SPA-Q: Structure-Preserving Adaptive Post-Training Quantization for Monocular Depth Estimation

Jaemin Choi[1*] **Jincheol Yang**[1*] **Nahyun Lim**[1*] **Yun-Seong Jeong**[1*]
Matti Zinke[2*] **Hyunwoo Yu**[1*] **Suk-Ju Kang**[1†]
[1]Department of Electronic Engineering, Sogang University
[2]Department of Computer Science, Sogang University
{jam0225, yjc3232, julialim5, jus8077, mattizinke, hyunwoo137, sjkang}@sogang.ac.kr

Abstract

Monocular depth estimation (MDE) has advanced rapidly with the emergence of foundation models such as Depth Anything. Their transformer-based architectures provide strong generalization across diverse scenes and domains, but also incur high computational and memory cost, making efficient deployment challenging. Post-Training Quantization (PTQ) provides an efficient and practical solution for model compression, yet low-bit PTQ remains challenging for MDE models. When applying PTQ to Depth Anything, we identify two key challenges: (i) independently minimizing the quantization error of the query and key projections fails to preserve the attention maps induced by their interaction, and (ii) quantization errors accumulate across layers, resulting in intermediate feature distribution shifts. To address these issues, we propose SPA-Q, a Structure-Preserving Adaptive PTQ framework for MDE, consisting of two main components. First, Attention-Preserving Calibration (APC) determines query and key quantization parameters by matching the full-precision attention distribution. Second, Channel-Wise Distribution Alignment (CWDA) learns channel-wise affine transformations to mitigate quantization-induced distribution shifts, and the learned parameters are absorbed into the weights after training. Experimental results show that SPA-Q consistently outperforms existing PTQ methods under 4-bit quantization, achieving an average 25.9% reduction in AbsRel and a 17.3% improvement in δ_1 across NYUv2 and KITTI datasets.

1 Introduction

Monocular depth estimation (MDE) is a fundamental task in computer vision, with applications in robotics [1, 2], autonomous driving [3, 4], and 3D scene understanding [5–7]. Recent advances in Vision Transformers (ViTs) [8–12] have significantly improved MDE, leading to strong foundation models such as MiDaS [13] and Depth Anything [14, 15]. 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 hinder real-world deployment in resource-constrained environments.

Model quantization [16–21] 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) [22–24] and post-training quantization (PTQ) [25–28]. QAT incorporates quantization effects during training and often achieves higher accuracy, but it requires additional training and access to training data. In contrast, PTQ applies quantization to a pre-trained model

*Equal contribution.
†Corresponding author.

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

without full retraining, making it suitable for practical deployment. Although PTQ has been widely studied for recognition models, its application to MDE remains underexplored. Unlike recognition tasks, depth estimation requires preserving fine-grained geometric and structural information in intermediate features, making it more sensitive to quantization-induced distortions.

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 [25–28] typically calibrate each quantizer by minimizing the discrepancy between full-precision and quantized outputs of each projection. 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 preservation of the resulting attention map. This issue is especially critical for MDE, as distorted attention maps can corrupt features aggregated by the decoder, thereby degrading scene geometry and object boundaries in depth predictions. **(ii) Quantization errors accumulate across layers and induce distribution shifts in intermediate features.** Since the DPT 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 full-precision counterparts.

To address these limitations, we propose **SPA-Q**, a novel **S**tructure-**P**reserving **A**daptive PTQ framework for transformer-based MDE models. SPA-Q is designed for architectures whose decoders aggregate multi-stage encoder features to produce dense depth predictions. It consists of two complementary components. **Attention-Preserving Calibration (APC)** determines the quantization parameters of the query and key projections by directly preserving the full-precision attention distribution instead of minimizing projection errors independently. In particular, APC minimizes the KL divergence between the full-precision and quantized attention distributions, thereby better capturing their discrepancy after softmax. **Channel-Wise Distribution Alignment (CWDA)** addresses distribution shifts in intermediate features through learnable channel-wise affine transformations, which are optimized with feature-level and depth-level objectives in an end-to-end manner. The learned affine parameters are then fused into the weights, introducing no additional inference overhead. Our contributions are summarized as follows:

- We identify two 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 map. Second, accumulated quantization errors induce distribution shift in intermediate features.
- We propose SPA-Q, a Structure-Preserving Adaptive PTQ method for MDE. SPA-Q introduces Attention-Preserving Calibration (APC), which calibrates query and key quantizers with an attention-level objective, and Channel-Wise Distribution Alignment (CWDA), a channel-wise alignment scheme for correcting intermediate feature distribution shifts.
- Extensive experiments on relative depth estimation benchmarks show that SPA-Q outperforms existing PTQ methods under 4-bit quantization across all evaluated settings.

2 Related Work

2.1 Monocular Depth Estimation Models

MDE [29, 30] refers to 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 [13] and Depth Anything [14, 15]. 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 [11]-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 [31] architecture, which aggregates multi-scale features using a transformer encoder and reconstructs them with a multi-scale decoder. We focus on Depth Anything v1 and v2 for evaluation, as they represent state-of-the-art MDE models with strong generalization performance.

2.2 Post-Training Quantization

PTQ [16–18] 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 [25] minimizes quantization error via block-wise reconstruction using a second-order approximation. By randomly dropping activation quantization during reconstruction, QDrop [26] mitigates overfitting and improves generalization. RepQ-ViT [27] further reduces inter-channel variation through scale reparameterization, enabling channel-wise calibration with tensor-wise inference. However, these methods have been primarily designed for and evaluated on classification tasks [8, 9, 32], and do not adequately address dense prediction tasks. In MDE, predictions rely on multi-scale intermediate features, making the task more sensitive to quantization errors. To address this limitation, QSCA [33] 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 PTQ method for MDE that preserves intermediate features required for dense depth prediction by maintaining attention maps and mitigating feature distribution shifts.

3 Method

3.1 Preliminaries

Uniform quantizer. A uniform quantizer [16–18] 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:

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

where the step size s and zero-point z are

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

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 [21] 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:

$$\bar{x} = \text{clip}\left(\left\lfloor -\log_2 \frac{x}{s} \right\rceil, 0, 2^b - 1\right), \quad \hat{x} = s \cdot 2^{-\bar{x}}. \tag{3}$$

The log2 quantizer defines quantization levels on an exponential scale, making it more effective for values concentrated near zero. Accordingly, we adopt the log2 quantizer for post-Softmax activations.

3.2 Motivation

To effectively quantize transformer-based MDE models, we analyze how low-bit quantization degrades representations used for dense depth prediction. In particular, we examine two sources of distortion: attention maps resulting from query-key interaction and intermediate feature distributions.

Mismatch between local quantization error minimization and attention preservation. 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 query and key projections, respectively, where N_q and N_k are the numbers of query and key tokens, and D is the input channel dimension. The query, key, and attention map are then given by:

$$\mathbf{Q} = X_Q W_Q, \qquad \mathbf{K} = X_K W_K, \qquad \mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right), \tag{4}$$

where $W_Q, W_K \in \mathbb{R}^{D \times d_h}$ are the corresponding projection weights, and d_h denotes the head dimension. This formulation shows that the attention map is determined by the query-key interaction through

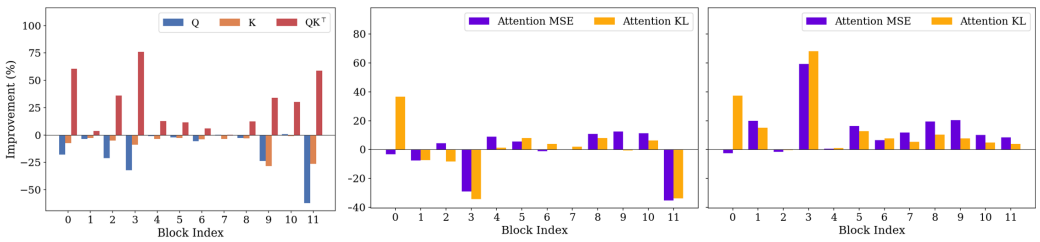


Figure 1: Comparison of calibration objectives for query-key quantization, where all improvements are measured relative to conventional local quantization error minimization. (a) MSE improvement from logits-level calibration. (b) Attention-level metric improvement from logits-level calibration. (c) Attention-level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

