BITROT: UNDERSTANDING QUANTIZATION FRAGILITY IN VISION TRANSFORMERS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) enables efficient deployment of Vision Transformers, but preserving clean ImageNet accuracy do not necessarily preserve performance under distribution shift. This is important in deployment settings where low-precision models may encounter corruptions, visual domain shifts, and naturally adversarial examples, yet the interaction between aggressive quantization and robustness under distribution shift remains underexplored. We systematically study this interaction across architectures, PTQ methods, pretraining strategies, model scales, activation precisions, calibration settings, and several forms of distribution shift. We find that W4A4 models can retain near-lossless in-distribution accuracy while suffering substantially larger degradation on shifted data, with the severity varying across distributions and model configurations. We further test whether this fragility is primarily explained by activation-statistics and calibration mismatch. Mixed calibration using samples from multiple shifted distributions provides only modest and inconsistent recovery, leaving a substantial gap to full precision. Motivated by these observations, we develop a post-hoc prediction-level model that treats quantization as a perturbation to full-precision classification margins and characterizes susceptibility through both the available margin and its sensitivity to accumulated quantization residuals. The resulting predictor tracks quantization-induced degradation across distributions, architectures, model scales, and activation bit-widths. At the sample level, examples with comparable full-precision margins but higher directional quantization sensitivity are consistently more likely to be misclassified after quantization. These results show that clean ImageNet accuracy alone can substantially underestimate the deployment cost of low-precision ViTs, and that robustness under distribution shift should be evaluated explicitly when assessing PTQ methods.

1 INTRODUCTION

Quantization of vision transformer (Dosovitskiy et al., 2021) models has been a popular choice for model compression relying on lower numerical precision to substantially reduce computational and memory footprint of these models, enabling easy deployment on edge and embedded platforms. Post-training quantization (PTQ) (Nagel et al., 2021) and Quantization-aware Training (QAT) (Esser et al., 2019) have been the go to strategies, with PTQ methods being preferred due to low cost compared to QAT's expensive training process. Recent ViT PTQ methods (Wu et al., 2025; 2024) report only modest clean-accuracy degradation at 4-bit precision, often approaching full-precision performance. When judged primarily by clean accuracy, these results suggest that aggressively quantized ViTs can be deployed with little loss in predictive performance.

However, clean ImageNet accuracy alone does not characterize how these models behave under deployment-time distribution shifts. To the best of our knowledge, PTQ methods have predominantly evaluated ImageNet pretrained ViTs on clean ImageNet data while using the same distribution for calibrating the quantized model. In practical deployment settings, a model can encounter a broader range of conditions such as corruptions, sensor noise, style renditions, and natural adversarial examples. How such shifts affect aggressively quantized ViTs remains comparatively underexplored. As shown in figure 1, we find that the near lossless performance on clean ImageNet at 4-bits

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

doesn't guarantee analogous performance under distribution shifts. The 4-bit models (W4A4) suffer substantial degradation on corruptions, style renditions and natural adversarial shifts from their full precision counterparts. Moreover, this degradation, as seen in figure 1, varies across datasets, model architectures, and PTQ methods.

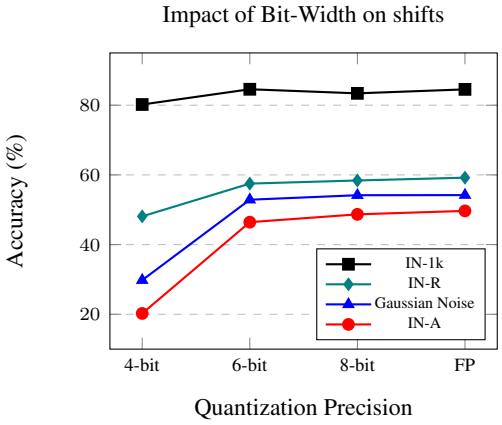


Figure 1: AdaLog accuracy across bitwidths on different datasets

Although prior work has examined the robustness and reliability of quantized models beyond clean accuracy, the interaction between quantization and distribution shift remains less well understood. RobustMQ (Xiao et al., 2023b) studies quantized CNNs under natural corruptions and reports relatively modest degradation at 4-bit precision. Yuan et al. (2023) studies PTQ reliability of CNNs from a worst-case group perspective, showing that individual classes can be considerably more sensitive to quantization than aggregate accuracy suggests and examining how calibration-set composition and bit-width affect this reliability. Recent work further shows that reduced precision can distort predictive confidence even when accuracy is largely preserved (Gu et al., 2026), while scaling-law studies in language models characterize how quantization degradation varies with model scale and numerical precision (Ayadi et al., 2026; Kumar et al., 2025). These works establish empirical trends, but provide more limited insight into why PTQ produces markedly different degradation across heterogeneous visual distribution shifts and model configurations, particularly for ViTs.

To investigate this interaction, we perform a systematic empirical study of quantized vision transformers specially at 4-bits under distribution shifts. We evaluate multiple ViT-specific PTQ methods together with quantization techniques adapted from language models, and study how quantization-induced degradation varies with architecture, model scale, activation precision, calibration data, and pretraining strategy. We find substantial degradation across the evaluated quantization and the severity varies across shifts, model configurations including pretraining. We test whether activation statistics and calibration data mismatch explains this fragility, but we find modest statistical differences and limited recovery from broader calibration data, suggesting that calibration mismatch alone is insufficient. Motivated by this, we model quantization as a perturbation to the full-precision classification margins and capture the prediction susceptibility via the interaction between the margins of the model and model's sensitivity to the accumulated quantization error. The resulting predictor tracks quantization fragility across distributions, architectures, model scales, activation bit-widths, and multiple PTQ methods. Our contributions are summarized as follows:

- **Systematic evaluation under distribution shift.** We evaluate post-training quantized Vision Transformers across multiple architectures, PTQ methods, and distribution shifts, showing that W4A4 models with near-lossless clean ImageNet accuracy can suffer substantially larger degradation under shift.
- **Characterization of quantization fragility.** We study the effects of model scale, activation precision, pretraining, calibration data, and target distribution, and find that activation-statistics and calibration mismatch alone do not explain the observed robustness loss.
- **Prediction-level model of susceptibility.** We model quantization as a perturbation to full-precision margins, capturing susceptibility through margin and sensitivity to accumulated quantization error. The predictor tracks fragility across distributions, architectures, scales, bit-widths, and PTQ methods, while higher-sensitivity samples are more likely to be misclassified even at comparable margins.

2 PRELIMINARIES

We briefly review the quantization topology of a standard Vision Transformer block. PTQ typically applies channel-wise quantization to the linear-layer weights (qkv, $proj$, $fc1$, and $fc2$) and

per-tensor quantization to intermediate activations, including linear inputs, attention logits and probabilities, and post-GELU features. These quantized tensors are part of core matrix multiplications, while layer normalization and residual operations remain in floating point. We next summarize the uniform and logarithmic quantizers briefly.

Uniform Quantization. Uniform methods map a continuous tensor X to a b-bit discrete grid using a step size s and integer zero-point Z. Asymmetric quantization calibrates via empirical bounds $[\alpha_{min}, \alpha_{max}]$, dictating $s = (\alpha_{max} - \alpha_{min})/(2^b - 1)$ and $Z = \text{clip}(\lfloor -\alpha_{min}/s \rceil, 0, 2^b - 1)$. The quantization process projects X to an integer q stored in hardware, which is subsequently dequantized during compute to a floating-point approximation $\hat{X}$:

$$q = \text{clip}\left(\left\lfloor \frac{X}{s} \right\rceil + Z, 0, 2^b - 1\right), \quad \hat{X} = s \cdot (q - Z)$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.

Logarithmic Quantization. Logarithmic quantization addresses the highly skewed, long-tailed distributions typical of post-softmax and post-GELU layers in ViTs. Instead of a linear grid, it maps inputs to discrete exponential levels, providing high resolution for small magnitudes. For a standard Log-2 quantizer, the continuous input X is quantized to an integer exponent q and dequantized via:

$$q = \text{clip}\left(\lfloor \log_2 |X| \rceil, q_{min}, q_{max}\right), \quad \hat{X} = \text{sgn}(X) \cdot 2^q$$

where $\text{sgn}(X)$ extracts the sign and $[q_{min}, q_{max}]$ bounds the representable hardware exponent range. Adalog (Wu et al., 2024), proposes an adaptive logarithmic quantizer for post-softmax and post-GeLU activations.

3 Robustness of 4-bit quantized vision transformers

In this section, we systematically evaluate the robustness of 4-bit quantized Vision Transformers under distribution shift. We show that W4A4 quantization can preserve much of the clean ImageNet accuracy but have a substantially larger degradation on distribution shift benchmarks, exposing an important limitation of evaluating tranformer PTQ methods only on clean in-distribution data. We study this effect across multiple quantization methods, model architectures, and distribution shifts, and also examine how it is influenced by calibration data, pretraining strategy, model scale, and activation bit-width. Together, these experiments characterize the disproportionate robustness of low-bit quantized vision transformers and motivate the explanatory analysis of the underexplored interaction between quantization and different distribution shifts.

