INVERSE AFFINE RECTIFICATION FOR LOW-BIT POST-TRAINING QUANTIZED VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit post-training quantization (PTQ) enables efficient deployment of vision transformers but can leave systematic distortions in the outputs of the resulting quantized model. We introduce Inverse Affine Rectification (InvAR), a simple post-hoc output rectification method for fixed quantized models that requires no rerunning of quantization optimization. InvAR uses paired logits from the full-precision (FP) and quantized (Q) models on a small set of unlabeled calibration images. The key idea is to model quantization-induced output distortion as a forward mapping from FP logits to Q logits. Rather than directly predicting FP logits from Q logits, InvAR estimates this distortion with a classwise affine model and analytically inverts it to rectify the quantized outputs. This distinction is important because, in the presence of residual variation, the least-squares Q-to-FP predictor generally differs from the inverse of the fitted FP-to-Q affine map. Using identical checkpoints and calibration images, vanilla InvAR and direct regression yield average Top-1 accuracy changes of +1.33 and -1.07 percentage points, respectively, across PTQ methods, ViT architectures, and bit widths. Direct inversion, however, can amplify estimation errors when calibration data are limited, especially for affine slopes near zero. InvAR therefore uses uncertainty-guided shrinkage to reduce classwise deviations from a shared affine reference before inversion. Once paired logits are collected, fitting is non-iterative and requires $\mathcal{O}(NC)$ operations for N calibration images and C classes. At inference, InvAR adds only one multiplication and one addition per prediction output. Experiments on ImageNet-1K and COCO demonstrate overall improvements across multiple PTQ methods, ViT and CNN architectures, bit widths, calibration budgets, and vision tasks. Our code will be publicly released.

1 INTRODUCTION

Model quantization reduces memory and inference costs by representing weights and activations at lower numerical precision. For vision transformers (ViTs) (Dosovitskiy et al., 2021; Liu et al., 2021a; Touvron et al., 2021), post-training quantization (PTQ) enables low-precision deployment using only a small unlabeled calibration set, without retraining the full model (Li et al., 2023; Wu et al., 2025a). Existing PTQ methods use scale reparameterization (Li et al., 2023) and reconstruction (Li et al., 2021), while second-order approaches incorporate Fisher or Hessian information to capture task sensitivity (Wu et al., 2025b;c; Hwang et al., 2026).

This leaves a complementary question: *once PTQ is complete, can classification accuracy be further improved by rectifying only the outputs of the fixed quantized model?* Such post-hoc rectification could improve existing quantized models without modifying their weights or quantization parameters and without rerunning PTQ.

A classwise affine map adjusts each logit's scale and offset as $z_c^{\mathrm{corr}} = a_c z_c^{\mathrm{q}} + b_c$, where z_c^{q} denotes the quantized (Q) logit for class c. For example, an approximate relation $z_c^{\mathrm{q}} \approx 0.5 z_c^{\mathrm{fp}} + 0.2$, with z_c^{fp} denoting the full-precision (FP) logit, suggests the inverse map $z_c^{\mathrm{corr}} = (z_c^{\mathrm{q}} - 0.2)/0.5 = 2z_c^{\mathrm{q}} - 0.4$. A natural baseline, direct Q-to-FP regression (Forward Affine), estimates a_c, b_c by predicting FP logits from paired Q logits and applies the resulting affine map at inference. However, prediction and distortion inversion are different statistical targets. Under the affine model and assumptions in

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.

Section 3.2, residual variation in the Q inputs attenuates the direct-regression slopes relative to the inverse distortion scales.

This distinction motivates *Inverse Affine Rectification* (InvAR), which fits the FP-to-Q relation and inverts the fitted affine component to rectify quantized outputs. We call this basic estimator *vanilla InvAR*. Using the same classwise affine form and paired calibration logits, direct regression yields smaller scale factors than vanilla InvAR (Figure 1(a)). Across the four configurations shown, vanilla InvAR yields positive mean Top-1 gains over the unrectified baselines, whereas direct regression yields negative gains (Figure 1(b)).

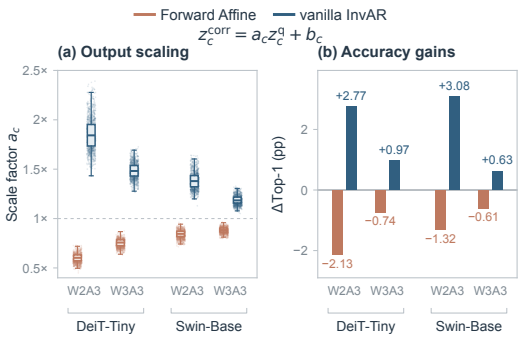

Figure 1: **Regression direction affects affine coefficients and classification accuracy.** Both panels show results for LS-ViT-quantized DeiT-Tiny and Swin-Base at W2A3 and W3A3, with both methods using the same calibration images. **(a)** Each point shows one class's deployment scale a_c, averaged over three seeds (1,000 classes). Boxes span the interquartile ranges, with medians marked inside; the dashed line marks $a_c = 1$. **(b)** Bars show three-seed mean paired ImageNet-1K Top-1 gains (pp) over the corresponding unrectified quantized checkpoints. Forward Affine fits Q-to-FP regression; vanilla InvAR fits FP-to-Q distortion and analytically inverts it. Zoom in to inspect individual points.

A remaining challenge is that inversion can become unstable when calibration data are limited, as estimation errors are amplified when fitted slopes approach zero. InvAR addresses this issue by shrinking uncertain classwise scale and mean deviations from a shared affine reference toward zero before inversion. Because the shared reference has a positive scale, shrinking all deviations to zero recovers a common inverse that preserves class rankings, while partial shrinkage retains more reliably estimated class-specific adjustments. This uncertainty-adaptive shrinkage mitigates severe accuracy degradation: with only 32 calibration images, InvAR achieves a mean Top-1 gain of 0.17 percentage points (pp) over the unrectified baselines, compared with -15.51 pp for vanilla InvAR (Section 4.5).

InvAR fits its parameters from unlabeled paired outputs using closed-form, non-iterative estimation, without backpropagation or model updates. The FP model is needed only for offline output collection; deployment stores two coefficients per class and adds one multiplication and one addition per class.

Reusing the original PTQ calibration images, InvAR consistently improves ImageNet-1K Top-1 accuracy across FIMA-Q Wu et al. (2025b), LS-ViT Hwang et al. (2026), and APHQ-ViT Wu et al. (2025c), with an average gain of 1.63 pp over the corresponding unrectified quantized baselines. These results cover seven DeiT, ViT, and Swin variants at W2A3, W2A4, and W3A3. For RepQ-ViT and AdaLog at W4A4, InvAR yields an average gain of 0.42 pp under their original small-sample calibration settings. Further experiments show gains on CNNs quantized with BRECQ and QDrop, as well as on COCO object detection and instance segmentation (Sections 4.2 and 4.2).

Our main contributions are as follows:

- We introduce InvAR, an output rectification method that estimates classwise affine distortion and applies its inverse to quantized outputs. Analysis distinguishes this target from direct Q-to-FP prediction; controlled comparisons show its advantage within the same affine form on the evaluated models.
- We develop uncertainty-adaptive shrinkage that mitigates severe accuracy degradation under limited calibration data by shrinking unreliable classwise deviations from a shared affine reference toward zero, without hyperparameter tuning.
- Experiments show low-bit accuracy gains across PTQ methods and ViT architectures, with positive mean gains across diverse calibration budgets. Further experiments demonstrate that InvAR also improves performance on CNN classification, COCO object detection, and instance segmentation.

2 Related Work

Model quantization. Quantization-aware training (QAT) incorporates simulated quantization during training (Jacob et al., 2018; Esser et al., 2020; Bhalgat et al., 2020; Li et al., 2022; Liu et al., 2023b; Liang et al., 2025; Xie et al., 2025). We focus on post-training quantization (PTQ), which quantizes pretrained models using a small calibration set (Banner et al., 2019; Zhao et al., 2019; Wang et al., 2020; Hubara et al., 2021; Frantar et al., 2022; Lee et al., 2023; Lin et al., 2023; Ma et al., 2023; Pham et al., 2025). PTQ methods have been developed for vision transformers (Liu et al., 2021b; Ma et al., 2024; Liu et al., 2023c; Frumkin et al., 2023; Moon et al., 2024; Zhong et al., 2026). A major line of PTQ minimizes layer- or block-level reconstruction errors (Nagel et al., 2020; Li et al., 2021; Wei et al., 2022), with recent ViT methods incorporating Hessian or Fisher information to account for the sensitivity of quantization perturbations (Wu et al., 2025c;b; Hwang et al., 2026). Complementary approaches improve quantizer design, scale reparameterization, and architecture-aware calibration (Lin et al., 2022; Yuan et al., 2022; Li et al., 2023; Wu et al., 2025a; Jiang et al., 2025). PD-Quant incorporates prediction-distribution matching into iterative quantization optimization (Liu et al., 2023a). InvAR instead operates after quantization is complete, rectifying the frozen model in seconds without rerunning hours of reconstruction; Section 4.3 evaluates its complementarity to prediction-distribution supervision.

