongest baseline on the Small models. This pattern is consistent with prior observations that activation outliers become more prominent as model scale increases (Darcet et al., 2024; Dettmers et al., 2022). The per-operator strategy provides the largest improvement where outliers are most prominent, which explains the stronger gains on Base models. On the Small variants under W4A6, the wider quantization range of 6-bit activations already absorbs most outlier-induced error, leaving less room for role-specific calibration to contribute. Under 4-bit quantization on the zero-shot benchmarks, ORA-Q maintains the same ordering observed on NYUv2 and KITTI across all three datasets.

Additional qualitative results under W4A6. We provide additional visual comparisons under W4A4, in Figure 5. Under W4A6, QSCA suffers from severe structural degradation, where object boundaries collapse and the depth ordering is no longer preserved. In contrast, ORA-Q produces depth maps that preserve the structural information of the scene, recovering edges and the relative depth ordering of objects.

Table 6. Fully 4-bit quantization zero-shot quantization results on Sintel (Butler et al., 2012), ETH3D (Schöps et al., 2017), and DIODE (Vasiljevic et al., 2019) for relative depth estimation. Baseline numbers are reproduced from QSCA (Yang et al., 2025). Best results in bold.

	Method	W/A	Depth Anything v1 (Yang et al., 2024a)				Depth Anything v2 (Yang et al., 2024b)			
			E. ViT-S		*E.* ViT-B		*E.* ViT-S		*E.* ViT-B	
			$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓
Sintel	FP	32/32	0.7304	0.2296	0.7539	0.2281	0.6974	0.2680	0.7111	0.2576
	MinMax	4/4	0.3217	0.4751	0.0069	54.2156	0.3101	0.4998	0.1727	6.1550
	Percentile	4/4	0.3585	0.4745	0.3654	0.4677	0.3772	0.4693	0.3237	0.4801
	BRECQ	4/4	0.3123	0.4787	0.3085	0.4820	0.3055	0.4816	0.3008	0.4821
	QDrop	4/4	0.3181	0.4991	0.3106	0.4867	0.3184	0.4948	0.3151	0.4882
	QSCA	4/4	0.3884	0.4531	0.4059	0.4781	0.3919	0.4474	0.4070	0.4679
	Ours	4/4	**0.4743**	**0.4068**	**0.5299**	**0.4545**	**0.5333**	**0.3898**	**0.5487**	**0.4047**
ETH3D	FP	32/32	0.9652	0.0584	0.9741	0.0513	0.9701	0.0548	0.9791	0.0467
	MinMax	4/4	0.5264	0.2754	0.2070	3.0313	0.5078	0.2872	0.0833	5.4311
	Percentile	4/4	0.6290	0.2186	0.5935	0.2380	0.6456	0.2148	0.5720	0.2491
	BRECQ	4/4	0.5082	0.2829	0.4874	0.2958	0.4962	0.2920	0.4870	0.2963
	QDrop	4/4	0.6069	0.2351	0.5373	0.2706	0.6186	0.2288	0.5341	0.2697
	QSCA	4/4	0.7332	0.1791	0.6309	0.2241	0.6791	0.1983	0.6298	0.2197
	Ours	4/4	**0.8156**	**0.1427**	**0.7991**	**0.1449**	**0.8451**	**0.1298**	**0.8001**	**0.1445**
DIODE	FP	32/32	0.9413	0.0753	0.9474	0.0745	0.9426	0.0721	0.9498	0.0701
	MinMax	4/4	0.6687	0.2110	0.1070	13.8008	0.6596	0.2154	0.1647	11.9492
	Percentile	4/4	0.7347	0.1862	0.6845	0.2038	0.7459	0.1830	0.7027	0.2019
	BRECQ	4/4	0.6755	0.2079	0.6464	0.2201	0.6538	0.2173	0.6441	0.2211
	QDrop	4/4	0.7275	0.1864	0.6819	0.2058	0.7279	0.1848	0.6777	0.2062
	QSCA	4/4	0.8033	0.1513	0.7099	0.1997	0.8135	0.1463	0.7094	0.1947
	Ours	4/4	**0.8498**	**0.1379**	**0.8757**	**0.1276**	**0.8588**	**0.1288**	**0.8791**	**0.1204**

Table 7. W4A6 quantization results on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012) for zero-shot relative depth estimation. Best results are highlighted in bold.