3.1 Experimental Setup

Evaluation Protocol. Each model is quantized once using clean ImageNet (Russakovsky et al., 2015) calibration data and evaluated without re-calibration or fine-tuning on any target distribution, unless specified. We use a fixed set of 32 ImageNet calibration samples for most methods, while reconstruction-based methods such as APHQ-ViT use 1024 samples following their original configuration. We additionally study a mixed-calibration setting in which calibration samples are drawn from the evaluation distributions. All experiments are run on an NVIDIA L40 GPU with 32 GB memory.

Datasets. We evaluate on ImageNet-1k (Russakovsky et al., 2015) as the in-distribution benchmark and three distribution-shift benchmarks capturing complementary forms of shift. From ImageNet-C (Hendrycks & Dietterich, 2019), we consider four at severity 5 corruptions: motion blur, fog, Gaussian noise, and contrast. ImageNet-R (Hendrycks et al., 2021a) evaluates robustness to artistic renditions, while ImageNet-A (Hendrycks et al., 2021b) contains naturally occurring adversarial examples. Together, these benchmarks cover corruption, domain, and naturally adversarial distribution shifts.

Models and Pretraining. Our main evaluation considers ViT (Dosovitskiy et al., 2021), DeiT (Touvron et al., 2021), and Swin (Liu et al., 2021). We additionally study ViT models across different scales and pretraining strategies and report additional results in the appendix A.2. For pretraining, we compare the standard supervised ViT-Base with CLIP (Radford et al., 2021) and MAE (He et al.,

2022) pretrained variants, focusing on how the quantization-induced accuracy drop changes across representations to check if their full precision robustness transfers to the quantized model. We also evaluate ViT, DeiT and Swin models with different sizes namely Small, Base and Large.

Quantization Methods and LLM Adaptation. We evaluate four vision-specific PTQ methods: RepQ-ViT (Li et al., 2023), ERQ (Zhong et al., 2025a), AdaLog (Wu et al., 2024), and APHQ-ViT (Wu et al., 2025)—together with three methods adapted from LLM quantization: GPTQ (Frantar et al., 2023), SmoothQuant (Xiao et al., 2023a), and QuaRot (Ashkboos et al., 2024). All methods are evaluated at W4A4 precision unless otherwise stated. For the LLM-derived methods, GPTQ provides the Hessian-aware weight-quantization backbone, while SmoothQuant and QuaRot respectively apply channel-wise rescaling and orthogonal rotations to reduce activation outliers. To adapt these methods to ViTs, we use asymmetric rather than symmetric weight quantization, incorporate RepQ-ViT's layer-normalization reparameterization, and replace the standard uniform post-softmax activation quantizer with a logarithmic quantizer better suited to the highly skewed ViT activation distribution.

3.2 Empirical Results

Robustness degradation across quantization methods and architectures. We evaluate ViT, DeiT, and Swin using four ViT-specific PTQ methods and three methods adapted from LLM quantization across four ImageNet-C corruptions, ImageNet-R, and ImageNet-A. Table 1 reports results for the Base models. Across methods, degradation on clean ImageNet is typically much smaller than under distribution shift. For example, AdaLog on ViT-Base drops by only 4.34 points on ImageNet-1K, compared with 24.38 points on Gaussian noise and 29.43 points on ImageNet-A. Even APHQ-ViT, a reconstruction-based method using a substantially larger calibration budget, limits the clean drop to 2.31 points but loses 18.65 and 21.43 points on the same shifts. Clean accuracy also does not reliably rank PTQ methods under shift. For example, ERQ retains lower ImageNet-1K accuracy than AdaLog on ViT-Base, yet performs better on several shifted distributions. Similarly, we find that no method is uniformly strongest across shifts. LLM methods QuaRot and GPTQ can outperform APHQ-ViT on individual natural shifts.

The magnitude of degradation also depends strongly on architecture and target distribution. Swin-Base, despite comparable full-precision robustness to ViT-Base, is more stable under several quantization methods. For example, GPTQ loses only 4.36 points on ImageNet-R and 10.09 points on ImageNet-A for Swin-Base, compared with 13.23 and 34.03 points for ViT-Base. In contrast, the smaller absolute drops observed for DeiT-Base on some shifts should be interpreted relative to its lower full-precision robustness, particularly on ImageNet-A. Overall, **W4A4 robustness depends on a strong interaction between the quantization method, model architecture, and target distribution, which is not captured by clean accuracy alone.**

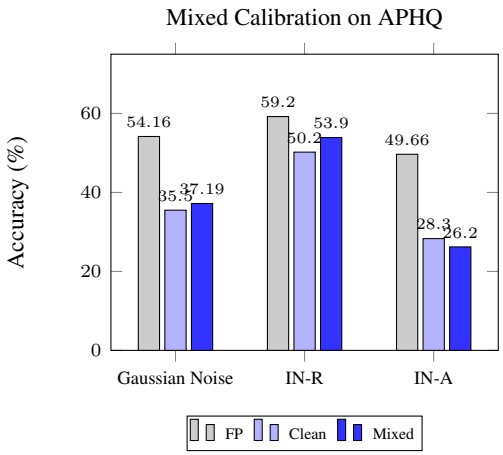


Figure 2: Effect of mixed calibration on APHQ-ViT performance.

Mixed calibration provides only limited recovery. A natural way to reduce this degradation is to expose the quantizer to shifted samples during calibration. We evaluate this on ViT-Base using a mixed calibration set with equal numbers of samples from ImageNet-1k, Gaussian Noise, ImageNet-R, and ImageNet-A. We take the expensive SOTA reconstruction-based method APHQ-ViT that jointly optimizes weights and quantization scales using a substantially larger calibration set of 1024 samples. As shown in Figure 2, mixed calibration yields only modest gains on some datasets and does not improve performance consistently across shifts. More importantly, a substantial gap to the full-precision model remains. **Thus, calibration mismatch contributes to the degradation, but broader calibration data alone is insufficient to recover the robustness lost under W4A4 quantization.**

Method	Model	IN1k	Motion Blur	Fog	Gaussian Noise	Contrast	INR	INA
Full Precision	ViT-Base	84.540	52.324	65.824	54.168	36.606	59.200	49.667
	DeiT-Base	81.800	35.134	62.924	45.680	49.074	44.890	27.110
	Swin-Base	85.270	49.400	67.112	48.960	50.658	59.310	51.720
ViT Quantization Methods								
RepQ-ViT	ViT-Base	73.016 / -11.52	16.132 / -36.19	24.288 / -41.54	14.418 / -39.75	0.578 / -36.03	41.170 / -18.03	9.080 / -40.59
	DeiT-Base	75.708 / -6.09	21.934 / -13.20	44.234 / -18.69	17.658 / -28.02	17.306 / -31.77	37.980 / -6.91	12.640 / -14.47
	Swin-Base	73.494 / -11.78	23.804 / -25.60	17.334 / -49.78	24.232 / -24.73	0.246 / -50.41	48.853 / -10.46	13.547 / -38.17
ERQ	ViT-Base	78.230 / -6.31	29.732 / -22.59	43.760 / -22.06	30.378 / -23.79	4.876 / -31.73	51.270 / -7.93	21.560 / -28.11
	DeiT-Base	79.066 / -2.73	29.572 / -5.56	51.604 / -11.32	33.096 / -12.58	16.148 / -32.93	42.343 / -2.55	17.080 / -10.03
	Swin-Base	83.290 / -1.98	36.364 / -13.04	55.856 / -11.26	37.456 / -11.50	32.440 / -18.22	55.407 / -3.90	34.960 / -16.76
AdaLog	ViT-Base	80.200 / -4.34	29.374 / -22.95	34.962 / -30.86	29.786 / -24.38	8.202 / -28.40	48.100 / -11.10	20.240 / -29.43
	DeiT-Base	78.030 / -3.77	25.752 / -9.38	53.986 / -8.94	26.288 / -19.39	26.818 / -22.26	37.697 / -7.19	14.040 / -13.07
	Swin-Base	82.470 / -2.80	39.310 / -10.09	49.644 / -17.47	30.068 / -18.89	35.450 / -15.21	52.250 / -7.06	31.360 / -20.36
APHQ-ViT	ViT-Base	82.228 / -2.31	39.374 / -12.95	50.836 / -14.99	35.518 / -18.65	18.922 / -17.68	50.470 / -8.73	28.240 / -21.43
	DeiT-Base	80.262 / -1.54	30.932 / -4.20	52.132 / -10.79	33.520 / -12.16	36.016 / -13.06	41.113 / -3.78	16.413 / -10.70
	Swin-Base	83.382 / -1.89	42.888 / -6.51	58.418 / -8.69	41.978 / -6.98	40.456 / -10.20	51.697 / -7.61	36.200 / -15.52
LLM-Based Quantization Methods								
GPTQ	ViT-Base	77.254 / -7.29	26.106 / -26.22	32.680 / -33.14	13.988 / -40.18	1.194 / -35.41	45.973 / -13.23	15.640 / -34.03
	DeiT-Base	77.210 / -4.59	24.446 / -10.69	50.276 / -12.65	28.084 / -17.60	7.704 / -41.37	39.673 / -5.22	15.600 / -11.51
	Swin-Base	83.502 / -1.77	42.074 / -7.33	51.362 / -15.75	39.286 / -9.67	39.474 / -11.18	54.953 / -4.36	41.627 / -10.09
SmoothQuant	ViT-Base	76.082 / -8.46	15.738 / -36.59	28.822 / -37.00	22.280 / -31.89	2.770 / -33.84	45.150 / -14.05	16.307 / -33.36
	DeiT-Base	78.250 / -3.55	22.520 / -12.61	43.758 / -19.17	35.344 / -10.34	11.294 / -37.78	39.480 / -5.41	16.053 / -11.06
	Swin-Base	83.396 / -1.87	39.298 / -10.10	50.318 / -16.79	40.272 / -8.69	27.830 / -22.83	53.727 / -5.58	41.373 / -10.35
QuaRot	ViT-Base	81.408 / -3.13	35.560 / -16.76	45.342 / -20.48	38.782 / -15.38	3.634 / -32.97	54.873 / -4.33	27.027 / -22.64
	DeiT-Base	80.670 / -1.13	29.878 / -5.26	53.488 / -9.44	34.096 / -11.58	21.306 / -27.77	43.840 / -1.05	21.653 / -5.46
	Swin-Base	–	–	–	–	–	–	–

