📌 PINNING VISUAL PATHS: PRESERVING VISUAL STRUCTURE FOR ULTRA-LOW-BIT VISION ENCODERS WITH TASK-INFORMED RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision encoders increasingly rely on low-bit inference to cut computation and memory. Post-training quantization (PTQ) therefore calibrates a low-bit student to sequentially reconstruct the block-wise activations of a full-precision teacher using a subset of images. Since uniform objectives like MSE spend limited low-bit capacity on reducing all feature errors equally, regardless of their task relevance, recent methods instead optimize the loss derived from the *gradient-curvature-objective* pipeline: they propagate quantization perturbations through a full-precision suffix (the full-precision layers following the quantized block) to obtain exact task-loss *gradients*, derive second-order *curvature* information from these gradients, and use the resulting Taylor approximation of the task loss as the optimization *objective*. However, these methods mainly attribute the improvement to more accurate prediction of the task loss, leaving the reliability of the estimation-and-optimization pipeline itself largely unexamined. The curvature inferred from sampled gradients may not match the Hessian required by the local Taylor expansion, and ultra-low-bit quantization produces perturbations far beyond the range where such a local approximation holds. More fundamentally, even removing this approximation and directly optimizing the exact KL divergence can overfit both the small calibration set and the full-precision suffix, while also introducing strong cross-batch variation in the optimization signal. This motivates a different objective: incorporating task information into reconstruction while keeping updates stable and reliable. We propose Pinning Visual Paths (PVP). It uses gradient magnitudes to weight block-output errors without fitting a full endpoint task-loss model, thereby avoiding unstable endpoint-KL update directions and overfitting the task-blind part of the KL update space. It adds local constraints on patch features and the next attention block's token aggregation to address the mismatch introduced by a quantized suffix. It first rotates and rescales activation coordinates to reduce outliers, then refines weight codes and activation scales to stabilize discrete optimization within a restricted update space that limits overfitting. Across seven ImageNet encoders, PVP improves mean W2A3 Top-1 accuracy by 5.31 percentage points over SOTA methods, and remains effective for self-supervised, vision-language, and downstream models. Packed kernels provide a geometric-mean 2.39× speedup and 2.99× memory reduction over BF16.

1 INTRODUCTION

Vision transformers now support multimodal assistants, embodied systems, and world models (Radford et al., 2021; Zhai et al., 2023; Liu et al., 2023a; Zitkovich et al., 2023; Assran et al., 2025), so their cost matters at the edge and in high-throughput inference. Vision PTQ has progressed at moderate precision (Yuan et al., 2022; Lin et al., 2022; Li et al., 2023), but ultra-low-bit accuracy remains suboptimal as intermediate-feature errors accumulate through the encoder. Consequently, visual modules are often kept at higher precision than language modules in deployed multimodal systems (Qwen Team, 2024; 2025; ggml-org contributors, 2026). The key challenge lies in compressing

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.

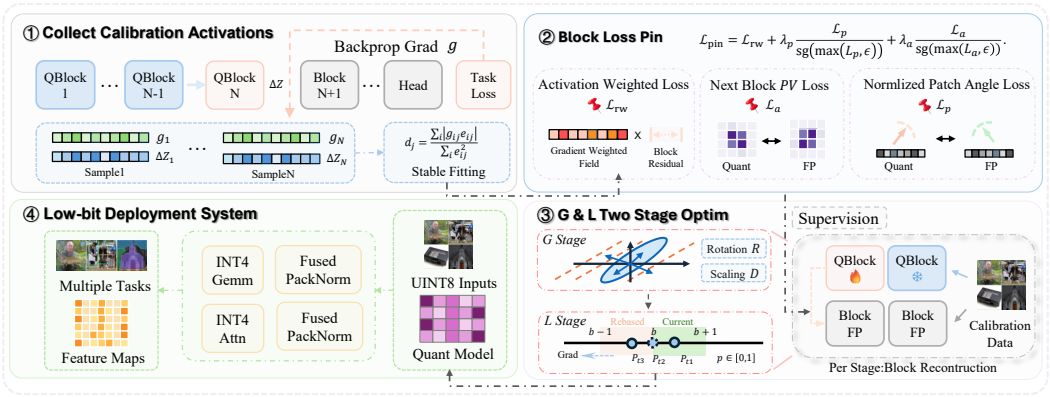


Figure 1: **Calibration-to-deployment pipeline of PVP.** (1) A small calibration set traces paired full-precision and quantized block activations, from which a stable task-informed activation field is fitted. (2) Activation-weighted residuals and two visual pins preserve patch directions and the next-block attention–value aggregate during block correction. (3) Gauge Alignment reshapes activation coordinates, while Lattice Pinning refines low-bit weight codes and activation scales. (4) The calibrated model is exported to packed kernels for diverse visual workloads.

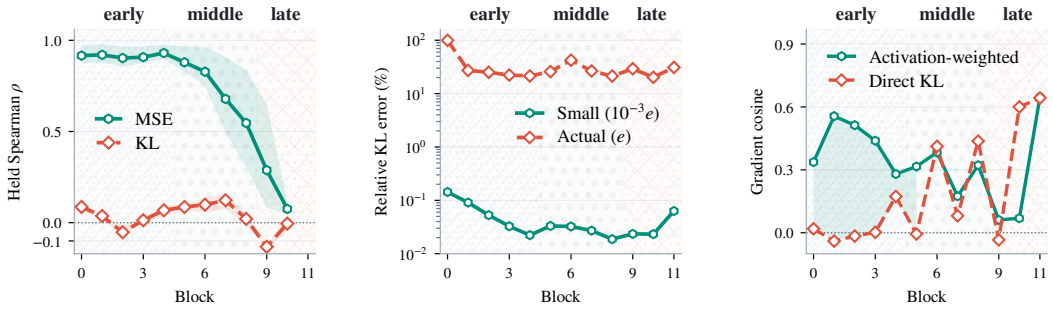
both weights and activations while preserving the visual structures essential for downstream tasks. Related work is summarized in Appendix C.

Recent reconstruction-based PTQ uses a small unlabeled set to fit a quantized student to a full-precision teacher. Block reconstruction is a practical compromise: it retains within-block compensation without exposing the whole network to few-shot overfitting (Li et al., 2021; Wei et al., 2022). MSE provides a stable baseline, but spends the limited low-bit capacity uniformly; task-aware methods therefore add Hessian- or Fisher-weighted corrections (Wu et al., 2025a; Hwang et al., 2026).

These methods motivate a natural but incomplete strategy: make task-loss prediction more accurate and optimize it directly. Our audit in Table 1 identifies three linked defects in this route. Gradients collected after perturbations are used to infer a curvature matrix and then predict a KL change, but the curvature mapping a perturbation to a gradient is not generally the curvature that integrates into the endpoint loss. The actual ultra-low-bit displacement can also exceed the radius in which either local approximation is reliable. Direct KL exposes a data-dependent update space: with few images, its local task metric has rank bounded by the observed task directions, leaving many parameter directions unconstrained. Different calibration subsets can therefore fit different solutions despite similar KL on the observed images, and their conditional gradients need not agree across batches. These gradients are also computed through a floating-point suffix although deployment uses a quantized suffix, so further optimization can sharpen the calibration fit while targeting a downstream computation that will change afterward. Finally, the available task-aware parameters reweight output errors but do not resolve the nonlinear activation geometry that creates outliers and moving quantization cuts. Together, these findings call for a field whose strength uses task sensitivity but whose direction is restored by the residual, local pins that survive suffix quantization, and a finite optimization space that also handles activation outliers and discrete code-cell boundaries.

We introduce *Pinning Visual Paths* (PVP) as a direct construction from this diagnosis. Its restoring field uses gradient magnitudes for task sensitivity but ties the direction to the teacher–student residual, avoiding batch-specific KL directions and overfitting in task-blind directions. Local patch and next-block attention–value pins preserve token geometry and the signal delivered to the changing suffix, addressing suffix mismatch. Gauge Alignment rotates and rescales activation coordinates to spread outliers and move quantization cuts. Lattice Pinning then updates nearby weight codes and activation log-scales through endpoint-preserving cell transitions, stabilizing discrete optimization while restricting directions exposed to few-shot overfitting. The resulting G–C–L pipeline follows the observed failure modes directly (Figure 1). Our contributions are:

- Through analysis and controlled experiments, we show that more accurate estimation of the Hessian-to-task-loss mapping does not necessarily improve ultra-low-bit ViT quantization, nor



(a) Reachable-code ranking (b) Finite-radius prediction (c) Batch update agreement

Figure 2: **Diagnostics of task-guided reconstruction mismatch.** (a) Held-out Spearman rank correlation on 128 legal W2 code moves for reconstruction weights induced by recent methods. (b) Exact-Hessian KL prediction error. (c) Cross-batch code/scale-gradient cosine.

does direct KL optimization. We attribute this gap to finite-radius approximation errors, curvature mismatch, quantized-suffix shifts, unstable few-shot gradients, and large empirical nullspaces, distinguishing task-loss fidelity from reliable calibration guidance.

- We propose PVP, a task-informed framework combining a gradient field with patch-direction and attention-aggregation constraints. *Gauge Alignment* reshapes feature coordinates for activation quantization, while *Lattice Pinning* jointly optimizes weight codes and activation scales.
- PVP improves mean Top-1 accuracy by 5.31 percentage points over the prior state of the art across diverse vision encoders and tasks. Its packed low-bit runtime achieves a geometric-mean 2.39× speedup and 2.99× peak-memory reduction over BF16.

2 Reconstruction, Task Approximations, and Update Fields

2.1 Problem formulation

Reconstruction-based PTQ fits quantized blocks sequentially to a full-precision teacher on a small unlabeled calibration set $\mathcal{C}$. A uniform b-bit quantizer uses $\hat{v} = s(k - z)$, where $k = \mathrm{clip}(\mathrm{round}(v/s) + z, 0, 2^b - 1)$, $s > 0$ is the scale, and z is the zero point.

Task-sensitive methods replace isotropic block-output reconstruction with a quadratic task-loss surrogate (Hwang et al., 2026; Wu et al., 2025a;b). Let $K(e)$ denote the teacher KL divergence evaluated through a full-precision suffix, and let $g = \nabla K(e)$. Using sampled gradients and perturbations to estimate curvature yields

$$e_l = \hat{y}_l - y_l, \qquad \rho_l(e_l) = \tfrac{1}{2} e_l^{\mathsf{T}} \widehat{H}_l e_l, \qquad \widehat{H}_l = \mathcal{A}_l(\{(e_{l,i}, g_{l,i})\}_{i=1}^m), \tag{1}$$

where $\mathcal{A}_l$ denotes the estimator and y_l denotes the outputs of block l. We call this the $g \to H \to \mathrm{KL}$ route *curvature-based task-loss prediction*. Recent methods pursue increasingly accurate curvature estimates, but our examination of the resulting loss predictions and direct KL optimization challenges the assumption. The information-budget and task-relevant-collision view underlying this limitation is formalized in Appendix F.

2.2 Does Fidelity to Task Loss Pay Off in Quantized Block Reconstruction?

2.2.1 Limits of Curvature-Based Task Loss Prediction

Behaviors along the $g \to H$ and $H \to \mathrm{KL}$ paths do not necessarily coincide. Rather than being an artifact of the estimation method, this divergence is fundamentally governed by quantization errors and local sensitivity. In low-bit regimes, actual quantization perturbations can exceed the radius of local validity, further amplifying this effect. Assuming twice differentiability, with $K(0) = 0$, $\nabla K(0) = 0$, and $H(u) = \nabla^2 K(u)$, the exact identities are illustrated in Fig. 3 (①–②).