	Method	W/A	Depth Anything v1 (Yang et al., 2024a)				Depth Anything v2 (Yang et al., 2024b)			
			E. ViT-S		*E.* ViT-B		*E.* ViT-S		*E.* ViT-B	
			$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓
NYUv2	FP	32/32	0.9720	0.0525	0.9791	0.0459	0.9736	0.0513	0.9770	0.0460
	MinMax	4/6	0.5632	0.2417	0.4973	0.2748	0.5324	0.2599	0.4738	0.2866
	Percentile	4/6	0.8837	0.1050	0.9071	0.0958	0.9196	0.0891	0.9355	0.0831
	BRECQ	4/6	0.5786	0.2337	0.5542	0.2462	0.5269	0.2611	0.5084	0.2697
	QDrop	4/6	0.6369	0.2071	0.6987	0.1823	0.7285	0.1700	0.7619	0.1560
	PTQ4ViT	4/6	0.6187	0.2166	0.6122	0.2201	0.5159	0.2642	0.6095	0.2275
	RepQ-ViT	4/6	0.8256	0.1304	0.8867	0.1058	0.8657	0.1131	0.9158	0.0918
	ERQ	4/6	0.9075	0.0945	0.9390	0.0766	0.9245	0.0874	0.9152	0.0914
	QwT	4/6	0.9189	0.0884	0.9089	0.0928	0.9323	0.0833	0.8944	0.1000
	QSCA	4/6	**0.9333**	**0.0810**	0.9441	0.0739	**0.9450**	**0.0757**	0.9468	0.0726
	Ours	4/6	0.9091	0.0933	**0.9471**	**0.0726**	0.9293	0.0844	**0.9581**	**0.0667**
KITTI	FP	32/32	0.9369	0.0818	0.9396	0.0804	0.9340	0.0832	0.9389	0.0814
	MinMax	4/6	0.4467	0.3290	0.3586	0.3769	0.3609	0.3744	0.3161	0.3982
	Percentile	4/6	0.8558	0.1269	0.7580	0.1687	0.8613	0.1283	0.7219	0.1866
	BRECQ	4/6	0.5223	0.2851	0.4035	0.3529	0.4320	0.3332	0.4093	0.3497
	QDrop	4/6	0.6052	0.2408	0.5897	0.2611	0.5351	0.2790	0.5268	0.3037
	PTQ4ViT	4/6	0.4446	0.3230	0.3488	0.3848	0.3931	0.3614	0.3308	0.3917
	RepQ-ViT	4/6	0.7888	0.1502	0.5610	0.2602	0.8293	0.1292	0.8109	0.1458
	ERQ	4/6	0.8629	0.1244	0.8705	0.1133	0.8770	0.1144	0.8685	0.1115
	QwT	4/6	0.8844	0.1152	0.7292	0.1759	0.8872	0.1169	0.7700	0.1550
	QSCA	4/6	**0.8857**	**0.1161**	0.8722	0.1174	**0.8893**	**0.1124**	0.8873	0.1067
	Ours	4/6	0.8557	0.1269	**0.8735**	**0.1121**	0.8765	0.1155	**0.8999**	**0.0996**

B. Additional Ablation Studies

B.1. Calibration efficiency.

Table 8 compares the cost of ORA-Q with existing PTQ methods on Depth Anything. ORA-Q completes calibration in approximately 55 seconds for ViT-Small and 100 seconds for ViT-Base on a single RTX 4090, making it $4\times$ faster than QSCA (Yang et al., 2025) and more than $40\times$ faster than block-reconstruction methods such as BRECQ (Li et al., 2021) and QDrop (Wei et al., 2022). ORA-Q adds only 11.4 KB and 16.8 KB of per-group scales and zero-points for ViT-Small and ViT-Base, respectively, whereas QSCA introduces 0.44 M and 1.77 M trainable compensation parameters. Despite this training-free, low-overhead design, ORA-Q achieves the highest δ_1 and the lowest AbsRel.

B.2. Pipeline correctness and calibration robustness

We verify that ORA-Q reproduces the diagnosis of Section 2 across all settings tested.

Match with manual diagnosis. On six model-dataset configurations (DAv1-B and DAv2-B on KITTI and NYU, DAv1-S on KITTI, DAv2-S on NYU), the automated classification produces the same per-operator strategy as the manual diagnosis. The quantized accuracy differs by less than $0.002\,\delta_1$ in all six cases.

Calibration set size As noted in Section 3, the diagnostic outcome is robust to the calibration set size. Table 9 confirms this. Across 16, 32, 50, and 100 calibration images, the structural and distributional analyses produce the same role assignment for every operator and therefore the same per-operator quantization strategy. The classification cost grows roughly linearly with the number of images, from 23 seconds at 16 images to 77 seconds at 100 images. We use 16 images throughout because the classification is already saturated and the calibration cost is minimized.

