Structure-Preserving Adaptive Post-Training Quantization for Monocular Depth Estimation

Jaemin Choi [1 *] **Jincheol Yang** [1 *] **Nahyun Lim** [1 *] **Yun-seong Jeong** [1 *] **Matti Alexander Zinke** [1 *]
Hyunwoo Yu [1 *] **Suk-Ju Kang** [1]

Abstract

Monocular Depth Estimation (MDE) foundation models such as Depth Anything achieve strong generalization across diverse scenes, but their high computational and memory costs hinder efficient deployment. Post-Training Quantization (PTQ) offers a practical compression strategy, yet low-bit PTQ for MDE remains challenging because quantizing query and key projections independently fails to preserve their induced attention maps, while accumulated quantization errors cause distribution shifts in intermediate features. To address these challenges, we propose **SPA-Q**, a **S**tructure-**P**reserving **A**daptive PTQ framework for MDE. SPA-Q introduces Attention-Preserving Calibration (APC), which calibrates quantization parameters by directly preserving the full-precision attention distributions, and Channel-Wise Distribution Alignment (CWDA), which mitigates quantization-induced feature distribution shifts through channel-wise affine transformations that are absorbed into the weights after training. Experiments on NYUv2 and KITTI show that SPA-Q consistently improves 4-bit quantization performance over existing PTQ methods, reducing AbsRel by 29.5% and improving δ_1 by 20.2% on average.

1. Introduction

Monocular depth estimation (MDE) is a fundamental task in computer vision, with applications in robotics (Dong et al., 2022; Wofk et al., 2019), autonomous driving (Weng & Kitani, 2019; Wang et al., 2019), and 3D scene understanding (Mildenhall et al., 2021; Kerbl et al., 2023). Recent advances in Vision Transformers (ViTs) (Dosovitskiy et al., 2020; Touvron et al., 2021; Caron et al., 2021; Oquab et al., 2023) have significantly improved MDE, leading to the development of strong foundation models such as MiDaS (Ranftl et al., 2020) and Depth Anything (Yang et al., 2024b;c). Depth Anything achieves strong zero-shot generalization through large-scale training with labeled and pseudo-labeled data, while leveraging multi-scale intermediate features for dense depth prediction. However, its large model size and high computational cost make it difficult to deploy in real-world, resource-constrained environments.

Model quantization (Gholami et al., 2022; Nagel et al., 2021; Wu et al., 2020) is an effective approach for reducing the memory and computational cost of deep neural networks. Existing quantization methods can be broadly divided into quantization-aware training (QAT) (Esser et al., 2019; Bhalgat et al., 2020; Choi et al., 2018) and post-training quantization (PTQ) (Li et al., 2021; Wei et al., 2022; Li et al., 2023). QAT incorporates quantization effects during training and often achieves high accuracy, but it requires additional training and access to training data. In contrast, PTQ quantizes a pre-trained model without full retraining, making it suitable for practical deployment. Although PTQ has been widely studied for recognition models, its application to MDE remains underexplored because depth estimation is particularly sensitive to distortions in fine-grained geometric and structural information carried by intermediate features.

In our study, we identify two key limitations of low-bit PTQ for transformer-based MDE models. **(i) Independent quantization error minimization for query and key projections is not well aligned with preserving the resulting attention distribution.** Existing PTQ methods (Li et al., 2021; Wei et al., 2022; Li et al., 2023) typically calibrate each quantizer by minimizing the discrepancy between the full-precision and quantized outputs of each projection independently. However, in self-attention, the attention map is determined by the query-key interaction passed through the softmax operation, rather than by the fidelity of each projection. Therefore, independently minimizing the quantization error of the query and key projections does not guarantee preserving the resulting attention map. This issue

[1]Department of Electronic Engineering, Sogang University, Seoul, South Korea.. Correspondence to: Suk-Ju Kang <sjkang@sogang.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

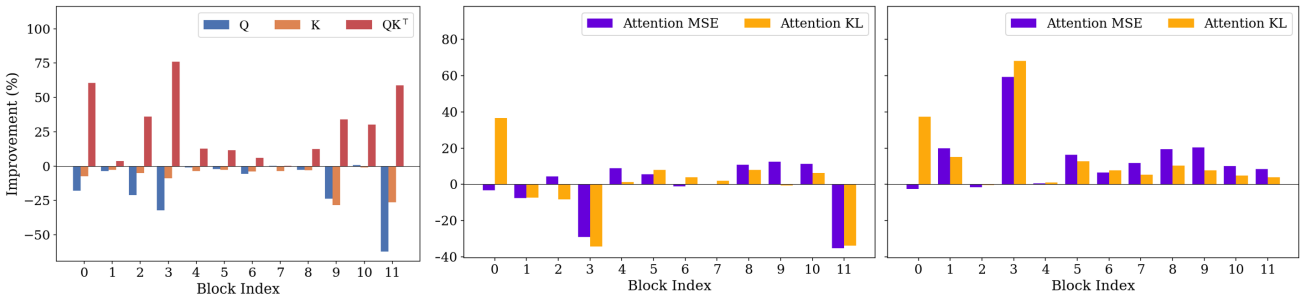


Figure 1. Comparison of calibration objectives for query-key quantization, where all improvements are measured relative to conventional local quantization error minimization. (a) MSE improvement from logits-level calibration. (b) Attention-level metric improvement from logits-level calibration. (c) Attention-level metric improvement from the proposed APC.

is especially critical for MDE, as distorted attention maps propagate to intermediate features and degrade the structural information delivered to the decoder. **(ii) Quantization errors accumulate across layers and induce distribution shifts in intermediate features.** Since the decoder fuses features from multiple encoder stages, even small shifts in feature distributions can propagate through the decoder and degrade depth prediction. In practice, low-bit quantization induces noticeable distribution shifts in intermediate features, making them increasingly mismatched with those of the full-precision model.