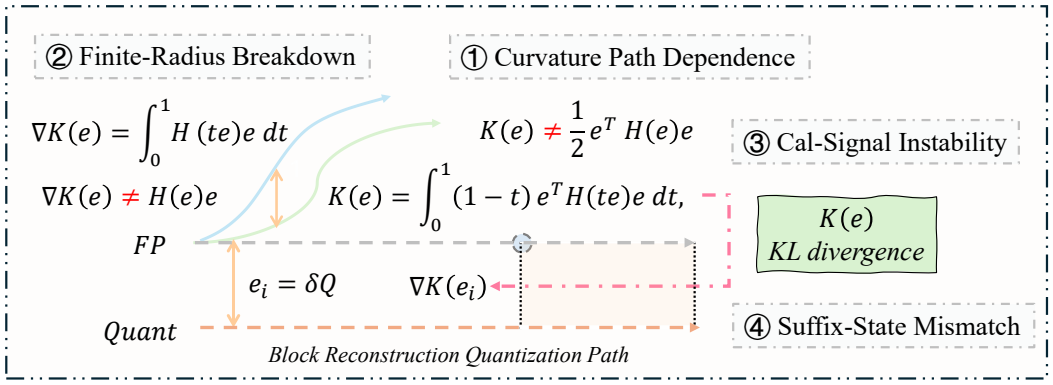


Figure 3: **Task-loss prediction and direct KL optimization in block reconstruction.** ①–②: curvature path dependence and finite-radius effects in Hessian-based KL prediction (Sec. 2.2.1). ③–④: calibration-signal instability and suffix-state mismatch in direct KL optimization (Sec. 2.2.2).

Table 1: **Block-loss audit at W2A3.** All rows use 1,024 ImageNet calibration images with a shared starting point. KL values are reported in units of 10^{-3}, and accuracy is ImageNet Top-1 (%). The held split is generated from the fit split using standard training-data augmentations.

Encoder	Bits	Steps/B	Loss	Fit ↓	Held ↓	H-F Δ	acc ↑
DeiT-T	W2A3	20K	PVP L_{rw}	0.959	1.179	0.220	44.95
DeiT-T	W2A3	20K	Block MSE	1.467	1.644	0.177	25.49
DeiT-T	W2A3	20K	FP suffix KL	2.282	2.641	0.359	2.792
ViT-S	W2A3	20K	PVP L_{rw}	2.480	2.700	0.220	55.13
ViT-S	W2A3	20K	Block MSE	3.375	3.536	0.161	33.82
ViT-S	W2A3	20K	FP suffix KL	6.513	7.298	0.785	0.622

The distinction matters: gradients average curvature uniformly along the path, while the loss gives more weight to its beginning. Fitting a matrix to these gradient responses therefore does not automatically justify the loss estimate $\frac{1}{2}e^\top He$. Replacing the path-dependent Hessian by $H(0)$ recovers both familiar approximations, but their accuracy depends on curvature variation over the residual radius, not just on the quality of Hessian estimation (Appendix B.6). Figure 2b isolates this issue by using the exact boundary Hessian and varying only the perturbation radius. The quadratic prediction is accurate near the teacher but deteriorates at the actual W2A3 residual radius, where the deployed perturbations lie outside the neighborhood in which the local approximation is reliable.

The ranking audit shows that their useful behavior does not require faithful prediction of the terminal KL. In early and middle blocks, the published reconstruction metrics can rank block-MSE changes well while remaining weak predictors of terminal-KL changes (Figure 2a). Their useful behavior need not come from faithful KL recovery. A fixed positive-semidefinite quadratic still has an MSE-like shrinkage bias: in residual space, its negative gradient cannot increase the residual norm to first order. This offers a plausible regularization role distinct from accurate task-loss estimation. The complete mechanism audit is collected in Appendix B; the executable metric-ranking results are detailed in Appendix B.1, and the finite-radius measurements are summarized in Appendix B.2. The non-identifiability of curvature from sampled gradient–perturbation pairs is proved in Appendix B.5, while the diagonal restoring choice is analyzed in Appendix B.3.

2.2.2 Why Neither KL nor MSE Suffices for Low-Bit Reconstruction

But does directly optimizing the KL—ostensibly the most faithful objective—resolve this issue? In practice, direct KL optimization performs poorly, particularly with small batches, and shows only occasional improvement with larger ones. This inconsistency prompts us to examine whether different mini-batches agree on the update direction. Ultimately, we find that direct optimization introduces distinct failure modes stemming from the calibration signal and downstream suffix drift (Fig. 3, ③–④). In the affected blocks, parameter gradients are often nearly orthogonal or opposed

along the KL path; activation weighting improves agreement in several early blocks (Figure 2c). Direct KL also transfers poorly to the held-out augmented views in Table 1.

Two properties help explain this behavior. First, KL constrains a block correction only through its downstream probability activation. For a smooth parameter surrogate with p_θ trainable coordinates, n images, and C classes, its local Gauss–Newton matrix satisfies $\mathrm{rank}(G_{\mathrm{KL}}) \le \min\{p_\theta, n(C-1)\}$. When the observed activation rank is small relative to the trainable dimension, many directions are invisible to this local task metric, and different batches can constrain different subspaces (Appendix F.1), which leads to calibration-sensitive updates. MSE, by penalizing the full block residual, also constrains representation changes that terminal KL may not see.

Second, the FP suffix defines a conditional objective that differs from deployment. A quantized suffix may already disagree with the teacher whose smooth local loss can consequently have a nonzero first-order term as well as different curvature (Appendix B.6). More optimization can therefore fit the calibration views more tightly while optimizing a suffix that will not be deployed.

KL divergence can be unstable, but naive MSE is also not sufficient in the ultra-low-bit regime, as observed here and in recent work Li et al. (2021); Wu et al. (2025a). Uniform residual penalties can overconstrain the limited representational capacity of the quantized model, making it difficult to enforce uniformly small errors across all channels. Such isotropic penalties can induce a steep and poorly conditioned optimization landscape. In contrast, weighted or relaxed objectives tolerate a controlled amount of quantization error in less critical directions, while still encouraging the quantized representation to remain close to the full-precision feature manifold along important directions. The complementary code-budget view is given in Appendix F.2.

3 Pinning Visual Paths

These insights reveal that an effective objective must capture task and architectural sensitivities, curtail error accumulation along long suffixes, and regularize internal representations against overfitting. To this end, PVP casts task sensitivity into reusable local constraints, pairs a task-weighted reconstruction field with dual visual losses, and optimizes activation geometry alongside weight codes. Below, we detail the pinned objective, our two-stage optimization, and packed execution.

3.1 Task-weighted restoring fields.

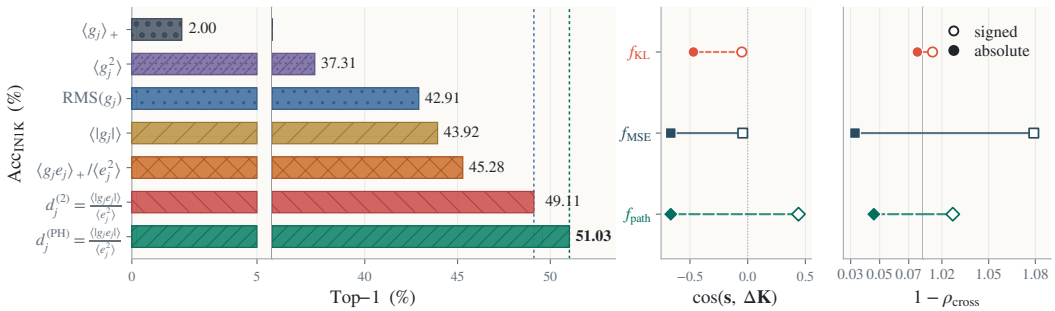
The batch disagreement in Sec. 2.2.2 changes how we use the task gradient. It is useful for measuring sensitivity, but its sign is not a stable reconstruction direction across calibration images. We therefore let the teacher–student residual determine the direction of motion and use task sensitivity only to set the correction strength. The resulting field retains the restoring direction of MSE while allocating more correction to task-relevant coordinates.

At boundary l, let $e_{l,i} = \hat{y}_{l,i} - y^\star_{l,i}$ be the student–teacher residual and $g_{l,i}$ the gradient of the task discrepancy with respect to the same activation. We obtain $g_{l,i}$ through the floating suffix, using teacher KL for classifiers and semantic discrepancies for representation encoders. Suppressing l, $|g_{ij}|$ measures local sensitivity, whereas $|g_{ij}e_{ij}|$ measures its first-order contribution at the error that is actually present. The latter is the quantity that can be compared across calibration images without allowing opposite signs to cancel.

To turn this sensitivity into a reconstruction weight, consider slightly shrinking coordinate j: $z_i(t) = e_i - te_{ij}u_j$, with u_j the coordinate unit vector. The sensitivity envelope $\mathcal{R}_i(z) = \sum_k |g_{ik}z_k|$ then decreases at rate $|g_{ij}e_{ij}|$. A weighted reconstruction energy $Q_d(z) = \frac{1}{2}\sum_k d_k z_k^2$ decreases at rate $d_j e_{ij}^2$. Matching these first-order energy changes gives

$$
\begin{gathered}
-\frac{\mathrm{d}}{\mathrm{d}t}\sum_i \mathcal{R}_i(z_i(t))\bigg|_{t=0} = \sum_i |g_{ij}e_{ij}|, \qquad -\frac{\mathrm{d}}{\mathrm{d}t}\sum_i Q_d(z_i(t))\bigg|_{t=0} = d_j \sum_i e_{ij}^2, \\
\sum_i |g_{ij}e_{ij}| = d_j \sum_i e_{ij}^2 \quad \Longrightarrow \quad d_j = \frac{\sum_i |g_{ij}e_{ij}|}{\sum_i e_{ij}^2}.
\end{gathered}
\tag{2}
$$

Thus d_j converts task sensitivity into correction strength per unit residual energy. Since $d_j \ge 0$, the reconstruction gradient $d_j e_{ij}$ keeps the residual's sign, so its negative always points back toward the



(a) Gradient-weighted comparison (b) Candidate-score stability

Figure 4: **Stable sensitivity and robust reconstruction.** (a) ViT-S W2A3 accuracy under alternative coordinate weights and potentials. (b) A single-block audit of candidate changes δ using KL, MSE, and path-gradient probes f. Open markers score $\sum_j f_j \delta_j$, whereas filled markers score $\sum_j |f_j \delta_j|$.

teacher activation. This is the distinction tested in Fig. 4(b): the signed candidate scores fluctuate across image splits, while the absolute scores provide a substantially more stable ordering even though their correlation with the signed KL change is weak. The experiment therefore supports the separation encoded by Eq. 2: magnitude sets the ruler, and the residual supplies the restoring direction.

A few large residuals can still dominate a quadratic loss. We use a pseudo-Huber potential to keep the same restoring direction while limiting their influence:

$$\rho_c(r) = c^2 \left(\sqrt{1 + (r/c)^2} - 1 \right), \qquad \mathcal{L}_{\text{rw}}(e_l) = \mathbb{E}_i \sum_j d_{l,j} \rho_{c_{l,i}}(e_{l,ij}), \tag{3}$$

where $c_{l,i} = 0.25 \max\{\text{RMS}(y^\star_{l,i}), 10^{-6}\}$. Small residuals receive an approximately linear correction; for large residuals, the gradient saturates smoothly, with bounded coordinate curvature. This keeps the local restoring behavior while preventing a few heavy-tailed coordinates from taking over the update. The endpoint comparison in Fig. 4(a) follows the same construction: among the quadratic choices, displacement-conditioned absolute weighting gives the strongest result, and replacing the quadratic potential by pseudo-Huber adds 1.92 Top-1 points. Thus the two design decisions are linked: the absolute first-order contribution determines where the field should act, and the pseudo-Huber curvature determines how strongly an unusually large residual may act. Appendix D.1 derives these properties; Appendix D.2 gives the calibration details.

3.2 Soft pins for local identity and aggregation

Long-suffix gradients vary with the calibration sample and suffix quantization. We therefore add two *soft pins* that preserve local visual structure while leaving room for weight–activation compensation: one anchors patch identity and the other anchors the next-block attention–value aggregate.