Affine compensation and output rectification. Quantization-error correction can act either inside the quantized network or on its final outputs. Internal methods intervene within the quantized network: bias correction compensates systematic quantization shifts (Banner et al., 2019; Nagel et al., 2019), Bias Compensation minimizes module output error through additive corrections (Gong et al., 2025), QwT predicts blockwise quantization residuals with linear compensation (Fu et al., 2025), and QwT-v2 fits channelwise affine maps from quantized to full-precision layer outputs (Tang et al., 2025). ERQ reduces internal quantization error by adjusting weights through closed-form ridge regression (Zhong et al., 2025).

Output rectification is complementary, operating on final predictions without modifying the quantized backbone. Vector scaling learns classwise affine maps from labeled validation data for confidence calibration (Guo et al., 2017). Dirichlet calibration applies an affine transformation to log-probabilities followed by softmax for multiclass probability calibration (Kull et al., 2019). CAT uses unlabeled paired logits to directly predict full-precision logits from quantized ones through input-dependent, cluster-specific affine maps (Zoljodi et al., 2025). InvAR instead fits a single input-independent classwise map by estimating and inverting full-precision-to-quantized affine distortion. This reverses the estimation target of direct quantized-to-full-precision regression, which generally does not recover the inverse of the underlying distortion even when both use the same classwise affine form (Section 3.2). With limited calibration data, InvAR stabilizes inversion by shrinking uncertain classwise deviations toward a shared affine reference whose inverse preserves class rankings.

3 Method

3.1 Problem Setup and Overview

Inverse Affine Rectification (InvAR) estimates and inverts affine distortion after post-training quantization (PTQ), keeping model weights and quantization parameters fixed. It uses paired outputs from a small set of unlabeled inputs and operates on corresponding output components; we use classification as the example.

Let the full-precision (FP) model f_{fp} and quantized (Q) model f_{q} each output logits for C classes. For a calibration set $\mathcal{D} = \{I_i\}_{i=1}^{N}$, we collect paired outputs $Z^{\mathrm{fp}}_{i,c} = f_{\mathrm{fp}}(I_i)_c$ and $Z^{\mathrm{q}}_{i,c} = f_{\mathrm{q}}(I_i)_c$, giving $Z^{\mathrm{fp}}, Z^{\mathrm{q}} \in \mathbb{R}^{N\times C}$. Each image provides a logit pair for every class, so each classwise fit uses all N pairs.

The rectification module has two input-independent parameter vectors $a, b \in \mathbb{R}^C$ and applies

$$Z^{\mathrm{corr}}(I) = a \odot f_{\mathrm{q}}(I) + b, \tag{1}$$

where $\odot$ denotes elementwise multiplication. During offline fitting, vanilla InvAR estimates FP-to-Q affine distortion and directly inverts it. The full method first shrinks uncertain classwise deviations

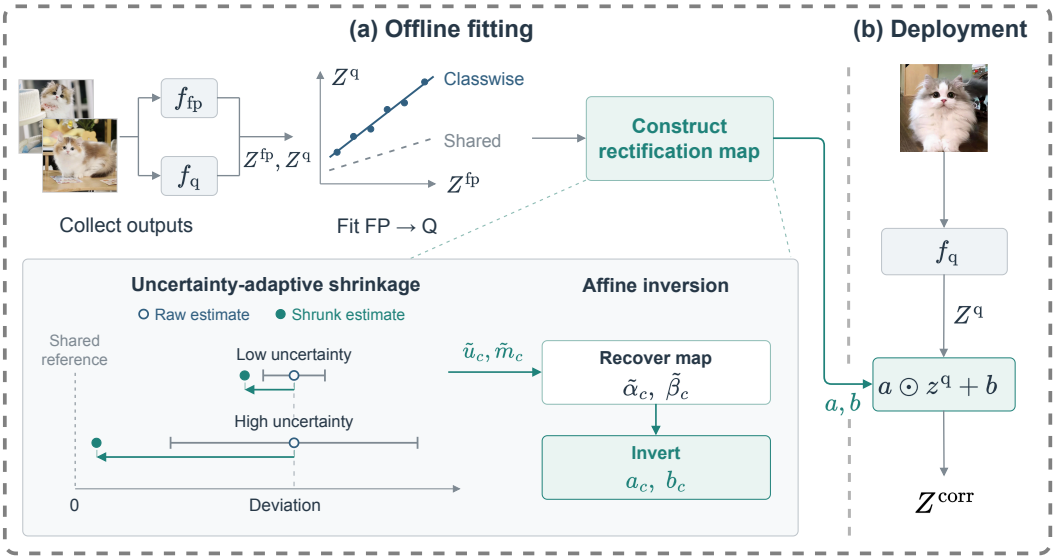

Figure 2: **Overview of InvAR. (a) Offline fitting.** Paired logits from the full-precision model f_{fp} (FP) and quantized model f_{q} (Q) are used to estimate classwise affine parameters by FP-to-Q regression (Section 3.2). A pooled fit across all classes provides the shared affine reference. HC3-based shrinkage reduces classwise deviations from this reference (Section 3.3); the affine parameters are then recovered and inverted to obtain (a, b). The schematic holds the raw deviation and signal variance fixed: greater uncertainty induces stronger shrinkage. Bars indicate ± 1 HC3-estimated standard error. **(b) Deployment.** The fixed coefficients (a, b) scale and shift the frozen Q model's logits classwise (Section 3.4).

toward a shared reference, then inverts the stabilized distortion to obtain a, b (Figure 2). Appendix E describes output pairing and adaptation to object detection.

3.2 Inverting Output Distortion

Affine distortion and inversion. Classification depends on the relative magnitudes of class logits. We describe changes in scale and offset using a positive-scale affine working model:

$$Z^{\text{q}}_{i,c} = \alpha_c Z^{\text{fp}}_{i,c} + \beta_c + \varepsilon_{i,c}, \qquad \alpha_c > 0, \tag{2}$$

where α_c and β_c are distortion parameters, and $\varepsilon_{i,c}$ is the residual unexplained by the affine relation. With known parameters, the inverse satisfies $(Z^{\text{q}}_{i,c} - \beta_c)/\alpha_c = Z^{\text{fp}}_{i,c} + \varepsilon_{i,c}/\alpha_c$, removing the affine distortion while retaining a scaled residual.

Prediction versus inversion. Direct Q-to-FP regression uses the same classwise affine form but minimizes the squared reconstruction error of FP outputs, following the fitting direction used by CAT (Zoljodi et al., 2025). Under the population relation in equation 2, assume $\text{Cov}(Z^{\text{fp}}_c, \varepsilon_c) = 0$ and finite variances $\sigma^2_{\text{fp},c} = \text{Var}(Z^{\text{fp}}_c) > 0$ and $\sigma^2_{\varepsilon,c} = \text{Var}(\varepsilon_c)$. Since $\text{Cov}(Z^{\text{q}}_c, Z^{\text{fp}}_c) = \alpha_c \sigma^2_{\text{fp},c}$ and $\text{Var}(Z^{\text{q}}_c) = \alpha_c^2 \sigma^2_{\text{fp},c} + \sigma^2_{\varepsilon,c}$, their ratio gives the population slope

$$\gamma_c = \frac{1}{\alpha_c} \frac{\alpha_c^2 \sigma^2_{\text{fp},c}}{\alpha_c^2 \sigma^2_{\text{fp},c} + \sigma^2_{\varepsilon,c}}. \tag{3}$$

Residual variance in the regression input attenuates the slope: $0 < \gamma_c < 1/\alpha_c$ when $\sigma^2_{\varepsilon,c} > 0$. This attenuation can vary across classes, so prediction and inversion can produce different rankings. Section 4.5 compares their classification performance using the same affine form and paired data.

Basic estimator. For classes with nonzero FP sample variance, ordinary least squares (OLS) fitting of equation 2 gives $(\hat{\alpha}_c, \hat{\beta}_c)$. With $\hat{\alpha}_c > 0$, vanilla InvAR applies $Z^{\text{corr}}_c(I) = (f_{\text{q}}(I)_c - \hat{\beta}_c)/\hat{\alpha}_c$. Appendix A provides OLS expressions and slope-floor conventions.

3.3 Uncertainty-Adaptive Estimation of Classwise Deviations

With limited calibration data, an underestimated slope can produce an excessively large inverse scale $1/\hat{\alpha}_c$. InvAR therefore stabilizes distortion estimates before inversion, using orthogonal residual shrinkage with HC3 variance estimates (HC3-ORS) to retain reliable classwise deviations.

Shared reference. We pool paired logits across classes and fit (α_0, β_0) by FP-to-Q OLS regression. For $\alpha_0 > 0$, its inverse applies the same positive scale and shift to every class, preserving rankings in exact arithmetic. It thus provides a shared endpoint for shrinking unreliable classwise deviations.