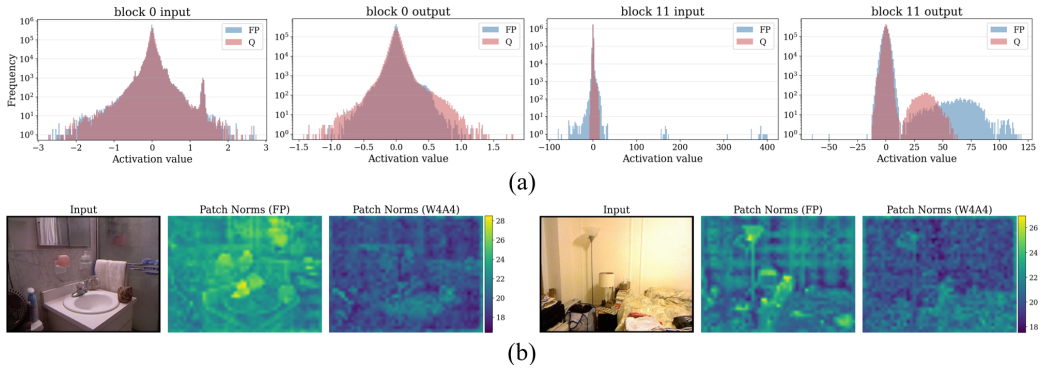


Figure 2: Quantization-induced distortion in intermediate features under 4-bit quantization (W4A4). (a) Distribution shift in intermediate features caused by accumulated quantization errors across layers. (b) Distortion of intermediate feature maps under W4A4, as illustrated by comparisons of patch norms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

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:

$$\theta_Q^\star = \arg\min_{\theta_Q} \left\| \mathbf{Q} - \hat{\mathbf{Q}}_{\theta_Q} \right\|_F^2, \qquad \theta_K^\star = \arg\min_{\theta_K} \left\| \mathbf{K} - \hat{\mathbf{K}}_{\theta_K} \right\|_F^2, \tag{5}$$

where $\theta_Q = \{s_Q, z_Q\}$ and $\theta_K = \{s_K, z_K\}$ denote the corresponding quantization parameter sets. This objective implicitly assumes that preserving $\mathbf{Q}$ and $\mathbf{K}$ individually is sufficient for preserving the resulting attention map, even though the attention map is produced by their interaction.

To verify this mismatch empirically, we replace the local quantization error minimization in Eq. (5) with a calibration objective defined on the pre-softmax attention logits:

$$\theta_Q^\star, \theta_K^\star = \arg\min_{\theta_Q, \theta_K} \left\| \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}} - \frac{\hat{\mathbf{Q}}_{\theta_Q} \hat{\mathbf{K}}_{\theta_K}^\top}{\sqrt{d_h}} \right\|_F^2. \tag{6}$$

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. However, reducing the MSE of the attention logits does not consistently improve the attention map under MSE and KL divergence, as shown in Figure 1 (b). This gap arises from the nonlinearity of the softmax. 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 discrepancy between the corresponding attention maps.

Feature distortion from accumulated quantization errors. Low-bit quantization progressively shifts intermediate feature distributions away from their full-precision counterparts, as illustrated

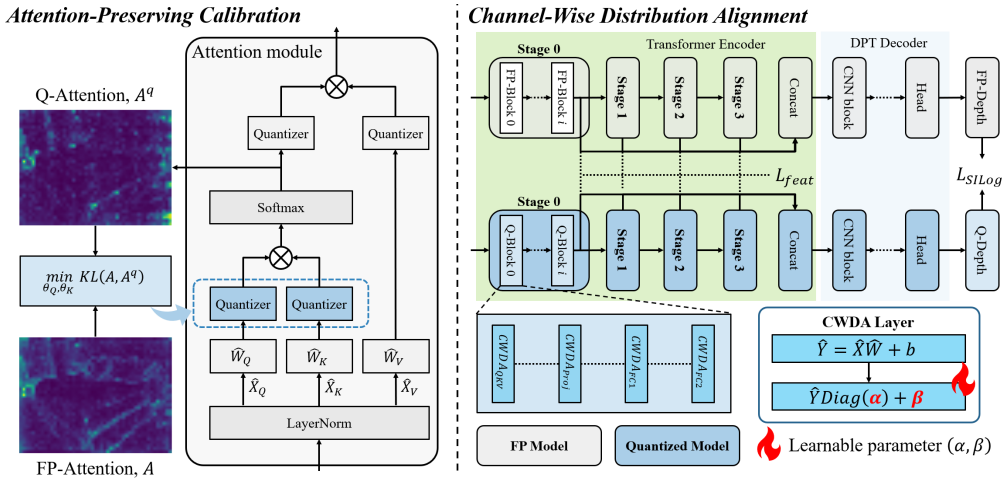


Figure 3: Overview of the SPA-Q framework. SPA-Q first applies Attention-Preserving Calibration (APC) to determine query and key quantization parameters by minimizing the row-wise KL divergence between full-precision and quantized attention maps. It then applies Channel-Wise Distribution Alignment (CWDA), where learnable affine parameters align quantized intermediate features with their full-precision counterparts using stage-wise feature maps from the transformer encoder and the depth prediction from the decoder. After optimization, the learned affine parameters are fused into the weights, introducing no additional inference overhead.

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 single quantization scale is shared across all channels, making the quantization error more sensitive to inter-channel variation [27, 34, 35]. As a result, the 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 MDE 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.

3.3 Attention-Preserving Calibration

As discussed in Sec. 3.2, local quantization error minimization for query and key projections is not well aligned with preserving the resulting attention map, and logits-level calibration is still insufficient for matching the attention distribution after softmax. Based on this observation, we propose **Attention-Preserving Calibration (APC)**, which determines the query and key quantization parameters by directly minimizing the discrepancy between the full-precision and quantized attention distributions. Unlike objectives defined on individual projections or pre-softmax attention logits, APC defines the calibration objective on the attention map. Since each row of the attention map forms a normalized probability distribution over tokens, KL divergence is more appropriate than Euclidean error for measuring discrepancies in probability mass and capturing distortions in the attention distribution. Accordingly, APC minimizes the row-wise KL divergence between the full-precision attention map $\mathbf{A}$ and the quantized attention map $\mathbf{A}^q$:

$$\theta_Q^\star, \theta_K^\star = \arg\min_{\theta_Q,\theta_K} \frac{1}{N_q} \sum_{i=1}^{N_q} D_{\mathrm{KL}}\left(\mathbf{A}_{i,:} \,\middle\|\, \mathbf{A}_{i,:}^q\right), \tag{7}$$

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 better aligns the quantized attention map with its full-precision counterpart than projection-wise or logits-level objectives. As a result, APC better maintains the structural information used to construct intermediate features for depth prediction.

3.4 Channel-Wise Distribution Alignment

To address the intermediate feature distribution shifts identified in Sec. 3.2, we propose **Channel-Wise Distribution Alignment (CWDA)**, which aligns dequantized output activations with their full-precision counterparts through learnable channel-wise affine transformations. 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 then given by:

$$\hat{\mathbf{Y}}^\star_{:,d} = \alpha_d \hat{\mathbf{Y}}_{:,d} + \beta_d, \quad d = 1, \ldots, D, \tag{8}$$

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. (8) becomes

$$\hat{\mathbf{Y}}^\star = \hat{\mathbf{Y}}\ \mathrm{Diag}(\boldsymbol{\alpha}) + \boldsymbol{\beta}. \tag{9}$$

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

$$\hat{\mathbf{Y}}^\star = \hat{\mathbf{X}}\hat{\mathbf{W}}\,\mathrm{Diag}(\boldsymbol{\alpha}) + \mathbf{b} \odot \boldsymbol{\alpha} + \boldsymbol{\beta} = \hat{\mathbf{X}}\hat{\mathbf{W}}^\star + \mathbf{b}^\star, \tag{10}$$

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.

End-to-end optimization. The DINOv2 [11] encoder and DPT [31] decoder jointly determine the final depth prediction in MDE. Quantization-induced distortion in intermediate features propagates through the network and ultimately affects the predicted depth map. To account for both representation-level and output-level effects, we optimize all CWDA parameters in an end-to-end manner using feature-level and depth-level objectives. To preserve the stage-wise features delivered to the decoder, we minimize the ℓ_1 distance between the full-precision and quantized stage features:

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left\| \mathbf{F}_s - \hat{\mathbf{F}}_s \right\|_1, \tag{11}$$

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. To preserve the output depth structure, we minimize the SILog loss [13] between full-precision and quantized depth maps:

$$\mathcal{L}_{\text{SILog}} = \frac{1}{n} \sum_{i=1}^{n} z_i^2 - \frac{1}{n^2} \left(\sum_{i=1}^{n} z_i \right)^2, \qquad z_i = \log \hat{d}_i - \log d_i, \tag{12}$$

where d_i and $\hat{d}_i$ denote the full-precision and quantized depth maps at pixel i, respectively, and n is the number of valid pixels. The final objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{feat}} + \lambda_{\text{SILog}} \mathcal{L}_{\text{SILog}}, \tag{13}$$

which is used to jointly optimize all CWDA parameters in the encoder and decoder.