Patch-direction pins. At a block boundary, each non-prefix token is a spatial feature vector. We align its normalized direction at the matched location:

$$\mathcal{L}_p = \mathbb{E}_{x,t \in \text{patches}} \left[1 - \cos\big(\text{LN}(\hat{y}_{l,t}), \text{LN}(y^\star_{l,t})\big) \right]. \tag{4}$$

This preserves the within-patch channel pattern while allowing shifts and positive rescaling to be absorbed elsewhere.

Aggregation pins. The remaining freedom affects how tokens exchange information. Rather than matching the attention map alone, we pin the delivered aggregate $Z = PV$ of the next transformer block through the full-precision suffix; this retains both routing and the value content transported through it.

We compute Z per head and compare all heads jointly after vectorization, before head concatenation and output projection:

$$\mathcal{L}_a = \mathbb{E}_x \left[1 - \cos(\text{vec}\, Z_q, \text{vec}\, Z_\star) + \gamma_e \left(\log\left(\text{RMS}\, Z_q / \text{RMS}\, Z_\star \right) \right)^2 \right]. \tag{5}$$

Table 2: **Top-1 accuracy of ViT-based models on ImageNet.** The metric is ImageNet-1K Top-1 (%). **Bold** indicates the best top-1 accuracy for each bit width. The W3A3 comparison with more methods is given in Appendix A.4.

Exp Setup		ViT		DeiT			Swin		Avg.
Method	Bits	Small	Base	Tiny	Small	Base	Small	Base	
FP	BF16	81.43	84.55	72.19	79.86	81.97	83.19	83.59	80.97
APHQ-ViT	W2A2	21.07	40.46	18.79	32.46	57.18	50.77	45.82	38.08
FIMA-Q	W2A2	21.64	48.35	21.78	38.27	59.70	58.06	63.93	44.53
LS-ViT	W2A2	22.62	40.66	18.25	39.18	58.14	60.41	66.65	43.70
Ours	W2A2	**33.33**	**54.05**	**23.98**	**44.25**	**60.87**	**61.68**	**66.78**	**49.28**
QDrop	W2A3	17.90	54.22	23.58	21.48	54.57	57.28	57.57	40.94
PD-Quant	W2A3	25.56	46.06	31.45	48.66	63.60	62.41	65.29	49.00
APHQ-ViT	W2A3	41.63	58.06	38.33	53.53	66.09	62.90	65.03	55.08
FIMA-Q	W2A3	38.82	61.15	36.93	52.57	66.24	59.57	63.15	54.06
LS-ViT	W2A3	50.09	65.65	43.02	60.51	72.06	71.78	74.47	62.51
Ours	W2A3	**59.54**	**72.41**	**50.13**	**65.81**	**74.68**	**74.67**	**77.52**	**67.82**
QDrop	W2A4	49.35	71.44	38.87	45.99	70.01	70.13	68.64	59.20
BRECQ	W2A4	45.20	67.18	46.31	61.20	69.72	69.21	70.36	61.31
PD-Quant	W2A4	45.12	69.82	46.86	61.26	71.99	71.51	73.03	62.80
APHQ-ViT	W2A4	56.04	72.85	49.79	65.02	73.07	71.75	72.64	65.88
FIMA-Q	W2A4	56.76	74.41	50.70	65.38	73.32	70.64	72.36	66.22
LS-ViT	W2A4	64.41	75.14	54.78	69.23	76.12	74.14	75.05	69.84
Ours	W2A4	**69.47**	**78.01**	**59.30**	**72.43**	**77.91**	**78.64**	**80.41**	**73.74**

The aggregate therefore allows routing and value errors to compensate while preserving the delivered feature. Its consumer-state interpretation and scale behavior are detailed in Appendices D.4 and D.5.

Joint fitting. The field and the two pins are combined as

$$\mathcal{L}_{\text{pin}} = \mathcal{L}_{\text{rw}} + \lambda_p \frac{\mathcal{L}_p}{\text{sg}(\max(L_p, \epsilon))} + \lambda_a \frac{\mathcal{L}_a}{\text{sg}(\max(L_a, \epsilon))}. \tag{6}$$

Here $\text{sg}(\cdot)$ denotes stop-gradient. Detached normalization gives the auxiliary terms comparable update budgets. Supporting pin and local-field derivations are given in Appendix D.3–D.6.

3.3 Geometry and lattice optimization

The pinned loss gives a desired correction, but a low-bit quantizer can reach it only through its available coordinates. Because few-shot calibration sets cannot reliably constrain full parameters as discussed above, we open two small and complementary update spaces. *Gauge Alignment* (G) changes only the activation rotation and scale, thereby moving the activation quantization cuts. *Lattice Pinning* (L) then changes only the neighbor rebase weight codes and log-domain activation scale; all other parameters stay fixed. This lets activation and weight errors compensate without turning reconstruction into an unconstrained full training. Let $u = (R, s)$ denote the G variables and $v = (q, a)$ the L variables. The two stages are

$$u^\star = \arg\min_{u \in \mathcal{G}} \mathcal{L}_{\text{pin}}(Q(u, v_0(u))), \qquad v^\star = \arg\min_{v \in \mathcal{C}(u^\star)} \mathcal{L}_{\text{pin}}(Q(u^\star, v)). \tag{7}$$

Here $v_0(u)$ is the initial code/scale fit after G, and $\mathcal{C}(u)$ is the small set of code/scale states allowed during L. Thus G supplies a better activation surface, while L supplies limited discrete freedom for the remaining quantization error. The complete handoff is formalized in Appendix E, with the operational procedure in Appendix E.1 and the corresponding settings listed in Appendix E.2.

4 Experiments

We evaluate PVP across three questions: (i) whether it preserves critical visual features under aggressive bit constraints, (ii) which design choices drive this retention, and (iii) whether these theoretical benefits translate into tangible runtime efficiency. Stage settings are detailed in Appendix E.2.

Table 3: **Modern encoder coverage across bit budgets.** ImageNet-1K Top-1 accuracy (%) across diverse structures and metrics. The paired Top-5 table is reported in Appendix 7.

Encoder	Metric	FP	W2A2	W2A3	W2A4	W3A3	W4A4
DINOv2-B	Prototype	64.146	29.108	52.172	56.788	57.398	62.016
DINOv2-B	Linear probe	82.298	54.838	74.950	78.754	77.252	80.538
DINOv3-B	Prototype	65.774	15.756	45.184	55.474	53.012	54.866
DINOv3-B	Linear probe	82.620	36.510	69.704	76.800	74.492	74.630
CLIP-L	Prototype	53.110	21.320	39.228	44.322	41.774	49.486
CLIP-L	Zero-shot	76.528	31.794	60.242	66.664	65.340	71.329
SigLIP-B	Prototype	65.096	34.846	50.624	55.952	55.012	61.118
SigLIP-B	Zero-shot	76.078	37.040	60.318	67.422	66.412	73.328
SigLIP-So	Prototype	74.196	47.466	65.004	69.256	67.604	71.444
SigLIP-So	Zero-shot	81.974	55.154	74.008	78.380	77.048	80.542
SigLIP2-So	Prototype	75.148	31.262	64.232	69.762	68.326	72.240
SigLIP2-So	Zero-shot	74.310	30.904	65.502	70.670	69.342	72.492

4.1 Accuracy across model families and bit budgets

Table 2 shows a consistent advantage across both flat and hierarchical encoders, yielding an average W2A3 improvement of 5.31 percentage points over prior methods. Notably, this gain is substantially more pronounced at W2A3 than at W2A4. This divergence confirms that PVP selectively concentrates reconstruction capacity on sensitive features and critical degrees of freedom precisely when activation budgets are severely constrained. Moreover, the sustained gains across varying bit budgets indicate that the pinned objective generalizes well beyond a specific precision regime.

While Top-1 classification demonstrates aggregate performance, it leaves open whether the underlying visual semantics are genuinely preserved. Table 3 evaluates quantized backbones to probe representation quality. Comparisons across methods are deferred to Appendix A.4.

4.2 Feature retention across tasks and datasets

Global classification can obscure local spatial degradation. To assess whether PVP retains the fine-grained spatial features essential for dense scene understanding, we evaluate detection and segmentation on COCO (Table 4). PVP consistently outperforms prior baseline on both bounding-box and mask AP. These simultaneous gains in localization and segmentation confirm that spatial fidelity is well-preserved across backbone stages, fulfilling the goal of the local visual pins.

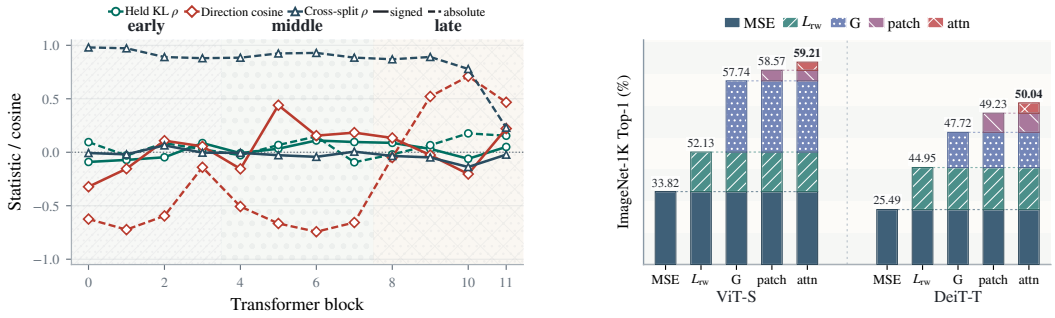
Beyond in-domain dense predictions, a robust low-bit representation should transfer gracefully to unseen distributions. We examine this on CIFAR-100 using frozen ImageNet-calibrated backbones (Appendix A.1). Under both geometry configurations, PVP exhibits competitive classification and retrieval performance. These complementary retrieval and classification metrics confirm that the learned quantization generalizes beyond the calibration distribution.

4.3 From low-bit representations to efficient inference

To bridge the gap between algorithm and deployment, we developed a whole-graph inference system with uint8 image input and packed low-bit kernels to facilitate efficient edge inference deployment. End-to-end W2A4 measurements across all models used by experiments yield a geometric-mean $2.39\times$ speedup and $2.99\times$ peak-memory reduction. Performance data are detailed in Fig. 6 and Table 11, while runtime configurations are described in Appendix A.6.

4.4 Component ablation of the reconstruction pipeline

Having established empirical performance gains, we now isolate the individual mechanisms driving PVP. As illustrated in Fig. 5(a), absolute scores maintain consistent cross-split rankings across depth, whereas signed KL changes show near-orthogonality and poor stability. These diagnostics



(a) Depth-wise PVP audit

(b) Cumulative component stack

Figure 5: **Component ablation of the pinned reconstruction.** (a) reports ranking correlation coefficient and candidate-score cosine against measured KL changes, and cross-split rank stability for signed and absolute path-based scores; (b) gives MSE, activation weighting, geometry, patch, and attention increments.

Table 4: **Multi-task evaluation on COCO.** Each cell reports bbox AP / mask AP on the complete 5K-image val2017 split. Prior is selected independently in each context as the strongest joint result among LS-ViT, FIMA-Q, APHQ-ViT, and GPTQ; superscripts L/F/A/G identify method.

Model context			W3A3		W2A4		W4A4	
Detector	Backbone	BF16	Prior	**Ours**	Prior	**Ours**	Prior	**Ours**
Mask	Swin-T	46.01/41.68	12.39/13.90^{L}	**12.52/13.92**	37.95/35.01^{L}	**41.08/37.53**	44.45/39.44^{A}	**44.63/39.74**
Mask	Swin-S	48.14/43.15	39.66/35.84^{A}	**43.07/38.72**	40.31/36.60^{L}	**43.47/39.05**	45.81/41.08^{A}	**46.18/41.39**
Cascade	Swin-T	50.46/43.76	45.67/39.73^{A}	**45.90/39.82**	44.49/38.69^{L}	**46.42/40.32**	48.06/41.53^{A}	**48.12/41.64**
Cascade	Swin-S	51.95/44.98	45.02/38.99^{A}	**47.26/40.88**	45.10/39.19^{L}	**47.70/41.24**	49.42/42.75^{A}	**50.04/43.27**

justify our design choice: using task magnitudes to allocate correction strength, while relying on the residual to dictate its update direction.