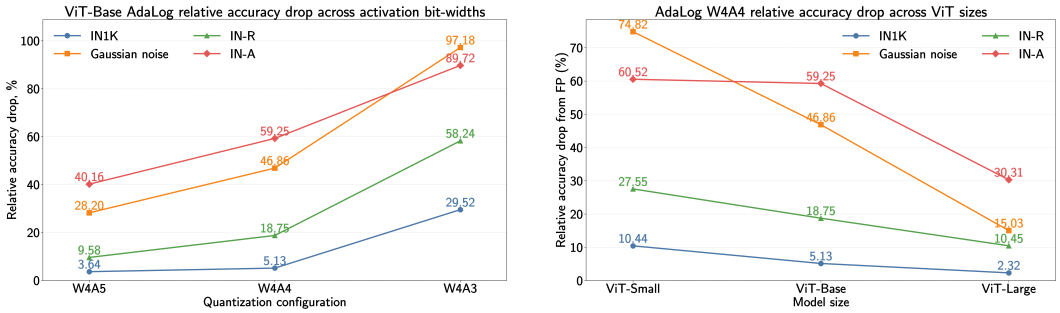
Table 1: 4-bit quantization performance across ImageNet-1k (IN1k), 4 ImageNet-C corruptions, ImageNet-R (INR), and ImageNet-A (INA) for ViT-Base, DeiT-Base, and Swin-Base. Results are reported as absolute validation accuracy and accuracy drop relative to the corresponding full-precision model (Val / Δ).

Pretraining can change susceptibility to quantization. We further examine whether quantization degradation depends on the model's pretraining strategy by comparing supervised ViT-Base with CLIP- and MAE-pretrained variants in Table 2. The clearest difference appears on ImageNet-R. At full precision, CLIP-IN1k substantially outperforms the supervised ViT-Base (68.65% vs. 59.20%), yet under ERQ W4A4 it incurs a much larger drop (17.82 vs. 7.93 points), reducing its quantized accuracy to 50.83%, slightly below the supervised model at 51.27%. A similar increase in quantization susceptibility is observed with AdaLog. CLIP also experiences substantial degradation on Gaussian noise, although its full-precision accuracy on this shift is already considerably lower. MAE exhibits a different robustness profile, with ERQ and AdaLog producing broadly similar degradation patterns. **Thus, stronger full-precision performance on a shifted distribution does not necessarily translate into greater robustness after quantization and pretraining can substantially change a model's susceptibility to the same low-precision setting.**

Method	Pretrain	IN1k	G. Noise	INR	INA
Full Precision	CLIP-IN1k	85.490	30.560	68.650	41.853
	CLIP-IN12k	86.196	39.884	64.697	50.120
	MAE-IN1k	82.772	46.282	47.660	32.442
ERQ (4-bit)	CLIP-1k	75.648 / -9.84	1.620 / -28.94	50.830 / -17.82	14.810 / -27.04
	CLIP-12k	75.854 / -10.34	3.240 / -36.64	48.330 / -16.37	21.850 / -28.27
	MAE-1k	77.124 / -5.648	14.712 / -31.570	43.117 / -4.543	13.773 / -18.669
AdaLog (4-bit)	CLIP-1k	79.768 / -5.72	3.228 / -27.33	50.530 / -18.12	17.827 / -24.03
	CLIP-12k	79.424 / -6.77	7.734 / -32.15	43.010 / -21.69	20.707 / -29.41
	MAE-1k	72.132 / -10.64	22.854 / -23.43	32.527 / -15.13	10.120 / -22.32

Table 2: 4-bit (W4A4) quantization performance of ViT-Base of IN1K finetuned models with CLIP and MAE pretraining. Accuracy is reported as absolute validation accuracy and drop compared to the respective full precision baseline (Val / Δ).

Effect of model scale and activation precision. Reducing activation precision from 5-bits to 3-bits sharply increases the relative accuracy degradation, and this increase is substantially larger under distribution shift than on ImageNet-1k. This degradation when having W4A4 is obscured in cur-



(a) (b)

Figure 3: (a) Relative accuracy drop of ViT-Base with AdaLog under different activation bit-widths (W4A5, W4A4, W4A3). (b) Relative accuracy drop of ViT model of different sizes.

rent PTQ works. Even at A4, sizable degradation remains, indicating that preserving robustness while keeping activations at 4 bits is challenging for current PTQ methods. Increasing the activation precision to A5 recovers a large fraction of the lost accuracy, but directly weakens the efficiency benefits of low-bit activation quantization. At fixed W4A4 precision, increasing ViT scale reduces the degradation but nevertheless has substantial drop. For example, the Gaussian-noise drop decreases from 74.82% for ViT-Small to 15.03% for ViT-Large. Similar trends with model size and precision have been reported in prior quantization scaling studies in language models, so we do not treat this behavior itself as a new scaling law. Rather, these controlled variations provide a useful test for the analysis in Section 4, which should account for how quantization fragility changes with both activation precision and model scale.

4 Modeling Quantization Fragility Under Distribution Shift

A natural explanation for the larger degradation under distribution shift could be that shifted inputs produce activation distributions that are less well matched to the quantization ranges calibrated on ImageNet. We therefore examine how these distributions change across the network. Across the quantized sites, the total variation distance between ImageNet and different distribution shift activation distributions remains small, ranging from approximately 0.002 to 0.05, refer 4. This indicates that the bulk of the activation distributions remains similar, even though differences can emerge in smaller regions and in the tails. These relatively modest distributional changes contrast with the substantially larger differences in quantization-induced degradation across datasets, suggesting that activation distribution shift alone is insufficient to explain the observed vulnerability. The effect of a quantization perturbation should also depend on the classification margin of the input and on the sensitivity of that margin to perturbations at the quantized layers. This motivates our analysis that considers the interaction between quantization error, classification margin, and local sensitivity.

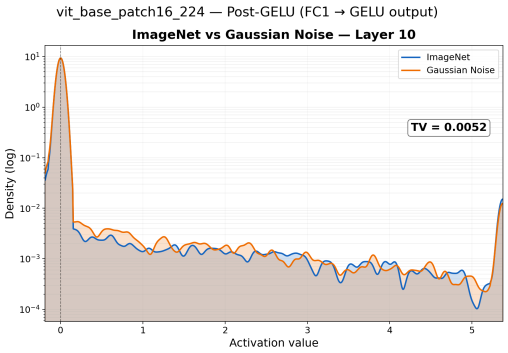


Figure 4: Activation densities for ImageNet and Gaussian-noise inputs at a representative post-GELU quantization site in an later transformer block

We therefore develop a post-hoc model for quantization-induced misclassification based on the interaction between three quantities: the full-precision classification margin, the magnitude of the quantization perturbation, and the sensitivity of the margin to this perturbation. We use *quantization-induced misclassification* to denote an input that is correctly classified by the full-precision model but becomes incorrectly classified after quantization. We subsequently evaluate the model on the same

in-distribution and shifted datasets considered in our experiments, including ImageNet, 4 ImageNet-C corruptions, ImageNet-R and ImageNet-A.

Let $f : \mathcal{X} \to \mathbb{R}^K$ denote the full-precision classifier, with logits $f(x) = (f_1(x), \ldots, f_K(x))$. We consider the subset of examples correctly classified by the full-precision model as $\mathcal{S} = \{(x, y) \in \mathcal{D} : \arg\max_k f_k(x) = y\}$. For $(x, y) \in \mathcal{S}$, let $k_1(x), \ldots, k_C(x)$ denote the C highest-scoring incorrect classes under the full-precision model. Their corresponding margins are $m_c(x) = f_y(x) - f_{k_c(x)}(x) > 0$ for $c = 1, \ldots, C$. For example, from the definition, $m_1(x)$ is the margin to the strongest full-precision competitor. Then according to this, a quantization-induced misclassification would occur when quantization causes atleast one competing margin to turn negative.