4 Experiments

4.1 Experimental Setup

Models, datasets, and evaluation. We use Depth Anything v1 [14] and v2 [15], built on the DINOv2 [11] encoder with ViT-S and ViT-B backbones, as representative foundation models for monocular depth estimation. We evaluate SPA-Q on five MDE benchmarks: NYUv2 [36], KITTI [37], ETH3D [38], DIODE [39], and Sintel [40]. These datasets correspond to the official zero-shot relative depth evaluation benchmarks used in Depth Anything. 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.

Table 1: Comparison of PTQ methods on NYUv2 [36], KITTI [37], ETH3D [38], DIODE [39], and Sintel [40] 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 [25]	4/4	0.2535	0.5395	0.2886	0.4692	0.2714	0.5042	0.2910	0.4646
	QDrop [26]	4/4	0.1742	0.7166	0.2334	0.5785	0.1773	0.7115	0.2332	0.5794
	RepQ-ViT [27]	4/4	0.1959	0.6639	0.2539	0.5410	0.2053	0.6385	0.2417	0.5634
	QSCA [33]	4/4	0.1377	0.8097	0.1875	0.6845	0.1361	0.8151	0.1875	0.6845
	SPA-Q (Ours)	4/4	**0.1090**	**0.8771**	**0.1021**	**0.8991**	**0.1081**	**0.8808**	**0.1001**	**0.9060**
KITTI	FP	32/32	0.0818	0.9369	0.0804	0.9396	0.0832	0.9340	0.0814	0.9389
	BRECQ [25]	4/4	0.3719	0.3522	0.3989	0.3160	0.3906	0.3344	0.3990	0.3175
	QDrop [26]	4/4	0.3934	0.3234	0.3855	0.3338	0.3620	0.3748	0.3412	0.4082
	RepQ-ViT [27]	4/4	0.3434	0.4159	0.3540	0.3915	0.3634	0.3897	0.3906	0.3287
	QSCA [33]	4/4	0.1874	0.7273	0.2365	0.6203	0.2067	0.6794	0.2296	0.6174
	SPA-Q (Ours)	4/4	**0.1622**	**0.7826**	**0.1840**	**0.7230**	**0.1807**	**0.7449**	**0.1696**	**0.7702**
ETH3D	FP	32/32	0.0584	0.9652	0.0513	0.9741	0.0548	0.9701	0.0467	0.9791
	BRECQ [25]	4/4	0.2829	0.5082	0.2958	0.4874	0.2920	0.4962	0.2963	0.4870
	QDrop [26]	4/4	0.2351	0.6069	0.2706	0.5373	0.2288	0.6186	0.2697	0.5341
	RepQ-ViT [27]	4/4	0.2195	0.6286	0.2491	0.5670	0.2215	0.6299	0.2606	0.5538
	QSCA [33]	4/4	0.1791	0.7332	0.2241	0.6309	0.1983	0.6791	0.2197	0.6298
	SPA-Q (Ours)	4/4	**0.1485**	**0.8071**	**0.1512**	**0.7943**	**0.1533**	**0.7958**	**0.1444**	**0.8140**
DIODE	FP	32/32	0.0753	0.9413	0.0745	0.9474	0.0721	0.9426	0.0701	0.9498
	BRECQ [25]	4/4	0.2079	0.6755	0.2201	0.6464	0.2173	0.6538	0.2211	0.6441
	QDrop [26]	4/4	0.1864	0.7275	0.2058	0.6819	0.1848	0.7279	0.2062	0.6777
	RepQ-ViT [27]	4/4	0.1746	0.7538	0.2016	0.6891	0.1800	0.7552	0.1999	0.6945
	QSCA [33]	4/4	0.1513	0.8033	0.1997	0.7099	0.1463	0.8135	0.1947	0.7094
	SPA-Q (Ours)	4/4	**0.1240**	**0.8653**	**0.1314**	**0.8595**	**0.1342**	**0.8524**	**0.1184**	**0.8901**
Sintel	FP	32/32	0.2296	0.7304	0.2281	0.7539	0.2680	0.6974	0.2576	0.7111
	BRECQ [25]	4/4	0.4787	0.3123	0.4820	0.3085	0.4816	0.3055	0.4821	0.3008
	QDrop [26]	4/4	0.4991	0.3181	0.4867	0.3106	0.4948	0.3184	0.4882	0.3151
	RepQ-ViT [27]	4/4	0.4722	0.3619	0.4642	0.3629	0.4693	0.3829	0.4838	0.3237
	QSCA [33]	4/4	0.4531	0.3884	0.4781	0.4059	0.4474	0.3919	0.4679	0.4070
	SPA-Q (Ours)	4/4	**0.4377**	**0.4318**	**0.3949**	**0.5382**	**0.4211**	**0.4225**	**0.3811**	**0.4892**

Implementation details. For fair comparison, we follow the experimental settings of prior work [33]. 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. We optimize all CWDA parameters in an end-to-end manner for one epoch using 5% of the training images without ground-truth depth, with $\lambda_{\text{SILog}} = 0.5$. 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.

4.2 Experimental Results

Quantitative results. As shown in Table 1, we compare SPA-Q with existing PTQ methods, including BRECQ [25], QDrop [26], RepQ-ViT [27], and QSCA [33], under 4-bit quantization on five zero-shot relative depth estimation benchmarks. Overall, our framework consistently achieves the best performance in both AbsRel and δ_1 across all evaluated datasets, outperforming all compared PTQ methods. For Depth Anything v1 with the ViT-B backbone, the improvement is substantial on the two representative benchmarks, NYUv2 [36] and KITTI [37]. On NYUv2, our framework reduces AbsRel from 0.1875 to 0.1021 and improves δ_1 from 0.6845 to 0.8991. On KITTI, AbsRel

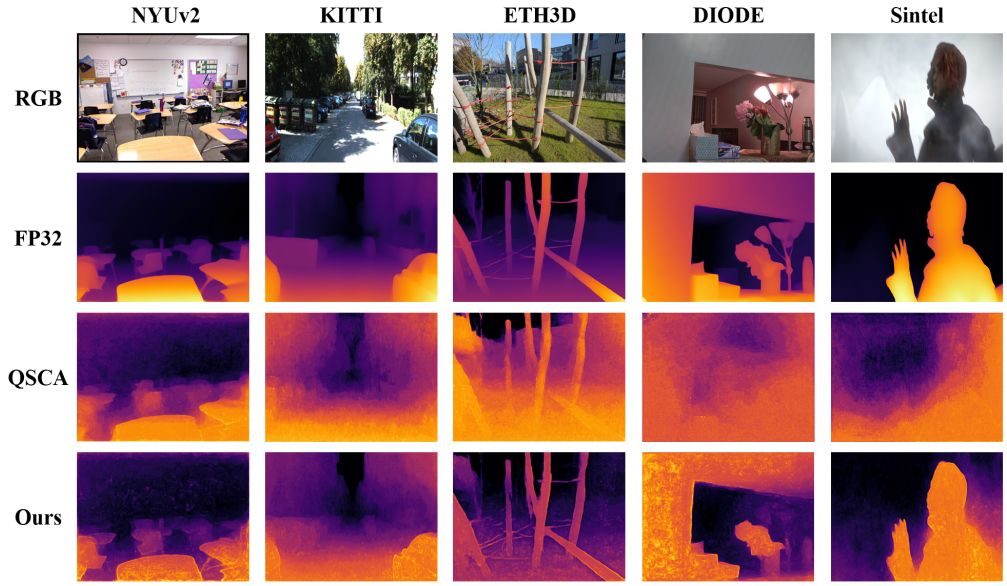


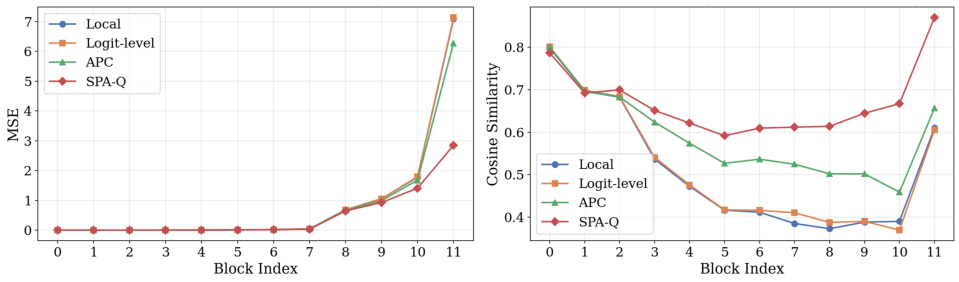
Figure 4: Qualitative comparison of 4-bit quantization results for Depth Anything with a ViT-S backbone on benchmarks: NYUv2 [36], KITTI [37], ETH3D [38], DIODE [39], and Sintel [40].

is further reduced from 0.2365 to 0.1840 and δ_1 is improved from 0.6203 to 0.7230. These results show that the proposed framework effectively preserves depth estimation performance under severe low-bit quantization in both indoor and outdoor settings. Consistent improvements are also observed on ETH3D [38], DIODE [39], and Sintel [40], where SPA-Q achieves the best results among the compared PTQ methods. In particular, the improvement remains consistent on Sintel, a challenging synthetic benchmark, indicating that the proposed framework generalizes well not only to indoor and outdoor scenes but also to synthetic domains under 4-bit quantization.