Scale and mean deviations. Let $\bar{Z}_c^{\mathrm{fp}}$ be the calibration FP mean and $x_{i,c} = Z_{i,c}^{\mathrm{fp}} - \bar{Z}_c^{\mathrm{fp}}$. Centering equation 2 gives

$$\begin{aligned} R_{i,c} &:= Z_{i,c}^{\mathrm{q}} - \alpha_0 Z_{i,c}^{\mathrm{fp}} - \beta_0 = u_c x_{i,c} + m_c + \varepsilon_{i,c}, \\ u_c &= \alpha_c - \alpha_0, \\ m_c &= u_c \bar{Z}_c^{\mathrm{fp}} + \beta_c - \beta_0. \end{aligned} \tag{4}$$

Here u_c measures the scale deviation from the shared reference, while m_c measures the output deviation at the FP mean. Centering makes the scale and constant design columns orthogonal, providing interpretable coordinates for separate shrinkage. Appendix A details their OLS estimates $(\hat{u}_c, \hat{m}_c)$ and covariance.

Estimating uncertainty. Classwise estimates differ in reliability. HC3 provides their marginal variance estimates $v_{u,c}$ and $v_{m,c}$, allowing heteroskedasticity and adjusting squared residuals for leverage (MacKinnon, 2013). We assume independent calibration images and zero conditional residual means, treating the fitted shared reference as fixed.

Uncertainty-adaptive retention. For each component, we use a zero-centered random-effects working model and estimate signal variance across classes by subtracting estimated noise from the mean squared deviations. Let $k \in \{u, m\}$, with $\hat{\theta}_{u,c} = \hat{u}_c$ and $\hat{\theta}_{m,c} = \hat{m}_c$. Over a nonempty set of valid classes $\mathcal{V}$ (Appendix A), we estimate

$$\hat{\tau}_k^2 = \left[\frac{1}{|\mathcal{V}|} \sum_{c \in \mathcal{V}} \left(\hat{\theta}_{k,c}^2 - v_{k,c} \right) \right]_+, \qquad [t]_+ = \max(t, 0). \tag{5}$$

We then apply moment-based empirical Bayes shrinkage (Xie et al., 2012) to valid classes:

$$g_{k,c} = \frac{\hat{\tau}_k^2}{\hat{\tau}_k^2 + v_{k,c}}, \qquad \tilde{u}_c = g_{u,c} \hat{u}_c, \qquad \tilde{m}_c = g_{m,c} \hat{m}_c. \tag{6}$$

When noise is small relative to the estimated signal, $g_{k,c}$ approaches one and retains the classwise deviation; larger relative noise shrinks the deviation toward the shared reference. When the estimated signal variance is zero, we set $g_{k,c} = 0$. The same HC3 variances enter both the noise subtraction in equation 5 and the retention factors in equation 6. Shrinkage strength requires no validation-set tuning. Appendix A details the moment-shrinkage motivation.

3.4 Rectification Map and Deployment

Shrinkage acts on distortion deviations before inversion. We recover the classwise distortion parameters and invert them to obtain the deployment coefficients in equation 1:

$$\begin{aligned} \tilde{\alpha}_c &= \alpha_0 + \tilde{u}_c, & \tilde{\beta}_c &= \beta_0 + \tilde{m}_c - \tilde{u}_c \bar{Z}_c^{\mathrm{fp}}, \\ a_c &= \frac{1}{\tilde{\alpha}_c}, & b_c &= -\frac{\tilde{\beta}_c}{\tilde{\alpha}_c}. \end{aligned} \tag{7}$$

The intercept adjustment preserves $\tilde{m}_c$ as the deviation at the FP mean after scale shrinkage. Full shrinkage gives the shared inverse; no shrinkage recovers vanilla InvAR when neither fallback nor slope clipping is triggered.

Offline computation comprises output collection and parameter fitting. For classification, collection requires FP and Q forward passes; subsequent fitting is non-iterative and takes $O(NC)$ arithmetic operations without backpropagation. Deployment stores $2C$ coefficients and adds one multiplication and one addition per class; the FP model, calibration images, and fitting statistics are no longer needed.

4 EXPERIMENTS

Table 1: **W3/A3 ImageNet-1K Top-1 accuracy (%).** † marks baselines reproduced using each PTQ method's official code; shaded rows apply rectification to the same checkpoints without additional calibration images. All local results average the same three seeds. v. InvAR denotes vanilla InvAR; bold marks the best accuracy within each reproduced group. Other results are reported in the respective method papers or LS-ViT (Hwang et al., 2026).

Method	Venue	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full precision	–	81.39	84.54	72.71	79.85	81.80	83.23	85.27
I&S-ViT	TPAMI'26	45.16	63.77	41.52	55.78	73.30	74.20	69.30
QDrop	ICLR'22	41.05	74.75	46.88	50.95	72.97	74.67	76.57
BRECQ	ICLR'21	54.33	66.62	49.27	63.72	72.82	75.20	77.64
PD-Quant	CVPR'23	54.73	73.13	53.79	67.52	75.93	76.62	78.66
APHQ-ViT	CVPR'25	63.17	76.31	55.42	68.76	76.31	76.10	78.14
FIMA-Q	CVPR'25	64.09	77.63	55.55	69.13	76.54	77.26	78.82
LS-ViT	CVPR'26	64.10	77.65	55.72	69.41	76.57	77.39	79.40
APHQ-ViT†	CVPR'25	64.80	76.52	56.90	69.34	76.58	76.68	78.49
+ v. InvAR	–	65.48	**77.02**	57.69	**70.39**	**77.03**	**77.29**	**79.22**
+ InvAR	–	**65.61**	76.93	**57.70**	70.33	76.94	77.21	79.05
FIMA-Q†	CVPR'25	62.83	76.99	55.33	69.10	76.32	76.68	78.51
+ v. InvAR	–	63.50	**77.51**	56.47	70.22	**76.78**	**77.20**	**79.32**
+ InvAR	–	**63.59**	77.42	**56.62**	**70.30**	76.76	77.14	79.15
LS-ViT†	CVPR'26	63.83	77.52	55.47	69.33	76.43	77.40	79.27
+ v. InvAR	–	64.48	**77.97**	56.44	70.23	**76.83**	**77.98**	**79.89**
+ InvAR	–	**64.57**	77.87	**56.62**	**70.33**	76.74	77.88	79.76

4.1 EXPERIMENTAL SETUP

Models and data. We evaluate InvAR on ImageNet-1K (Russakovsky et al., 2015) using models quantized with the second-order PTQ methods FIMA-Q (Wu et al., 2025b), APHQ-ViT (Wu et al., 2025c), and LS-ViT (Hwang et al., 2026), following their respective model and low-bit evaluation settings. The main evaluation covers seven architectures: ViT-Small/Base (Dosovitskiy et al., 2021), DeiT-Tiny/Small/Base (Touvron et al., 2021), and Swin-Small/Base (Liu et al., 2021a). We test W2/A3, W2/A4, W3/A3, and W4/A4, yielding 84 method–model–bit-width configurations; W/A denotes weight/activation bits. To evaluate applicability beyond second-order ViT quantization, we include RepQ-ViT (Li et al., 2023) and AdaLog (Wu et al., 2025a) on the same seven models at W4/A4. For CNNs, we use BRECQ (Li et al., 2021) and QDrop (Wei et al., 2022) on ResNet-18/50 (He et al., 2016) and MobileNetV2 (Sandler et al., 2018) at W2/A3 and W3/A3. We report ImageNet-1K validation Top-1 accuracy averaged over three random seeds, with standard deviations provided in the appendix. All classification experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU.

Calibration and comparisons. We reproduce each classification PTQ method using its public code and settings with three seeds per configuration, comparing rectification methods on the same checkpoints with weights and quantization parameters fixed. Vanilla InvAR is the unshrunk comparison for second-order ViTs and CNNs; ForwardAffine directly regresses FP logits on Q logits to test fitting direction.

By default, InvAR fits paired full-precision and quantized logits collected from the original unlabeled PTQ calibration images: $N = 1024$ for second-order ViTs and CNNs, and $N = 32$ for

RepQ-ViT and AdaLog. The budget study refits rectification parameters on the same fixed checkpoints using shared paired logits, varying N from 32 to 3072; budgets above 1024 include additional calibration images (Appendix B.1).

4.2 Classification results and task extensions

Reusing the original 1024 calibration images, InvAR improves Top-1 accuracy by an average of 1.63 percentage points (pp) across 63 W3/A3, W2/A4, and W2/A3 configurations. Gains are positive in all 189 seed-level paired comparisons and average 1.77, 1.57, and 1.56 pp for models quantized with FIMA-Q, LS-ViT, and APHQ-ViT, respectively. Vanilla InvAR (v. InvAR in the tables) achieves a slightly higher mean gain of 1.74 pp over the same 63 configurations, showing that inverse affine rectification is effective even without shrinkage at this budget. Section 4.5 examines the role of shrinkage as the calibration budget varies.

Table 2: **ImageNet-1K Top-1 accuracy (%) at W2/A4 and W2/A3.** † marks our reproduced baselines; shaded rows show the two InvAR variants. v. InvAR denotes vanilla InvAR; bold marks the best accuracy within each reproduced group. Other results are from LS-ViT (Hwang et al., 2026); dashes indicate results not reported there.