First-order effect of quantization. We start by considering quantization at the linear sites qkv, $proj$, $fc1$, and $fc2$ of each transformer block. At site ℓ, we denote $H^{(\ell)}(x)$ and $W^{(\ell)}$ to be full-precision activation and weight tensors. Let quantization errors for quantizers $Q_A^{(l)}$ and $Q_W^{(l)}$ be

$$E_A^{(\ell)}(x) = Q_A^{(\ell)}(H^{(\ell)}(x)) - H^{(\ell)}(x), \qquad E_W^{(\ell)} = Q_W^{(\ell)}(W^{(\ell)}) - W^{(\ell)}. \tag{1}$$

The effect of these errors also depends on how strongly the margin responds to perturbations at each quantized site. We capture this local response for competitor c through the margin derivatives

$$G_A^{(\ell,c)}(x) = \frac{\partial m_c(x)}{\partial H^{(\ell)}(x)}, \qquad G_W^{(\ell,c)}(x) = \frac{\partial m_c(x)}{\partial W^{(\ell)}}. \tag{2}$$

These derivatives allow us to propagate the local quantization perturbations to the output margin. We expand the quantized margin around the full-precision trajectory with Taylor approximation

$$m_c^Q(x) = m_c(x) + \Delta m_c^{\text{lin}}(x) + r_c(x), \tag{3}$$

where

$$\Delta m_c^{\text{lin}}(x) = \sum_\ell \langle G_A^{(\ell,c)}(x), E_A^{(\ell)}(x) \rangle_F + \sum_\ell \langle G_W^{(\ell,c)}(x), E_W^{(\ell)} \rangle_F, \quad S_Q(x) \equiv -\Delta m_1^{\text{lin}}(x). \tag{4}$$

Here, $r_c(x)$ collects higher-order effects and interactions among perturbations. We refer to $S_Q(x)$ as the *directional quantization sensitivity*. Large positive values of $S_Q(x)$ indicates stronger predicted contraction of the margin to the strongest FP competitor. To estimate the probability of misclassification, we next characterize the quantization errors through their first two moments. Let $\widetilde{\mathbf{e}}(x)$ denote the surrogate error vector obtained by concatenating the activation and weight perturbations,

$$\mathbb{E}[\widetilde{\mathbf{e}}(x)] = \boldsymbol{\mu}_e(x), \qquad \text{Cov}[\widetilde{\mathbf{e}}(x)] = \boldsymbol{\Sigma}_e(x). \tag{5}$$

Let $\mathbf{g}_c(x)$ denote the vectorized margin derivatives corresponding to $\widetilde{\mathbf{e}}(x)$. These error moments give the margin's mean and covariance

$$\beta_c(x) = \mathbf{g}_c(x)^\top \boldsymbol{\mu}_e(x), \qquad \Sigma_{m,cd}(x) = \mathbf{g}_c(x)^\top \boldsymbol{\Sigma}_e(x) \mathbf{g}_d(x). \tag{6}$$

For our quantizer, $\boldsymbol{\mu}_e$ and $\boldsymbol{\Sigma}_e$ are instantiated from the clipping and rounding components of the quantization error. The details of the computation are in Appendix A.5.

Misclassification Model. The first-order margin perturbation aggregates many coordinate-level quantization errors. We approximate this aggregate effect with a moment-matched Gaussian motivated by a Lindeberg CLT style argument,

$$\widetilde{\boldsymbol{\Delta m}}(x) \sim \mathcal{N}(\boldsymbol{\beta}(x), \boldsymbol{\Sigma}_m(x)), \tag{7}$$

where $\boldsymbol{\beta}(x)$ captures the mean margin shift induced by clipping and $\boldsymbol{\Sigma}_m(x)$ the modeled covariance due to rounding. This is a surrogate for the perturbation distribution and does not assume stochastic quantized execution.

To relate this perturbation to prediction failure, let $\sigma_c(x) = \sqrt{\Sigma_{m,cc}(x)}$ denote the predicted standard deviation of the perturbation to margin c. We normalize each effective margin by this perturbation scale,

$$\rho_c(x) = \frac{m_c(x) + \beta_c(x)}{\sigma_c(x)}. \tag{8}$$

The dependence between perturbations to competing margins c and d is captured by their correlation,

$$R_{cd}(x) = \frac{\Sigma_{m,cd}(x)}{\sigma_c(x)\sigma_d(x)}. \tag{9}$$

For a single competitor, a quantization-induced error occurs when its perturbed margin crosses zero, giving

$$\pi_c(x) = \Pr[m_c^Q(x) \le 0] \approx 1 - \Phi(\rho_c(x)), \tag{10}$$

where Φ denotes the standard normal CDF. Since any competing class may cause a misclassification, we consider the top C full-precision competitors jointly:

$$\pi_C(x) = 1 - \Phi_C(\boldsymbol{\rho}(x); \mathbf{R}(x)), \tag{11}$$

where $\boldsymbol{\rho}(x) = (\rho_1(x), \ldots, \rho_C(x))$ and $\mathbf{R}(x)$ is the corresponding $C \times C$ correlation matrix. $\Phi_C(\cdot; \mathbf{R})$ denotes the CDF of a zero-mean C-dimensional Gaussian with correlation matrix $\mathbf{R}$. This gives a truncated approximation over the top C competitors, with $C = K - 1$ recovering the full event under the surrogate.

Finally, averaging over the set $\mathcal{S}$ of samples correctly classified in full precision gives the predicted conditional quantization-induced misclassification rate

$$\widehat{P}_{\text{mis}}^{(C)} = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \pi_C(x). \tag{12}$$

Assuming that quantization rarely corrects samples that are misclassified by the full-precision model, we use the *relative accuracy drop* as an empirical measure of quantization-induced degradation,

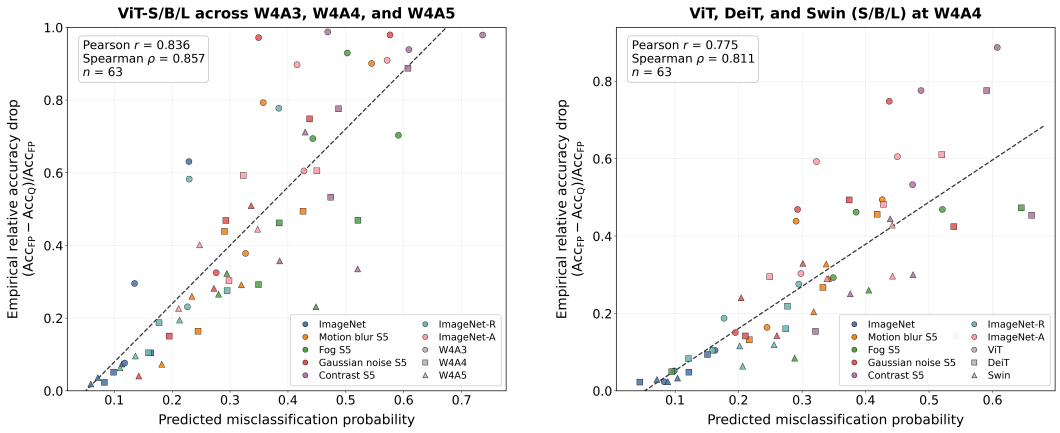
$$\Delta_{\text{rel}} = \frac{\text{Acc}_{\text{FP}} - \text{Acc}_{\text{Q}}}{\text{Acc}_{\text{FP}}}, \tag{13}$$

where Acc_{FP} and Acc_{Q} denote the full-precision and quantized accuracies, respectively.

Model instantiation. We instantiate the model using a calibrated uniform quantizer with MSE-selected clipping ranges following an ACIQ-style empirical search (Banner et al., 2019).Since Section 3 shows that disproportionate degradation persists across diverse PTQ methods, this controlled setting allows us to isolate the relationship between quantization-induced margin perturbations, margin sensitivity, and robustness without introducing additional method-specific optimization procedures. We then compute the activation and weight error moments and use them to estimate the conditional probability of quantization-induced misclassification. Implementation details are provided in Appendix A.5.

Configuration-level validation. We evaluate whether the model captures changes in quantization fragility across controlled configurations. First, varying ViT model scale and activation precision across W4A3, W4A4, and W4A5 yields substantial agreement between predicted misclassification probability and empirical relative accuracy drop in equation 13 across 63 model distribution conditions in Figure 5a, with Pearson $r = 0.836$ and Spearman $\rho = 0.857$. We then vary architecture and model scale at W4A4 across ViT, DeiT, and Swin, again over 63 conditions in Figure 5b, obtaining $r = 0.775$ and $\rho = 0.811$. Together, these results indicate that the model captures substantial variation in quantization susceptibility across distributions, architectures, model scales, and activation precisions rather than simply separating evaluation domains. We further repeat both validation settings with ERQ and observe similar positive correlations, indicating that the relationship is not specific to AdaLog.