Qualitative results. To validate the effectiveness of the proposed framework, we present qualitative comparisons of predicted depth maps in Figure 4. Compared with QSCA [33], SPA-Q produces qualitatively better depth maps with improved structural consistency and sharper object boundaries across diverse indoor, outdoor, and synthetic scenes. For NYUv2 [36], it better captures the indoor layout, including the desk and chair boundaries, while providing clearer separation between foreground objects and distant regions. On KITTI [37], it produces smoother depth transitions across the road scene, with better separation between nearby vehicles and the background. Similarly, on ETH3D [38] and DIODE [39], thin structures such as wooden poles and ropes, as well as the boundary of the wall opening and object silhouettes, are well recovered, whereas QSCA largely fails to preserve them. Sintel [40] is particularly challenging due to its domain gap from real-world imagery and rendering effects such as atmospheric haze and motion blur. While QSCA collapses into a nearly radial depth pattern that loses the character entirely, SPA-Q recovers its contours, including the raised hand, consistent with the full-precision prediction. These results demonstrate that the proposed framework consistently preserves depth structure across diverse domains, including indoor, outdoor, and synthetic scenes.

4.3 Ablation Study

Comparison of calibration objectives. We conduct an ablation study on NYUv2 [36] and KITTI [37] 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 changing from 0.1908 to 0.1839 and δ_1 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 final depth performance. In contrast, replacing the logits-level objective with APC further improves performance, reducing AbsRel to 0.1444 and improving δ_1 to 0.7990. Finally, combining APC with



(a) (b)

Figure 5: Block-wise feature analysis under 4-bit quantization for Depth Anything v1 [14] 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.

Table 2: Ablation study of calibration objectives under 4-bit quantization for Depth Anything v1 with a ViT-B backbone on NYUv2 [36] and KITTI [37].

Dataset	Method	AbsRel ↓	δ_1 ↑
NYUv2	Local	0.1908	0.6836
	Logits-level	0.1839	0.7013
	APC	0.1444	0.7990
	SPA-Q	**0.1021**	**0.8991**
KITTI	Local	0.3241	0.4366
	Logits-level	0.3002	0.4828
	APC	0.2490	0.5789
	SPA-Q	**0.1840**	**0.7230**

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.

Encoder	Method	Model size (MB)	Params (M)	BOPs (G)
ViT-S	FP	94.55	24.79	85742
	QSCA [33]	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 [33]	60.02	99.24	5092
	SPA-Q*	51.84	97.65	4493
	SPA-Q	51.16	97.47	4270

CWDA yields the best performance, reducing AbsRel to 0.1021 and improving δ_1 to 0.8991. KITTI shows a similar trend, where logits-level calibration provides only limited gains, while APC and the full SPA-Q framework lead to clear performance improvements.

Block-wise comparison of feature similarity. We further evaluate whether SPA-Q preserves encoder features for dense depth prediction across transformer blocks. Since the decoder reconstructs depth maps from encoder features, preserving their fidelity is important for maintaining depth structure under quantization. As shown in Figure 5 (a), SPA-Q consistently achieves the lowest MSE across most blocks, with the advantage becoming more pronounced in deeper blocks. Figure 5 (b) shows that SPA-Q yields the highest cosine similarity to the full-precision features throughout the encoder, indicating better block-wise feature preservation. These results demonstrate that APC and CWDA jointly preserve the encoder features required for depth prediction by improving attention-aware calibration and mitigating accumulated feature distortion.

Efficiency analysis. We further compare model size, parameter count, and bit operations (BOPs) of SPA-Q and QSCA [33] 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 after training, so the fused model restores the parameter count to that of the original quantized network. These results show that SPA-Q improves depth estimation accuracy without introducing additional inference overhead.

5 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, while accumulated quantization errors induce distribution shifts in intermediate features. To address these issues, we proposed SPA-Q, a structure-preserving adaptive PTQ framework that combines APC and CWDA to preserve attention distributions and mitigate feature distortion. Experimental results on multiple zero-shot depth estimation benchmarks showed that SPA-Q consistently outperforms existing PTQ methods under 4-bit quantization.

Limitations & Future Works. Although the proposed method consistently outperforms existing PTQ baselines under 4-bit quantization and yields improved qualitative results, the quantized depth maps remain visibly less clean than their full-precision counterparts. We conjecture that this limitation stems from the restricted set of depth values representable under low-bit quantization, which is more discrete than in full precision and can reduce the smoothness of predicted depth maps. Our study is limited to MDE with Depth Anything v1 and v2, partly due to GPU memory constraints for larger foundation models. Extending SPA-Q to larger foundation models and other depth estimation settings, such as stereo and video depth estimation, remains an important direction for future work.

Acknowledgements

This work was supported by 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] Xingshuai Dong, Matthew A Garratt, Sreenatha G Anavatti, and Hussein A Abbass. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16940–16961, 2022.

[2] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019.

[3] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

[4] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. 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*, pages 8445–8453, 2019.

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65 (1):99–106, 2021.

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023.

[7] Jincheol Yang, Matti Zinke, Beoungwoo Kang, Hyung Uk Cho, Hyunyoung Choi, SangGu Lee, and Suk-Ju Kang. Efficient monocular depth-based physical distance measurement for low-depth scales. *IEEE Transactions on Industrial Informatics*, 2025.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[12] 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.

[13] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 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, 2022.

[14] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024.

[15] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[16] 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.

[17] 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.

[18] 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.

[19] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3009–3018. IEEE, 2019.

[20] 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.

[21] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021.

[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] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

[26] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022.

[27] 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.

[28] 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.

[29] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[30] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[31] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[32] 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*, pages 10012–10022, 2021.

[33] Jincheol Yang, Jaemin Choi, Matti Zinke, and Suk-Ju Kang. Qsca: Quantization with self-compensating auxiliary for monocular depth estimation. *Advances in Neural Information Processing Systems*, 38: 31444–31465, 2026.

[34] 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.

[35] Tao Jiang, Yucheng Jiang, Xiwen Yao, Gong Cheng, and Junwei Han. Uq-vit: Harmonizing extreme activations with hardware-friendly uniform quantization in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22354–22362, 2026.

[36] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[37] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

[38] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 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*, pages 3260–3269, 2017.

[39] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[40] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[41] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2810–2819, 2019.

[42] 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.

[43] 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.

[44] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019.

[45] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pages 75602–75637, 2025.

[46] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 608–617. IEEE, 2020.

[47] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *European Conference on Computer Vision*, pages 331–348. Springer, 2018.

