Rotation Symmetry in Vision Quantization: The Objective Function is the Bottleneck

Jaewoo Park [1] **Jihae Lee** [1]

Abstract

Rotation-based post-training quantization (PTQ) is widely used to suppress activation outliers in large language models. In vision classification, however, rotations learned with the standard task loss perform worse than random rotations on ViT-family models.

We evaluate rotation learning on the Stiefel manifold across five architectures (DeiT-S, Swin-T, ViT-S, ResNet-50, MobileNetV2). Replacing the task loss with Final KD or Block Output MSE eliminates this reversal in all five models.

The gain is small on ViT-family models at $+0.3\sim0.7$pp. On MobileNetV2, where rotation is restricted to conv1$\times$1, task-loss accuracy of 18.79% rises to a 4-seed KD accuracy of $67.11\pm0.29\%$, a gap of about $+48$pp that exposes a strong architecture heterogeneity. The result suggests that the gain from rotation symmetry is determined more by *the choice of optimization objective* than by the symmetry itself.

1. Introduction

Post-Training Quantization (PTQ) is a practical path to compress a pretrained model to low precision without retraining. Most Vision PTQ work corrects errors that arise *after* quantization, but grid correction alone is insufficient when the weight distribution is itself unfriendly to quantization. Orthogonal rotation tackles the problem *before* quantization as a weight-space symmetry: applying R with $R^\top R = I$ to a linear layer $y = Wx$ leaves the output unchanged via $WR^\top \cdot RX = WX$ while redistributing energy across channels and breaking up outliers.

This symmetry has been put to good use in the LLM domain. QuaRot (Ashkboos et al., 2024) disperses outliers without learning via a fixed Hadamard transform, SpinQuant (Liu et al., 2024) learns R on the Stiefel manifold and achieves "learned $>$ random," and DartQuant (Shao et al., 2025) speeds up learning with the Whip loss. Vision classification models differ from LLMs in several ways: they mix Linear, Conv, and depthwise Conv layers, their quantization sensitivity varies sharply by architecture, and their gradient flows from a single softmax output. Whether the LLM finding that learning beats random rotation transfers to vision is therefore an open question.

We ask: **Is learned rotation always more effective than random rotation for vision classification, and if not, what causes the failure and how can it be fixed?** Across five ImageNet classification models at W4A8 and W6A6, we report:

- **Task loss is not reliable across architectures.** On three ViT-family models, rotations learned with the task loss are consistently worse than random ones. A loss-landscape analysis on the Stiefel manifold links this to high Quant Only accuracy: the landscape flattens, and the STE gradient signal weakens.
- **The objective function controls the gap.** Replacing the task loss with Final Output KD (Hinton et al., 2015) or Block Output MSE (Li et al., 2021) removes the failure across all five models. We interpret this as soft targets providing a steadier gradient signal on a flat Stiefel landscape.
- **The objective gap is large even under a constrained architecture.** On MobileNetV2, where only conv1$\times$1 layers can be rotated, task-loss accuracy of 18.79% rises to a 4-seed Final KD accuracy of $67.11\pm0.29\%$, recovering 92.0% of the FP baseline 72.94%.

2. Method

2.1. Rotation-Based Quantization

Orthogonal rotation reshapes the weight distribution into a quantization-friendly form without changing the floating-point output. For a linear layer $y = Wx$, applying R gives $WR^\top \cdot RX = WX$, so the output before quantization is

[1]Xiilab Inc., Seoul, South Korea. Correspondence to: Jaewoo Park <jw.park@xiilab.com>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

unchanged. After quantization, each layer becomes:

$$Y = Q_w(WR^\top) \cdot Q_a(RX) \tag{1}$$

Q_w and Q_a are b-bit symmetric uniform quantizers for weights and activations; weight scales are per-channel and chosen to minimize MSE. The transform preserves total energy ($\|WR^\top\|_F = \|W\|_F$) but spreads it across channels, breaking up channel-concentrated outliers and lowering kurtosis. A flatter distribution makes per-channel scales more effective and reduces the quantization error.