	W2/A4							W2/A3						
Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
QDrop	49.35	71.44	38.87	45.99	70.01	70.13	68.64	17.90	54.22	23.58	21.48	54.57	57.28	57.57
BRECQ	45.20	67.18	46.31	61.20	69.72	69.21	70.36	–	–	–	–	–	–	–
PD-Quant	45.12	69.82	46.86	61.26	71.99	71.51	73.03	25.56	46.06	31.45	48.66	63.60	62.41	65.29
APHQ-ViT	56.04	72.85	49.79	65.02	73.07	71.75	72.64	41.63	58.06	38.33	53.53	66.09	62.90	65.03
FIMA-Q	56.76	74.41	50.70	65.38	73.32	70.64	72.36	38.82	61.15	36.93	52.57	66.24	59.57	63.15
LS-ViT	58.48	74.91	52.12	65.66	73.78	73.89	74.95	42.05	62.89	39.28	53.86	67.06	65.77	67.73
FIMA-Q†	57.30	74.67	51.02	65.38	73.61	71.17	72.36	39.01	61.62	37.34	52.61	66.68	60.48	63.88
+ v. InvAR	58.73	**75.94**	52.90	67.71	**74.74**	**72.82**	**74.50**	41.69	**64.31**	40.00	57.50	**68.57**	**63.70**	**67.40**
+ InvAR	**58.83**	75.68	**52.92**	**67.78**	74.52	72.66	74.11	**41.86**	64.01	**40.40**	**57.82**	68.37	63.40	66.82
LS-ViT†	58.89	75.16	51.09	65.59	73.70	74.06	74.98	41.48	63.41	37.68	53.79	67.14	66.04	67.85
+ v. InvAR	60.24	**76.37**	52.95	**67.98**	**74.81**	**75.46**	**76.74**	43.66	**65.60**	40.45	58.27	**68.99**	**68.71**	**70.93**
+ InvAR	**60.32**	76.16	**52.99**	67.91	74.66	75.30	76.39	**43.73**	65.16	**40.84**	**58.37**	68.71	68.39	70.30
APHQ-ViT†	56.67	73.28	50.26	64.75	73.03	72.03	72.97	42.72	60.25	38.76	53.65	66.84	63.13	65.71
+ v. InvAR	58.11	**74.46**	**52.13**	**66.91**	**74.17**	**73.54**	**75.00**	44.95	**62.98**	41.33	57.68	**68.84**	**66.32**	**68.75**
+ InvAR	**58.21**	74.28	52.07	66.81	73.97	73.38	74.53	**45.02**	62.48	**41.37**	**57.71**	68.50	65.91	68.11

InvAR's gains become larger at lower bit-widths: mean gains rise from 0.68 pp at W3/A3 to 1.50 pp at W2/A4 and 2.72 pp at W2/A3. These gains extend to strong baselines: for DeiT-S at W2/A3, LS-ViT has the highest accuracy among the three reproduced PTQ baselines, and InvAR further improves it from 53.79% to 58.37% (+4.58 pp). Tables 1 and 2 report results at W3/A3 and W2/A4–W2/A3, respectively; results for all four bit-widths and paired-gain statistics appear in Table 7.

Beyond second-order PTQ. On RepQ-ViT and AdaLog models at W4/A4, InvAR reuses the original 32 calibration images and improves accuracy by an average of 0.42 pp across 14 configurations, with positive gains in all 42 seed-level comparisons (Appendix B, Table 6). These gains extend output rectification beyond models produced by second-order PTQ while retaining each method's original calibration budget.

CNN classification. For CNNs quantized with BRECQ and QDrop, InvAR improves mean accuracy in all 12 configurations at $N = 1024$, averaging 1.45 pp overall: 2.52 pp for BRECQ and 0.39 pp for QDrop. The same classwise rectification therefore benefits convolutional and transformer architectures without modifying their internal layers. Table 8 compares the unrectified baselines, vanilla InvAR, and InvAR.

Object detection and instance segmentation. To examine applicability beyond image-level classification, we evaluate object detection and instance segmentation on COCO (Lin et al., 2014).

Bounding-box and segmentation scores improve in all six Mask R-CNN (He et al., 2017) configurations; Cascade Mask R-CNN (Cai & Vasconcelos, 2021) scores improve slightly or remain unchanged at the reported precision. Appendix E provides the adaptation and results.

4.3 Complementarity with output-aware reconstruction

InvAR remains effective after output-aware reconstruction. We augment LS-ViT's blockwise reconstruction with prediction-distribution (PD) (Liu et al., 2023a) or ranking supervision. PD matches the teacher's terminal prediction distribution through a KL objective; ranking supervision encourages teacher-consistent ordering between its top-five classes and lower-scoring classes. After reconstruction, we freeze each checkpoint and fit InvAR (Appendix D). InvAR improves all nine W2/A3 checkpoints under each scheme across DeiT-S, ViT-S, and Swin-S, with three-seed mean Top-1 gains of 1.56–2.83 pp for PD and 2.32–3.73 pp for ranking (Table 3). All nine W3/A3 checkpoints under each scheme also improve (Table 10). These gains compare each checkpoint before and after rectification, supporting InvAR's complementarity to both output-aware objectives.

Table 3: **Complementarity and offline cost of InvAR at W2/A3 ($N = 1024$).** ImageNet-1K Top-1 (%) and paired gains (pp) are reported as mean ± std over three seeds. T_{rec} denotes reconstruction time; ΔT estimates output collection and InvAR fitting time on RTX 3090.

Model	Reconstruction	Before InvAR	After InvAR	Gain (pp)	T_{rec} (h)	ΔT (s)
DeiT-S	LS-ViT†	53.79 ± 0.24	58.37 ± 0.04	$+4.58 \pm 0.21$	2.38	2.34
	LS-ViT + PD	55.87 ± 0.20	58.70 ± 0.20	$+2.83 \pm 0.15$	4.15	
	LS-ViT + ranking	54.88 ± 0.37	58.62 ± 0.28	$+3.73 \pm 0.24$	4.14	
ViT-S	LS-ViT†	41.48 ± 0.82	43.73 ± 0.47	$+2.25 \pm 0.47$	2.32	2.38
	LS-ViT + PD	45.20 ± 0.59	46.76 ± 0.23	$+1.56 \pm 0.36$	4.11	
	LS-ViT + ranking	41.19 ± 0.72	43.72 ± 0.60	$+2.53 \pm 0.47$	4.08	
Swin-S	LS-ViT†	66.04 ± 0.16	68.39 ± 0.22	$+2.35 \pm 0.08$	4.28	5.90
	LS-ViT + PD	67.41 ± 0.19	69.08 ± 0.11	$+1.68 \pm 0.11$	13.93	
	LS-ViT + ranking	66.30 ± 0.45	68.63 ± 0.52	$+2.32 \pm 0.20$	13.85	

4.4 Efficiency of post-hoc rectification

InvAR adds only a few seconds of post-hoc optimization overhead to an already quantized model (Table 3). It reuses the same 1024 unlabeled calibration images and requires only paired-output collection followed by non-iterative parameter fitting, without backpropagation, model updates, or rerunning PTQ. Across DeiT-S, ViT-S, and Swin-S, this entire offline procedure takes an estimated 2.34–5.90 s on RTX 3090. For ImageNet-1K, deployment stores 2,000 float32 coefficients (7.81 KiB) and adds 1,000 multiplications and 1,000 additions per image.

4.5 Ablation studies and analysis

We examine the fitting direction, the behavior of the reliability-enhanced procedure with limited calibration data, and the choice of variance estimator.

Fitting direction matters. ForwardAffine directly fits Q-to-FP, whereas vanilla InvAR fits FP-to-Q and inverts the map. Both use the same checkpoints, 1024 original calibration images, and classwise affine deployment form, without shrinkage. Vanilla InvAR achieves higher three-seed mean accuracy in all 84 configurations across three PTQ methods, seven ViT architectures, and four bit-widths. Its mean gain over unrectified baselines is +1.33 pp, compared with -1.07 pp for ForwardAffine, which reduces mean accuracy in every configuration (Table 4). This controlled comparison supports distortion inversion over direct regression.

Table 4: **Regression direction ($N = 1024$).** Mean Top-1 gain (pp).

Method	W2/A3	W2/A4	W3/A3	W4/A4	All
F.Affine	−2.18	−1.10	−0.83	−0.16	−1.07
v. InvAR	**+2.88**	**+1.63**	**+0.70**	**+0.10**	**+1.33**

Reliability enhancement with limited calibration data. Across 84 configurations and 12 calibration budgets from 32 to 3072 images, InvAR achieves positive overall mean gains over the unrectified baselines (Figure 3). At $N = 32$, it mitigates vanilla InvAR's severe degradation, yielding +0.17 pp versus -15.51 pp. At $N = 256$, InvAR achieves +0.69 pp versus -0.07 pp, with nonnegative three-seed mean gains in all 84 configurations. At larger calibration budgets, both methods continue to improve accuracy over the unrectified baselines: InvAR and vanilla InvAR achieve mean gains of 1.25 and 1.33 pp at $N = 1024$, respectively, increasing to 1.57 and 1.59 pp at $N = 3072$. The complete procedure thus broadens the usable calibration range of inverse rectification, mitigating small-budget degradation while maintaining accuracy gains at larger budgets. We next isolate the variance estimator's contribution within the shrinkage construction.