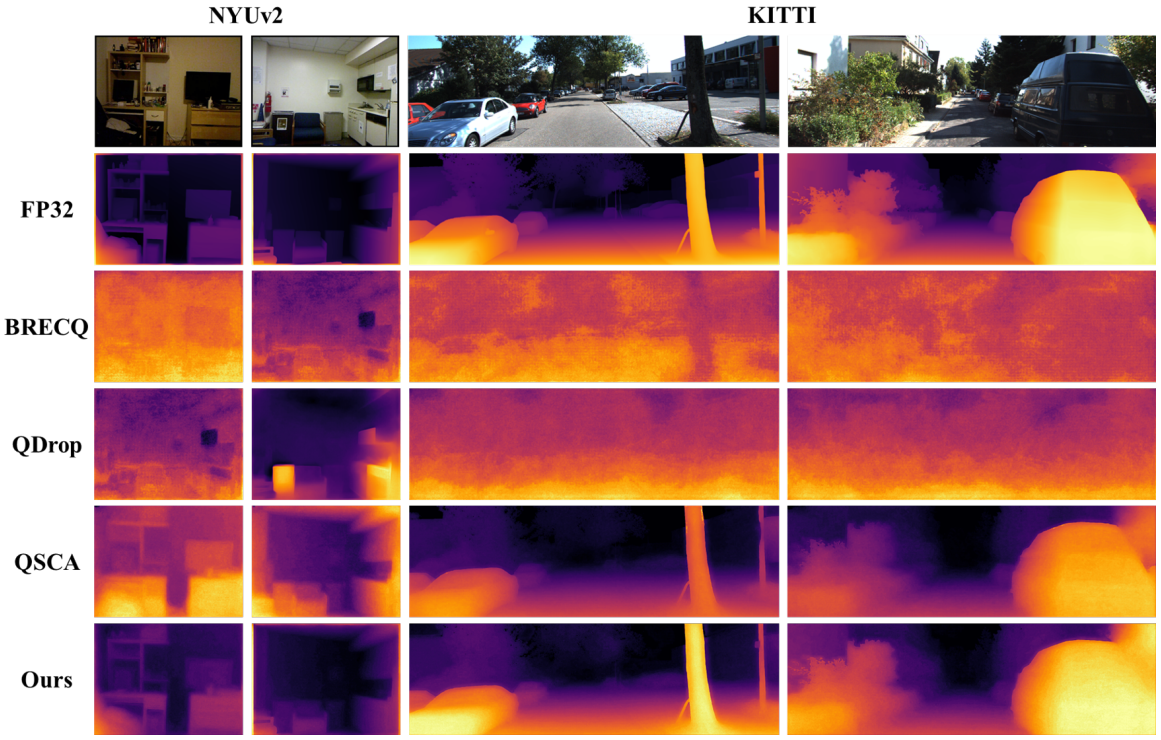


Figure 5. Qualitative comparison on NYUv2 under W4A6. QSCA suffers from structural degradation, while ORA-Q preserves edges and depth ordering.

Table 8. Efficiency and accuracy comparison on Depth Anything. Calibration times are measured on a single RTX 4090.

Backbone	Method	Params	Time	$\delta_1 \uparrow$	AbsRel↓
ViT-S	BRECQ (Li et al., 2021)	24.79 M	∼3157 s	0.5395	0.2535
	QDrop (Wei et al., 2022)	24.79 M	∼3678 s	0.7166	0.1742
	QSCA (Yang et al., 2025)	25.23 M	∼210 s	0.8097	0.1377
	ORA-Q (Ours)	24.80 M	∼55 s	**0.8224**	**0.1331**
ViT-B	BRECQ (Li et al., 2021)	97.47 M	∼4069 s	0.4692	0.2886
	QDrop (Wei et al., 2022)	97.47 M	∼5820 s	0.5785	0.2334
	QSCA (Yang et al., 2025)	99.24 M	∼411 s	0.6845	0.1875
	ORA-Q (Ours)	97.49 M	∼100 s	**0.8526**	**0.1218**

Table 9. Effect of calibration set size on the diagnostic outcome. Across all sizes, the role assignment is identical for every operator, so the quantization strategy is unchanged. We report fc2 and the decoder convolutions as the two representative cases since they sit on opposite sides of the threshold $\tau = 0.15$.

# Images	Classification time	fc2 role	conv role	Strategy match
16	23 s	clip-tolerant	clip-sensitive	✓
32	37 s	clip-tolerant	clip-sensitive	✓
50	53 s	clip-tolerant	clip-sensitive	✓
100	77 s	clip-tolerant	clip-sensitive	✓