Conv2d layers are rotated the same way after a channel-wise reshape. Depthwise convolutions with groups$>$1 are excluded because their channels do not interact. The rotation target depends on the architecture: DeiT-S, Swin-T, and ViT-S rotate only `linear` layers, ResNet-50 uses `all` to rotate Linear, Conv1$\times$1, and Conv3$\times$3 layers, and MobileNetV2 rotates only `conv1x1` layers. At inference, $W' = WR^\top$ is pre-absorbed into the weights and adds no runtime cost. The activation-side RX vanishes between adjacent layers that share the same R because $R^\top R = I$, as in our per-layer setup. Per-layer sharing limits the scope to a single layer, so a small dense RX matmul can remain at layer boundaries; this leftover absorbs into a Hadamard transform or the previous Linear weight and its measured cost is negligible. A TensorRT INT8 benchmark shows a latency change of within $\pm 0.2\%$.

2.2. Optimization on the Stiefel Manifold

The orthogonality constraint $R^\top R = I$ defines the Stiefel manifold $\mathrm{St}(n, n)$. We optimize R on this manifold with SGDG (Liu et al., 2024) and the Cayley retraction:

$$R_{t+1} = \left(I + \tfrac{\alpha}{2}W\right)^{-1}\left(I - \tfrac{\alpha}{2}W\right)R_t \tag{2}$$

$W = -W^\top$ is the skew-symmetric matrix derived from the Riemannian gradient and α is a cosine-annealed learning rate. The model weights are frozen; only R is trained.

2.3. Objective Functions

Task Loss. The standard objective minimizes the cross-entropy of the quantized model with the Straight-Through Estimator (STE). This works when quantization error is large enough that the STE gradient carries a useful signal, but it breaks down when Quant Only accuracy is already high: the loss landscape on the Stiefel manifold flattens and the gradient signal weakens.

Final Output KD. We learn R using the FP model's logits as the teacher signal:

$$\mathcal{L}_{\mathrm{KD}} = T^2 \cdot \mathrm{KL}\left(\sigma\left(\frac{z_{\mathrm{FP}}}{T}\right) \middle\| \sigma\left(\frac{z_Q}{T}\right)\right) \tag{3}$$

z_{FP} and z_Q are the logits of the FP and the quantized-with-rotation model, and T is the temperature. An ablation gives T=8 as the best value, which we use throughout. A hard one-hot task target concentrates the gradient on the correct class, while the KD soft target spreads gradient mass over all classes, reducing variance on a flat landscape.

Table 1. W4A8 ImageNet Top-1 (%) under per-layer sharing. **Bold**: best learning objective per model. The KD column reports Mean$\pm$Std over 4 seeds; other columns use seed 42 only. FP = full-precision baseline, QO = Quant Only, O-MSE = Block Output MSE.

Model	FP	QO	Rand.	Task	O-MSE	**KD**
DeiT-S	79.86	77.16	78.98	77.47	79.14	**79.31$\pm$0.01**
Swin-T	81.38	79.81	80.43	79.17	80.69	**80.75$\pm$0.10**
ViT-S	81.38	72.86	79.66	78.04	**80.37**	80.21$\pm$0.15
ResNet-50	76.13	64.02	74.15	73.53	75.24	**75.45$\pm$0.09**
MobileNetV2	72.94	11.62	16.64	18.79	65.09	**67.11$\pm$0.29**

Block Output MSE. We minimize the MSE between the block outputs of the FP and quantized models, $\sum_b \|f_b^Q(X) - f_b^{\mathrm{FP}}(X)\|_2^2$. Aligning intermediate representations rather than the final output avoids the gradient cancellation of multi-component losses and supplies block-level supervision.

3. Experiments

3.1. Experimental Setup

Models and dataset. We evaluate five ImageNet-1K (Deng et al., 2009) classifiers on the 50K validation set: DeiT-S (Touvron et al., 2021), Swin-T (Liu et al., 2021), ViT-S (Dosovitskiy et al., 2020), ResNet-50 (He et al., 2016), and MobileNetV2 (Sandler et al., 2018).

Quantization setup. The default is W4A8 symmetric uniform quantization. Weight scales are per-channel and chosen by minimizing MSE over 80 candidates; activation scales are per-tensor.