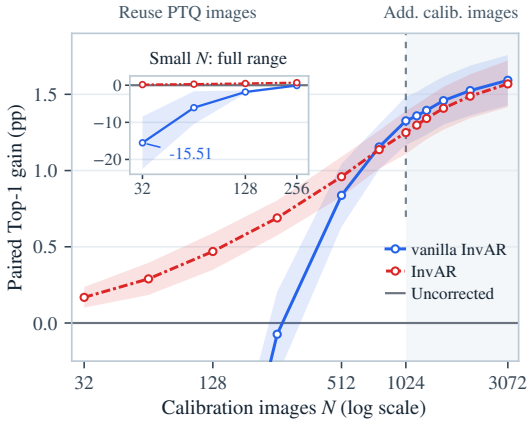

Figure 3: **Calibration budget.** Mean Top-1 gains over uncorrected checkpoints (84 configurations); bands: RMS of within-configuration three-seed SDs. "Add. calib. images" are extra fitting images beyond $N = 1024$, with checkpoints fixed. Inset: full range for $N = 32$–256.

Variance estimation within shrinkage. Underestimated variances can inflate signal moments and weaken shrinkage, retaining noisy deviations before inversion. We replace HC3 (MacKinnon, 2013) with homoskedastic estimates in both signal moments and shrinkage factors, holding all other steps fixed. On LS-ViT-quantized DeiT-Small at W3/A3, we test $N \in \{32, 64, 128\}$ with three checkpoint seeds, sharing calibration samples and validation logits. HC3 improves accuracy in all nine paired conditions, whereas the homoskedastic variant reduces it; HC3's mean advantage decreases from 2.97 to 1.09 pp as N increases (Table 5). Its gain SD is approximately 0.02 pp at each budget, versus 0.13–0.31 pp for the homoskedastic variant. These results support HC3 within InvAR: shrinkage alone does not ensure positive gains in these settings. Appendix C gives estimator formulas and the paired protocol.

Table 5: **Variance-estimator ablation.** LS-ViT, DeiT-S, W3/A3. Top-1 gains (pp) over unrectified baselines, reported as mean ± sample SD over three seeds. Only the variance estimator changes; HC3 (ours) denotes InvAR. Δ is its advantage over the homoskedastic variant.

N	Homoskedastic	HC3 (ours)	Δ
32	$-2.80_{\pm 0.13}$	$+0.17_{\pm 0.02}$	$+2.97_{\pm 0.15}$
64	$-1.68_{\pm 0.31}$	$+0.27_{\pm 0.02}$	$+1.95_{\pm 0.32}$
128	$-0.69_{\pm 0.24}$	$+0.40_{\pm 0.02}$	$+1.09_{\pm 0.26}$

5 Conclusion

We introduced InvAR, a lightweight post-hoc rectification method for fixed post-training quantized models. Its central principle is to estimate the affine distortion introduced by quantization in the FP-to-Q direction and then invert the fitted relation, rather than directly regressing quantized outputs toward full-precision ones. Our controlled comparisons show that estimating and inverting distortion yields higher classification accuracy than direct regression, even though both approaches use the same classwise affine form at inference. InvAR further stabilizes inverse estimation under limited calibration data through HC3-guided shrinkage toward a ranking-preserving shared reference. Across low-bit ViT architectures and PTQ methods, InvAR consistently improves ImageNet-1K accuracy while reusing the original unlabeled calibration images, with additional gains on CNN classification, object detection, and instance segmentation. Because fitting is non-iterative and deployment requires only a small classwise affine map, InvAR provides a lightweight complement to existing PTQ pipelines without modifying quantized weights or rerunning quantization.

REFERENCES

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. LSQ+: Improving low-bit quantization through learnable offsets and better initialization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2978–2985, 2020.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021.

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. In *International Conference on Learning Representations (ICLR)*, 2020.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4475–4488, 2022.

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.

Natalia Frumkin, Dibakar Gope, and Diana Marculescu. Jumping through local minima: Quantization in the loss landscape of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16978–16988, 2023.

Minghao Fu, Hao Yu, Jie Shao, Junjie Zhou, Ke Zhu, and Jianxin Wu. Quantization without tears. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4462–4472, 2025.

Cheng Gong, Haoshuai Zheng, Mengting Hu, Zheng Lin, Deng-Ping Fan, Yuzhi Zhang, and Tao Li. Minimize quantization output error with bias compensation. *CAAI Artificial Intelligence Research*, 4:9150036, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4466–4475, 2021.

Hyunha Hwang, Xuan Truong Nguyen, and Hyuk-Jae Lee. LS-ViT: Least-squares Hessian based block reconstruction for low-bit post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33588–33597, 2026.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713, 2018.

Runqing Jiang, Ye Zhang, Longguang Wang, Pengpeng Yu, and Yulan Guo. AIQViT: Architecture-informed post-training quantization for vision transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):17635–17643, 2025.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jung Hyun Lee, Jeonghoon Kim, Se Jung Kwon, and Dongsoo Lee. FlexRound: Learnable rounding based on element-wise division for post-training quantization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 18913–18939, 2023.

Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-ViT: Accurate and fully quantized low-bit vision transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34451–34463, 2022.

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

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. 17227–17236, 2023.

Guang Liang, Xinyao Liu, and Jianxin Wu. GPLQ: A general, practical, and lightning QAT method for vision transformers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 7033–7057, 2025.

Chen Lin, Bo Peng, Zheyang Li, Wenming Tan, Ye Ren, Jun Xiao, and Shiliang Pu. Bit-shrinking: Limiting instantaneous sharpness for improving post-training quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16196–16205, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693, pp. 740–755, 2014.

Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. FQ-ViT: Post-training quantization for fully quantized vision transformer. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pp. 1173–1179, 2022.

Jiawei Liu, Lin Niu, Zhihang Yuan, Dawei Yang, Xinggang Wang, and Wenyu Liu. PD-Quant: Post-training quantization based on prediction difference metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24427–24437, 2023a.

Shih-Yang Liu, Zechun Liu, and Kwang-Ting Cheng. Oscillation-free quantization for low-bit vision transformers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 21813–21824, 2023b.

Yijiang Liu, Huanrui Yang, Zhen Dong, Kurt Keutzer, Li Du, and Shanghang Zhang. NoisyQuant: Noisy bias-enhanced post-training activation quantization for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20321–20330, 2023c.

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, 2021a.

Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28092–28103, 2021b.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Xuefeng Xiao, Rui Wang, Shilei Wen, Xin Pan, Fei Chao, and Rongrong Ji. Solving oscillation problem in post-training quantization through a theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7950–7959, 2023.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. Outlier-aware slicing for post-training quantization in vision transformer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 33811–33825, 2024.

James G. MacKinnon. Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer New York, 2013.

Jaehyeon Moon, Dohyung Kim, Junyong Cheon, and Bumsub Ham. Instance-aware group quantization for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16132–16141, 2024.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1325–1334, 2019.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7197–7206, 2020.

Cuong Pham, Anh Dung Hoang, Cuong C. Nguyen, Trung Le, Dinh Phung, Gustavo Carneiro, and Thanh-Toan Do. MetaAug: Meta-data augmentation for post-training quantization. In *Computer Vision – ECCV 2024*, volume 15085, pp. 236–252, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations (ICLR)*, pp. 45472–45496, 2024.

Ningyuan Tang, Minghao Fu, Hao Yu, and Jianxin Wu. QwT-v2: Practical, effective and efficient post-training quantization. arXiv:2505.20932, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 10347–10357, 2021.

Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9847–9856, 2020.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhuguanyu Wu, Jiaxin Chen, Hanwen Zhong, Di Huang, and Yunhong Wang. AdaLog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *Computer Vision – ECCV 2024*, volume 15085, pp. 411–427, 2025a.

Zhuguanyu Wu, Shihe Wang, Jiayi Zhang, Jiaxin Chen, and Yunhong Wang. FIMA-Q: Post-training quantization for vision transformers by Fisher information matrix approximation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14891–14900, 2025b.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9686–9695, 2025c.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38087–38099, 2023.

Weiying Xie, Zihan Meng, Jitao Ma, Wenjin Guo, Haowei Li, Haonan Qin, Leyuan Fang, and Yunsong Li. Allowing oscillation quantization: Overcoming solution space limitation in low bit-width quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24615–24624, 2025.

Xianchao Xie, S. C. Kou, and Lawrence D. Brown. SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization. In *Computer Vision – ECCV 2022*, volume 13672, pp. 191–207, 2022.

Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Christopher De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7543–7552, 2019.

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, 2025.

Yunshan Zhong, Jiawei Hu, Mingbao Lin, Mengzhao Chen, and Rongrong Ji. I&S-ViT: An inclusive & stable method for post-training ViTs quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1063–1080, 2026.

Ali Zoljodi, Radu Timofte, and Masoud Daneshtalab. Cat: Post-training quantization error reduction via cluster-based affine transformation. arXiv:2509.26277v1, 2025.

A METHOD DERIVATIONS AND IMPLEMENTATION

Shared reference. Let $\bar{Z}^{\mathrm{fp}} = (NC)^{-1} \sum_{i,c} Z^{\mathrm{fp}}_{i,c}$ and $\bar{Z}^{\mathrm{q}} = (NC)^{-1} \sum_{i,c} Z^{\mathrm{q}}_{i,c}$. The pooled OLS reference is

$$\alpha_0 = \frac{\sum_{i,c}(Z^{\mathrm{fp}}_{i,c} - \bar{Z}^{\mathrm{fp}})(Z^{\mathrm{q}}_{i,c} - \bar{Z}^{\mathrm{q}})}{\sum_{i,c}(Z^{\mathrm{fp}}_{i,c} - \bar{Z}^{\mathrm{fp}})^2}, \qquad \beta_0 = \bar{Z}^{\mathrm{q}} - \alpha_0 \bar{Z}^{\mathrm{fp}}. \tag{8}$$

Pooling defines a reference without assuming independence of the NC entries; its fully shared inverse preserves class rankings when $\alpha_0 > 0$.