To address these limitations, we propose **SPA-Q**, a **S**tructure-**P**reserving **A**daptive PTQ framework for MDE. SPA-Q is a novel PTQ framework for transformer-based depth estimation models whose the decoder aggregates multi-stage encoder features for dense prediction. It consists of two complementary components. **Attention-Preserving Calibration (APC)** calibrates the query and key projections by directly preserving the full-precision attention distribution, rather than minimizing the quantization error of the two projections independently. In particular, APC minimizes the KL divergence between the full-precision and quantized attention distributions, thereby better accounting for their discrepancy in the normalized distribution after softmax. **Channel-Wise Distribution Alignment (CWDA)** addresses distribution shifts in intermediate features through learnable channel-wise affine transformations, and optimizes them with a two-step depth-aware optimization that separately aligns encoder features and decoder depth predictions. After training, the learned parameters are fused into the weights and biases. Our contributions are summarized as follows:

- We identify two key limitations of low-bit PTQ for transformer-based MDE models. First, independent quantization error minimization for query-key projections fails to preserve the resulting attention distribution. Second, accumulated quantization errors induce distribution shifts in intermediate features.
- We propose SPA-Q, a Structure-Preserving Adaptive PTQ framework for MDE. SPA-Q consists of Attention-Preserving Calibration (APC) for directly preserving the full-precision attention distribution during calibration and Channel-Wise Distribution Alignment (CWDA) for mitigating quantization-induced distribution shifts through learnable channel-wise affine transformations.
- Extensive experiments on relative depth estimation benchmarks show that SPA-Q consistently outperforms existing PTQ methods under 4-bit quantization across all evaluated settings.

2. Method

2.1. Attention-Preserving Calibration

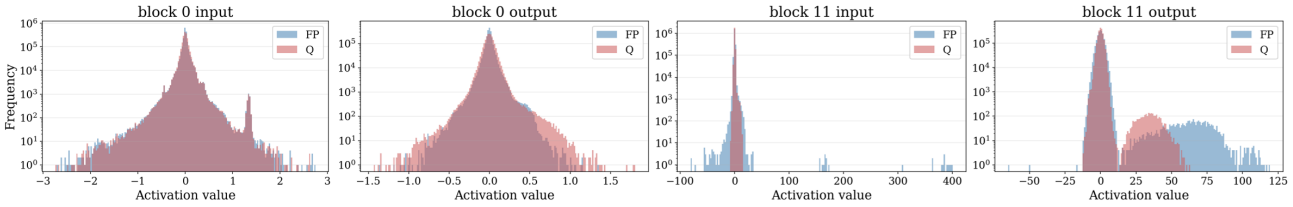
Local quantization error minimization fails to preserve query-key interaction. Existing PTQ methods typically calibrate the query and key projections by minimizing local quantization error. Let $X_Q \in \mathbb{R}^{N_q \times D}$ and $X_K \in \mathbb{R}^{N_k \times D}$ denote the input features to the two projections. The query, key, and attention map are then given by

$$\mathbf{Q} = X_Q W_Q,\ \mathbf{K} = X_K W_K,\ \mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right), \tag{1}$$

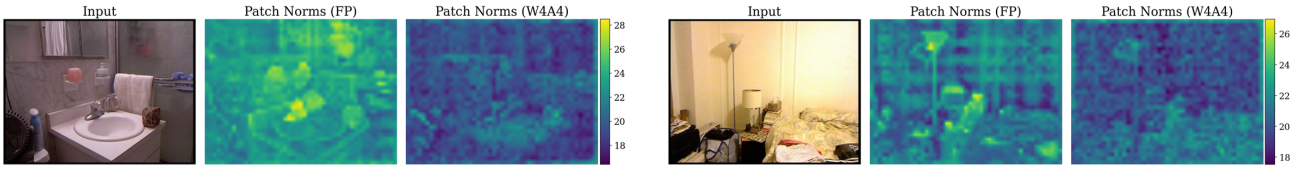
where $W_Q, W_K \in \mathbb{R}^{D \times d_h}$ are the corresponding projection weights. This formulation shows that the attention map is determined by the query-key interaction through the pre-softmax logits $\mathbf{Q}\mathbf{K}^\top/\sqrt{d_h}$, rather than by the fidelity of each projection alone. Nevertheless, conventional PTQ minimizes the local quantization error of the two projections independently:

$$\theta_Q^\star = \arg\min_{\theta_Q} \left\| \mathbf{Q} - \hat{\mathbf{Q}}_{\theta_Q} \right\|_2, \theta_K^\star = \arg\min_{\theta_K} \left\| \mathbf{K} - \hat{\mathbf{K}}_{\theta_K} \right\|_2, \tag{2}$$

where $\theta_Q = \{s_Q, z_Q\}$ and $\theta_K = \{s_K, z_K\}$ denote the corresponding quantization parameter sets. This objective



(a) Distribution shift in intermediate features caused by accumulated quantization errors across layers.



(b) Intermediate feature map distortion under W4A4 quantization.

Figure 2. Quantization-induced distortion in intermediate features under 4-bit quantization (W4A4). (a) Distribution shift in intermediate features caused by accumulated quantization errors across layers. (b) Distortion of intermediate feature maps under W4A4, as illustrated by comparisons of patch norms with the full-precision model.

implicitly assumes that preserving the two projections individually is sufficient for preserving the attention map, even though the attention map is produced by their interaction.

To verify this misalignment empirically, we replace the local quantization error minimization in Eq. (2) with a calibration objective defined on the pre-softmax attention logits:

$$\theta_Q^\star, \theta_K^\star = \arg\min_{\theta_Q,\theta_K} \left\| \mathbf{Q}\mathbf{K}^\top - \hat{\mathbf{Q}}_{\theta_Q}\hat{\mathbf{K}}_{\theta_K}^\top \right\|_2 . \tag{3}$$

As shown in Figure 1 (a), although this logits-level objective often degrades the MSE of $\mathbf{Q}$ and $\mathbf{K}$ individually, it consistently improves the MSE of the attention logits $\mathbf{Q}\mathbf{K}^\top$. This observation reveals a fundamental mismatch between local quantization error minimization and query-key interaction: optimizing the two projections individually does not necessarily preserve their interaction. This result motivates defining the calibration objective on the attention logits rather than preserving the two projections independently.