Training setup. We use 1,000 calibration images, 100 epochs, SGDG with an initial learning rate of 1.5 under a cosine schedule, and seed 42. Unless noted otherwise, each layer has its own R (per-layer sharing).

3.2. Objective Function Comparison

ViT family: learned $<$ random under task loss. On all three ViT-family models, task-loss rotations underperform random rotations: -1.51pp on DeiT-S, -1.26pp on Swin-T, and -1.62pp on ViT-S. The 4-seed run on DeiT-S below shows the gap is 20 times the standard deviation, so this is not a seed effect. Both alternative objectives close the gap. Final KD beats random at 79.31% on DeiT-S and 80.79% on Swin-T; on ViT-S the best is Output MSE at 80.37%. ViTs sit at a Quant Only accuracy of 77$\sim$80%, where the loss landscape is too flat for the STE gradient of the task

loss to find a useful descent direction. We analyze this in Section 3.3.

ResNet-50: same pattern as the ViTs. ResNet-50 has Quant Only accuracy 64.02% (84% FP recovery), placing it in the same flat regime as the ViTs, and task-learned 73.53% trails random 74.15%. Both alternative objectives recover the gap: O-MSE reaches 75.24% and Final KD reaches **75.42%** (99.1% of FP).

MobileNetV2: KD gives the largest gain under constrained rotation. Depthwise convolutions are skipped, so only conv1×1 layers are rotated. Under this restriction, task loss reaches 18.79% and random rotation 16.64%. The STE signal weakens along the long depthwise→pointwise→head path. Final KD provides direct logit-level supervision and reaches **67.11±0.29%** over 4 seeds, a gain of about +48pp over the task loss. The absolute accuracy is still −5.83pp below FP, the limit of this hard-to-quantize regime, but the std of 0.29 confirms that the +48pp gap is reproducible. Block Output MSE reaches **65.09%**, within 1.92pp of KD. Non-rotated layers add noise to the supervision signal, but the score still beats task 18.79% and random 16.64% by a wide margin, so both alternative objectives work even in a depthwise-separable model.

Summary. On every model, the alternative objectives beat task loss and match or beat random rotation. The size of the gain tracks FP recovery. For the three ViTs and ResNet-50 it is a modest +0.3∼1.9pp; for MobileNetV2, restricted to conv1×1 rotation, it reaches +48pp. ViT-S is the only model where O-MSE 80.37% beats KD 80.22%, by 0.15pp. The absence of a distillation token may make block-level feature alignment a more direct supervision signal than logit matching; Section 4 discusses this in more detail.

3.3. Loss Landscape Analysis

We sample two random tangent directions at the learned R^*, project back to the Stiefel manifold, and evaluate the loss on a 40×40 grid. Figure 1 shows the result.

Across $N = 12$ data points (4 models × 3 sharing strategies), the Spearman correlation between STE gradient norm and the learned−random gap is $\rho = 0.760$ ($p = 0.004$). A stronger gradient signal goes with a larger gain from learning. The KD soft target spreads gradient mass over all classes and reduces the directional effect of quantization noise; on DeiT-S, KD std of 0.02 is much smaller than task std of 0.07, consistent with this picture. Section 4 considers alternative explanations. T=8 is optimal, but going from T=1 to T=8 adds only +0.19pp on DeiT-S, so the gap is driven by KD itself, not by temperature tuning.

Multi-seed verification. Table 2 shows the "Task learned < Random" gap on all three ViTs is at least 10σ, not seed noise. Final KD reverses the gap on every model: +0.30pp on DeiT-S, +0.20pp on Swin-T, and +0.46pp on ViT-S. KD std (0.02∼0.15) is smaller than or comparable to task std (0.07∼0.13), so the optimization is also more stable.