Building on this metric, the cumulative ablation in Fig. 5(b) examines how individual losses interact. Task-weighted reconstruction contributes the bulk of the initial gain, geometry alignment further elevates this baseline, and local visual pins contribute consecutive accuracy boosts. These step-wise improvements confirm that preserving patch identity and next-block aggregation imposes complementary geometric constraints beyond block-level fidelity alone.

Finally, we link these macroscopic gains to microscopic optimization behavior. Following stage G, activation tails contract, enabling subsequent scale trajectories to stabilize substantially earlier at representative layers (Appendix A.7)—a clear signal of improved conditioning for stage L. This controlled experiment confirms that crossing code-cell boundaries provides an active degree of freedom that effectively complements the distribution reshaping performed in stage G. The coupled-error interpretation of this compensation is given in Appendix B.4.

5 Conclusion

Our analysis shows that curvature-based task approximations are locally informative but can lose predictive fidelity due to limited perturbation coverage, finite quantization radius, and calibration-dependent updates. These observations motivate a balance between task awareness and stable correction fields over small calibration sets. We therefore propose PVP, a reconstruction-based PTQ framework built on shared visual paths that translates task-level endpoint signals into coupled block corrections via task-informed activation weights and short-horizon visual pins, stabilizing the finite-sample correction field along the reachable quantization trajectory.

AI USE STATEMENT

Generative AI was used solely to assist with the organization and formatting of early manuscript drafts. It was not used for method conception, experimental implementation, or the substantive writing of the paper.

ETHICS STATEMENT

This work studies model compression using standard computer vision benchmarks. We do not use human-subject data or introduce a new dataset. The main foreseeable impact is reduced inference memory and compute, although the environmental cost of the calibration sweeps should also be considered.

REPRODUCIBILITY STATEMENT

An anonymized supplementary package will include the exact coordinate, calibration, reconstruction, and terminal-evaluation configurations; per-cell manifests; and fresh-process evaluation scripts.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi: 10.48550/arXiv.2506.09985. URL https://arxiv.org/abs/2506.09985.

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*, 2023. URL https://arxiv.org/abs/2210.17323.

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.

ggml-org contributors. Quantize: Multimodal components. llama.cpp documentation, 2026. URL https://github.com/ggml-org/llama.cpp/blob/master/tools/quantize/README.md. Accessed 2026-09-09.

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, June 2026.

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*, 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.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.01721.

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*, pp. 1173–1179, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023a. URL https://arxiv.org/abs/2304.08485.

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, 2023b. URL https://arxiv.org/abs/2212.07048.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.16406.

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 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 2020.

Qwen Team. Qwen2-VL-7B-Instruct-AWQ: Model configuration. Hugging Face model repository, 2024. URL https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct-AWQ/blob/main/config.json. Accessed 2026-09-09.

Qwen Team. Qwen2.5-VL-7B-Instruct-AWQ: Model configuration. Hugging Face model repository, 2025. URL https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct-AWQ/blob/main/config.json. Accessed 2026-09-09.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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*, 2022.

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

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, 2025a. doi: 10.1109/CVPR52734.2025.01387.

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, 2025b. doi: 10.1109/CVPR52734.2025.00905.

Lianwei Yang, Haisong Gong, Haokun Lin, Yichen Wu, Caifeng Shan, Zhenan Sun, and Qingyi Gu. DopQ-ViT: Towards distribution-friendly and outlier-aware post-training quantization for vision transformers. *arXiv preprint arXiv:2408.03291*, 2024. URL https://arxiv.org/abs/2408.03291.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, pp. 27168–27183, 2022. URL https://arxiv.org/abs/2206.01861.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://arxiv.org/abs/2303.15343. IEEE Xplore paginates this paper as 11941–11952; CVF Open Access uses 11975–11986.

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. doi: 10.1109/TPAMI.2025.3610466.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A Additional Experimental Results and Deployment

This appendix collects the extended empirical results in one sequence: cross-dataset and multi-task evaluations, modern-encoder readouts, precision optimizer studies, geometry/lattice dynamics, and measured deployment efficiency. The tables and audits use the same model and metric conventions as the main experiments.

A.1 Cross-dataset classification and retrieval

Models calibrated on ImageNet are evaluated on CIFAR-100 with a frozen backbone. The paired classification and retrieval readouts provide a compact check that the feature distribution and neighborhood structure remain useful beyond the ImageNet readout.

Table 5: **Cross-dataset evaluation on CIFAR-100.** Aggregate results over seven encoders and four bit budgets. The frozen backbone is evaluated with a 50-epoch linear probe and 10K-query cosine retrieval. Classification is reported as Top-1/Top-5 (%), and retrieval as recall (%) and mAP@100 (%).

Optimization stages				Classification (%)		Retrieval (%)			
G stage	G steps	L stage	L steps	Top-1 ↑	Top-5 ↑	R@1 ↑	R@5 ↑	R@10 ↑	mAP@100 ↑
PVP	3K	PVP	5K	**53.02**	**81.83**	**49.32**	**72.45**	**80.93**	**13.50**
PVP	3K	LS-ViT	5K	52.38	81.05	48.24	71.29	79.73	12.75
PVP	3K	APHQ-ViT	5K	48.60	77.79	43.19	66.16	75.32	10.76
PVP	3K	FIMA-Q	5K	51.38	80.26	47.35	70.17	78.83	12.28
Identity	–	PVP	5K	**62.84**	**83.76**	**45.94**	**69.87**	**78.98**	**11.76**
Identity	–	LS-ViT	5K	62.56	83.58	45.64	69.63	78.63	11.30
Identity	–	APHQ-ViT	5K	57.06	78.51	39.83	63.78	73.43	9.03
Identity	–	FIMA-Q	5K	61.23	82.30	44.70	68.71	77.82	10.86

A.2 A100 CALIBRATION EFFICIENCY

The shorter calibration schedule reaches stronger endpoints at lower wall-clock cost. Table 6 reports the complete A100 timing comparison for four representative models; entries are wall-clock minutes on a single NVIDIA A100 GPU.

Table 6: **A100 calibration time.** Wall-clock GPU time for PTQ calibration on four representative models, measured in minutes on a single NVIDIA A100 GPU.

	Calibration time (min)						Relative
Encoder	APHQ-ViT	FIMA-Q	LS-ViT	**Ours G+L**	Qdrop-MSE (5K)	Qdrop-MSE (20K)	Ours vs. LS-ViT Speedup
DeiT-T	64.5	69.6	65.8	**15.5**	14.16	48.63	**4.25**×
ViT-S	125.8	108.1	87.7	**23.2**	13.51	71.67	**3.78**×
DeiT-S	123.4	114.4	100.4	**22.9**	14.27	73.41	**4.38**×
ViT-B	375.3	337.9	201.4	**47.8**	14.07	152.30	**4.21**×

A.3 MODERN-ENCODER TOP-5 ENDPOINTS

Table 7 complements the Top-1 results in Table 3 with prototype, zero-shot, and frozen-linear-probe readouts across the modern encoders and five bit budgets. Each quantized Top-5 value uses the same configuration as its corresponding Top-1 entry.

Table 7: **Modern encoder Top-5 accuracy.** ImageNet-1K Top-5 accuracy (%) for the prototype, zero-shot, and frozen-linear-probe readouts in Table 3; FP denotes the corresponding floating-point readout, and a dash marks an incomplete endpoint.

Encoder	Metric	FP	W2A2	W2A3	W2A4	W3A3	W4A4
DINOv2-B	Prototype	86.864	53.936	77.630	81.526	81.666	85.180
DINOv2-B	Linear probe	95.848	78.804	92.702	94.536	94.034	95.168
DINOv3-B	Prototype	87.832	35.000	72.012	80.752	78.834	79.988
DINOv3-B	Linear probe	96.492	61.206	89.864	93.922	92.630	92.760
CLIP-L	Prototype	75.046	43.122	64.504	68.958	66.276	72.951
CLIP-L	Zero-shot	95.120	62.784	87.950	91.330	90.440	93.435
SigLIP-B	Prototype	87.518	62.202	77.454	81.564	80.708	85.188
SigLIP-B	Zero-shot	94.386	67.256	86.538	90.500	90.030	93.184
SigLIP-So	Prototype	92.192	74.262	87.254	89.562	88.940	90.994
SigLIP-So	Zero-shot	96.704	83.154	93.938	95.614	95.214	96.348
SigLIP2-So	Prototype	92.282	57.756	86.450	89.798	89.032	91.028
SigLIP2-So	Zero-shot	87.550	54.854	83.882	86.268	85.816	86.956

A.4 W3A3 CLASSIFICATION COMPARISON

We compare with representative reconstruction, activation-aware, and W3A3 PTQ baselines Li et al. (2021); Wei et al. (2022); Liu et al. (2023b); Wu et al. (2025b;a); Hwang et al. (2026); Yuan et al. (2022); Li et al. (2023); Frantar et al. (2023); Liu et al. (2025); Lin et al. (2024); Zhong et al. (2026); Yang et al. (2024); Wu et al. (2024). Top-1 and Top-5 readouts use the same model columns and reporting conventions as the main table.

A.5 LATTICE POLICY AND FITTING-OBSERVATION ABLATIONS

The matched optimizer comparison in Table 10 holds the transported initialization, W2A3 codebooks, reconstruction protocol, and fitting budget fixed. Its mean $+0.372$-point change localizes the effect of selective endpoint-preserving rebase relative to soft-code optimization. The four per-model differences show where the extra cell freedom contributes.

Table 8: **W3A3 classification.** Top-1 (%, upper) and Top-5 (%, lower) on ImageNet-1K. Methods retain their source protocols and all model columns are reported where results are available. UQ and BR denote uniform quantization and block reconstruction, respectively.

Method	UQ	BR	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B	Mean
BF16	–	–	81.43	84.55	72.19	79.86	81.97	83.19	83.59	80.97
PTQ4ViT	×	×	0.10	0.10	3.50	0.10	31.06	28.69	20.13	11.95
RepQ-ViT	×	×	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
GPTQ	✓	×	1.17	0.28	1.72	5.89	27.95	27.24	11.14	10.77
SpinQuant	✓	×	12.93	1.25	24.15	46.43	64.48	–	–	–
DuQuant	✓	×	0.22	3.24	4.01	13.05	49.44	–	–	–
I&S-ViT	×	×	45.16	63.77	41.52	55.78	73.30	74.20	69.30	60.43
DopQ-ViT	×	×	54.72	65.76	44.71	59.26	74.91	74.77	69.63	63.39
QDrop	✓	✓	41.05	74.75	46.88	50.95	72.97	74.67	76.57	62.55
BRECQ	✓	✓	54.33	66.62	49.27	63.72	72.82	75.20	77.64	65.66
PD-Quant	✓	✓	54.73	73.13	53.79	67.52	75.93	76.62	78.66	68.63
AdaLog	×	×	62.50	76.73	48.78	63.24	74.99	76.01	78.82	68.72
APHQ-ViT	✓	✓	63.17	76.31	55.42	68.76	76.31	76.10	78.14	70.60
FIMA-Q	✓	✓	64.09	77.63	55.55	69.13	76.54	77.26	78.82	71.29
LS-ViT	✓	✓	67.40	77.79	57.03	71.05	77.86	77.69	80.05	72.70
Ours	✓	✓	**71.30**	**78.23**	**60.38**	**73.24**	**78.20**	**79.58**	**80.87**	**74.54**

Table 9: **Precision ablation of quantized modules at ultra-low-bit budgets.** ImageNet-1K Top-1 (%). QK, PV, and Head input denote activation bitwidths, while the other quantizers remain at W2A2. The four row groups show the baseline, attention precision changes, head-only refits, and head refits after attention recalibration. All values are presented after 5K L-stage steps.