Logits-level calibration does not necessarily yield a more accurate attention map. Although the logits-level objective in Eq. (3) often reduces the MSE of the attention logits, this improvement does not consistently improve the attention map under MSE and KL divergence, as shown in Figure 1 (b). This indicates that the attention logits do not provide a sufficient objective for matching the attention distribution. This gap arises from the nonlinearity of the softmax operation. The attention logits are unnormalized real-valued scores, whereas the attention map is a normalized probability distribution after softmax. Consequently, minimizing the Euclidean error of the attention logits is not equivalent to minimizing the difference between the corresponding attention distributions.

Motivated by the above observations, we propose **Attention-Preserving Calibration (APC)**, which calibrates quantization parameters directly with respect to the attention map. Unlike objectives defined on individual projections or pre-softmax attention logits, APC adopts a distribution-aware objective. Since the attention map is a normalized probability distribution after softmax, its discrepancy should be measured with a distribution-sensitive objective. KL divergence is more appropriate than Euclidean error because it measures discrepancies in probability mass, thereby better capturing structural changes in the attention distribution. APC determines the quantization parameters by minimizing the row-wise KL divergence between the full-precision attention map $\mathbf{A}$ and the quantized attention map $\mathbf{A}^q$:

$$\theta_Q^\star, \theta_K^\star = \arg\min_{\theta_Q,\theta_K} \frac{1}{N_q}\sum_{i=1}^{N_q} D_{\mathrm{KL}}\left(\mathbf{A}_{i,:} \,\middle\|\, \mathbf{A}_{i,:}^q\right), \tag{4}$$

where $\mathbf{A}^q = \mathrm{softmax}(\hat{\mathbf{Q}}_{\theta_Q}\hat{\mathbf{K}}_{\theta_K}^\top/\sqrt{d_h})$. As shown in Figure 1 (c), the proposed APC objective consistently reduces both the MSE and KL divergence between the full-precision and quantized attention maps across most blocks. This demonstrates that attention-level calibration more effectively preserves the attention structure than projection-wise or logits-level objectives. As a result, APC better preserves the structural information used to construct intermediate features for depth prediction.

2.2. Channel-Wise Distribution Alignment

Accumulated quantization errors progressively distort intermediate features. Low-bit quantization progressively shifts intermediate feature distributions away from their full-precision counterparts, as illustrated in Figure 2 (a). This distortion arises because the quantized output of each layer is fed into the next layer, causing quantization errors to accumulate across layers. Under tensor-wise quantization, this accumulation can become more severe because a

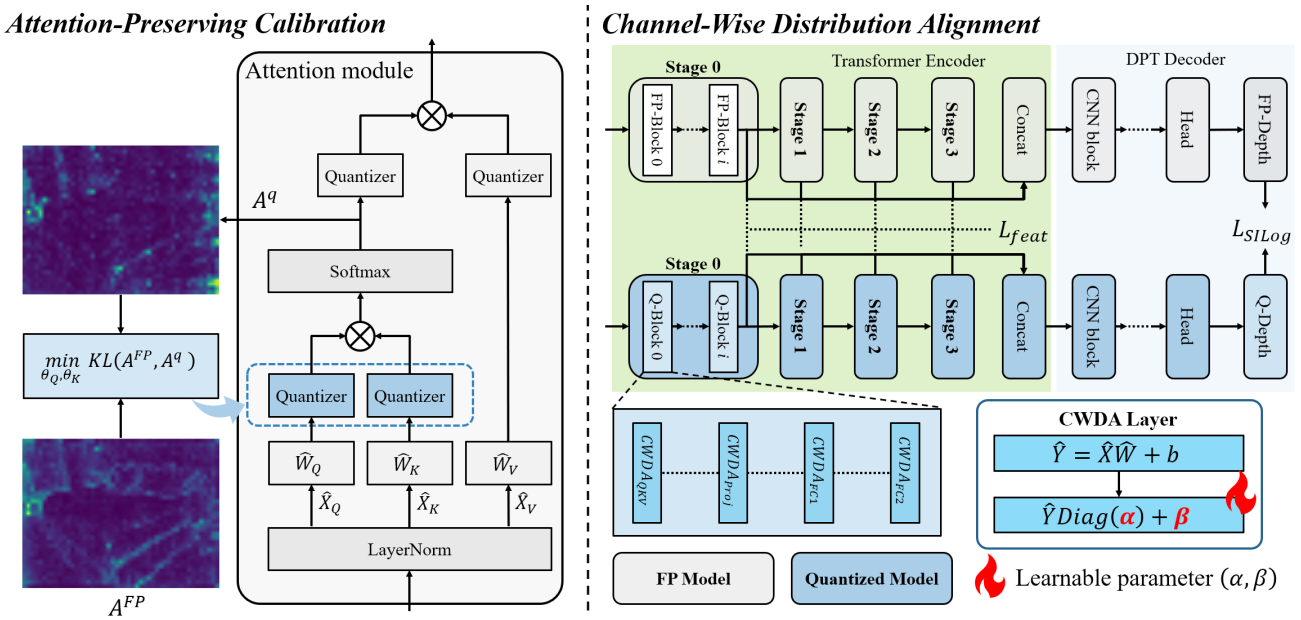


Figure 3. Overview of the SPA-Q framework. Attention-Preserving Calibration determines quantization parameters for query and key activations by minimizing row-wise KL divergence between the full-precision and quantized attention maps. In Channel-Wise Distribution Alignment, channel-wise affine parameters are learned using full-precision stage-wise feature maps from the transformer encoder and the full-precision depth map produced by the DPT decoder.

single quantization scale is shared across all channels, making the quantization error more sensitive to inter-channel variation (Li et al., 2023; Yang et al., 2024a; Jiang et al., 2026). As a result, the resulting distribution shift can be further amplified, distorting not only feature statistics but also the corresponding feature maps, as shown in Figure 2 (b). This issue is especially critical in transformer-based monocular depth estimation models, where the decoder relies on stage-wise features from multiple encoder blocks. Once these distorted features are propagated to the decoder, the quality of depth prediction deteriorates.