Quantization sensitivity provides information beyond the FP margin. The margins capture only part of quantization susceptibility. We therefore ask whether the first-order directional sensitivity in Eq. 4 can distinguish samples that have similar full-precision margins but respond differently to the realized quantization perturbation. To test this, we group FP-correct samples into narrow margin ranges and, within each group, compare samples with low and high sensitivity. We then measure their empirical quantization-induced misclassification rates using the AdaLog quantized checkpoint. Across all three architectures and four datasets, the high-sensitivity groups consistently exhibit higher misclassification rates despite having comparable FP margins (Table 3), with increases



(a) (b)

Figure 5: Scatter plot and correlation between empirical relative drop and misclassification probability predicted by the model with (a) different model sizes and activation bitwidths and (b) across different model sizes and architectures at W4A4 for 4 datasets.

Model	Dataset	Low-Sens.	High-Sens.	Δ
ViT-B	ImageNet	2.33%	16.28%	+13.95
	Gaussian Noise	38.89%	65.93%	+27.04
	ImageNet-R	14.64%	28.21%	+13.57
	ImageNet-A	53.60%	71.20%	+17.60
DeiT-B	ImageNet	1.71%	11.95%	+10.24
	Gaussian Noise	30.87%	46.09%	+15.22
	ImageNet-R	17.05%	25.35%	+8.29
	ImageNet-A	45.38%	57.69%	+12.31
Swin-B	ImageNet	3.49%	7.67%	+4.19
	Gaussian Noise	29.46%	52.70%	+23.24
	ImageNet-R	11.38%	18.28%	+6.90
	ImageNet-A	31.92%	45.38%	+13.46

Table 3: Empirical quantization-induced misclassification rates for samples with comparable full-precision margins but low and high first-order quantization sensitivity. Δ denotes the increase in misclassification rate for the high-sensitivity group, in percentage points.

ranging from 4.19 to 27.04 percentage points. We additionally observe that Swin-B exhibits lower overall directional sensitivity than ViT-B, consistent with its smaller empirical degradation under several shifts in section 3.2. **Together, these results show that first-order quantization sensitivity captures susceptibility that is not explained by the floating-point margin alone and also tracks broader differences in quantization fragility across architectures.**

5 Conclusion

We show that preserving clean ImageNet accuracy after W4A4 post-training quantization does not guarantee robustness under distribution shift. Across architectures, PTQ methods, model scales, and visual shifts, we find that the degradation varies substantially. We also find that using diverse calibration data to tackle the modest activation statistics mismatch brings limited recovery, failing to close the gap. In order to explain this interaction, we develop a post-hoc model based on quantization-induced perturbations to full-precision margins. The model tracks configuration-level fragility, and for samples with comparable margins, we find that higher directional quantization sensitive samples are more likely to be misclassified. These results suggest that low-bit robustness depends not only on clean accuracy or calibration mismatch, but also on how strongly quantization perturbations contract decision margins.

AI USE STATEMENT

For this work, we used generative AI tools to aid implementing methods, mainly in coding and reviewing for any bugs in implementation. We have not used generative AI tools to formulate or develop research ideas, methodology, theoretical frameworks and for data analysis. We have also used generative AI to polish and edit parts of the paper to improve readability and make it more coherent. We have reviewed all AI-assisted code and AI-assisted polishing of the paper. We take responsibility for the text and code generated with AI's assistance.

ETHICS STATEMENT

This work studies the robustness of post-training quantized Vision Transformers under distribution shift and analyzes factors that contribute to quantization-induced performance degradation. We do not identify any ethical concerns arising from this work, as it does not involve human subjects, personal data, sensitive attributes, or deployment in high-risk application domains.

REPRODUCIBILITY STATEMENT

We inform the experimental settings, datasets, model architectures, calibration procedures, and evaluation protocols needed to reproduce our results. The study relies on publicly available datasets, pretrained models and code repositories. We have given details regarding the implementation in the appendix.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Sirine Ayadi, Sándor Daróczi, Stephan Günnemann, and Bertrand Charpentier. Reliability scaling laws for quantized large language models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=UUBijehMQO.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in neural information processing systems*, 32, 2019.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13166–13175. IEEE, 2020.

Albert Catalan-Tatjer, Niccolo Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. In *International Conference on Learning Representations*, volume 2026, pp. 64294–64319, 2026.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020.

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

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

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiawei Gu, Fengyuan Nie, Hao Tang, and Yanpeng Sun. Precision-induced miscalibration: Understanding and correcting confidence distortion in quantized neural networks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=zCkFbxKeF5.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021b.

Saqib Javed, Hieu Le, and Mathieu Salzmann. Qt-dog: Quantization-aware training for domain generalization. *arXiv preprint arXiv:2410.06020*, 2024.

Jiacheng Jiang, Yuan Meng, Chen Tang, Han Yu, Qun Li, Zhi Wang, and Wenwu Zhu. Quantization meets ood: Generalizable quantization-aware training from a flatness perspective. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11937–11946, 2025.

Tanishq Kumar, Zachary Ankner, Benjamin Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, volume 2025, pp. 71833–71864, 2025.

Qun Li, Yuan Meng, Chen Tang, Jiacheng Jiang, and Zhi Wang. Investigating the impact of quantization on adversarial robustness. *arXiv preprint arXiv:2404.05639*, 2024.

Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, pp. 154–170. Springer, 2022.

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 (ICCV)*, pp. 17181–17190, 2023.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. FQ-ViT: Post-training quantization for fully quantized vision transformer. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1173–1179, 2022.

Shih-Yang Liu, Zechun Liu, and Kwang-Ting Cheng. Oscillation-free quantization for low-bit vision transformers. In *International conference on machine learning*, pp. 21813–21824. PMLR, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

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

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 (ICML)*, 2021.

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

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

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023a.

Yisong Xiao, Aishan Liu, Tianyuan Zhang, Haotong Qin, Jinyang Guo, and Xianglong Liu. Robustmq: benchmarking robustness of quantized models. *Visual Intelligence*, 1(1):30, 2023b.

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 (ECCV)*, pp. 191–207, 2022.

Zhihang Yuan, Jiawei Liu, Jiaxiang Wu, Dawei Yang, Qiang Wu, Guangyu Sun, Wenyu Liu, Xinggang Wang, and Bingzhe Wu. Benchmarking the reliability of post-training quantization: a particular focus on worst-case performance. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. URL https://openreview.net/forum?id=iKBdmOVJ5T.

Yunshan Zhong, You Huang, Jiawei Hu, Yuxin Zhang, and Rongrong Ji. Towards accurate post-training quantization of vision transformers via error reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2676–2692, 2025a.

Yunshan Zhong, Yuyao Zhou, Yuxin Zhang, Wanchen Sui, Shen Li, Yong Li, Fei Chao, and Rongrong Ji. Semantic alignment and reinforcement for data-free quantization of vision transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12479–12490. IEEE, 2025b.

A Appendix

A.1 Related Works

Post-Training Quantization for Vision Transformers Quantization for vision models has been studied under both quantization-aware training (QAT) and post-training quantization (PTQ). QAT simulates low-precision arithmetic during training or fine-tuning (Liu et al., 2023; Dong et al., 2020), whereas PTQ quantizes a pretrained model using a comparatively small calibration set and avoids costly retraining. This makes PTQ particularly attractive for deployment, although ViTs introduce challenges absent or less pronounced in CNNs, including power-law post-Softmax distributions and heavy-tailed post-GELU activations (Lin et al., 2022; Yuan et al., 2022).

Recent ViT-specific PTQ methods has addressed these effects through increasingly sophisticated quantization schemes. PTQ4ViT (Yuan et al., 2022) and FQ-ViT (Lin et al., 2022) introduced early transformer-specific quantization strategies, while RepQ-ViT (Li et al., 2023) uses reparameterization to decouple calibration-friendly quantization from efficient inference. AdaLog (Wu et al., 2024) adapts logarithmic quantization to better accommodate post-Softmax and post-GELU activations, while ERQ (Zhong et al., 2025a) uses a ridge-regression formulation to compensate for activation and weight quantization errors. More recently, APHQ-ViT (Wu et al., 2025) employs Hessian-guided block-wise reconstruction inspired by (Dong et al., 2020) to improve low-bit quantization accuracy, at the cost of a larger calibration budget and additional reconstruction overhead. A separate line of data-free or zero-shot quantization methods removes the need for real calibration data by synthesizing inputs from model-derived priors, with ViT-specific approaches exploiting patch similarity, inter-patch structure, and attention statistics to construct more informative calibration samples (Cai et al., 2020; Zhong et al., 2025b; Li et al., 2022).

Despite substantial progress in preserving full-precision performance at aggressive bit-widths, ViT PTQ methods are evaluated predominantly through clean ImageNet-1K accuracy with calibration data drawn from the same distribution. Consequently, strong clean-data retention does not establish whether the resulting quantized models preserve their full-precision behaviour under distribution shift, which is the focus of our study.