QK	PV	Head	ViT-S	DeiT-T
2	2	2	23.756	14.036
2	3	2	27.506	16.440
2	4	2	28.314	17.434
3	2	2	26.942	15.308
3	3	2	29.938	17.958
4	2	2	27.304	15.604
4	4	2	31.428	18.614
2	2	4	24.192	14.996
2	2	8	25.114	16.140
4	4	4	32.148	19.804
4	4	8	32.900	21.128

Table 10: **Matched weight-lattice policy.** W2A3 Top-1 (%) under a shared geometry initialization and fitting budget.

Model	AdaRound	Plain Fixed rounding	Lattice transitions	Δ (pp)
ViT-S	62.729	62.830	63.154	+0.324
ViT-B	73.523	73.280	73.869	+0.589
DeiT-T	50.451	50.662	51.026	+0.364
Swin-S	74.379	74.528	74.737	+0.209
Mean	65.271	65.325	65.697	+0.372

Table 11: **Packed encoder latency and efficiency.** End-to-end batch-eight CUDA Graph measurements on RTX 3090 (rev117 snapshot). Latencies are milliseconds per batch; larger is better for all ratios. The final row is the geometric mean across all 13 encoders.

Encoder	BF16 (ms)	W2A4 (ms)	Speedup (×)	Peak (×)	Storage (×)
DeiT-T	1.283	0.697	1.84	4.70	2.61
DeiT-S	2.452	1.299	1.89	3.34	2.54
DeiT-B	7.101	2.871	2.47	2.78	2.54
ViT-S	2.467	1.314	1.88	3.34	2.54
ViT-B	7.114	2.863	2.48	2.78	2.54
Swin-S	6.808	3.661	1.86	3.30	3.27
Swin-B	9.965	4.618	2.16	2.97	2.93
CLIP-L	54.807	21.849	2.51	2.85	3.47
SigLIP2-SO	104.164	37.769	2.76	2.65	3.00
DINOv3-B	9.615	4.163	2.31	2.98	4.33
SigLIP-B	9.566	2.798	3.42	2.44	2.72
SigLIP-SO	34.274	10.532	3.25	2.88	3.06
DINOv2-B	45.337	15.578	2.91	2.45	3.05
Geometric mean	–	–	2.39	2.99	2.93

A.6 Per-encoder deployment measurements

The systems benchmark compares BF16 and packed encoders at batch eight using CUDA Graph replay with an NVIDIA RTX 3090, CUDA Toolkit 12.9, and PyTorch 2.13. The kernel is implemented using CUTLASS. Table 11 reports measured BF16 and W2A4 latency for every encoder together with their ratios. Peak memory and storage are reported for the materialized packed endpoints. The runtime also exposes raw uint8 ingress for real-time edge-device deployment, with preprocessing parameters folded into the patch-embedding rows. The corresponding latency–memory geometry is visualized in Fig. 6.

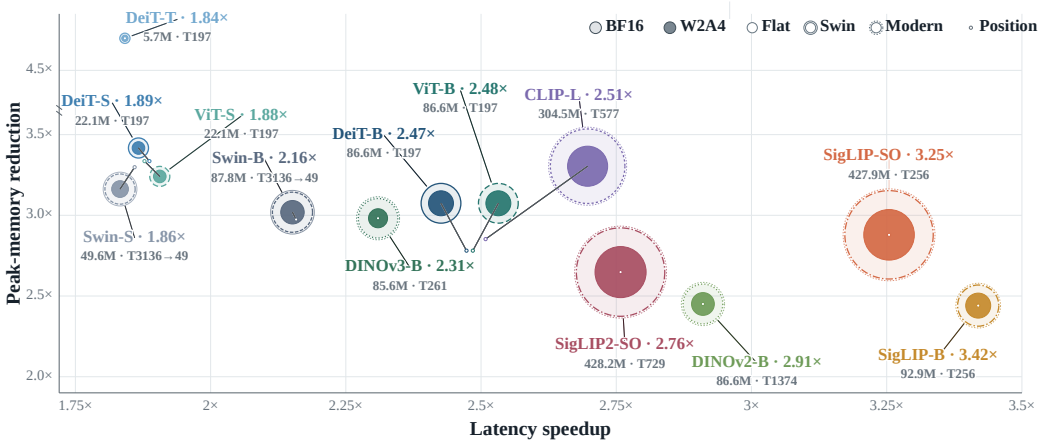
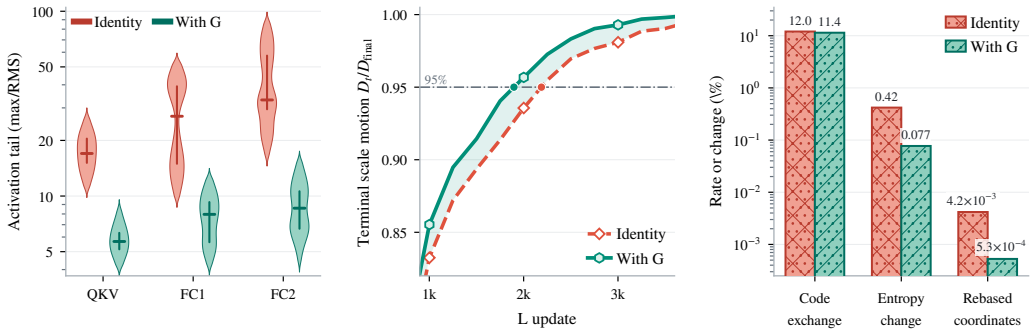


Figure 6: **Inference system memory and latency.** Image preprocessing is fused into patch embedding using raw uint8 images for full-graph execution on edge devices. Bubble size represents resident memory. Detailed latency metrics are reported in Table 11.

A.7 Geometry alignment and lattice dynamics observations

This appendix section expands the compact cross-reference in the main experiments. Geometry alignment contracts activation tails across QKV, FC1, and FC2, while the subsequent lattice stage reaches its terminal scale regime earlier. The matched traces also show that code-cell exchanges can coexist with nearly unchanged aggregate entropy and endpoint-preserving rebase rates.



(a) Layerwise tail redistribution (b) L settling demand (c) Lattice reallocation

Figure 7: **Activation geometry and lattice dynamics.** Identity and geometry-aligned coordinates are compared in two matched audits. (a) Log-domain max/RMS distributions across the 12 DeiT-T blocks; vertical segments span the interquartile range and thick horizontal marks denote medians. (b) At a representative late MLP site in ViT-S W2A3, normalized scale motion reaches its terminal regime earlier and with less total movement from the G chart, alongside a 5.46-point endpoint gain. (c) Pointwise codes exchange cells while aggregate entropy and endpoint-preserving rebase rates remain nearly unchanged.

B Mechanism Audits and Task-Approximation Analysis

This appendix records the mechanism checks that delimit the method's operating range. We first evaluate published reconstruction metrics on reachable low-bit decisions, then separate estimator structure from finite-radius extrapolation and update consistency.

B.1 Estimator-induced reconstruction rankings on reachable low-bit proposals

We instantiate the executable reconstruction forms induced by three prior second-order estimators: the average perturbation-Hessian estimator of APHQ-ViT, the dynamic Fisher approximation of FIMA-Q, and the least-squares Hessian estimator of LS-ViT. The resulting reconstruction weights were checked against the corresponding public implementations. To compare the metrics rather than their different quantizers and optimizers, we freeze one DeiT-T W2A3 endpoint and use an identical decision set: at each audited attention-output projection, every proposal changes one native integer code by one legal adjacent step while retaining its scale and zero point. Metric statistics are estimated on 64 images and then frozen; a disjoint set of 64 images measures the hard block-MSE and terminal teacher-KL changes for the same 128 proposals. No held image enters estimator fitting. Figure 2a reports the scan over Blocks 0–10; Table 12 retains three representative depths for the exact per-metric values.

The contrast is the main observation: a metric may be highly faithful to its local reconstruction target without reliably ordering finite task changes. It also does not imply that the prior metrics are uniformly uninformative. In a complementary source-equation audit across ViT-S, DeiT-S, and Swin-S at W2A3, W2A4, and W3A3, the score induced by the LS-ViT least-squares Hessian estimator attains mean held terminal-KL Spearman $-.019$, $-.090$, and $.622$ at early, middle, and late blocks, respectively. Thus task correspondence is depth-conditioned, which is the narrower claim used by our motivation. Full-model endpoint comparisons for all three prior methods appear in Table 2.

B.2 The true second-order radius of the hard residual

For $p_T(z) = \mathrm{softmax}(F_{>l}(z)/T)$, define

$$K_T(\alpha) = \mathrm{KL}(p_T(z) \,\|\, p_T(z + \alpha\delta)), \qquad Q_T(\alpha) = \frac{\alpha^2}{2}\delta^\top H_T(0)\delta. \tag{8}$$

At sufficiently small α, the exact zero-point Hessian predicts the KL value accurately; the realized hard residual occupies a substantially larger finite-radius regime. The full-block summary in Table 13 gives ranges over blocks 0–11 (each entry is the median of the per-image relative errors).

Table 12: **Estimator-induced ranking fidelity on reachable W2 decisions.** Held Spearman correlation between each induced reconstruction score and the measured change after a legal adjacent code move. High local agreement at early/middle depth verifies that the probe preserves the intended reconstruction signal; terminal-KL ranking is weak at all three depths.

Depth (block)	Metric form	Block MSE	Terminal KL
Early (0)	APHQ-ViT	.973	.107
	FIMA-Q	.875	.068
	LS-ViT	.900	.084
Middle (5)	APHQ-ViT	.966	.100
	FIMA-Q	.844	.090
	LS-ViT	.828	.069
Late (10)	APHQ-ViT	.122	.030
	FIMA-Q	.050	$-.048$
	LS-ViT	.054	.002

Table 13: Summary of exact-Hessian value prediction across all 12 blocks. The hard residual is evaluated at $\alpha = 1$; $\alpha = .001$ is a radial diagnostic in representation space rather than a quantizer parameter update.

	Block-median error range		Images above 10% error	
Temperature	$\alpha = .001$	$\alpha = 1$	$\alpha = .001$	$\alpha = 1$
$T = 1$	.078–.409%	54.56–93.32%	0–1 of 32	29–32 of 32
$T = 20$	.019–.143%	20.20–99.58%	0 of 32	22–31 of 32

At $\alpha = .001$, the T1 error ranges from 0.078% to 0.409% and the T20 error from 0.019% to 0.143%. At the actual hard residual ($\alpha = 1$), the corresponding ranges are 54.56–93.32% (T1) and 20.20–99.58% (T20). The threshold is an operational diagnostic; the mathematical radius remains dependent on the selected tolerance and block.

For blocks 0, 6, and 11 under the same current-block condition, the small-radius counts above 10% error at $\alpha = .001$ are respectively $0/32$, $0/32$, and $0/32$ for T1 and T20. At $\alpha = 1$, T1 has $32/32$ exceedances in all three blocks, while T20 has $31/32$, $29/32$, and $28/32$. The cumulative-quantized-prefix condition is a separate perturbation source and is reported separately from this current-block audit. Detailed path identities and their proof are given in Appendix B.6.

B.3 Why the activation ruler is diagonal

The current activation ruler uses a coordinate-wise, displacement-conditioned activation gain and a pseudo-Huber reconstruction potential. Its regression, bounded influence, and radial-pull properties are given in Appendix D.1. The design audit also compared coordinate-wise and pooled-activation variants on held reachable proposals. The pooled activation improved late-block ranking but did not transfer to the early block, while the coordinate-wise ruler remained competitive at the late block. Together with the weak identifiability of cross-coordinate sign/covariance relations on a small calibration set, this motivates omitting pooled cross-coordinate coupling from the final field. This internal model-selection observation is separate from the published-estimator comparison in Table 12.