To mitigate this distortion, we propose **Channel-Wise Distribution Alignment (CWDA)**, which applies an affine transformation to each channel of the dequantized output activation. CWDA can be applied to output activations from both fully connected and convolutional layers. For notational simplicity, we represent the dequantized output activation as $\hat{\mathbf{Y}} \in \mathbb{R}^{N\times D}$, where N denotes the number of tokens or flattened spatial positions, and D denotes the number of channels. The aligned output $\hat{\mathbf{Y}}^\star$ is given by

$$\hat{\mathbf{Y}}^\star_{:,d} = \alpha_d \hat{\mathbf{Y}}_{:,d} + \beta_d, \quad d = 1, \ldots, D, \tag{5}$$

where α_d and β_d are learnable scaling and shifting parameters for the d-th channel, optimized on training set. Defining $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_D]^\top$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_D]^\top$, Eq. (5) becomes

$$\hat{\mathbf{Y}}^\star = \hat{\mathbf{Y}}\ \mathrm{Diag}(\boldsymbol{\alpha}) + \boldsymbol{\beta}. \tag{6}$$

Since the affine transformation is applied independently to each output channel, it can be absorbed into the weight and bias of the preceding transformation:

$$\hat{\mathbf{Y}}^\star = \hat{\mathbf{X}}\hat{\mathbf{W}}\,\mathrm{Diag}(\boldsymbol{\alpha}) + \mathbf{b} \odot \boldsymbol{\alpha} + \boldsymbol{\beta} = \hat{\mathbf{X}}\hat{\mathbf{W}}^\star + \mathbf{b}^\star, \tag{7}$$

where $\hat{\mathbf{W}}^\star = \hat{\mathbf{W}}\,\mathrm{Diag}(\boldsymbol{\alpha})$ and $\mathbf{b}^\star = \mathbf{b} \odot \boldsymbol{\alpha} + \boldsymbol{\beta}$. Therefore, CWDA introduces no additional inference overhead after training.

Two-step depth-aware optimization. The DINOv2 encoder and DPT decoder play different roles in MDE. The encoder is responsible for producing structurally informative intermediate features, whereas the decoder aggregates multi-scale features to generate the depth map. Because the objectives of the two modules are different, jointly optimizing all CWDA parameters with a single objective can entangle feature preservation and depth reconstruction. We therefore adopt a two-step optimization strategy that separately trains the encoder and the decoder according to their respective roles.

Step 1: Encoder CWDA via stage feature alignment. We optimize the CWDA parameters in the encoder to preserve the stage-wise features delivered to the decoder. Since these intermediate features provide the structural information used for subsequent multi-scale decoding, their distortion directly degrades depth estimation. To reduce this distortion, we minimize the ℓ_1 distance between the full-precision and quantized stage features:

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \left\| \mathbf{F}_s - \hat{\mathbf{F}}_s \right\|_1, \tag{8}$$

Table 1. Comparison of PTQ methods on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2013) for zero-shot relative depth estimation under 4-bit quantization. *E.* denotes the encoder backbone used in the MDE architectures.

Dataset	Method	W/A	Depth Anything v1				Depth Anything v2			
			E. ViT-S		*E.* ViT-B		*E.* ViT-S		*E.* ViT-B	
			AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
NYUv2	FP	32/32	0.0525	0.9720	0.0459	0.9791	0.0513	0.9736	0.0460	0.9770
	BRECQ (Li et al., 2021)	4/4	0.2535	0.5395	0.2886	0.4692	0.2714	0.5042	0.2910	0.4646
	QDrop (Wei et al., 2022)	4/4	0.1742	0.7166	0.2334	0.5785	0.1773	0.7115	0.2332	0.5794
	RepQ-ViT (Li et al., 2023)	4/4	0.1959	0.6639	0.2539	0.5410	0.2053	0.6385	0.2417	0.5634
	QSCA (Yang et al.)	4/4	0.1377	0.8097	0.1875	0.6845	0.1361	0.8151	0.1875	0.6845
	SPA-Q (Ours)	4/4	**0.1049**	**0.8900**	**0.1044**	**0.8957**	**0.1002**	**0.8991**	**0.0971**	**0.9109**
KITTI	FP	32/32	0.0818	0.9369	0.0804	0.9396	0.0832	0.9340	0.0814	0.9389
	BRECQ (Li et al., 2021)	4/4	0.3719	0.3522	0.3989	0.3160	0.3906	0.3344	0.3990	0.3175
	QDrop (Wei et al., 2022)	4/4	0.3934	0.3234	0.3855	0.3338	0.3620	0.3748	0.3412	0.4082
	RepQ-ViT (Li et al., 2023)	4/4	0.3434	0.4159	0.2539	0.5410	0.3634	0.3897	0.3906	0.3287
	QSCA (Yang et al.)	4/4	0.1874	0.7273	0.2365	0.6203	0.2067	0.6794	0.2296	0.6174
	SPA-Q (Ours)	4/4	**0.1595**	**0.7801**	**0.1639**	**0.7806**	**0.1699**	**0.7859**	**0.1609**	**0.7920**

where $\mathcal{S}$ denotes the set of stage indices whose outputs are fed to the decoder, and $\mathbf{F}_s$, $\hat{\mathbf{F}}_s \in \mathbb{R}^{N\times C}$ denote the full-precision and quantized feature maps at stage s, respectively. After optimizing this objective, we freeze the encoder-side CWDA parameters and proceed to the second stage.

Step 2: Decoder CWDA via depth-aware optimization. With the encoder-side CWDA parameters frozen, we optimize the CWDA parameters in the DPT head using a depth-level objective. We therefore minimize the SILog loss (Ranftl et al., 2020) between the full-precision and quantized depth maps:

$$\mathcal{L}_{\text{SILog}} = \frac{1}{n}\sum_{i=1}^{n} z_i^2 - \frac{\lambda}{n^2}\left(\sum_{i=1}^{n} z_i\right)^2, \; z_i = \log \hat{d}_i - \log d_i, \tag{9}$$

where $\hat{d}_i$ and d_i denote the full-precision and quantized depth maps at pixel i, respectively, and n is the number of valid pixels. This step allows the decoder to adapt to the aligned quantized features while preserving the depth structure of the full-precision model.