Quantization under Robustness, Reliability, and Distribution Shift. Prior work has examined the reliability of quantized models beyond clean accuracy, although much of it focuses on settings different from ours. Recent studies of adversarial robustness show that the effect of quantization depends strongly on the quantization and robustness pipeline, and that apparent robustness can vary with the attack and training procedure (Li et al., 2024). Beyond adversarial attacks, RobustMQ (Xiao et al., 2023b) evaluates QAT-based quantized CNNs under natural corruptions and systematic noise, finding comparatively modest degradation at 4-bit precision under natural corruptions, while Yuan et al. (2023) studies PTQ reliability under distribution shift, calibration variation, and data noise. Precision-induced miscalibration further shows that reduced precision can alter predictive confidence even when accuracy is largely preserved (Gu et al., 2026).

Related work has also examined the interaction between quantization and generalization. QT-DoG (Javed et al., 2024) uses QAT-induced perturbations as an implicit regularizer to improve domain generalization, with evaluation focused primarily on CNN backbones and only limited ViT experiments. FQAT (Jiang et al., 2025) similarly studies the effect of QAT on OOD performance and introduces a flatness-oriented training strategy, again evaluating mainly CNN-based models. In parallel, quantization reliability has been studied in language models. Precision-aware scaling laws characterize how PTQ degradation varies with model scale, training data, and numerical precision (Kumar et al., 2025), while reliability scaling laws examine uncertainty, calibration, and robustness under character- and word-level input perturbations (Ayadi et al., 2026). More recently, Catalan-Tatjer et al. (2026) show that PTQ robustness in large language models is strongly influenced by training dynamics, including the learning-rate schedule and other training hyperparameters. These studies identify important factors governing quantization behaviour, but largely focus on QAT-based domain generalization, CNN architectures, or aggregate reliability trends in language models.

For Vision Transformers, the behaviour of aggressively post-training-quantized models under heterogeneous visual distribution shifts remains comparatively less characterized, particularly at the level of individual predictions. Our work addresses this gap by studying pretrained ViTs across cor-

ruptions, renditions, and naturally adversarial examples, and by relating quantization fragility to the interaction between full-precision margins and quantization-induced margin sensitivity.

LLM Quantization techniques Quantization for large language models (LLMs) has developed largely in parallel to ViT quantization, with distinct methods and evaluation conventions. Key PTQ approaches include LLM.int8() (Dettmers et al., 2022), which handles activation outliers via mixed-precision decomposition; SmoothQuant (Xiao et al., 2023a), which migrates quantization difficulty from activations to weights via mathematically equivalent scaling; GPTQ (Frantar et al., 2023), which applies layer-wise second-order weight updates; and AWQ (Lin et al., 2024), which protects salient weights based on activation magnitude. Unlike ViTs, since LLMs are trained on web-scale data, LLM quantization is assessed across a notably broader suite: perplexity on WikiText-2 and C4, alongside diverse zero-shot and few-shot downstream tasks spanning reasoning, mathematics, and language understanding. This evaluation regime implicitly tests for the generalization capability of the quantized LLMs.

A.2 Additional Empirical Quantization Results

Table 4 provides the full results underlying our analyses of model scale and activation precision. We evaluate AdaLog across nine Vision Transformer configurations spanning ViT, DeiT, and Swin at Small, Base, and Large scales, with weight precision fixed at 4 bits and activation precision varied across A3, A4, and A5. We use AdaLog for this controlled study because it provides a strong accuracy while avoiding costly reconstruction.

For each configuration, we report absolute accuracy together with the relative degradation from its corresponding full-precision model, $(\mathrm{Acc}_{\mathrm{FP}} - \mathrm{Acc}_{\mathrm{Q}})/\mathrm{Acc}_{\mathrm{FP}} \times 100$. These results further show that clean-data degradation alone is a poor indicator of quantization fragility under shift and provide the empirical targets used in our configuration-level susceptibility analysis.

A.3 Attention Entropy

We additionally examine CLS-token attention entropy 6 across early, middle, and late layers. Although entropy profiles vary somewhat across datasets, their overall structure remains similar between ImageNet and the shifted distributions, and the FP-to-W4A4 entropy changes do not reflect the large variation in robustness degradation observed across shifts. This suggests that changes in attention concentration alone are insufficient to explain quantization fragility, motivating our model based on margins and quantization sensitivity.

A.4 Additional Model validation

Model validation with ERQ We validate our model along two settings in a similar way as 4 with ERQ quantization method as well across 36 conditions (on 4 datasets) in each setting in figures 7b and 7a. In the first setting of varying ViT size scale and activation precision across W4A3, W4A4 and W4A5, we get a strong correlation with Pearson $r = 0.720$ and Spearman $\rho = 0.780$ in 36 conditions (4-datasets). In the second setting, we vary architecture and model size at W4A4 across ViT, DeiT, and Swin we again obtain a strong correlation with Pearson $r = 0.726$ and Spearman $\rho = 0.811$. This shows that our model can track the quantization susceptibility of models across different datasets and PTQ methods.

A.5 Implementation of Misclassification-Probability Estimation

We instantiate the model in Section 4 by quantizing the inputs and weights of `attn.qkv`, `attn.proj`, `mlp.fc1`, and `mlp.fc2`. Activation ranges are calibrated from 512 ImageNet samples using an ACIQ-style empirical search Banner et al. (2019). For each site and bit-width, we search percentile-based clipping intervals and select the range minimizing empirical quantization error. Weight ranges are selected similarly per output channel.

For each quantized tensor, we separate the quantization error into clipping and rounding components. The clipping error determines the sample-specific mean $\boldsymbol{\mu}_e(x)$, while the rounding moments are computed directly from the calibrated quantization ranges. These quantities define $\boldsymbol{\mu}_e(x)$ and $\boldsymbol{\Sigma}_e(x)$ used in the main formulation.