The controlled ablation in Fig. 4 is consistent with two measured properties of the fitted field. First, $g_j e_j < 0$ for 46.8–49.5% of observations in three audited blocks; averaging signed products before nonnegative clipping leaves only 59.2–76.4% of coordinates supported. The displacement-conditioned gain retains action intensity without exporting these sample-dependent signs. Second, activation residuals are heavy-tailed: across four representative blocks their kurtosis is 8.2–21.8, and the largest magnitude is 18.2–44.3 times the RMS. Pseudo-Huber therefore preserves the local quadratic pull while bounding the influence of the tail. These diagnostics explain the paired endpoint gains; they are not used as an additional model-selection signal.

B.4 Coupled block errors and attainable cancellation

For a linear reader $Y = XW$, an activation chart and a weight chart with errors E_X, E_W satisfy the exact identity

$$\widehat{Y} - Y = E_X W + X E_W + E_X E_W. \tag{9}$$

For a multi-block suffix, the analogous expansion is a sum of first-order block activations plus pairwise and higher-order cross terms. In particular, for first-order block activations $a_1, \dots, a_m$ and a PSD output metric M,

$$\left\| \sum_{j=1}^{m} a_j \right\|_M^2 = \sum_{j=1}^{m} \|a_j\|_M^2 + 2 \sum_{1 \le j < k \le m} \langle a_j, a_k \rangle_M, \qquad \langle u, v \rangle_M = u^\top M v. \tag{10}$$

The cross terms can be negative or positive, exposing attainable cancellation and amplification. Joint local fitting retains these terms; the selected Transformer block granularity is an empirical deployment–conditioning choice.

B.5 Curvature non-identifiability from gradient–perturbation pairs

Let $E = [e_1, \dots, e_n] \in \mathbb{R}^{d \times n}$ collect boundary perturbations and suppose that the observed pairs satisfy $g_i = H e_i$ for a PSD local curvature matrix $H \succeq 0$. If $\operatorname{rank}(E) < d$, choose a nonzero $v \in \ker(E^\top)$. For every $\alpha \ge 0$,

$$H_\alpha = H + \alpha v v^\top \succeq 0, \qquad H_\alpha e_i = H e_i = g_i.$$

Thus all gradient–perturbation observations are identical for the continuum of curvatures $\{H_\alpha\}$, while the quadratic energy in the unseen direction changes by $\alpha (v^\top e)^2$. The pair data identify the restriction of curvature to $\operatorname{span}(E)$; full recovery requires an additional structural assumption, such as a known low-rank model or probes spanning the boundary space.

Proof. Since $v^\top e_i = 0$, $v v^\top e_i = 0$ for every observed pair. A sum of PSD matrices is PSD, proving the first claim; the second follows by substitution. For a test perturbation e, the quadratic difference is $e^\top (H_\alpha - H) e = \alpha (v^\top e)^2$. A fitted activation ruler therefore uses its chosen structure to extend sampled sensitivities to new displacements. The empirical rank of a Gauss–Newton matrix is a local coverage diagnostic at the matched reference. Under a fixed few-shot cache, its spectrum records only the task directions exposed by those images, leaving the remaining update space to the optimizer's implicit bias.

B.6 Finite KL paths and the gradient–loss mismatch

Let ℓ be a matched-suffix KL (or any C^2 scalar loss), let y be the current boundary state, and let e be a boundary displacement. Write $g_0 = \nabla \ell(y)$ and $H_t = \nabla^2 \ell(y + te)$. The exact path identities are

$$\nabla \ell(y + e) - g_0 = \int_0^1 H_t e \, \mathrm{d}t, \tag{11}$$

$$\begin{aligned} \ell(y + e) - \ell(y) &= \int_0^1 \nabla \ell(y + te)^\top e \, \mathrm{d}t \\ &= g_0^\top e + \int_0^1 (1 - t) e^\top H_t e \, \mathrm{d}t. \end{aligned} \tag{12}$$

The gradient change integrates the Hessian with uniform path weight, whereas the curvature contribution to the loss carries the triangular weight $1 - t$. A single endpoint Hessian therefore represents a local approximation to the finite path. A single endpoint gradient–perturbation pair summarizes one action along that path; additional probes resolve its variation.

If $\nabla^2 \ell$ is M-Lipschitz on the segment and $H_0 = \nabla^2 \ell(y)$, Taylor's theorem with integral remainder gives

$$\left| \ell(y + e) - \ell(y) - g_0^\top e - \tfrac{1}{2} e^\top H_0 e \right| \le \frac{M}{6} \|e\|^3. \tag{13}$$

Proof. Subtract $g_0^\top e + \frac{1}{2} e^\top H_0 e$ from the second identity in Eq. 12. The absolute remainder is at most $\int_0^1 (1-t) M t \|e\|^3 \, \mathrm{d}t = M\|e\|^3/6$. For KL, H_t is the pulled-back Fisher matrix plus suffix-dependent second-derivative terms; both vary when the prefix and the hard suffix move. Consequently, the endpoint KL remains task-aligned, while its small-sample optimization field depends on the full finite path.

For completeness, replacing the floating suffix $F^\star_{>l}$ by its quantized counterpart $F^q_{>l}$ changes the reference point. At the teacher boundary $y_l^\star$, a smooth local expansion gives

$$F^q_{>l}(y_l^\star + e) - F^\star_{>l}(y_l^\star) = b_l + J_l^q e + O(\|e\|^2), \qquad b_l = F^q_{>l}(y_l^\star) - F^\star_{>l}(y_l^\star). \tag{14}$$

The generally nonzero offset b_l introduces a linear term in a teacher-matched output discrepancy and changes its local curvature.

C Related Work

Quantization structure in vision transformers. Vision-transformer PTQ has identified precision-sensitive structures in normalization, attention probabilities, and activation distributions. FQ-ViT develops power-of-two scaling and logarithmic integer softmax to support fully quantized LayerNorm and attention operations (Lin et al., 2022). PTQ4ViT introduces twin uniform quantization for post-softmax and post-GELU activations, together with Hessian-guided scale selection (Yuan et al., 2022). RepQ-ViT uses scale reparameterization to reconcile distribution-aware calibration with simpler inference quantizers, addressing inter-channel variation and highly nonuniform attention values (Li et al., 2023). AdaLog further adapts the logarithmic base to activation distributions and uses bias reparameterization to extend logarithmic quantization to post-GELU activations (Wu et al., 2024). Beyond quantizer design, Evol-Q examines the jagged loss landscape induced by quantization-scale changes and combines evolutionary search with contrastive calibration (Frumkin et al., 2023). Together, these studies show how activation structure and quantizer parameterization shape the difficulty of low-bit ViT calibration.

Reconstruction-based PTQ with limited calibration data. Reconstruction-based PTQ uses a small calibration set to recover intermediate responses of a pretrained model. AdaRound learns weight-rounding decisions through layer-wise reconstruction, while BRECQ reconstructs entire blocks to account for cross-layer dependencies (Nagel et al., 2020; Li et al., 2021). QDrop randomly bypasses activation quantization during reconstruction, encouraging solutions that generalize across quantization perturbations (Wei et al., 2022). PD-Quant incorporates network prediction differences into quantization parameter optimization and adjusts calibration activation distributions to mitigate overfitting (Liu et al., 2023b). For large language transformers, ZeroQuant uses layer-wise knowledge distillation to recover accuracy with limited calibration cost (Yao et al., 2022).

A related line of work refines the task-sensitive metric used to weight reconstruction errors. BRECQ derives Hessian-weighted block reconstruction and uses a diagonal Fisher approximation for tractable optimization (Li et al., 2021). APHQ-ViT estimates an average perturbation Hessian from gradient responses (Wu et al., 2025b), while FIMA-Q connects reconstruction to local KL geometry through a diagonal-plus-low-rank Fisher approximation (Wu et al., 2025a). LS-ViT fits the relation $g \simeq H\Delta z$ by least squares, addressing limitations of averaging and independence assumptions in representative Hessian estimation (Hwang et al., 2026). Our work examines how the fidelity of these task-loss approximations relates to the reliability of calibration updates under finite low-bit perturbations and limited data. We develop a complementary activation-path formulation that combines calibration-set task statistics with patch-direction and next-block attention–value constraints, coupling task-informed error allocation with local representation preservation.

D Implementation Details and Supporting Derivations

After the empirical appendices, this module records the implementation details needed to reproduce the method: the activation field, its calibration protocol, visual constraints, local block fields, and stage configuration.

D.1 TASK-WEIGHTED RESTORING FIELDS: DERIVATION AND INTERPRETATION

This section expands the construction introduced in Sec. 3. The derivation has three linked steps. A suffix supplies a local task action, the boundary displacement supplies the restoring direction, and a low-capacity field fit converts the observed action into a coefficient that can be reused for the current block. The resulting field is a task-weighted reconstruction metric; it is not intended to recover a unique Hessian of the end-to-end loss.

From a suffix action to a local field. Fix a block boundary and suppress its index. For calibration sample i, let $e_i = \hat{y}_i - y_i^\star$ be the current quantization displacement and let g_i be the gradient of the suffix objective with respect to the same boundary activation. A first-order suffix model for a small boundary change v_i is $g_i^\top v_i$. If the change is regularized by a Euclidean trust region, the elementary proximal model is

$$m_i(v_i) = g_i^\top v_i + \frac{1}{2\eta}\|v_i\|_2^2, \qquad \eta > 0. \tag{15}$$

We restrict the restoring proposal to $v_i = -\eta P e_i$, where P is a shared symmetric positive semidefinite field. Substitution and completion of squares give

$$\arg\min_{P\in\mathcal{C}} \sum_i m_i(-\eta P e_i) = \arg\min_{P\in\mathcal{C}} \frac{\eta}{2}\sum_i \|g_i - P e_i\|_2^2, \tag{16}$$

up to the constant $\frac{\eta}{2}\sum_i \|g_i\|_2^2$. Thus the field fit comes from matching a local task action to a displacement-conditioned restoring proposal. The restriction to a shared P is essential: it asks which part of the activation can be reused across calibration images, instead of storing one sample-specific update per image.

The full signed regression in Eq. 16 is too expressive for a small calibration cache. Signs of g_{ij} can change with the image, the quantized suffix, and the current code cell, while the desired restoring sign is already supplied by e_{ij}. PVP therefore retains the coordinate activation magnitude and uses the displacement for the direction. This absolute diagonal ruler is a deliberate restricted model, rather than the exact solution of the unrestricted signed regression. Equivalently, for a candidate displacement z define the worst signed first-order activation over its coordinatewise uncertainty box,

$$\mathcal{R}_i(z) := \max_{|\delta_j|\le|z_j|} |g_i^\top \delta| = \sum_j |g_{ij}|\,|z_j|. \tag{17}$$

The maximization exposes $|g_{ij}|$ as a reusable coordinate sensitivity while removing the sample-specific sign from the shared coefficient.

Why displacement-conditioned weighting appears. The natural shared surrogate is the nonnegative diagonal potential

$$Q_d(z) = \frac{1}{2}\sum_j d_j z_j^2, \qquad d_j \ge 0. \tag{18}$$

Its negative gradient $-d \odot z$ points toward the teacher boundary. To calibrate its strength, contract only coordinate j of sample i by $z_i(t) = e_i - t e_{ij} u_j$. At the current point, the envelope and quadratic potential decrease at rates

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_i(z_i(t))\Big|_{t=0} &= |g_{ij}|\,|e_{ij}|, \\ -\frac{\mathrm{d}}{\mathrm{d}t}Q_d(z_i(t))\Big|_{t=0} &= d_j e_{ij}^2. \end{aligned} \tag{19}$$

Matching the aggregate rates over the cache gives, with $S_j = \sum_i e_{ij}^2$,

$$d_j = \begin{cases} \dfrac{\sum_i |g_{ij}|\,|e_{ij}|}{S_j}, & S_j > \tau_{\mathrm{sup}}, \\ 0, & S_j \le \tau_{\mathrm{sup}}, \end{cases} \qquad \tau_{\mathrm{sup}} = 10^{-12}. \tag{20}$$

The same expression is the nonnegative scalar regression

$$d_j = \arg\min_{d\ge 0} \sum_i \big(|g_{ij}| - d|e_{ij}|\big)^2 \quad \text{when } S_j > \tau_{\mathrm{sup}}. \tag{21}$$

Indeed, the unconstrained stationary point is the ratio in Eq. 20; it is nonnegative, so the nonnegative constraint leaves it unchanged. The support test is the operational meaning of insufficient displacement support: a coordinate whose accumulated squared displacement does not exceed 10^{-12} is not promoted to a shared activation coefficient.