Centered OLS estimates. Let $\bar{Z}^{\mathrm{fp}}_c$ and $\bar{Z}^{\mathrm{q}}_c$ denote classwise calibration means, and $Q_c = \sum_i x^2_{i,c}$. Since $\sum_i x_{i,c} = 0$, the design columns $(x_{i,c}, 1)$ are orthogonal, and for $Q_c > 0$ the OLS estimates in equation 4 are

$$\hat{u}_c = \frac{\sum_i x_{i,c} R_{i,c}}{Q_c}, \qquad \hat{m}_c = \frac{1}{N} \sum_i R_{i,c}. \tag{9}$$

The unshrunk distortion estimates are $\hat{\alpha}_c = \alpha_0 + \hat{u}_c$ and $\hat{\beta}_c = \bar{Z}^{\mathrm{q}}_c - \hat{\alpha}_c \bar{Z}^{\mathrm{fp}}_c$. The residual $e_{i,c} = R_{i,c} - \hat{u}_c x_{i,c} - \hat{m}_c$ is the ordinary classwise OLS residual.

HC3 marginal variances. With leverage $h_{i,c} = N^{-1} + x^2_{i,c}/Q_c$ and finite weights $w_{i,c} = e^2_{i,c}/(1 - h_{i,c})^2$, the HC3 marginal variances (MacKinnon, 2013) are

$$v_{u,c} = \frac{\sum_i w_{i,c} x^2_{i,c}}{Q^2_c}, \qquad v_{m,c} = \frac{\sum_i w_{i,c}}{N^2}. \tag{10}$$

Under the independent-image and zero conditional residual mean assumptions in Section 3.3, these estimates treat the shared reference as fixed.

Moment shrinkage. The working model $\theta \sim \mathcal{N}(0, \tau^2)$, $\hat{\theta} \mid \theta \sim \mathcal{N}(\theta, v)$ with known $\tau^2, v > 0$ gives posterior mean $\tau^2 \hat{\theta}/(\tau^2 + v)$ and $\mathbb{E}[\hat{\theta}^2] = \tau^2 + v$, motivating equation 5–equation 6 (Xie et al., 2012). InvAR uses a moment-based empirical-Bayes plug-in with HC3 marginal variances and separate signal estimates for u and m. The valid set $\mathcal{V}$ contains classes with $Q_c > 10^{-12}$, finite deviation estimates, and finite nonnegative marginal variance estimates.

Unshrunk reference. Vanilla InvAR replaces the fitted slope by $\max(\hat{\alpha}_c, 10^{-3})$ *before* computing the intercept $\bar{Z}^{\mathrm{q}}_c - \hat{\alpha}_c \bar{Z}^{\mathrm{fp}}_c$ and the inverse.

B CLASSIFICATION PROTOCOLS AND COMPLETE RESULTS

B.1 EVALUATION AND CALIBRATION PROTOCOLS

All classification experiments evaluate ImageNet-1K Top-1 on the full 50,000-image validation set. Both variants rectify the same frozen checkpoint. By default, fitting reuses the images used for PTQ: 1024 for FIMA-Q, LS-ViT, APHQ-ViT, BRECQ, and QDrop, and the complete original 32-image set for RepQ-ViT and AdaLog. All checkpoints are our reproductions using the public implementations. CNN inference and preprocessing follow BRECQ/QDrop. Vanilla InvAR uses the slope floor specified in Appendix A.

Budget sampling. For budgets 32, 64, 128, 256, 512, 768, and 1024, each seed deterministically orders its original 1024-image pool without replacement and takes nested prefixes. For total budgets 1152, 1280, 1536, 2048, and 3072, we retain this pool and add nested prefixes of 2048 training images, excluding original images by path and content hash. Only the rectification parameters are refitted, and both variants share paired logits.

B.2 COMPLETE CLASSIFICATION RESULTS

Tables 7 and 8 report absolute accuracies for both variants and paired InvAR gains for all 84 ViT and 12 CNN configurations.

Table 6 reports results for the 14 RepQ-ViT and AdaLog configurations at W4/A4, using their original 32 calibration images.

Table 6: **InvAR on non-second-order PTQ methods.** ImageNet-1K Top-1 (%) at W4/A4 with $N = 32$: baseline / **+InvAR**. Accuracies are three-seed means; paired gains (+InvAR minus baseline, pp) are mean ± sample standard deviation across seeds.

Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
RepQ-ViT	64.93 / **65.88**	72.53 / **73.27**	57.47 / **58.15**	68.27 / **68.98**	75.57 / **75.76**	79.14 / **79.47**	72.95 / **73.77**
Gain (pp)	+0.96±0.11	+0.74±0.15	+0.68±0.16	+0.71±0.11	+0.19±0.05	+0.33±0.07	+0.82±0.26
AdaLog	72.72 / **73.24**	79.89 / **79.94**	63.26 / **63.65**	72.37 / **72.60**	77.96 / **78.00**	81.01 / **81.12**	82.67 / **82.74**
Gain (pp)	+0.53±0.12	+0.05±0.06	+0.39±0.11	+0.23±0.03	+0.03±0.04	+0.11±0.01	+0.07±0.02

Table 7: **Full ViT results with the vanilla InvAR reference.** ImageNet-1K Top-1 accuracy (%) at $N = 1024$.

W/A	Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
FIMA-Q								
2/3	Baseline	39.01±0.41	61.62±0.60	37.34±0.26	52.61±0.37	66.68±0.12	60.48±0.57	63.88±0.69
	+ vanilla InvAR	41.69±0.19	**64.31**±0.51	40.00±0.32	57.50±0.41	**68.57**±0.15	**63.70**±0.31	**67.40**±0.36
	+ InvAR	**41.86**±0.13	64.01±0.54	**40.40**±0.43	**57.82**±0.25	68.37±0.08	63.40±0.20	66.82±0.45
	InvAR gain (pp)	+2.85±0.31	+2.39±0.18	+3.05±0.24	+5.22±0.14	+1.69±0.05	+2.91±0.37	+2.94±0.24
2/4	Baseline	57.30±0.33	74.67±0.18	51.02±0.58	65.38±0.06	73.61±0.18	71.17±0.13	72.36±0.21
	+ vanilla InvAR	58.73±0.35	**75.94**±0.14	52.90±0.32	67.71±0.27	**74.74**±0.21	**72.82**±0.21	**74.50**±0.14
	+ InvAR	**58.83**±0.34	75.68±0.11	**52.92**±0.36	**67.78**±0.26	74.52±0.25	72.66±0.07	74.11±0.15
	InvAR gain (pp)	+1.54±0.15	+1.00±0.12	+1.91±0.27	+2.40±0.21	+0.91±0.18	+1.48±0.10	+1.75±0.19
3/3	Baseline	62.83±0.20	76.99±0.08	55.33±0.21	69.10±0.26	76.32±0.13	76.68±0.17	78.51±0.26
	+ vanilla InvAR	63.50±0.28	**77.51**±0.06	56.47±0.23	70.22±0.15	**76.78**±0.04	**77.20**±0.18	**79.32**±0.28
	+ InvAR	**63.59**±0.25	77.42±0.14	**56.62**±0.16	**70.30**±0.20	76.76±0.06	77.14±0.20	79.15±0.24
	InvAR gain (pp)	+0.76±0.07	+0.43±0.07	+1.29±0.11	+1.20±0.13	+0.44±0.11	+0.46±0.06	+0.64±0.14
4/4	Baseline	76.43±0.09	82.95±0.09	66.92±0.04	76.71±0.02	80.19±0.11	81.75±0.08	83.45±0.05
	+ vanilla InvAR	76.51±0.17	**83.02**±0.07	67.04±0.06	76.80±0.05	**80.32**±0.02	**81.81**±0.07	**83.55**±0.02
	+ InvAR	**76.52**±0.11	82.99±0.06	**67.09**±0.02	**76.81**±0.04	80.28±0.05	81.79±0.10	83.53±0.02
	InvAR gain (pp)	+0.09±0.05	+0.04±0.03	+0.17±0.04	+0.10±0.05	+0.09±0.07	+0.04±0.07	+0.08±0.04
LS-ViT								
2/3	Baseline	41.48±0.82	63.41±0.33	37.68±0.17	53.79±0.24	67.14±0.10	66.04±0.16	67.85±0.40
	+ vanilla InvAR	43.66±0.44	**65.60**±0.14	40.45±0.15	58.27±0.14	**68.99**±0.09	**68.71**±0.31	**70.93**±0.31
	+ InvAR	**43.73**±0.47	65.16±0.19	**40.84**±0.16	**58.37**±0.04	68.71±0.08	68.39±0.22	70.30±0.42
	InvAR gain (pp)	+2.25±0.47	+1.75±0.16	+3.16±0.09	+4.58±0.21	+1.56±0.09	+2.35±0.08	+2.45±0.10
2/4	Baseline	58.89±0.20	75.16±0.14	51.09±0.17	65.59±0.17	73.70±0.19	74.06±0.31	74.98±0.20
	+ vanilla InvAR	60.24±0.15	**76.37**±0.20	52.95±0.08	**67.98**±0.18	**74.81**±0.12	**75.46**±0.21	**76.74**±0.19
	+ InvAR	**60.32**±0.14	76.16±0.17	**52.99**±0.06	67.91±0.16	74.66±0.16	75.30±0.23	76.39±0.21
	InvAR gain (pp)	+1.43±0.16	+1.00±0.04	+1.89±0.13	+2.32±0.01	+0.96±0.05	+1.24±0.09	+1.41±0.06
3/3	Baseline	63.83±0.18	77.52±0.21	55.47±0.32	69.33±0.04	76.43±0.20	77.40±0.08	79.27±0.14
	+ vanilla InvAR	64.48±0.22	**77.97**±0.19	56.44±0.29	70.23±0.11	**76.83**±0.13	**77.98**±0.04	**79.89**±0.07
	+ InvAR	**64.57**±0.18	77.87±0.16	**56.62**±0.34	**70.33**±0.08	76.74±0.11	77.88±0.02	79.76±0.07
	InvAR gain (pp)	+0.74±0.01	+0.35±0.05	+1.15±0.03	+1.01±0.08	+0.31±0.10	+0.48±0.10	+0.50±0.07
4/4	Baseline	76.51±0.07	83.03±0.08	66.79±0.08	76.74±0.09	80.31±0.09	81.80±0.03	83.55±0.03
	+ vanilla InvAR	76.56±0.02	**83.08**±0.08	66.98±0.10	76.83±0.09	**80.37**±0.13	81.87±0.04	**83.69**±0.06
	+ InvAR	**76.60**±0.03	83.07±0.10	**67.04**±0.09	**76.92**±0.06	80.35±0.09	**81.88**±0.06	83.64±0.05
	InvAR gain (pp)	+0.08±0.07	+0.04±0.02	+0.25±0.01	+0.18±0.05	+0.04±0.01	+0.08±0.04	+0.09±0.08