Config.	Model	IN-Val	Motion	Fog	Gauss.	Contrast	IN-R	IN-A
FP	ViT-S	81.380	40.470	44.800	31.848	16.488	46.050	26.747
	ViT-B	84.540	52.324	65.824	53.874	36.606	59.207	49.667
	ViT-L	85.838	60.014	61.896	59.508	39.852	64.347	55.293
	DeiT-S	79.856	29.788	50.722	33.100	39.436	42.240	18.680
	DeiT-B	81.980	35.134	62.924	45.682	49.074	44.890	27.107
	DeiT-L	84.778	52.210	73.380	59.736	63.576	56.473	48.480
	Swin-S	83.316	44.578	61.540	46.904	47.880	54.837	43.427
	Swin-B	85.274	49.400	67.112	48.958	50.658	59.310	51.720
	Swin-L	86.318	55.956	70.516	50.806	55.314	63.747	61.200
W4A3	ViT-S	30.048 (↓63.1)	4.014 (↓90.1)	3.152 (↓93.0)	0.662 (↓97.9)	0.350 (↓97.9)	10.243 (↓77.8)	2.427 (↓90.9)
	ViT-B	59.582 (↓29.5)	10.838 (↓79.3)	19.550 (↓70.3)	1.526 (↓97.2)	0.460 (↓98.7)	24.720 (↓58.2)	5.107 (↓89.7)
	ViT-L	79.280 (↓7.6)	37.322 (↓37.8)	18.936 (↓69.4)	40.178 (↓32.5)	2.442 (↓93.9)	49.483 (↓23.1)	21.853 (↓60.5)
	DeiT-S	18.252 (↓77.1)	1.320 (↓95.6)	2.384 (↓95.3)	0.344 (↓99.0)	0.270 (↓99.3)	7.167 (↓83.0)	1.613 (↓91.4)
	DeiT-B	62.738 (↓23.5)	10.702 (↓69.5)	22.844 (↓63.7)	12.814 (↓71.9)	0.878 (↓98.2)	24.460 (↓45.5)	5.587 (↓79.4)
	DeiT-L	77.158 (↓9.0)	29.786 (↓42.9)	56.306 (↓23.3)	31.808 (↓46.8)	25.312 (↓60.2)	43.280 (↓23.4)	16.973 (↓65.0)
	Swin-S	72.894 (↓12.5)	13.882 (↓68.9)	29.784 (↓51.6)	15.270 (↓67.4)	8.470 (↓82.3)	38.480 (↓29.8)	10.853 (↓75.0)
	Swin-B	75.252 (↓11.8)	15.818 (↓68.0)	36.158 (↓46.1)	14.090 (↓71.2)	11.446 (↓77.4)	45.513 (↓23.3)	13.693 (↓73.5)
	Swin-L	77.302 (↓10.4)	18.796 (↓66.4)	42.670 (↓39.5)	17.960 (↓64.6)	11.174 (↓79.8)	48.667 (↓23.7)	20.947 (↓65.8)
W4A4	ViT-S	72.886 (↓10.4)	20.474 (↓49.4)	24.114 (↓46.2)	8.020 (↓74.8)	1.856 (↓88.7)	33.363 (↓27.6)	10.560 (↓60.5)
	ViT-B	80.200 (↓5.1)	29.374 (↓43.9)	34.962 (↓46.9)	29.786 (↓46.6)	8.202 (↓77.6)	48.100 (↓18.8)	20.240 (↓59.2)
	ViT-L	83.848 (↓2.3)	50.188 (↓16.4)	43.792 (↓29.2)	50.564 (↓15.0)	18.620 (↓53.3)	57.620 (↓10.5)	38.533 (↓30.3)
	DeiT-S	72.354 (↓9.4)	16.162 (↓45.7)	26.644 (↓47.5)	16.768 (↓49.3)	8.922 (↓77.4)	33.003 (↓21.9)	7.280 (↓61.0)
	DeiT-B	78.030 (↓4.8)	25.752 (↓26.7)	53.986 (↓14.2)	26.288 (↓42.5)	26.818 (↓45.4)	37.697 (↓16.0)	14.040 (↓48.2)
	DeiT-L	82.864 (↓2.3)	45.402 (↓13.0)	69.798 (↓4.9)	51.274 (↓14.2)	53.892 (↓15.2)	51.777 (↓8.3)	34.173 (↓29.5)
	Swin-S	80.880 (↓2.9)	30.020 (↓32.7)	56.406 (↓8.3)	31.452 (↓32.9)	35.884 (↓25.1)	48.483 (↓11.6)	24.827 (↓42.8)
	Swin-B	82.470 (↓3.3)	39.310 (↓20.4)	49.644 (↓26.0)	41.978 (↓14.3)	35.450 (↓30.0)	52.250 (↓11.9)	36.400 (↓29.6)
	Swin-L	84.296 (↓2.3)	39.666 (↓29.1)	67.400 (↓4.4)	38.580 (↓24.1)	30.552 (↓44.8)	59.687 (↓6.4)	43.493 (↓28.9)
W4A5	ViT-S	75.376 (↓7.4)	28.676 (↓29.1)	30.328 (↓32.3)	15.620 (↓51.0)	4.748 (↓71.2)	37.090 (↓19.5)	14.853 (↓44.5)
	ViT-B	81.460 (↓3.6)	38.714 (↓26.0)	50.588 (↓23.1)	38.892 (↓27.8)	24.336 (↓33.5)	53.527 (↓9.6)	29.720 (↓40.2)
	ViT-L	84.244 (↓1.9)	55.646 (↓7.3)	45.438 (↓26.6)	57.082 (↓4.1)	25.602 (↓35.8)	60.280 (↓6.3)	42.773 (↓22.6)
	DeiT-S	57.240 (↓28.3)	8.098 (↓72.8)	16.130 (↓68.2)	9.876 (↓70.2)	0.790 (↓98.0)	25.783 (↓39.0)	5.400 (↓71.1)
	DeiT-B	78.674 (↓4.0)	28.658 (↓18.4)	45.936 (↓27.0)	39.738 (↓13.0)	24.500 (↓50.1)	40.160 (↓10.5)	18.227 (↓32.8)
	DeiT-L	83.768 (↓1.2)	48.258 (↓7.6)	72.084 (↓1.8)	56.020 (↓6.2)	55.172 (↓13.2)	54.447 (↓3.6)	40.587 (↓16.3)
	Swin-S	82.166 (↓1.4)	37.594 (↓15.7)	55.186 (↓10.3)	38.636 (↓17.6)	40.226 (↓16.0)	51.783 (↓5.6)	33.253 (↓23.4)
	Swin-B	83.862 (↓1.7)	39.354 (↓20.3)	59.232 (↓11.7)	40.750 (↓16.8)	41.054 (↓19.0)	55.933 (↓5.7)	39.960 (↓22.7)
	Swin-L	85.264 (↓1.2)	47.126 (↓15.8)	68.240 (↓3.2)	43.180 (↓15.0)	44.094 (↓20.3)	62.023 (↓2.7)	51.680 (↓15.6)

Table 4: Accuracy (%) across ImageNet and different distribution shifts. Values in parentheses denote the relative accuracy drop (%) with respect to the corresponding full-precision (FP) model, computed as $(\mathrm{Acc}_{FP} - \mathrm{Acc}_{Q})/\mathrm{Acc}_{FP} \times 100$.

For each target dataset, we evaluate up to 2000 images and retain only samples correctly classified by the full-precision model. For each sample, we select the top-C incorrect full-precision competitors. We take top-$C = 3$. To get the gradients, we then backpropagate each margin through the full-precision model to obtain $G_A^{(\ell,c)}(x)$ and $G_W^{(\ell,c)}(x)$.

Using these derivatives, the clipping errors determine the margin-space mean $\boldsymbol{\beta}(x)$, while the rounding moments determine $\boldsymbol{\Sigma}_m(x)$. Cross-competitor covariance is retained when constructing $\boldsymbol{\Sigma}_m(x)$. For the sensitivity analysis, we separately compute the realized first-order margin shift in Eq. 4 and obtain $S_Q(x)$.

Finally, we construct $\boldsymbol{\rho}(x)$ and $\mathbf{R}(x)$ as defined in Section 4. The joint Gaussian CDF is evaluated numerically using 4096 Gaussian samples. For numerical stability, $\mathbf{R}(x)$ is symmetrized and projected to a valid correlation matrix before evaluation. The resulting probabilities are averaged over the full-precision-correct samples.

A.6 Effect of Mixed Calibration Samples

We attempted to increase the number of mixed calibration samples to check if the performance of the quantized model increases. Here, we show the results of QuaRot, the next best method compared to APHQ. With GPTQ as the base quantizer for QuaRot, we took 512 samples to calibrate with equal number of samples from ImageNet, INC-Gaussian Noise, INR and INA. We display the change in performance when we increase the number of mixed calibration samples from 256 to 2048 giving equal number of samples for each dataset. We can observe from table 5 that increasing the number of total mixed calibration samples do not give meaningful improvements to their performance.

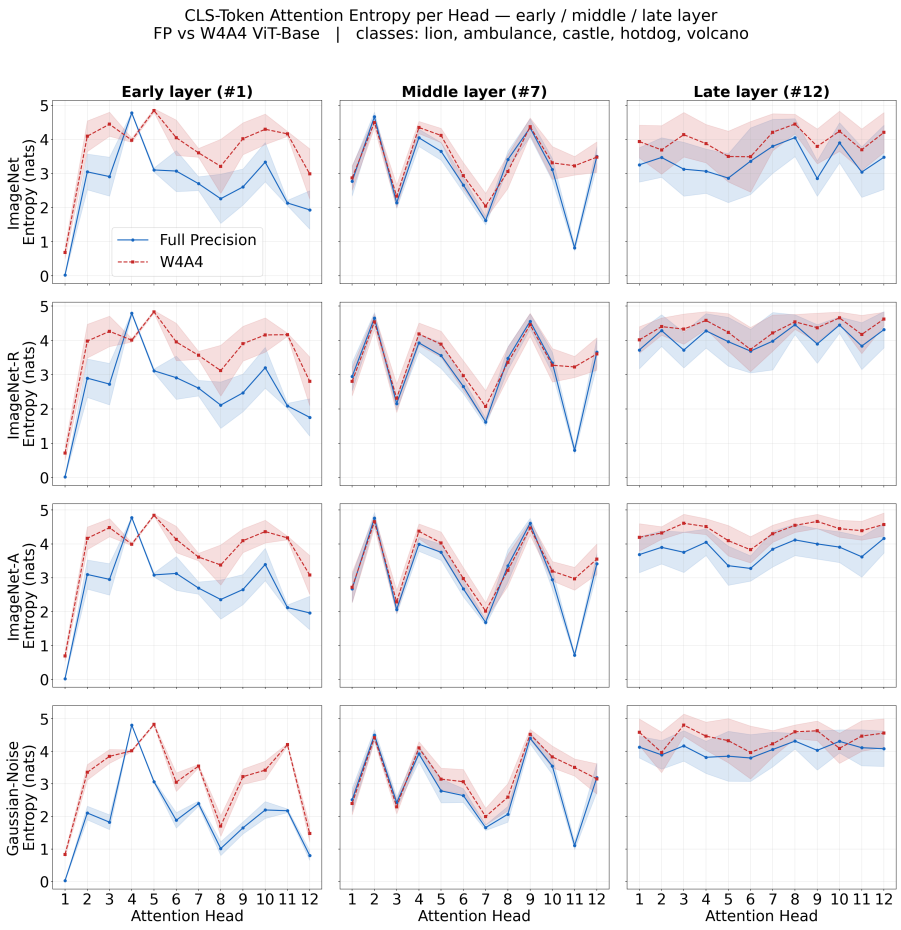


Figure 6: Plot of the CLS-token attention entropy across early, middle and late layers of ViT-Base model in IID and distribution shifts.