Appendix

A Extended Comparison with Additional PTQ Baselines

Table 4: Extended comparison with additional PTQ baselines on NYUv2 [36] and KITTI [37] 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
	MinMax [16]	4/4	0.2728	0.5024	1.5735	0.1972	0.2815	0.4873	2.3712	0.1102
	Percentile [41]	4/4	0.2006	0.6542	0.2522	0.5430	0.1935	0.6675	0.2543	0.5379
	BRECQ [25]	4/4	0.2535	0.5395	0.2886	0.4692	0.2714	0.5042	0.2910	0.4646
	QDrop [26]	4/4	0.1742	0.7166	0.2334	0.5785	0.1773	0.7115	0.2332	0.5794
	PTQ4ViT [42]	4/4	0.2393	0.5693	0.2294	0.5895	0.2735	0.5034	0.2694	0.5183
	RepQ-ViT [27]	4/4	0.1959	0.6639	0.2539	0.5410	0.2053	0.6385	0.2417	0.5634
	ERQ [28]	4/4	0.1785	0.7062	0.2182	0.6126	0.1751	0.7140	0.2883	0.4705
	QwT [43]	4/4	0.1407	0.8007	0.2024	0.6486	0.1400	0.8050	0.1992	0.6589
	QSCA [33]	4/4	0.1377	0.8097	0.1875	0.6845	0.1361	0.8151	0.1875	0.6845
	SPA-Q (Ours)	4/4	**0.1090**	**0.8771**	**0.1021**	**0.8991**	**0.1081**	**0.8808**	**0.1001**	**0.9060**
KITTI	FP	32/32	0.0818	0.9369	0.0804	0.9396	0.0832	0.9340	0.0814	0.9389
	MinMax [16]	4/4	0.3770	0.3441	1.9612	0.2058	0.3938	0.3423	4.4358	0.0832
	Percentile [41]	4/4	0.3418	0.4099	0.3876	0.3327	0.3668	0.3780	0.3932	0.3275
	BRECQ [25]	4/4	0.3719	0.3522	0.3989	0.3160	0.3906	0.3344	0.3990	0.3175
	QDrop [26]	4/4	0.3934	0.3234	0.3855	0.3338	0.3620	0.3748	0.3412	0.4082
	PTQ4ViT [42]	4/4	0.3439	0.4106	0.3923	0.3251	0.3308	0.4187	0.3972	0.3200
	RepQ-ViT [27]	4/4	0.3434	0.4159	0.2539	0.5410	0.3634	0.3897	0.3906	0.3287
	ERQ [28]	4/4	0.3178	0.4847	0.3528	0.4241	0.3176	0.4616	0.3490	0.4066
	QwT [43]	4/4	0.1802	0.6862	0.2417	0.5867	0.1951	0.6941	0.2539	0.5346
	QSCA [33]	4/4	0.1874	0.7273	0.2365	0.6203	0.2067	0.6794	0.2296	0.6174
	SPA-Q (Ours)	4/4	**0.1622**	**0.7826**	**0.1840**	**0.7230**	**0.1807**	**0.7449**	**0.1696**	**0.7702**

Table 4 provides an extended comparison with additional PTQ baselines on NYUv2 [36] and KITTI [37] under 4-bit quantization. In addition to the baselines reported in the main paper, we include MinMax [16], Percentile [41], PTQ4ViT [42], ERQ [28], and QwT [43] to further evaluate the robustness of SPA-Q across different calibration and reconstruction strategies. Across Depth Anything v1 [14] and v2 [15] with both ViT-S and ViT-B backbones, SPA-Q achieves the best performance in all evaluated settings. On NYUv2, SPA-Q consistently improves both AbsRel and δ_1 over all compared methods. The improvement is more pronounced for ViT-B backbones, where preserving attention distributions and aligning intermediate feature distributions are especially important under low-bit quantization. On KITTI, SPA-Q also achieves the best results across all model variants, showing that the proposed framework remains effective in outdoor driving scenes. Although QwT provides a strong baseline in some KITTI settings, SPA-Q still achieves lower AbsRel and higher δ_1, demonstrating the benefit of explicitly preserving depth-relevant intermediate representations.

Overall, these extended results confirm that the gains of SPA-Q are not limited to the main PTQ baselines. Instead, SPA-Q remains consistently effective against a broader set of PTQ methods, further supporting the importance of attention-preserving calibration and channel-wise distribution alignment for quantized monocular depth estimation.

Table 5: Extended comparison with additional PTQ baselines on NYUv2 [36] and KITTI [37] for zero-shot relative depth estimation under W4A6 and W4A8 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
	MinMax [16]	4/6	0.2417	0.5632	0.2748	0.4973	0.2599	0.5324	0.2866	0.4738
	Percentile [41]	4/6	0.1050	0.8837	0.0958	0.9071	0.0891	0.9196	0.0831	0.9355
	BRECQ [25]	4/6	0.2337	0.5786	0.2462	0.5542	0.2611	0.5269	0.2697	0.5084
	QDrop [26]	4/6	0.2071	0.6369	0.1823	0.6987	0.1700	0.7285	0.1560	0.7619
	PTQ4ViT [42]	4/6	0.2166	0.6187	0.2201	0.6122	0.2642	0.5159	0.2275	0.6095
	RepQ-ViT [27]	4/6	0.1304	0.8256	0.1058	0.8867	0.1131	0.8657	0.0918	0.9158
	ERQ [28]	4/6	0.0945	0.9075	0.0766	0.9390	0.0874	0.9245	0.0914	0.9152
	QwT [43]	4/6	0.0884	0.9189	0.0928	0.9089	0.0833	0.9323	0.1000	0.8944
	QSCA [33]	4/6	0.0810	0.9333	0.0739	0.9441	0.0757	0.9450	0.0726	0.9468
	SPA-Q (Ours)	4/6	**0.0732**	**0.9473**	**0.0623**	**0.9597**	**0.0735**	**0.9485**	**0.0647**	**0.9603**
	MinMax [16]	4/8	0.1522	0.7740	0.2253	0.5934	0.1345	0.8187	0.2507	0.5465
	Percentile [41]	4/8	0.0861	0.9233	0.0648	0.9588	0.0798	0.9375	0.0604	0.9621
	BRECQ [25]	4/8	0.1182	0.8533	0.2152	0.6166	0.0999	0.8927	0.2193	0.6095
	QDrop [26]	4/8	0.0700	0.9504	0.0599	0.9627	**0.0651**	**0.9592**	0.0601	0.9621
	PTQ4ViT [42]	4/8	0.2001	0.6544	0.1877	0.6837	0.2596	0.5246	0.2190	0.6328
	RepQ-ViT [27]	4/8	0.1066	0.8833	0.0864	0.9259	0.1066	0.8792	0.0788	0.9367
	ERQ [28]	4/8	0.0812	0.9333	0.0608	0.9632	0.0809	0.9355	0.0597	0.9641
	QwT [43]	4/8	0.0769	0.9393	0.0789	0.9348	0.0663	0.9564	0.0808	0.9300
	QSCA [33]	4/8	0.0682	0.9551	0.0657	0.9542	0.0679	0.9549	0.0599	0.9635
	SPA-Q (Ours)	4/8	**0.0664**	**0.9574**	**0.0560**	**0.9672**	**0.0651**	0.9591	**0.0562**	**0.9682**
KITTI	FP	32/32	0.0818	0.9369	0.0804	0.9396	0.0832	0.9340	0.0814	0.9389
	MinMax [16]	4/6	0.3290	0.4467	0.3769	0.3586	0.3744	0.3609	0.3982	0.3161
	Percentile [41]	4/6	0.1269	0.8558	0.1687	0.7580	0.1283	0.8613	0.1866	0.7219
	BRECQ [25]	4/6	0.2851	0.5223	0.3529	0.4035	0.3332	0.4320	0.3497	0.4093
	QDrop [26]	4/6	0.2408	0.6052	0.2611	0.5897	0.2790	0.5351	0.3037	0.5268
	PTQ4ViT [42]	4/6	0.3230	0.4446	0.3848	0.3488	0.3614	0.3931	0.3917	0.3308
	RepQ-ViT [27]	4/6	0.1502	0.7888	0.2602	0.5610	0.1292	0.8293	0.1458	0.8109
	ERQ [28]	4/6	0.1244	0.8629	0.1133	0.8705	0.1144	0.8770	0.1115	0.8685
	QwT [43]	4/6	**0.1152**	0.8844	0.1759	0.7292	0.1169	0.8872	0.1550	0.7700
	QSCA [33]	4/6	0.1161	**0.8857**	0.1174	0.8722	**0.1124**	**0.8893**	0.1067	0.8873
	SPA-Q (Ours)	4/6	0.1247	0.8849	**0.1023**	**0.8907**	0.1684	0.7892	**0.0994**	**0.9003**
	MinMax [16]	4/8	0.3486	0.4215	0.3275	0.4845	0.3529	0.4237	0.3327	0.4376
	Percentile [41]	4/8	0.1111	0.8740	0.1263	0.8564	0.1051	0.8892	0.1045	0.9003
	BRECQ [25]	4/8	0.3532	0.4070	0.3317	0.4513	0.3194	0.4575	0.3239	0.4493
	QDrop [26]	4/8	0.3071	0.4883	0.2754	0.5422	0.2713	0.5368	0.2688	0.5616
	PTQ4ViT [42]	4/8	0.3224	0.4580	0.3813	0.3550	0.3538	0.3909	0.3969	0.3205
	RepQ-ViT [27]	4/8	0.1375	0.8145	0.2227	0.6436	0.1212	0.8431	0.1136	0.8683
	ERQ [28]	4/8	0.1164	0.8801	0.1005	0.8937	0.1040	0.8881	0.0988	0.9043
	QwT [43]	4/8	0.1044	0.8970	0.1698	0.7362	0.1061	**0.9017**	0.1328	0.8228
	QSCA [33]	4/8	**0.1019**	0.9022	0.1045	0.8963	**0.1020**	0.9010	0.0931	**0.9161**
	SPA-Q (Ours)	4/8	0.1072	**0.9035**	**0.0904**	**0.9148**	0.1057	0.8990	**0.0920**	0.9135

We further evaluate SPA-Q on additional MDE models, MiDaS-BEiT-B-384 [13], DPT-L-384 [31], and ZoeDepth [30], as shown in Table 6. Models with BEiT backbones, namely MiDaS and ZoeDepth, suffer more severe degradation under 4-bit quantization than DPT-L. Nevertheless, SPA-Q consistently outperforms the state-of-the-art QSCA [33] across all models and datasets.