Continued on next page

Table 7 continued

W/A	Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
APHQ-ViT								
2/3	Baseline	42.72±0.48	60.25±0.69	38.76±0.50	53.65±0.23	66.84±0.32	63.13±0.61	65.71±0.47
	+ vanilla InvAR	44.95±0.28	**62.98**±0.58	41.33±0.09	57.68±0.12	**68.84**±0.28	**66.32**±0.37	**68.75**±0.38
	+ InvAR	**45.02**±0.31	62.48±0.55	**41.37**±0.28	**57.71**±0.15	68.50±0.28	65.91±0.40	68.11±0.36
	InvAR gain (pp)	+2.30±0.17	+2.23±0.15	+2.61±0.24	+4.06±0.22	+1.66±0.19	+2.77±0.21	+2.40±0.24
2/4	Baseline	56.67±0.30	73.28±0.29	50.26±0.11	64.75±0.15	73.03±0.12	72.03±0.16	72.97±0.35
	+ vanilla InvAR	58.11±0.31	**74.46**±0.24	**52.13**±0.08	**66.91**±0.14	**74.17**±0.13	**73.54**±0.18	**75.00**±0.30
	+ InvAR	**58.21**±0.34	74.28±0.19	52.07±0.19	66.81±0.18	73.97±0.17	73.38±0.11	74.53±0.28
	InvAR gain (pp)	+1.54±0.27	+0.99±0.10	+1.80±0.13	+2.06±0.07	+0.94±0.05	+1.35±0.09	+1.56±0.09
3/3	Baseline	64.80±0.27	76.52±0.20	56.90±0.25	69.34±0.22	76.58±0.08	76.68±0.24	78.49±0.26
	+ vanilla InvAR	65.48±0.29	**77.02**±0.12	57.69±0.14	**70.39**±0.11	**77.03**±0.09	**77.29**±0.12	**79.22**±0.14
	+ InvAR	**65.61**±0.28	76.93±0.18	**57.70**±0.15	70.33±0.16	76.94±0.07	77.21±0.14	79.05±0.27
	InvAR gain (pp)	+0.81±0.05	+0.41±0.10	+0.80±0.15	+0.99±0.15	+0.37±0.01	+0.53±0.15	+0.56±0.04
4/4	Baseline	76.30±0.06	82.39±0.07	66.83±0.04	76.68±0.13	80.22±0.06	81.54±0.06	83.29±0.14
	+ vanilla InvAR	76.42±0.15	**82.46**±0.09	66.94±0.11	**76.81**±0.16	**80.28**±0.06	**81.64**±0.04	**83.42**±0.06
	+ InvAR	**76.45**±0.12	82.44±0.07	**66.95**±0.07	76.77±0.16	80.26±0.06	81.62±0.07	83.40±0.10
	InvAR gain (pp)	+0.16±0.08	+0.05±0.03	+0.12±0.04	+0.09±0.04	+0.04±0.01	+0.08±0.00	+0.11±0.05

Table 8: **Unshrunk and reliability-aware rectification on CNNs.** ImageNet-1K Top-1 accuracy (%) at $N = 1024$.

W/A	Method	ResNet-18	ResNet-50	MobileNetV2
BRECQ				
2/3	Baseline	62.09±0.05	63.92±0.20	14.69±1.55
	+ vanilla InvAR	**62.52**±0.06	64.99±0.35	19.11±1.12
	+ InvAR	**62.52**±0.07	**65.05**±0.27	**20.14**±1.29
	InvAR gain (pp)	+0.43±0.09	+1.13±0.08	+5.45±0.39
3/3	Baseline	65.93±0.04	69.06±0.04	27.29±2.12
	+ vanilla InvAR	**66.23**±0.09	**69.60**±0.13	33.95±1.85
	+ InvAR	**66.23**±0.07	**69.60**±0.09	**34.54**±1.81
	InvAR gain (pp)	+0.30±0.03	+0.54±0.10	+7.24±0.32
QDrop				
2/3	Baseline	61.33±0.12	66.12±0.08	40.35±0.36
	+ vanilla InvAR	61.62±0.17	**66.76**±0.08	40.88±0.36
	+ InvAR	**61.67**±0.19	66.66±0.08	**41.17**±0.32
	InvAR gain (pp)	+0.34±0.08	+0.54±0.15	+0.82±0.08
3/3	Baseline	65.70±0.04	71.30±0.11	54.58±0.07
	+ vanilla InvAR	65.67±0.07	**71.52**±0.09	54.78±0.10
	+ InvAR	**65.74**±0.06	**71.52**±0.15	**54.97**±0.07
	InvAR gain (pp)	+0.04±0.02	+0.22±0.05	+0.39±0.06

Accuracy and paired InvAR gains are mean ± sample standard deviation across three seeds; gains are in pp. Evaluation and implementation details are in Appendix B.1.

C VARIANCE ESTIMATION AT SMALL CALIBRATION BUDGETS

Table 9 reports the full accuracy statistics for the variance-estimator ablation in Section 4.5. We use LS-ViT-quantized DeiT-Small at W3/A3 with same checkpoint as the previous reported. For each checkpoint, calibration sets of $N \in \{32, 64, 128\}$ are deterministic nested of the original 1024-image PTQ calibration set. The homoskedastic variant replaces the HC3 variances with homoskedastic estimates in both the signal-moment estimate and the shrinkage factors.

Homoskedastic variance estimates. Using the centered regression in Appendix A, let $e_{i,c} = R_{i,c} - \hat{u}_c x_{i,c} - \hat{m}_c$ and $Q_c = \sum_{i=1}^{N} x_{i,c}^2$. Treating the shared reference as fixed, the homoskedastic

variant estimates

$$\hat{\sigma}_c^2 = \frac{\sum_{i=1}^{N} e_{i,c}^2}{\max(N-2,1)}, \qquad v_{u,c}^{\mathrm{hom}} = \frac{\hat{\sigma}_c^2}{\max(Q_c, \epsilon_{\mathrm{num}})}, \qquad v_{m,c}^{\mathrm{hom}} = \frac{\hat{\sigma}_c^2}{N}. \tag{11}$$

The residual degrees of freedom are $N-2$ because the classwise regression fits a slope and an intercept; all tested budgets have $N > 2$. The protected denominators follow the numerical conventions in Appendix A; for valid classes, $Q_c > \epsilon_{\mathrm{num}}$. Homoskedasticity assumes a common residual variance across calibration images within each class, while allowing this variance to differ across classes. We substitute $v_{u,c}^{\mathrm{hom}}$ and $v_{m,c}^{\mathrm{hom}}$ for the HC3 variances in both equation 5 and equation 6, recomputing the signal moments and shrinkage factors for this variant.

Paired statistics. We first compute the accuracy difference between HC3 and the homoskedastic variant for each checkpoint and then report its mean and sample standard deviation across the three seeds. These standard deviations reflect variation across the checkpoint seeds and their calibration samples. The observed accuracy advantage establishes the effect of the variance-estimator choice.