3. Experiments

3.1. Experimental Setup

Models, datasets, and evaluation. We use Depth Anything v1 (Yang et al., 2024b) and v2 (Yang et al., 2024c), built on the DINOv2 encoder (Oquab et al., 2023) with ViT-S and ViT-B backbones, as the baseline models. Depth Anything is chosen as a representative foundation model for monocular depth estimation. We evaluate SPA-Q on five MDE benchmarks: NYUv2 (Silberman et al., 2012), KITTI (Geiger et al., 2013), ETH3D (Schops et al., 2017), DIODE (Vasiljevic et al., 2019), and Sintel (Butler et al., 2012). These datasets correspond to the official zero-shot relative depth evaluation benchmarks used in Depth Anything (Yang et al., 2024b;c). For evaluation, we report standard depth estimation metrics, including the absolute relative error (AbsRel) and threshold accuracy δ_1. AbsRel measures the mean relative difference between predicted and reference depth values, while δ_1 measures the percentage of predictions satisfying a threshold ratio of 1.25.

Implementation details. For fair comparison, we follow the experimental settings of prior work (Yang et al.). We adopt the widely used percentile method for calibration, using channel-wise quantization for weights and tensor-wise quantization for activations. The calibration set is constructed by randomly sampling 16 images from the training set of NYUv2 for indoor scenes or KITTI for outdoor scenes. The CWDA parameters are initialized such that all elements of $\boldsymbol{\alpha}$ are set to 1 and all elements of $\boldsymbol{\beta}$ are set to 0. Following the proposed two-step depth-aware optimization, we optimize the CWDA parameters using 5% of the training set for one epoch at each step. We use the Adam optimizer with a learning rate of 1×10^{-4} and no weight decay. All experiments are conducted with a batch size of 1 on a single RTX 6000 Ada Generation GPU.

3.2. Experimental Results

Quantitative results. As shown in Table 1, we compare SPA-Q with existing PTQ methods, including BRECQ (Li et al., 2021), QDrop (Wei et al., 2022), RepQ-ViT (Li et al., 2023), and QSCA (Yang et al.), under 4-bit quantization

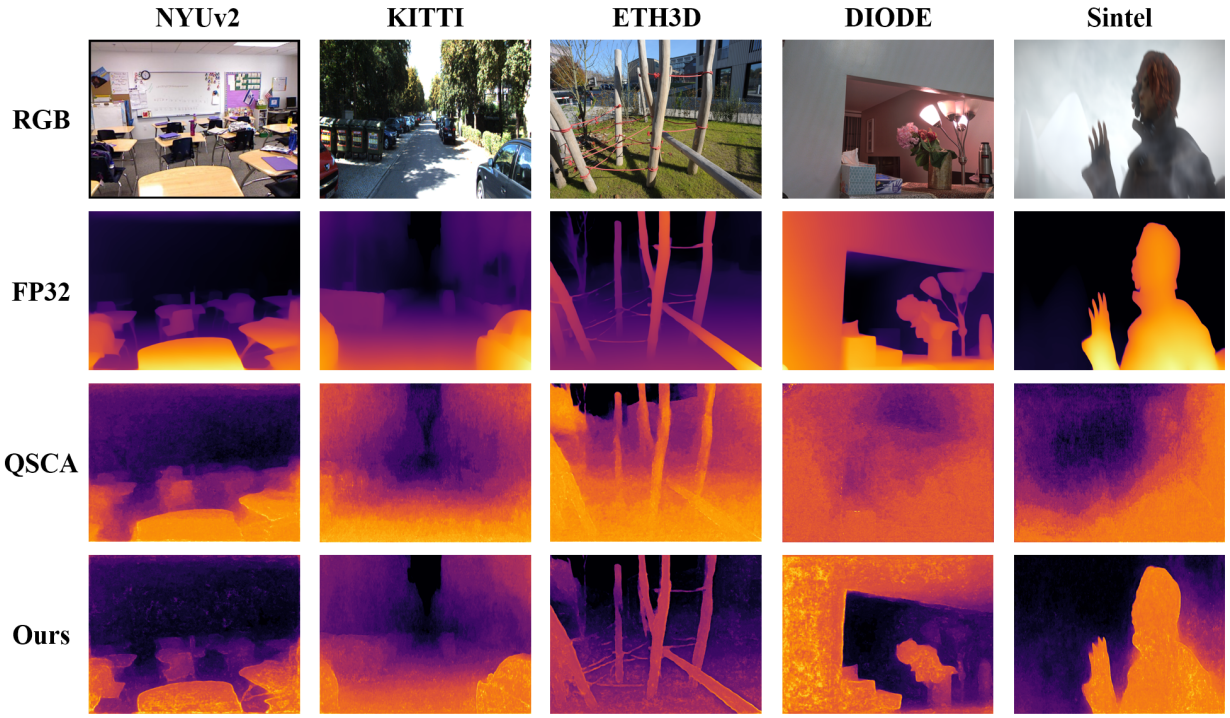


Figure 4. Qualitative comparison of 4-bit quantization results for Depth Anything with a ViT-S backbone on benchmarks: NYUv2 (Silberman et al., 2012), KITTI (Geiger et al., 2013), ETH3D (Schops et al., 2017), DIODE (Vasiljevic et al., 2019), and Sintel (Butler et al., 2012).

on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2013). These two benchmarks are used as representative indoor and outdoor zero-shot relative depth estimation datasets. Overall, SPA-Q consistently achieves the best performance in both AbsRel and δ_1 across Depth Anything v1 and v2 with ViT-S and ViT-B backbones. For example, the improvement is particularly pronounced for Depth Anything v1 with the ViT-B backbone. On NYUv2, SPA-Q reduces AbsRel from 0.1875 to 0.1044 and improves δ_1 from 0.6845 to 0.8957. On KITTI, SPA-Q reduces AbsRel from 0.2365 to 0.1639 and improves δ_1 from 0.6203 to 0.7806. These results demonstrate that the proposed framework effectively preserves depth estimation performance under severe low-bit quantization in both indoor and outdoor settings. Additional quantitative results on ETH3D, DIODE, and Sintel are provided in Appendix D.