B Generalization Across MDE Models

Table 6: Comparison of PTQ methods on NYUv2 [36] and KITTI [37] with additional MDE architectures, MiDaS-BEiT-B-384 [13], DPT-L-384 [31], and ZoeDepth [30], under 4-bit quantization.

Dataset	Method	W/A	MiDaS-BEiT-B-384		DPT-L-384		ZoeDepth (Metric)	
			AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
NYUv2	FP	32/32	0.0663	0.9568	0.0947	0.9112	0.0755	0.9544
	RepQ-ViT [27]	4/4	0.2603	0.5254	0.1394	0.8139	0.4949	0.1033
	QSCA [33]	4/4	0.2253	0.5998	0.1380	0.8192	0.3969	0.2054
	SPA-Q (Ours)	4/4	**0.1803**	**0.7081**	**0.1276**	**0.8425**	**0.3051**	**0.3894**
KITTI	FP	32/32	0.1928	0.7076	0.1079	0.8858	0.0589	0.9664
	RepQ-ViT [27]	4/4	0.3761	0.3580	0.1661	0.7824	0.4650	0.1529
	QSCA [33]	4/4	0.3740	0.3628	0.1650	0.7810	0.2701	0.5237
	SPA-Q (Ours)	4/4	**0.3250**	**0.4383**	**0.1427**	**0.8379**	**0.2551**	**0.6171**

To examine whether SPA-Q generalizes beyond Depth Anything, we further evaluate it on three additional MDE models with different backbones and prediction targets: MiDaS-BEiT-B-384 [13] and DPT-L-384 [31] for relative depth estimation, and ZoeDepth [30] for metric depth estimation. As shown in Table 6, SPA-Q achieves the best AbsRel and δ_1 for all three models on both NYUv2 [36] and KITTI [37] under 4-bit quantization, consistently outperforming RepQ-ViT [27] and QSCA [33]. These results indicate that SPA-Q is not tailored to a specific architecture and is applicable to a broad range of MDE models, including those for metric depth estimation.

C Evaluation on Structural Metrics

Table 7: Comparison of edge, boundary, and planarity metrics for Depth Anything v1 [14] under W4A4 and W4A6 quantization. Edge F1 [44], Boundary F1 [45], and ORD [46] are evaluated on NYUv2 [36], and planarity errors (PE) are evaluated on iBims-1 [47].

Encoder	Method	W/A	Depth Anything v1				
			Boundary & Ordinal Accuracy			Planarity Errors (cm / °)	
			Edge F1@0.5 ↑	Boundary F1 ↑	ORD ↓	$\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}$ ↓	$\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}$ ↓
ViT-S	FP	32/32	0.1862	0.0438	0.1177	1.3337	4.4868
	RepQ-ViT [27]	4/4	0.1025	0.0018	0.5405	11.4530	35.0064
	QSCA [33]	4/4	**0.1292**	0.0126	0.2956	**10.5740**	24.1218
	SPA-Q (Ours)	4/4	0.1286	**0.0161**	**0.2423**	13.1100	**19.8219**
	RepQ-ViT [27]	4/6	0.1251	0.0174	0.2795	5.4846	14.5692
	QSCA [33]	4/6	**0.1915**	**0.0475**	**0.1627**	**4.0663**	9.8937
	SPA-Q (Ours)	4/6	0.1907	0.0436	**0.1627**	4.2679	**9.0219**
ViT-B	FP	32/32	0.1991	0.0477	0.1005	1.0963	4.1992
	RepQ-ViT [27]	4/4	0.1011	0.0011	0.7412	14.1688	44.1632
	QSCA [33]	4/4	0.1069	0.0018	0.4732	14.0001	34.3514
	SPA-Q (Ours)	4/4	**0.1111**	**0.0091**	**0.2441**	**13.6755**	**20.1182**
	RepQ-ViT [27]	4/6	0.1172	0.0180	0.2418	4.1241	15.4422
	QSCA [33]	4/6	0.1658	0.0328	0.1672	5.6829	11.1425
	SPA-Q (Ours)	4/6	**0.1938**	**0.0404**	**0.1377**	**3.7100**	**7.7550**

We further report structural metrics to evaluate how well each method preserves the geometric structure of depth maps, including boundary and ordinal accuracy metrics (Edge F1, Boundary F1, and ORD) and planarity errors, as shown in Table 7. SPA-Q achieves the best performance on most metrics under both W4A4 and W4A6 quantization, indicating that it effectively preserves fine depth structures such as object boundaries, relative depth ordering, and planar surfaces.

D Block-wise Feature Similarity Analysis

Figure 6 and Figure 7 provide an extended block-wise feature similarity analysis under 4-bit quantization across Depth Anything v1 [14] and v2 [15] on NYUv2 [36] and KITTI [37]. For each setting, we compare the MSE and cosine similarity between full-precision and quantized encoder features across transformer blocks. Across all settings, SPA-Q consistently yields lower feature MSE and higher cosine similarity than the compared calibration objectives, especially in deeper blocks.

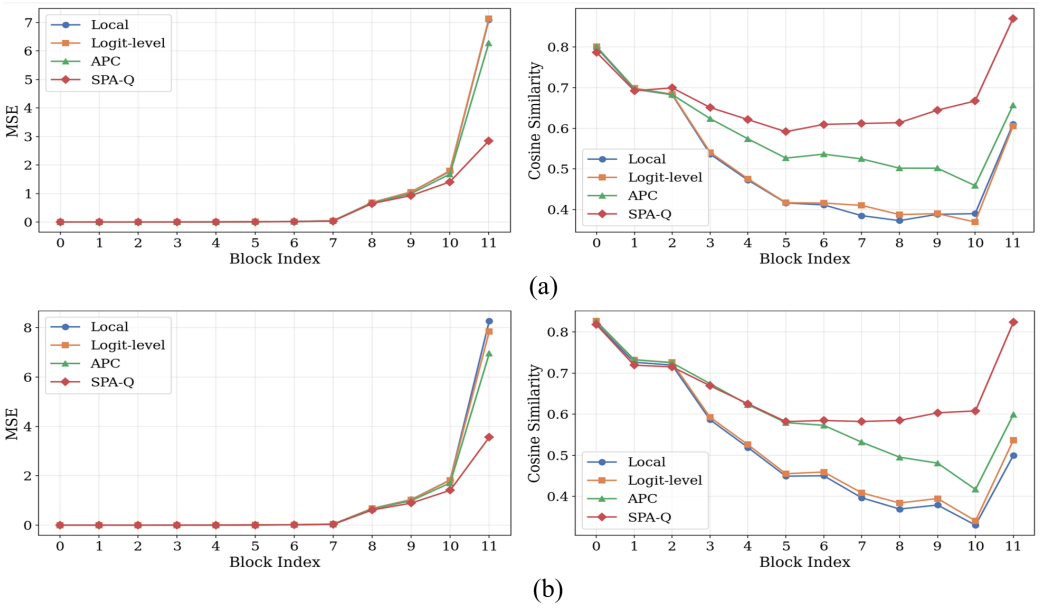


Figure 6: Block-wise feature similarity analysis under 4-bit quantization on Depth Anything v1. (a) and (b) show results on NYUv2 [36] and KITTI [37]. We compare the MSE and cosine similarity between full-precision and quantized encoder features across transformer blocks.