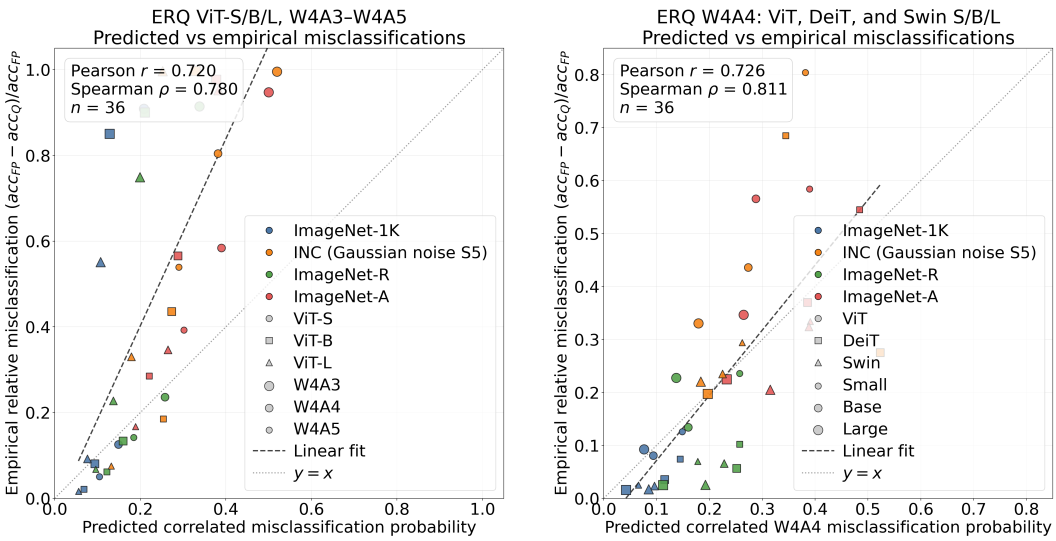
Algorithm 1 Quantization-induced misclassification probability estimation

Require: FP model f, calibrated quantizers Q, dataset $\mathcal{D}$, competitors C

Retain FP-correct samples $\mathcal{S}$
for $(x, y) \in \mathcal{S}$ **do**
  Select the top-C FP competitors and compute margins $m_c(x)$
  for $c = 1, \ldots, C$ **do**
    Backpropagate $m_c(x)$ to obtain $G_A^{(\ell,c)}(x)$ and $G_W^{(\ell,c)}(x)$
  end for
  Compute clipping and rounding errors statistics from the calibrated quantizers
  Construct $\boldsymbol{\mu}_e(x)$ and $\boldsymbol{\Sigma}_e(x)$
  Propagate them to obtain $\boldsymbol{\beta}(x)$ and $\boldsymbol{\Sigma}_m(x)$
  Compute $\boldsymbol{\rho}(x)$ and $\mathbf{R}(x)$
  Evaluate $\pi_C(x) = 1 - \Phi_C(\boldsymbol{\rho}(x); \mathbf{R}(x))$
end for
return $\widehat{P}_{\text{mis}}^{(C)} \leftarrow |\mathcal{S}|^{-1} \sum_{(x,y)\in\mathcal{S}} \pi_C(x)$

A.7 Activation Statistics

Here, we analyze how the activation value statistics differ between different data distributions. We check the overlap between ImageNet-C (Gaussian Noise), ImageNet-R and ImageNet-A with ImageNet at early, middle and later layer of the network. We can observe a significant overlap in the



(a) (b)

Figure 7: Scatter plot and correlation between empirical relative drop and misclassification probability predicted by the model with (a) different model sizes and activation bitwidths and (b) across different model sizes and architectures at W4A4 for 4 datasets.

Method	Calibration	Samples	IN	Gaussian Noise	IN-R	IN-A
QuaRot	Clean	1024	81.342	38.782	52.867	25.227
QuaRot	Mixed	256	81.202 (−0.140)	39.212 (+0.430)	52.987 (+0.120)	25.187 (−0.040)
QuaRot	Mixed	512	81.264 (−0.078)	38.852 (+0.070)	53.327 (+0.460)	24.640 (−0.587)
QuaRot	Mixed	1024	81.232 (−0.110)	38.458 (−0.324)	52.940 (+0.073)	25.227 (+0.000)
QuaRot	Mixed	2048	81.308 (−0.034)	38.976 (+0.194)	53.000 (+0.133)	25.520 (+0.293)

Table 5: QuaRot W4A4 performance under clean and mixed calibration. Values in parentheses denote the absolute accuracy difference (percentage points) relative to clean calibration with 1024 samples.

activation densities of ImageNet (clean) and other distribution shifts. We also calculate the total variation distance and show in figures 8 and 9.

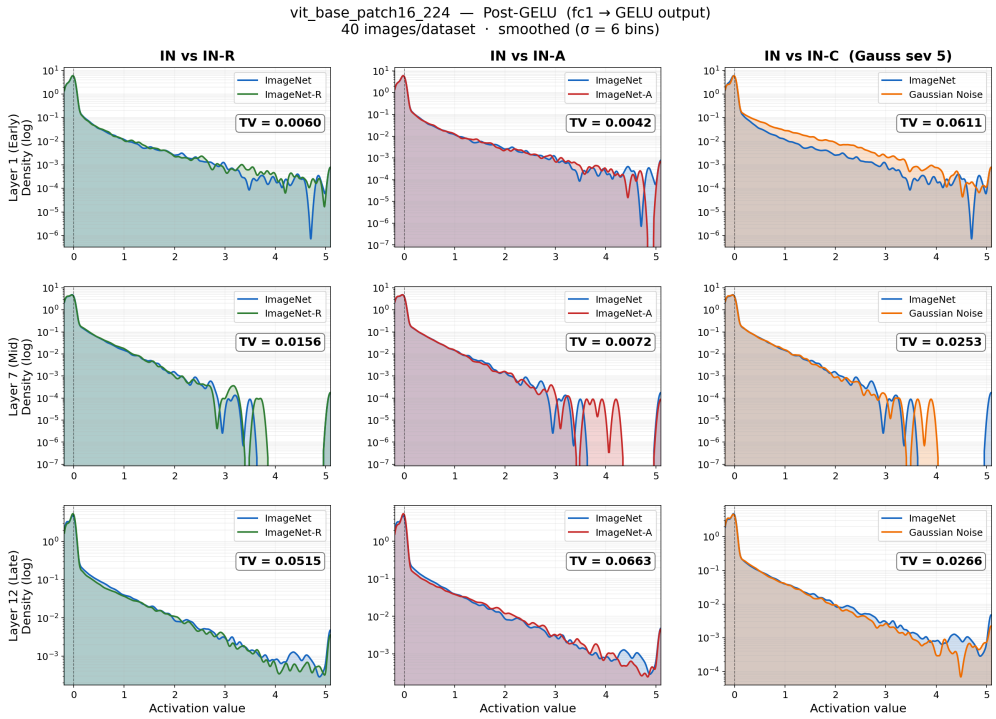


Figure 8: Plots of Post-GeLU Activation density distribution overlap between IID (ImageNet) with other distribution shifts

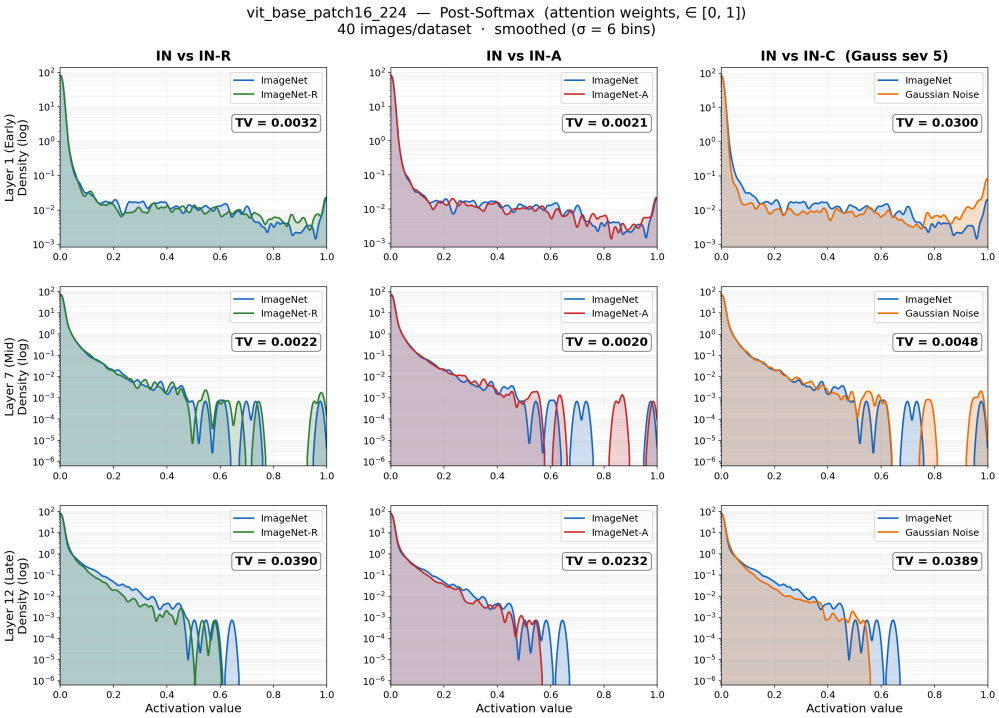


Figure 9: Plots of Post-Softmax Activation density distribution overlap between IID (ImageNet) with other distribution shifts