Qualitative results. To further validate the generality of SPA-Q, we present qualitative comparisons of predicted depth maps on five benchmarks in Figure 4. Compared with QSCA (Yang et al.), SPA-Q produces depth maps with better structural consistency and sharper object boundaries across indoor, outdoor, and synthetic scenes. For NYUv2 (Silberman et al., 2012) and DIODE (Vasiljevic et al., 2019), SPA-Q better preserves indoor layouts and foreground-object boundaries. In KITTI (Geiger et al., 2013) and ETH3D (Schops et al., 2017), it produces more coherent depth structures in outdoor scenes, including clearer separation between nearby objects and background regions. For Sintel (Butler et al., 2012), SPA-Q better maintains the silhouette of the foreground character, indicating that the proposed method generalizes well to synthetic scenes as well as real-world indoor and outdoor datasets.

4. Conclusion

In this paper, we addressed low-bit post-training quantization for MDE. We identified two key limitations of existing PTQ methods. Independently minimizing quantization error for query and key projections does not adequately preserve the resulting attention distribution, and accumulated quantization errors induce distribution shift in intermediate features. To address these issues, we proposed SPA-Q, a structure-preserving adaptive PTQ framework consisting of Attention-Preserving Calibration (APC) and Channel-Wise Distribution Alignment (CWDA). Together, these two components preserve both attention structure and intermediate feature distributions under quantization. Experimental results on multiple zero-shot depth estimation benchmarks showed that SPA-Q consistently outperforms existing PTQ methods under 4-bit quantization.

Acknowledgements

This research was supported by 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, 50%) 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, 50%).

Impact Statement

This work aims to improve the practical deployment of monocular depth estimation models by reducing the computational and memory cost of foundation models through post-training quantization. More efficient depth estimation can broaden the accessibility of such models in resource-constrained environments, including mobile, embedded, and robotic systems. In this sense, the proposed method reduces the computational and memory demands of vision systems that rely on dense depth prediction. Monocular depth estimation is also deployed in domains such as robotics and autonomous driving, where unreliable predictions can directly affect system behavior. Although our method improves low-bit quantized performance, a noticeable gap to full precision still remains, especially in challenging cases. This suggests that reliability remains an important consideration when deploying quantized depth estimation models in real-world systems. Overall, we believe the primary impact of this work is positive, as it advances efficient model deployment of low-bit quantized depth estimation.

References

Bhalgat, Y., Lee, J., Nagel, M., Blankevoort, T., and Kwak, N. 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*, pp. 696–697, 2020.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., and Müller, M. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pp. 611–625. Springer, 2012.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Choi, J., Wang, Z., Venkataramani, S., Chuang, P. I.-J., Srinivasan, V., and Gopalakrishnan, K. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Dong, X., Garratt, M. A., Anavatti, S. G., and Abbass, H. A. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16940–16961, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Esser, S. K., McKinstry, J. L., Bablani, D., Appuswamy, R., and Modha, D. S. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

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.

Jiang, T., Jiang, Y., Yao, X., Cheng, G., and Han, J. Uq-vit: Harmonizing extreme activations with hardware-friendly uniform quantization in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22354–22362, 2026.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 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. *arXiv preprint arXiv:2102.05426*, 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*, pp. 17227–17236, 2023.

Lin, Y., Zhang, T., Sun, P., Li, Z., and Zhou, S. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

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.

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van Baalen, M., and Blankevoort, T. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

Ranftl, R., Bochkovskiy, A., and Koltun, V. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Schops, T., Schonberger, 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 conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

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.

Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., and Weinberger, K. Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8445–8453, 2019.

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.

Weng, X. and Kitani, K. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

Wofk, D., Ma, F., Yang, T.-J., Karaman, S., and Sze, V. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6101–6108. IEEE, 2019.

Wu, H., Judd, P., Zhang, X., Isaev, M., and Micikevicius, P. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602*, 2020.

Yang, J., Choi, J., Zinke, M. A., and Kang, S.-J. Qsca: Quantization with self-compensating auxiliary for monocular depth estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yang, L., Gong, H., Lin, H., Wu, Y., Sun, Z., and Gu, Q. Dopq-vit: Towards distribution-friendly and outlier-aware post-training quantization for vision transformers. *arXiv preprint arXiv:2408.03291*, 2024a.

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*, pp. 10371–10381, 2024b.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024c.

Appendix

A. Related Work

A.1. Monocular Depth Estimation Models

MDE (Lee et al., 2019; Bhat et al., 2023) is the task of predicting a dense depth map from a single image, where each pixel reflects the 3D structure of the scene. Recent advances driven by large-scale data and transformer-based architectures have significantly improved depth prediction, leading to models such as MiDaS (Ranftl et al., 2020) and Depth Anything (Yang et al., 2024b;c). MiDaS improves generalization by training on multiple datasets with varying scales and distributions, reducing dependence on a single dataset. Building on this, Depth Anything adopts a DINOv2 (Oquab et al., 2023)-pretrained ViT encoder. Depth Anything v1 leverages both labeled and unlabeled data, while v2 further improves performance by incorporating large-scale synthetic and pseudo-labeled data. Both MiDaS and Depth Anything are typically built on the DPT (Ranftl et al., 2021) architecture, which aggregates multi-scale features using a transformer encoder and reconstructs them with a multi-scale decoder. In this work, we focus on Depth Anything v1 and v2 for evaluation, as they represent state-of-the-art MDE models with strong generalization performance.