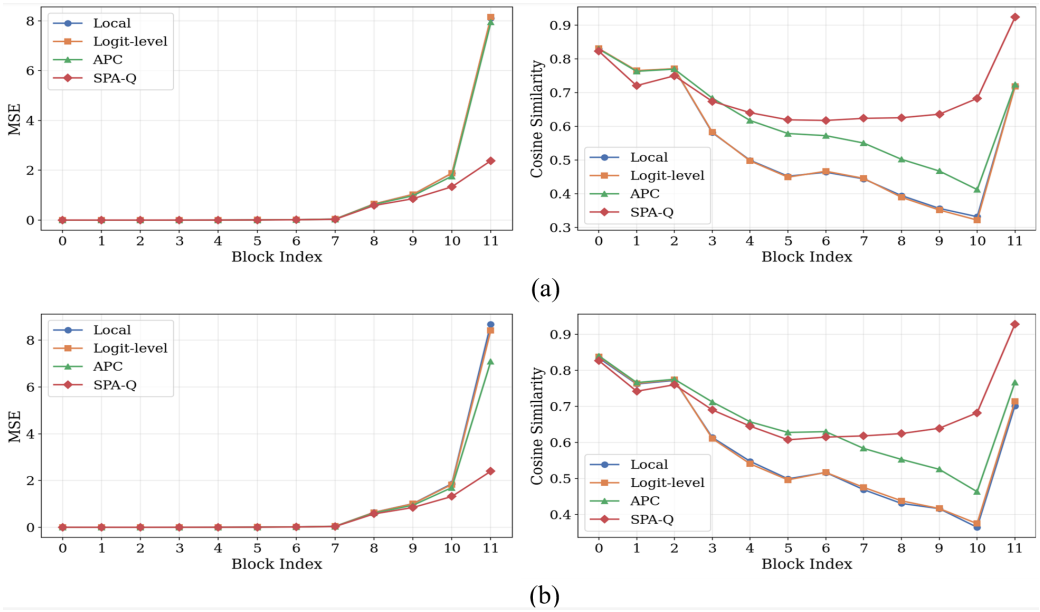


Figure 7: Block-wise feature similarity analysis under 4-bit quantization on Depth Anything v2. (a) and (b) show results on NYUv2 [36] and KITTI [37]. We compare the MSE and cosine similarity between full-precision and quantized encoder features across transformer blocks.

E Qualitative Analysis of Feature Maps

Figures 8, 9 visualize feature maps on NYUv2 [36], while Figures 10 and 11 show the corresponding results on KITTI [37]. We compare the full-precision model, naive W4A4 quantization, and SPA-Q. Naive W4A4 quantization substantially distorts the spatial activation patterns of intermediate features, which the decoder relies on for dense depth prediction. In contrast, SPA-Q yields feature maps with higher fidelity to their full-precision counterparts, preserving spatial layouts, object regions, and boundary-related responses. These results show that SPA-Q mitigates feature distortion caused by low-bit quantization across different Depth Anything models, backbones, and datasets.

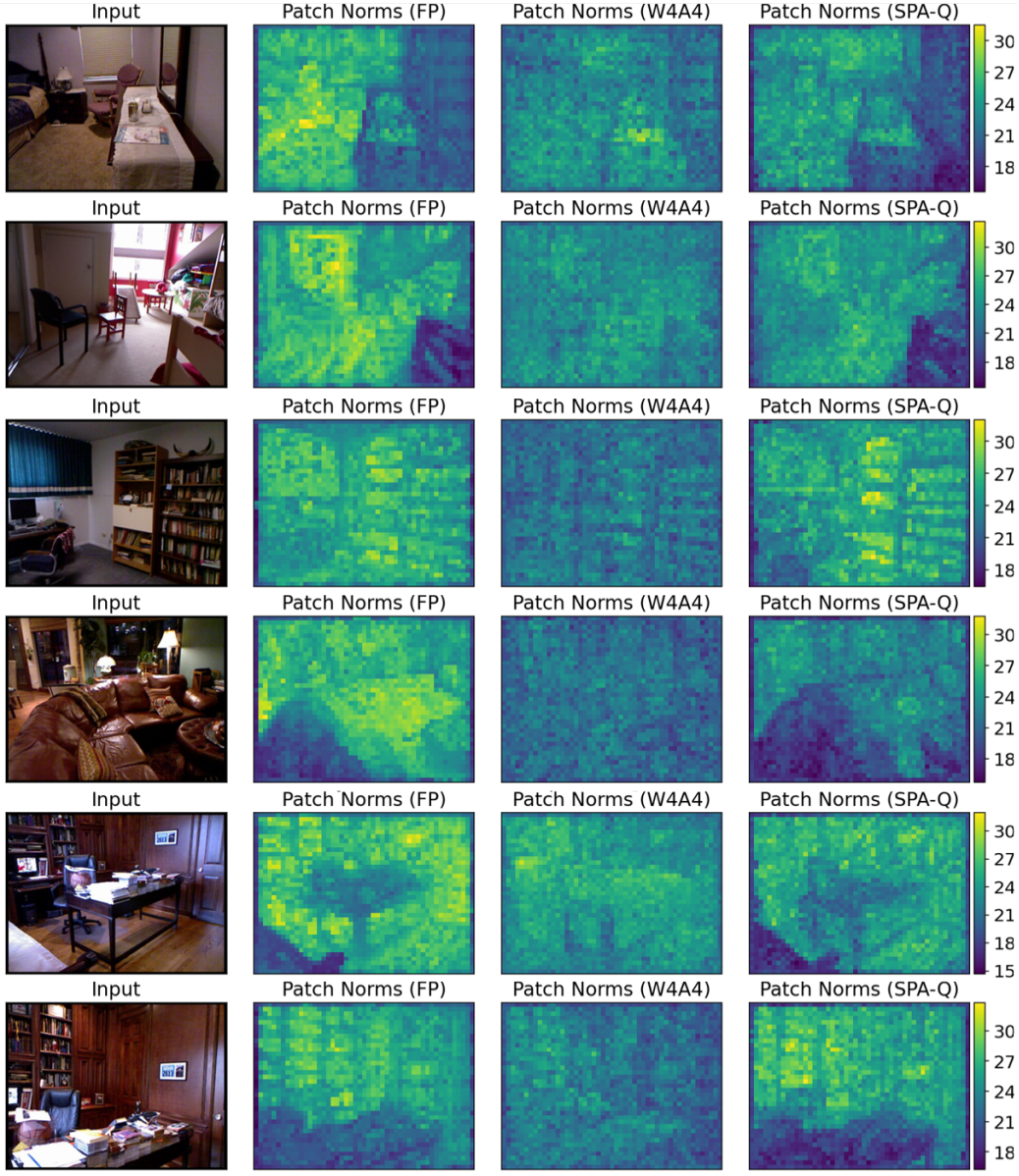


Figure 8: Feature map visualization for Depth Anything v2 with a ViT-S backbone on NYUv2 [36] under 4-bit quantization. We compare the full-precision model, W4A4 quantization, and SPA-Q.