For $e_{ij} \neq 0$, define the sample ratio $a_{ij} = |g_{ij}|/|e_{ij}|$. Then

$$d_j = \sum_{i:e_{ij}\neq 0} \omega_{ij} a_{ij}, \qquad \omega_{ij} = \frac{e_{ij}^2}{S_j}, \qquad \sum_i \omega_{ij} = 1. \tag{22}$$

The coefficient is therefore an energy-weighted average of local activation sensitivity per unit displacement. Large, reliably observed displacements receive more weight, while coordinates that are nearly unchanged contribute little. This is the sense in which the ruler is displacement-conditioned: the task signal sets correction strength only where the calibration path actually exposes a boundary direction.

Robust potential and restoring property. The quadratic potential is used only as the local calibration reference. The optimization loss uses the coordinate-separable pseudo-Huber potential

$$\rho_c(r) = c^2 \left(\sqrt{1 + (r/c)^2} - 1 \right), \qquad \mathcal{L}_{\mathrm{rw}}(e) = \frac{1}{n} \sum_i \ell_i(e_i), \qquad \ell_i(e_i) = \sum_j d_j \rho_{c_i}(e_{ij}), \tag{23}$$

where $c_i = 0.25 \max\{\mathrm{RMS}(y_i^\star), 10^{-6}\}$ and the RMS is over all non-batch token/spatial and channel coordinates. During reconstruction, d_j and c_i are detached calibration statistics. Differentiating one coordinate yields the restoring field and its tangent:

$$\begin{aligned} F_{ij}(e) &:= \frac{\partial \ell_i}{\partial e_{ij}} = \frac{d_j e_{ij}}{\sqrt{1 + (e_{ij}/c_i)^2}}, \\ \frac{\partial F_{ij}}{\partial e_{ij}} &= d_j \left[1 + (e_{ij}/c_i)^2\right]^{-3/2} \geq 0. \end{aligned} \tag{24}$$

For $|e_{ij}| \ll c_i$, $F_{ij} = d_j e_{ij} + O(e_{ij}^3/c_i^2)$, so the fitted activation acts as a local task-weighted quadratic field. For large residuals, $|F_{ij}| \leq d_j c_i$; the potential keeps the direction of the correction but limits the influence of a single boundary outlier. The field is inward in the radial sense because

$$\frac{1}{n} \sum_i \langle e_i, F_i(e_i) \rangle = \frac{1}{n} \sum_i \sum_j \frac{d_j e_{ij}^2}{\sqrt{1 + (e_{ij}/c_i)^2}} \geq 0. \tag{25}$$

Consequently $-F_i$ always has a nonpositive radial component with respect to e_i. The statement is local to the boundary field; it does not claim that a finite optimizer step monotonically improves the terminal task loss.

From boundary field to quantizer update. Let θ denote the current weight-code and activation coordinates and let $B_i = \partial \hat{y}_i / \partial \theta$ be the local activation differential. A differentiable parameter step induced by the field is $\delta\theta = -\eta n^{-1} \sum_i B_i^\top F_i$. The corresponding first-order change of the restoring objective is

$$\delta \mathcal{L}_{\mathrm{rw}} = -\eta \left\| \frac{1}{n} \sum_i B_i^\top F_i \right\|_2^2 \leq 0, \tag{26}$$

for the shared parameter update, with B_i, d_j, c_i frozen in this infinitesimal calculation. For an individual sample the local activation operator $B_i B_i^\top \succeq 0$ is positive semidefinite; shared parameters also couple activations across samples through $B_i B_k^\top$. In practice the quantizer uses relaxed or straight-through differentials and an optimizer state, so the implemented step can differ from this infinitesimal identity. The identity explains the division of labor: F_i supplies a bounded activation-weighted restoring signal, while the quantizer differential decides which weight and activation movements can realize it.

D.2 CALIBRATION PROTOCOL AND ESTIMATOR CONTRACT

The derivation in Appendix D.1 defines the field; this section records how it is instantiated during block reconstruction. For each block boundary l, a calibration pass caches boundary observations

$$(e_{l,i}, g_{l,i}), \qquad e_{l,i} = \hat{y}_{l,i} - y^{\star}_{l,i},$$

from the current quantized block state. The suffix action is evaluated in the same boundary coordinates. When both floating and already-quantized suffix states are available, their observations are combined before aggregation.

The coefficients $d_{l,j}$ are computed once from the cached pairs and then detached (stop-gradient) for the reconstruction updates of that block. The support rule is operational: with $S_{l,j} = \sum_i e^2_{l,ij}$ and $\tau_{\text{sup}} = 10^{-12}$, coordinates with $S_{l,j} \le \tau_{\text{sup}}$ are assigned $d_{l,j} = 0$; supported coordinates use the estimator in Eq. 20. Thus neither a ridge term nor an identity floor is silently introduced for an unobserved displacement direction.

For sample i, the transition scale is likewise detached and set to

$$c_{l,i} = 0.25 \max\{\text{RMS}(y^{\star}_{l,i}), 10^{-6}\},$$

where RMS reduces over every non-batch activation coordinate: spatial or token position and channel depth. All positive activation, pin, and task-loss normalizers are detached before they are used to form an update. This makes the reported mixture weights describe the actual update-time loss budget, while keeping the field fit fixed during the local optimizer trajectory.

At the end of a block update, the hard endpoint is materialized and the next block receives a fresh set of boundary observations. A final endpoint pass recomputes the activation statistics after the last chart transport. This refresh rule is important because both the suffix context and the quantizer chart change as reconstruction proceeds; the protocol never reuses a field across a changed block state.

The moving-chart effect follows from the quantizer itself. For a scalar affine quantizer $Q_{s,z}$ and a coordinate map U_α, define the transported calibration distortion $C(\alpha, \phi) = n^{-1} \sum_i \|Q_\phi(U_\alpha x_i) - U_\alpha x_i\|^2$. Then

$$q_\phi(U_\alpha x) \neq U_\alpha q_\phi(x), \qquad B(\alpha, \phi_0) = C(\alpha, \phi_0) - \min_\phi C(\alpha, \phi) \ge 0, \tag{27}$$

where C is calibration distortion after transport. Away from rounding boundaries, the fitted chart changes smoothly; with the chart held fixed, a boundary crossing changes a code and can create a finite error jump (the population objective is typically continuous but kinked when the input law has no atoms). Alternating coordinate updates with fresh chart fitting minimizes B locally, while the endpoint fit repeats calibration after final materialization. This staging makes chart estimation follow the transported distribution. Inside one fixed code cell, an interior optimum $\phi^*(\alpha)$ varies smoothly by the implicit-function theorem, and retaining the old chart incurs a nonnegative second-order excess in the transport displacement. At a rounding boundary the assignment changes and the smooth derivative is replaced by a piecewise-smooth path with a possible endpoint jump. This is why the observer is refit after geometry updates and after hard materialization.

The fine-cell optimizer is separate from the activation objective. Let P_t denote the diagonal Adam preconditioner and let g_R, g_Q be reconstruction and rounding gradients. The balancer tracks $r_t = \text{RMS}(P_t g_R)$ and $q_t = \text{RMS}(P_t g_Q)$, then moves a bounded multiplier toward a target ratio $w_t \propto r_t/(q_t + \epsilon)$. It controls the preconditioned reconstruction/rounding ratio as an optimizer factor; λ_p and λ_a allocate the separate pin-activation budget.

Endpoint-preserving rebase lemma. Let $c = b + p$ with $b \in \mathbb{Z}$ and $p \in [0, 1]$, and let $k = b + \mathbf{1}\{p \ge 1/2\}$ be the hard code, where $\mathbf{1}\{\cdot\}$ is the indicator function. The transitions $(b, 0) \mapsto (b-1, 1)$ and $(b, 1) \mapsto (b + 1, 0)$ preserve both c and k for mutable-base (unwrapped), unclipped lattice sites; they therefore preserve the deployed quantized tensor while changing the continuous chart. Optimizer moments are retained or reset according to the declared policy. In log-scale coordinates $\rho = \log s$, the chain rule gives $\partial L/\partial \rho = s\, \partial L/\partial s$, so a step in ρ has a relative-scale interpretation.

Equivalently, let the hard endpoint map be $\Pi(b, p, s) = s\,[b + \mathbf{1}\{p \ge 1/2\} - z]$ for an affine quantizer with zero point z. At $p = 0$ and $p = 1$, the two transitions in the lemma leave both c and

Π unchanged. Hence any deployment-only objective $L_{\mathrm{hard}}(\Pi)$ is exactly invariant at the instant of a rebase; only the future relaxed Jacobian and optimizer state can change. This separates an endpoint-preserving chart operation from an objective improvement and explains why hard-code churn can change while the current reconstruction scalar does not.

Suffix-recursion proposition. If $\delta y_k = J_k \delta y_{k-1} + \xi_k + r_k$ with $\|r_k\| \le M_k \|\delta y_{k-1}\|^2/2$ and $\Phi_{k\to L} = J_L \cdots J_{k+1}$, induction gives $\delta y_L = \Phi_{l\to L} e_l + \sum_{k>l} \Phi_{k\to L}\xi_k + \sum_{k>l} \Phi_{k\to L} r_k$. Here ξ_k denotes the intrinsic quantization error injected by layer k (at the matched prefix), while r_k is the nonlinear remainder; an empty product is the identity. If the endpoint loss has PSD output second-order form H_o and readout Jacobian R, the boundary Gauss–Newton metric is $G_l = \Phi_{l\to L}^\top R^\top H_o R \Phi_{l\to L} \succeq 0$; omitted terms are cross-layer covariance and nonlinear remainders. The proof is direct induction and the chain rule. If $\|J_k\| \le \gamma_k$ and $\|\xi_k + r_k\| \le b_k$, the omitted tail is bounded by $\sum_{k>l}(\prod_{m=k+1}^{L} \gamma_m) b_k$; this finite-horizon bound explains why an explicit one-hop probe remains useful alongside a longer-range suffix activation.

D.3 Rulers, pins, and local activation coverage

A sample defines the depth-indexed path $\Gamma_\theta^q(x) = (\hat{y}_0(x), \ldots, \hat{y}_L(x))$ in $\prod_{l=0}^{L} \mathcal{Y}_l$, where $\mathcal{Y}_l = \mathbb{R}^{N_l \times d_l}$ can change with depth. A single deployed parameter setting θ determines the paths of all images. Teacher and student states are compared at matched boundaries in the aligned coordinates of the current legal chart. The rope picture depicts this depth-indexed sequence and its activation constraints.