A.2. Post-Training Quantization

PTQ (Gholami et al., 2022; Nagel et al., 2021; Wu et al., 2020) has been widely studied as an efficient approach for compressing neural networks without requiring full retraining. Most existing PTQ methods are designed for recognition tasks such as image classification. BRECQ (Li et al., 2021) minimizes quantization error via block-wise reconstruction using second-order approximation. By randomly dropping activation quantization during reconstruction, QDrop (Wei et al., 2022) mitigates overfitting and improves generalization. RepQ-ViT (Li et al., 2023) further reduces inter-channel variation through scale reparameterization, enabling channel-wise calibration with tensor-wise inference. However, these methods are primarily evaluated on classification tasks (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021) and do not adequately address dense prediction problems. In MDE, predictions rely on multi-scale intermediate features, making the task more sensitive to quantization errors. To address this limitation, QSCA (Yang et al.) introduces lightweight auxiliary modules for residual error compensation. Despite its effectiveness, this approach incurs additional parameters and computational overhead, which is undesirable for efficient deployment. In this work, we propose a structure-aware PTQ method for monocular depth estimation that preserves intermediate representations under quantization.

B. Preliminaries

Uniform quantizer. A uniform quantizer (Gholami et al., 2022; Nagel et al., 2021; Wu et al., 2020) maps full-precision values to evenly spaced discrete levels. Given an input x and bit-width b, the quantized integer $\bar{x}$ and the dequantized value $\hat{x}$ are computed as

$$\bar{x} = \text{clip}\left(\left\lfloor \frac{x}{s} \right\rceil + z, 0, 2^b - 1\right), \quad \hat{x} = s \cdot (\bar{x} - z), \tag{10}$$

where the step size s and zero-point z are

$$s = \frac{\max(x) - \min(x)}{2^b - 1}, \quad z = \left\lfloor \frac{-\min(x)}{s} \right\rceil, \tag{11}$$

and $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, while $\text{clip}(\cdot, 0, 2^b - 1)$ constrains the quantized integer to the valid range $[0, 2^b - 1]$.

Log2 quantizer. As a basic form of logarithmic quantizer, the log2 quantizer (Lin et al., 2021) maps inputs to quantization levels defined on a base-2 logarithmic scale. Compared to uniform quantization, it allocates more quantization levels to small values, making it better suited for highly skewed or long-tailed distributions such as post-Softmax activations. Given an input x and step size s, the quantized value $\hat{x}$ is computed as:

$$\bar{x} = \text{clip}\left(\left\lfloor -\log_2 \frac{x}{s} \right\rceil, 0, 2^b - 1\right), \quad \hat{x} = s \cdot 2^{-\bar{x}}. \tag{12}$$

Unlike uniform quantization, the log2 quantizer defines quantization levels on an exponential scale, making it more effective for values concentrated near zero.

C. Ablation Study

Table 2. Ablation study of calibration objectives under 4-bit quantization for Depth Anything v1 with a ViT-B backbone on NYUv2 and KITTI.

Dataset	Method	AbsRel ↓	δ_1 ↑
NYUv2	Local	0.1908	0.6836
	Logit-level	0.1839	0.7013
	APC	0.1444	0.7990
	SPA-Q	**0.1044**	**0.8957**
KITTI	Local	0.3241	0.4366
	Logit-level	0.3002	0.4828
	APC	0.2490	0.5789
	SPA-Q	**0.1639**	**0.7806**

Table 3. Comparison of model size, parameter count, and bit operations (BOPs) under 4-bit quantization. SPA-Q* denotes the model before fusing the learned CWDA parameters into the weights and biases.

Encoder	Method	Model size (MB)	Params (M)	BOPs (G)
ViT-S	FP	94.55	24.79	85742
	QSCA (Yang et al.)	16.27	25.23	1544
	SPA-Q*	14.50	24.87	1457
	SPA-Q	14.16	24.79	1340
ViT-B	FP	371.82	97.47	273280
	QSCA (Yang et al.)	60.02	99.24	5092
	SPA-Q*	51.84	97.65	4493
	SPA-Q	51.16	97.47	4270

Comparison of calibration objectives. We conduct an ablation study on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2013) to examine the effect of the proposed calibration objective under 4-bit quantization for Depth Anything v1 with a ViT-B backbone. As shown in Table 2, on NYUv2, replacing local quantization error minimization with a logits-level objective yields only a marginal improvement, with AbsRel decreasing from 0.1908 to 0.1839 and δ_1 increasing from 0.6836 to 0.7013. This result suggests that directly considering query-key interaction at the logits level is still insufficient to consistently improve depth estimation performance. In contrast, APC substantially improves over logits-level calibration, reducing AbsRel to 0.1444 and increasing δ_1 to 0.7990, which indicates that directly matching the attention distribution is a more effective calibration strategy. Finally, combining APC with CWDA yields the best results, with AbsRel further reduced to 0.1044 and δ_1 improved to 0.8957, indicating that feature distribution alignment further improves performance over calibration alone.

Block-wise comparison of feature similarity. We further evaluate the similarity between full-precision and quantized features across transformer blocks under 4-bit quantization. As shown in Figure 5 (a), local and logits-level calibration exhibit similar MSE trends across blocks, indicating that improving the attention logits alone does not substantially improve feature preservation. In contrast, APC consistently lowers the MSE relative to both local and logits-level calibration, while SPA-Q achieves the lowest MSE across nearly all blocks, with the gap becoming more pronounced in deeper blocks. Figure 5 (b) shows a similar pattern in cosine similarity. APC yields consistently higher cosine similarity than both local and logits-level calibration, and SPA-Q further improves it, achieving the highest similarity across most blocks. These results indicate that APC improves calibration by better preserving the attention map, while CWDA further improves the preservation of intermediate representations across layers.