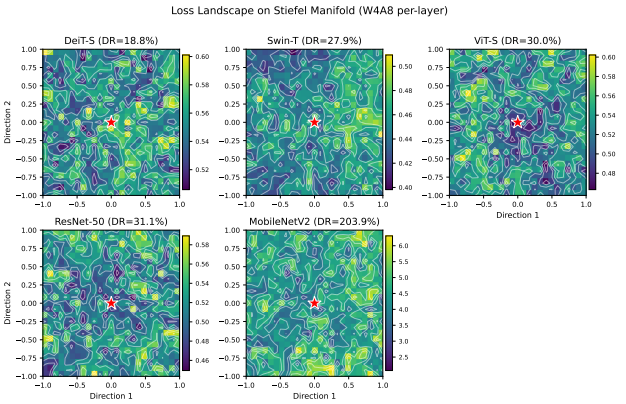


Figure 1. Loss landscape on the Stiefel manifold (W4A8 per-layer). MobileNetV2, which is restricted to conv1×1 rotation, exhibits the largest dynamic range, whereas the three high-recovery ViTs and ResNet-50 lie in a flat region, with ResNet-50 at 44.9% versus DeiT-S at 28% (a factor of 1.6). On flat landscapes the STE gradient is dominated by noise rather than signal.

Table 2. W4A8 per-layer multi-seed results on three ViT-family models with 4 seeds (42/0/123/777). σ = —Task−Random— / pooled std.

Model	Random	Task	KD	$\sigma_{\text{T-R}}$
DeiT-S	78.98±0.03	77.59±0.07	79.28±0.02	**20**
Swin-T	80.55±0.02	79.20±0.11	80.75±0.10	**12**
ViT-S	79.75±0.06	78.23±0.13	80.21±0.15	**11**

Comparison with prior work. Table 3 compares us with recent Vision PTQ methods at W4A8 Top-1. Our rotation-only method beats AdaLog (Wu et al., 2024) and RepQ-ViT (Li et al., 2023) but trails IGQ-ViT (Moon et al., 2024) and FIMA-Q (Wu et al., 2025), both of which optimize the quantization grid. Rotation only reshapes the input distribution and leaves the grid alone, so the two directions are complementary. A separate experiment that combines rotation with block-wise reconstruction (Li et al., 2021) pushes RF-DETR W4A8 above its FP baseline.

W6A6 cross-validation. At W6A6 per-tensor, the three ViTs and ResNet-50 again show Task close to Random, matching the W4A8 pattern. ResNet-50 Task 74.27% beats Random 73.67% by only +0.60pp. Final KD beats Task by +0.19∼+1.68pp on DeiT-S, Swin-T, ResNet-50, and MobileNetV2; ViT-S is the only exception at −0.20pp. The KD gain shrinks at W6A6 relative to W4A8, in line with the view that higher bit-widths leave a flatter landscape and less room for learning.

W4A4 preliminary. At W4A4 per-tensor, KD is again the

Table 3. Comparison with prior Vision PTQ methods (W4A8, Top-1 %). †layer-wise sharing yields 79.40; per-layer sharing yields 79.31.

Method	Type	DeiT-S	Swin-T
AdaLog (Wu et al., 2024)	log quant	78.42	—
RepQ-ViT (Li et al., 2023)	reparam	—	80.72
IGQ-ViT (Moon et al., 2024)	grid opt	79.47	—
FIMA-Q (Wu et al., 2025)	Fisher recon	79.50	**81.02**
Ours†	rotation	79.40	80.79

best objective on all three ViTs. KD reaches **71.13%** on DeiT-S, ahead of Task 66.14%, Random 67.26%, and O-MSE 68.46%; on Swin-T it reaches **77.05%**, and on ViT-S **63.00%**. These numbers use a per-layer configuration and a single seed. Absolute accuracy varies with FP recovery, but the KD>Task pattern matches W4A8 and W6A6. On DeiT-S, adding group-wise quantization at group size 64 raises KD further to 72.79%, so grid optimization and rotation are complementary. The claim here is the consistency of the loss-objective axis; W4A4 multi-seed verification is left for the extended version.

4. Discussion

Relation to SpinQuant and QuaRot. SpinQuant (Liu et al., 2024) reports that on LLMs the task loss already makes learned rotation beat random. The same objective does not always do so on Vision classifiers. The difference is quantization difficulty. At W4A4, LLMs degrade heavily and the loss landscape is steep, but many Vision models keep a high Quant Only accuracy at W4A8 and the landscape is flat: DeiT-S has W4A8 Quant Only 77.16%, close to its FP accuracy 79.86%. The QuaRot view that "random rotation suffices" is partly consistent with these easy-quantization models, but our KD results show that, with the right objective, learned rotation can still beat random on the same models.