Table 9: **Variance-estimator ablation.** LS-ViT, DeiT-S, W3/A3. Top-1 (%) and paired differences (pp), reported as mean ± sample SD over three seeds. Only the variance estimator changes; HC3 (ours) denotes InvAR. Δ is its advantage over the homoskedastic variant.

N	Baseline	Homoskedastic	HC3 (ours)	Δ
32	$69.33_{\pm 0.04}$	$66.53_{\pm 0.15}$	$69.49_{\pm 0.03}$	$+2.97_{\pm 0.15}$
64	$69.33_{\pm 0.04}$	$67.64_{\pm 0.29}$	$69.59_{\pm 0.05}$	$+1.95_{\pm 0.32}$
128	$69.33_{\pm 0.04}$	$68.64_{\pm 0.26}$	$69.72_{\pm 0.04}$	$+1.09_{\pm 0.26}$

D Output-aware reconstruction and rectification

Table 10 supplements Section 4.3 with W3/A3 results for PD- and ranking-augmented LS-ViT.

Table 10: **Rectification after output-aware reconstruction at W3/A3.** ImageNet-1K Top-1 (%) before and after InvAR at $N = 1024$. Accuracies and within-checkpoint gains (pp) are three-seed means ± sample SD. W2/A3 results and offline costs are reported in Table 3.

Model	Reconstruction	Before InvAR	After InvAR	Paired gain (pp)
DeiT-S	LS-ViT + PD	69.59 ± 0.02	70.29 ± 0.06	$+0.71 \pm 0.04$
	LS-ViT + ranking	69.51 ± 0.21	70.41 ± 0.16	$+0.89 \pm 0.06$
ViT-S	LS-ViT + PD	64.37 ± 0.38	64.96 ± 0.41	$+0.59 \pm 0.02$
	LS-ViT + ranking	63.56 ± 0.14	64.54 ± 0.11	$+0.98 \pm 0.13$
Swin-S	LS-ViT + PD	77.54 ± 0.07	77.90 ± 0.02	$+0.36 \pm 0.09$
	LS-ViT + ranking	77.43 ± 0.12	77.92 ± 0.14	$+0.49 \pm 0.02$

Setting	Value
Steps / minibatch per unit	20,000 / 32 images
Units	14 for DeiT/ViT; 29 for Swin-S
Optimizer / learning rates	Adam; 10^{-3} rounding, 4×10^{-5} scales
Scale schedule	Cosine decay
Rounding regularizer	Weight 0.01; 20% warmup; exponent 20 to 2
Reconstruction loss	LS-ViT Hessian term ($T = 20$); head KL ($T = 1$)
QDrop probability	0.5

Original stem/head bit-width exceptions are retained, including eight-bit patch-embedding convolution activations. There is no early stopping or validation-based checkpoint selection. Weights are hard-rounded and the final model is reloaded before evaluation and InvAR fitting.

Prediction-distribution objective. We add PD-Quant's KL-based prediction-difference objective (Liu et al., 2023a) to local LS-ViT reconstruction, without PD-Quant's BatchNorm-based distribution correction.[1] Let t_i be the detached teacher logits and s_i the current unit's logits after the frozen suffix, for image i in a minibatch of size $B = 32$. With $K = 1000$ classes, define $p_{ic} = \mathrm{softmax}(t_i/T)_c$ and $q_{ic} = \mathrm{softmax}(s_i/T)_c$. The added term is

$$\mathcal{L}_{\mathrm{PD}} = \frac{1}{\lambda_r B} \sum_{i=1}^{B} \sum_{c=1}^{K} p_{ic} \log \frac{p_{ic}}{q_{ic}}, \qquad T = 4, \quad \lambda_r = 0.1. \tag{12}$$

Ranking objective. For image i, let $\mathcal{T}_i$ be the teacher's five highest-scoring classes and let $s_{b,i}$ be the terminal logits obtained from current unit b through its frozen suffix. The teacher defines the ordered pairs

$$\begin{aligned} \mathcal{P}_i &= \{(c,d) : c \in \mathcal{T}_i,\ d \in \{1,\dots,K\},\ t_{i,c} > t_{i,d}\}, \\ \mathcal{L}_{\mathrm{rank},b} &= \frac{1}{B} \sum_{i=1}^{B} \frac{1}{|\mathcal{P}_i|} \sum_{(c,d)\in\mathcal{P}_i} \mathrm{softplus}(-(s_{b,i,c} - s_{b,i,d})). \end{aligned} \tag{13}$$

We use temperature one without a margin or teacher-margin weighting.

Before updating unit b, four seed-fixed probe batches measure the reconstruction and ranking gradient norms on the same unit output $y_b^{(j)}$. The ranking weight is fixed throughout that unit's reconstruction:

$$\lambda_b = 0.1 \, \frac{\sum_{j=1}^{4} \|\nabla_{y_b^{(j)}} \mathcal{L}_{\mathrm{rec},b}^{(j)}\|_2}{\sum_{j=1}^{4} \|\nabla_{y_b^{(j)}} \mathcal{L}_{\mathrm{rank},b}^{(j)}\|_2}, \qquad \mathcal{L}_b = \mathcal{L}_{\mathrm{rec},b} + \mathcal{L}_{\mathrm{round},b} + \lambda_b \mathcal{L}_{\mathrm{rank},b}. \tag{14}$$

Here $\mathcal{L}_{\mathrm{rec},b}$ is the original LS-ViT reconstruction loss, excluding the rounding regularizer.

E Extending rectification to object detection

E.1 Paired output collection

We collect paired full-precision and quantized outputs on 256 unlabeled images. For Mask R-CNN (He et al., 2017), the quantized model generates proposals, and both models evaluate the same regions of interest using their respective feature maps to obtain paired class logits and raw box offsets. For Cascade Mask R-CNN (Cai & Vasconcelos, 2021), both models follow the quantized model's original RoI refinement trajectory; we collect the three-stage average logits and the final-stage box offsets. Detector weights and quantization parameters remain fixed.

For inverse affine rectification of the paired output components, we follow InvAR's distortion estimation, uncertainty-adaptive shrinkage, and inversion described in Sections 3.2–3.4.

E.2 Detection and instance segmentation results

All object detection and instance segmentation experiments were conducted on a single NVIDIA A100 GPU with 40 GB of memory.

Table 11 reports COCO results for InvAR on reproduced LS-ViT, FIMA-Q, and APHQ-ViT models using Mask R-CNN and Cascade Mask R-CNN.

F Limitations

InvAR requires paired full-precision and quantized outputs for offline fitting. These outputs can be collected within a PTQ workflow while the full-precision source model remains available. Applying InvAR to a separately obtained quantized checkpoint requires access to its full-precision counterpart or previously collected paired outputs. Our default classification protocol reuses the original

[1] The objective follows the official implementation.

Table 11: **Object detection and instance segmentation.** Reproduced baseline **/ +InVAR** scores on Mask R-CNN (Mask) and Cascade Mask R-CNN (Cascade); bbox and segm denote bounding boxes and segmentation.

Method	Mask-T		Mask-S		Cascade-T		Cascade-S	
	bbox	segm	bbox	segm	bbox	segm	bbox	segm
LS-ViT	39.1 / **41.6**	37.6 / **39.0**	43.6 / **44.8**	40.8 / **41.4**	48.5 / **48.6**	42.4 / **42.5**	50.0 / **50.1**	43.5 / **43.5**
FIMA-Q	39.2 / **41.5**	37.4 / **38.8**	43.5 / **44.4**	40.8 / **41.0**	48.3 / **48.4**	42.1 / **42.2**	49.9 / **49.9**	43.6 / **43.6**
APHQ-ViT	38.6 / **41.6**	37.6 / **39.2**	43.3 / **44.6**	40.6 / **41.2**	48.7 / **48.7**	42.5 / **42.6**	50.0 / **50.1**	43.4 / **43.4**

PTQ calibration images without increasing the calibration data budget. Structured prediction further requires pairing under matched prediction conditions, such as identical regions of interest in our detection extension, which entails task-specific alignment (Appendix E).

The classification rectifier intentionally uses an input-independent classwise affine map to model scale and offset distortion. This restricted form enables closed-form estimation, uncertainty-guided shrinkage of classwise deviations, analytic inversion, and deployment with only one multiplication and one addition per class. It does not explicitly capture input-dependent nonlinear distortion or cross-class interactions, so residual discrepancies may remain after inversion. More expressive rectifiers could capture such effects, but extending the estimate-then-invert principle beyond affine maps while preserving stable inversion, reliable uncertainty estimation, and lightweight fitting and inference remains an interesting direction.

Uncertainty-guided shrinkage entails a bias–variance trade-off. Our budget study shows that the complete InvAR procedure substantially mitigates performance degradation when calibration data are scarce, whereas vanilla InvAR can achieve slightly higher mean accuracy at larger calibration budgets. Further refining the uncertainty estimates and shrinkage rule may improve large-budget accuracy while preserving InvAR's small-sample reliability.

G AI USE STATEMENT

We used generative AI tools only to refine the language, clarity, and fluency of this manuscript. The human authors developed the research ideas, designed the experiments, and conducted all data analyses; generative AI was not used to generate research ideas or analyze experimental results. We take full responsibility for the final content of this work, including all text, claims, and artifacts prepared with AI assistance.