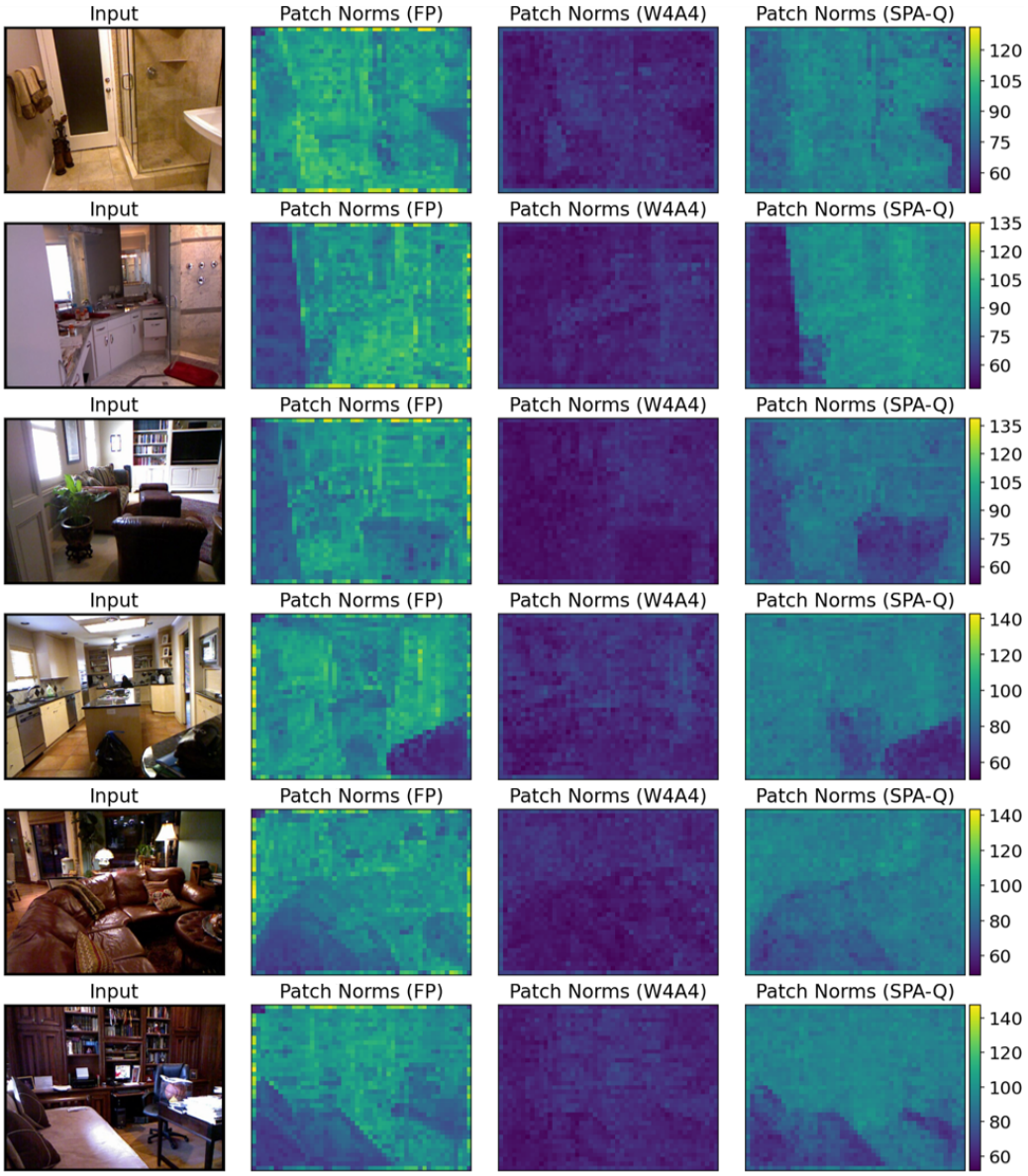


Figure 9: Feature map visualization for Depth Anything v1 with a ViT-B backbone on NYUv2 [36] under 4-bit quantization. We compare the full-precision model, W4A4 quantization, and SPA-Q.

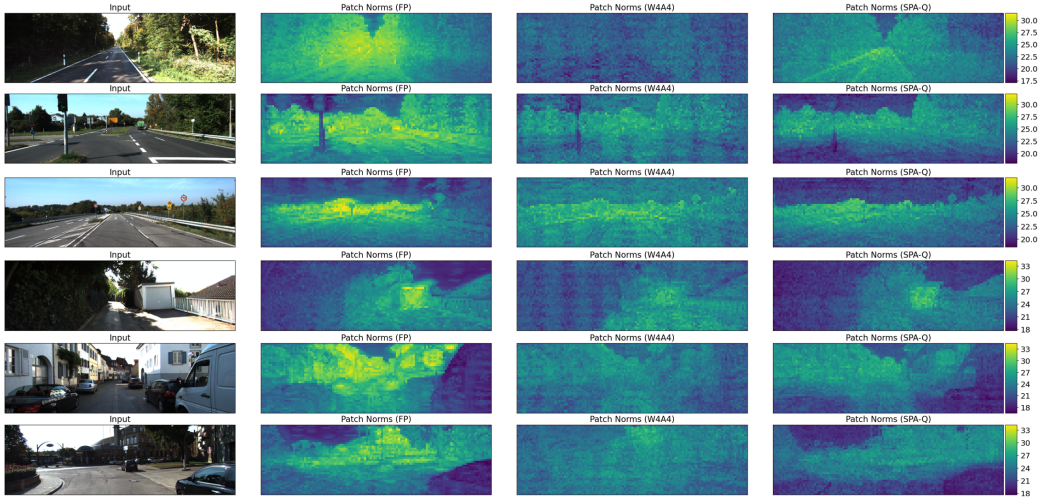


Figure 10: Feature map visualization for Depth Anything v2 with a ViT-S backbone on KITTI [37] under 4-bit quantization. We compare the full-precision model, W4A4 quantization, and SPA-Q.

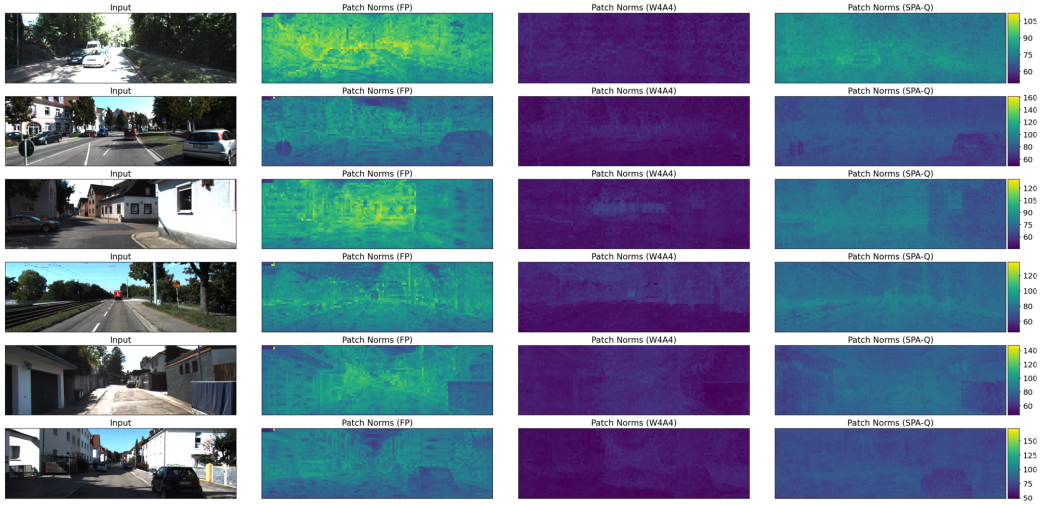


Figure 11: Feature map visualization for Depth Anything v1 with a ViT-B backbone on KITTI [37] under 4-bit quantization. We compare the full-precision model, W4A4 quantization, and SPA-Q.

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 main claims in the abstract and introduction match the paper's contributions and experimental results. The paper claims that SPA-Q addresses attention-preservation and intermediate feature distribution shift in low-bit PTQ for MDE, and these claims are supported by the method description and experiments in Secs. 1, 3, and 4.

   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:

   Justification: The paper explicitly discusses limitations in the Limitations & Future Works section. In particular, it notes that a noticeable gap to full precision still remains under severe low-bit quantization, that quantized depth maps can appear less smooth due to the restricted set of representable depth values, and that the current study is limited to monocular depth estimation rather than stereo or video depth estimation.

   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 include formal theoretical results such as theorems, lemmas, propositions, or proofs. It presents objective formulations and empirical analyses, but no claims requiring formal proof.

   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 main experiments are conducted on publicly available models and datasets, and the paper specifies the model variants, datasets, evaluation metrics, calibration protocol, optimization objectives, and training details needed for reproduction. Since the evaluation is performed on open benchmarks with standard settings, the reported results can be reproduced by other researchers.

   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 evaluates the proposed method on publicly available datasets and provides the experimental details needed for reproduction, but it does not release implementation code, reproduction scripts, or an anonymized repository link in the current submission.

   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 specifies the main experimental details needed to understand the reported results, including the evaluated model variants and backbones, datasets and evaluation metrics, calibration set construction, quantization setting, CWDA initialization, optimization strategy, loss weighting, optimizer, learning rate, batch size, and hardware configuration.

   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 does not report formal measures of statistical significance. However, the experiments follow the standard PTQ evaluation protocol used in prior work, and all methods are compared under the same quantization setting, calibration procedure, and evaluation benchmarks.

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 reports that all experiments are conducted on a single RTX 6000 Ada Generation GPU, and also provides model size, parameter count, and BOPs for the evaluated models.

   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: We have reviewed the NeurIPS Code of Ethics, and this work conforms to it. The paper studies model quantization for monocular depth estimation using standard public benchmarks and does not involve human subjects, personal data collection, or high-risk model release.

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 manuscript does not include a dedicated broader impacts discussion because this work focuses on a PTQ method for existing MDE models, rather than introducing a new dataset, deployed system, or application. Its societal impacts are mainly indirect and tied to the underlying MDE applications, while the paper focuses on model compression and evaluation under standard benchmarks.

    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: This paper does not release new datasets, models, or other assets that require special safeguards for responsible use. Therefore, this question is not applicable to the current work.

    - 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, models, and prior methods, and properly credits their creators through citations. We respect the corresponding licenses and terms of use for all existing assets used in this work.

    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 new datasets, code, models, or other assets. Therefore, this question is not applicable to the current work.

    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. Therefore, this question is not applicable to the current work.

    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 human subjects research or crowdsourcing. Therefore, IRB approval or an equivalent review is not applicable to this work.

    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 were used only for writing assistance in preparing the manuscript. They were not used in the core methodology, experiments, or analysis, and therefore did 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.