Why KD works on the Stiefel manifold: a hypothesis. On a flat Stiefel landscape, the STE gradient $\nabla_R \mathcal{L}_{\text{CE}}$ is dominated by quantization-discontinuity noise, and the descent direction becomes unstable. We hypothesize that the KD soft target spreads gradient mass over all classes instead of the argmax, lowering gradient variance and stabilizing the trajectory. On DeiT-S the KD multi-seed std of 0.02% is about 3.5 times smaller than the task std of 0.07%, consistent with this hypothesis but not a causal proof. Another possibility is implicit regularization toward the FP model's representation. Separating STE noise, manifold geometry, and loss curvature is left for future work. The takeaway is that, on a constrained manifold, the *quality of the learning signal* matters as much as the symmetry parameterization.

Limitations. (1) The loss-landscape analysis uses a 2D projection of a high-dimensional manifold; a stronger claim would need a full curvature analysis such as the Hessian spectrum. (2) All five models are verified across 4 seeds: the three ViTs in Table 2, and ResNet-50 V1 and MobileNetV2 at 75.45±0.09 and 67.11±0.29. All numbers come from a single calibration set. (3) All models are small to medium; ViT-L/H is not verified. (4) The Spearman correlation ρ=0.760 rests on N=12 data points and is empirical, not causal. (5) Two exceptions on ViT-S—O-MSE beating KD by +0.15pp at W4A8 and Task beating KD by 0.20pp at W6A6—do not fit the landscape-flatness hypothesis. We tentatively link them to differences in the informativeness of block outputs caused by the CLS-only architecture and the absence of a distillation token. A label-smoothing baseline that disentangles whether the gap arises from the soft target or from FP teacher matching, and a ViT-B/L verification, are left for the extended version. (6) The MobileNetV2 +48pp gap is measured at 67.11±0.29% over 4 seeds, providing a robustness margin at the std-0.29 level, but a full sensitivity sweep over learning rate and calibration data is left for the extended version.

5. Conclusion

In Vision classification PTQ, the gain from orthogonal rotation symmetry depends more on the choice of optimization objective than on the symmetry itself. When Quant Only accuracy is high, the task loss fails because the Stiefel landscape is flat. Final Output KD and Block Output MSE deliver a steadier gradient signal on all five architectures and recover useful rotations.

The effect is uneven across architectures and is most extreme on MobileNetV2 with conv1×1-only rotation. There task and random essentially collapse to 18.79% and 16.64%, while KD reaches 67.11±0.29% over 4 seeds. This contrast highlights both the importance of objective-function choice and the architecture heterogeneity that surrounds it. The pattern repeats at W6A6 on four of the five models.

The takeaway: when exploiting a weight-space symmetry, *how* the symmetry is optimized is a design variable on par with *which* symmetry is used, and the same objective-function reasoning is worth applying to symmetries beyond rotation.

Acknowledgements

This work was supported by the Technological Innovation R&D Program (RS-2024-00470370) funded by the Ministry of SMEs and Startups (MSS, Korea).

References

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Li, Z., Xiao, J., Yang, L., and Gu, Q. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17227–17236, 2023.

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

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024.

Moon, J., Kim, D., Cheon, J., and Ham, B. Instance-aware group quantization for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16132–16141, 2024.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Shao, Y., Chen, Y., Wang, P., Yu, J., Lin, J., Yao, Y., Wei, Z., and Cheng, J. Dartquant: Efficient rotational distribution calibration for llm quantization. *arXiv preprint arXiv:2511.04063*, 2025.

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

Wu, Z., Chen, J., Zhong, H., Huang, D., and Wang, Y. Adalog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *European Conference on Computer Vision*, pp. 411–427. Springer, 2024.

Wu, Z., Wang, S., Zhang, J., Chen, J., and Wang, Y. Fimaq: Post-training quantization for vision transformers by fisher information matrix approximation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14891–14900, 2025.