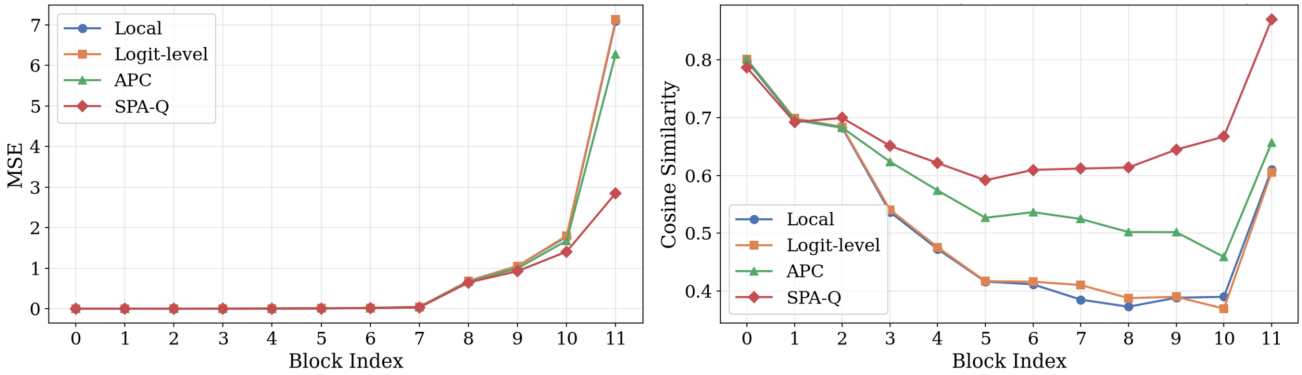


Figure 5. Block-wise feature analysis under 4-bit quantization for Depth Anything v1 with a ViT-B backbone across transformer blocks. (a) Mean squared error (MSE) between full-precision and quantized features. (b) Cosine similarity between full-precision and quantized features.

Efficiency analysis. We further compare model size, parameter count, and bit operations (BOPs) of SPA-Q and QSCA (Yang et al.) under 4-bit quantization for Depth Anything v1 with ViT-S and ViT-B backbones. The efficiency results in Table 3 are reported on the NYUv2 dataset. As shown in Table 3, SPA-Q is more efficient than QSCA in both model size and computational cost. For ViT-S, the fused SPA-Q model reduces model size from 16.27 MB to 14.16 MB and BOPs from 1544 G to 1340 G. For ViT-B, SPA-Q also yields a smaller model size than QSCA, decreasing it from 60.02 MB to 51.16 MB after fusion. Although SPA-Q before fusion introduces additional channel-wise affine parameters, these parameters are absorbed into the weights and biases after training, so the fused model has the same parameter count as the original quantized network. These results show that SPA-Q improves depth estimation accuracy without introducing additional inference overhead.

D. Additional Quantitative Results

We provide additional quantitative results on ETH3D (Schops et al., 2017), DIODE (Vasiljevic et al., 2019), and Sintel (Butler et al., 2012) to further evaluate the zero-shot generalization of SPA-Q under 4-bit quantization. Table 4 compares SPA-Q with existing PTQ methods (Li et al., 2021; Wei et al., 2022; Li et al., 2023; Yang et al.) for both Depth Anything v1 and v2 with ViT-S and ViT-B backbones. SPA-Q consistently achieves the best performance across datasets and model configurations, demonstrating that the proposed method generalizes well across a broader range of zero-shot depth estimation benchmarks.

Table 4. Additional comparison of PTQ methods on ETH3D (Schops et al., 2017), DIODE (Vasiljevic et al., 2019), and Sintel (Butler et al., 2012) for zero-shot relative depth estimation under 4-bit quantization. *E.* denotes the encoder backbone used in the MDE architectures.

Dataset	Method	W/A	Depth Anything v1				Depth Anything v2			
			E. ViT-S		*E.* ViT-B		*E.* ViT-S		*E.* ViT-B	
			AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
ETH3D	FP	32/32	0.0584	0.9652	0.0513	0.9741	0.0548	0.9701	0.0467	0.9791
	BRECQ (Li et al., 2021)	4/4	0.2829	0.5082	0.2958	0.4874	0.2920	0.4962	0.2963	0.4870
	QDrop (Wei et al., 2022)	4/4	0.2351	0.6069	0.2706	0.5373	0.2288	0.6186	0.2697	0.5341
	RepQ-ViT (Li et al., 2023)	4/4	0.2195	0.6286	0.2491	0.5670	0.2215	0.6299	0.2606	0.5538
	QSCA (Yang et al.)	4/4	0.1791	0.7332	0.2241	0.6309	0.1983	0.6791	0.2197	0.6298
	SPA-Q (Ours)	4/4	**0.1396**	**0.8267**	**0.1529**	**0.8005**	**0.1378**	**0.8385**	**0.1311**	**0.8441**
DIODE	FP	32/32	0.0753	0.9413	0.0745	0.9474	0.0721	0.9426	0.0701	0.9498
	BRECQ (Li et al., 2021)	4/4	0.2079	0.6755	0.2201	0.6464	0.2173	0.6538	0.2211	0.6441
	QDrop (Wei et al., 2022)	4/4	0.1864	0.7275	0.2058	0.6819	0.1848	0.7279	0.2062	0.6777
	RepQ-ViT (Li et al., 2023)	4/4	0.1746	0.7538	0.2016	0.6891	0.1800	0.7552	0.1999	0.6945
	QSCA (Yang et al.)	4/4	0.1513	0.8033	0.1997	0.7099	0.1463	0.8135	0.1947	0.7094
	SPA-Q (Ours)	4/4	**0.1258**	**0.8723**	**0.1276**	**0.8783**	**0.1213**	**0.8795**	**0.1206**	**0.8882**
Sintel	FP	32/32	0.2296	0.7304	0.2281	0.7539	0.2680	0.6974	0.2576	0.7111
	BRECQ (Li et al., 2021)	4/4	0.4787	0.3123	0.4820	0.3085	0.4816	0.3055	0.4821	0.3008
	QDrop (Wei et al., 2022)	4/4	0.4991	0.3181	0.4867	0.3106	0.4948	0.3184	0.4882	0.3151
	RepQ-ViT (Li et al., 2023)	4/4	0.4722	0.3619	0.4642	0.3629	0.4693	0.3829	0.4838	0.3237
	QSCA (Yang et al.)	4/4	0.4531	0.3884	0.4781	0.4059	0.4474	0.3919	0.4679	0.4070
	SPA-Q (Ours)	4/4	**0.4324**	**0.4908**	**0.3788**	**0.5321**	**0.4084**	**0.4851**	**0.4115**	**0.4711**