Soft pins as activation residuals. A pin is a triple $(\psi_{i,l}^q, t_{i,l}, d_i)$: an observer, a detached teacher target, and a discrepancy. For a fixed calibration batch, stack all boundary states in y_l and absorb sample and token averaging into the observer embeddings. The implemented direction and energy terms admit the residual form

$$r_i(y_l) = \psi_{i,l}^q(y_l) - t_{i,l}, \qquad L_i = \tfrac{1}{2}\|r_i(y_l)\|^2, \qquad i \in \{p, a\}. \tag{28}$$

For a nonzero patch token, use the embedding $n(\mathrm{LN}(u))$, where $n(z) = z/\|z\|$; for one attention activation $z = \mathrm{vec}([A_h V_h]_{h=1}^H)$, measured before head concatenation and the attention output projection, use $[n(z); \sqrt{2\gamma_e}\log \mathrm{rms}(z)]$. The identity $\|n(u) - n(v)\|^2/2 = 1 - \cos(u, v)$ gives Eqs. 4–5, including the energy coefficient. These identities describe the nondegenerate, unclamped region of the implemented norms. The teacher target uses the floating counterpart of each observer. In particular, a quantized attention consumer can have a nonzero residual even at the floating boundary.

The activation-weighted ruler has its own exact residual embedding,

$$[\chi_{\mathrm{rw}}(e)]_{ij} = \mathrm{sign}(e_{ij})\sqrt{2 d_j \rho_{c_i}(e_{ij})}, \qquad \mathcal{L}_{\mathrm{rw}}(e) = \tfrac{1}{2}\|\chi_{\mathrm{rw}}(e)\|^2. \tag{29}$$

The value at zero is defined by continuity; locally $[\chi_{\mathrm{rw}}(e)]_{ij} = \sqrt{d_j} e_{ij} + O(e_{ij}^3)$. Let $A_{\mathrm{rw}} = D_e \chi_{\mathrm{rw}}$ and $K_{\mathrm{rw}} = A_{\mathrm{rw}}^\top A_{\mathrm{rw}} \succeq 0$. This exact residual Gramian is distinct from the derivative of the restoring field,

$$R_{\mathrm{rw}}(e) = \mathrm{Diag}\Big(d_j [1 + (e_{ij}/c_i)^2]^{-3/2}\Big) \succeq 0. \tag{30}$$

The former describes centered residual coverage; the latter is the exact boundary Jacobian of $\nabla_e \mathcal{L}_{\mathrm{rw}}$. They agree at zero and have the same kernel, namely coordinates outside the supported diagonal field. Calibration-sample blocks and their averaging are implicit below.

The local activation operator. Choose a current reference y_l^0 and freeze the chart, consumer parameters, teacher targets, ruler statistics, and positive denominators c_p, c_a. Let $e_l^0 = y_l^0 - y_l^\star$ and $C_i = D_y \psi_{i,l}^q(y_l^0)$. For quantized consumers, C_i denotes the selected relaxed/STE activation differential; smooth expansions use ordinary derivatives of the specified activation relaxation. The weighted residual bundle and its boundary derivative are

$$r_l(y) = \begin{bmatrix} \chi_{\mathrm{rw}}(y - y_l^\star) \\ \sqrt{\lambda_p/c_p}\, r_p(y) \\ \sqrt{\lambda_a/c_a}\, r_a(y) \end{bmatrix}, \qquad \mathcal{O}_l = \begin{bmatrix} A_{\mathrm{rw}} \\ \sqrt{\lambda_p/c_p}\, C_p \\ \sqrt{\lambda_a/c_a}\, C_a \end{bmatrix}. \tag{31}$$

Writing $K_i = C_i^\top C_i$ gives the positive local field tangent

$$R_l(e_l^0) = R_{\mathrm{rw}}(e_l^0) + \frac{\lambda_p}{c_p} K_p + \frac{\lambda_a}{c_a} K_a \succeq 0, \tag{32}$$

while the centered residual bundle has Gramian $\mathcal{O}_l^\top \mathcal{O}_l = K_{\mathrm{rw}} + \lambda_p K_p/c_p + \lambda_a K_a/c_a$. Here B_l is the chosen relaxed or straight-through boundary differential with respect to the shared coordinates θ_l. It defines the affine surrogate $e_l^0 + B_l \delta\theta_l$ and the activation differential $\mathcal{T}_l = \mathcal{O}_l B_l$. A genuine smooth relaxation supplies its ordinary Taylor expansion; a straight-through rule supplies the differential used by the calibration algorithm. Hardened endpoints provide the discrete evaluation.

What the pins add. Combining the exact ruler tangent with the Gauss–Newton forms of the pin residuals gives the local parameter-space pullback

$$G_l^{\mathrm{pin}} = B_l^\top R_l(e_l^0) B_l = B_l^\top \left(R_{\mathrm{rw}}(e_l^0) + \frac{\lambda_p}{c_p} C_p^\top C_p + \frac{\lambda_a}{c_a} C_a^\top C_a \right) B_l \succeq 0. \tag{33}$$

Thus each pin adds a residual-Jacobian Gram contribution on reachable directions observed by its activation differential. This is a local Gauss–Newton statement. It does not assert that the complete Hessian of a cosine objective is PSD away from a matched activation: residual-weighted second derivatives also enter that Hessian. Shared parameters and the finite lattice determine which locally observed activations can be fitted jointly.

Proposition (reachable activation coverage). For positive weights,

$$\ker R_l = \ker R_{\mathrm{rw}}(e_l^0) \cap \ker C_p \cap \ker C_a. \tag{34}$$

If the nonzero reachable subspace $\mathcal{S}_l = \operatorname{range} B_l$ meets this kernel only at zero, then $\mu_l = \min_{v \in \mathcal{S}_l, \|v\|=1} v^\top R_l v > 0$ and

$$\tfrac{1}{2}(B_l \delta\theta_l)^\top R_l (B_l \delta\theta_l) \geq \tfrac{1}{2} \mu_l \|B_l \delta\theta_l\|^2.$$

For a smooth teacher-matched relaxation, this extends to the local activation potential with the usual higher-order remainder.

Proof. Each summand in R_l is PSD; the pseudo-Huber diagonal is strictly positive exactly on the supported coordinates. This gives the kernel intersection. The compact unit sphere of $\mathcal{S}_l$ then has a positive minimum Rayleigh quotient. Substitution of $B_l \delta\theta_l$ gives the inequality. □

A new pin enlarges reachable coverage when its derivative observes a direction in the previous reachable kernel. Positive coefficients retain the same kernel intersection and tune its conditioning. Shared coordinates couple the constraints supplied by different calibration images.

Proposition (affine pin-fitting floor). Let $r_l^0 = r_l(y_l^0)$ and fix the affine activation model $r_l^0 + \mathcal{T}_l \delta\theta_l$. With $\mathcal{T}_l^\dagger$ denoting the Moore–Penrose pseudoinverse,

$$\min_{\delta\theta_l \in \mathbb{R}^{p_l}} \tfrac{1}{2} \|r_l^0 + \mathcal{T}_l \delta\theta_l\|^2 = \tfrac{1}{2} \|(I - \mathcal{T}_l \mathcal{T}_l^\dagger) r_l^0\|^2. \tag{35}$$

Restricting the displacement to an admissible local cell or trust region can raise this optimum.

Proof. Decompose r_l^0 into its orthogonal projections onto $\operatorname{range} \mathcal{T}_l$ and its orthogonal complement. The choice $\delta\theta_l = -\mathcal{T}_l^\dagger r_l^0$ cancels the first component; the second component is unchanged by every feasible affine activation. □

This exact statement concerns the affine surrogate. For a smooth residual model with nonzero r_l^0, the true energy differs from the affine energy by $O(\|r_l^0\| \|\delta\theta_l\|^2 + \|\delta\theta_l\|^3)$ locally. Its Hessian contains $\mathcal{T}_l^\top \mathcal{T}_l$ and the residual-weighted second derivatives of the composed activation residual. The PSD Gramian describes directional coverage; the affine residual retains the current teacher–student offset.

D.4 Consumer state and pin geometry

Write $Z_r(y)$ for the next-block attention–value aggregate under fixed consumer state $r \in \{\mathrm{fp}, q\}$. Its teacher target is $Z_\star(y_l^\star)$. Using $e_l = \hat{y}_l - y_l^\star$, a smooth activation relaxation gives

$$Z_r(y_l^\star + e_l) - Z_\star(y_l^\star) = b_{a,r} + J_{a,r} e_l + O(\|e_l\|^2), \qquad b_{a,r} = Z_r(y_l^\star) - Z_\star(y_l^\star). \tag{36}$$

A shared floating consumer gives $b_{a,\mathrm{fp}} = 0$; a quantized consumer includes its own baseline displacement. For hard quantization, the selected relaxation/STE differential defines the affine search model. For the observed student and teacher attention tensors, the exact identity is

$$Z_q - Z_\star = (A_q - A_\star)V_\star + A_\star(V_q - V_\star) + (A_q - A_\star)(V_q - V_\star). \tag{37}$$

Here q marks the student activation, with consumer precision fixed by the comparison. The three terms identify route, value-source, and interaction changes. Observing AV measures their delivered sum and can use the attention output boundary of a fused operator.

For a patch with nonzero centered norm, let $C = I - \mathbf{1}\mathbf{1}^\top/d$, $u = Cy^\star$, and $n = u/\|u\|$. The centered, normalized direction has derivative $(C - nn^\top)/\|u\|$, giving

$$1 - \cos(\mathrm{LN}(y^\star + e), \mathrm{LN}(y^\star)) = \frac{e^\top(C - nn^\top)e}{2\|u\|^2} + O(\|e\|^3).$$

This statement uses affine-free channel normalization and the nondegenerate norm region; common shift and positive rescaling preserve its direction.

D.5 Exact scale property of detached normalization

The activation coefficients have a simple invariance that is useful for cross-layer comparison. For $a > 0$ with both $L(\theta) > \epsilon$ and $aL(\theta) > \epsilon$, define $\widetilde{L} = L/\operatorname{sg}(L)$. Then

$$\nabla_\theta \left[\frac{aL(\theta)}{\operatorname{sg}(aL(\theta))}\right] = \frac{\nabla_\theta L(\theta)}{L(\theta)} = \nabla_\theta \log L(\theta). \tag{38}$$

Thus multiplying an activation loss by a positive, parameter-independent unit conversion leaves its backward field unchanged above the floor. Below the floor the field is $a\nabla L/\epsilon$, so the invariance is intentionally broken by the numerical safeguard. Equation equation 38 also shows why λ_p and λ_a define a normalized direct-loss budget, while the actual optimizer-step share still depends on gradient angles and preconditioning.

D.6 Local fields and finite-resolution approximation

The target of path pinning is an activation-weighted approximation,

$$\min_{F^q \in \mathcal{F}_\mathbf{b}} \mathbb{E}_{x\sim\mathcal{D}} \sum_l d_l\big(\Psi_l^q(\hat{y}_l(x)), \Psi_l^\star(y_l^\star(x))\big), \tag{39}$$

where $\mathcal{F}_\mathbf{b}$ fixes topology and storage precision, and Ψ_l specifies the chosen activations. Coordinate discovery allocates the chart; local fields fit its codes and scales using cached boundaries and selected activation horizons.

An indexed family of local fields. At calibration state k, define the surrogate field for block l by

$$v_{l,k} = -P_{l,k}\left[\widetilde{\nabla}_{\theta_l}\mathcal{L}_{\mathrm{rw,l,k}} + \sum_{i\in\{p,a\}} \frac{\lambda_i \widetilde{\nabla}_{\theta_l} L_{i,l,k}}{\max\{L_{i,l,k}, \epsilon\}}\right], \tag{40}$$

where $P_{l,k} \succeq 0$ is a frozen preconditioner and $\widetilde{\nabla}$ denotes the chosen backward differential. The block field is the indexed family $\{v_{l,k}\}_l$; each member lives in its block's coordinate space. The schedule activates these fields in their respective spaces and refreshes their activation statistics. The stateful optimizer additionally carries momentum, scale, and rebase state.

If the losses are differentiable with exact backward derivatives and all teacher and consumer statistics are fixed, the unpreconditioned field is the negative gradient of

$$\Phi_l(\theta_l) = \mathcal{L}_{\mathrm{rw,l}}(\theta_l) + \sum_i \lambda_i \rho_\epsilon(L_{i,l}(\theta_l)), \qquad \rho_\epsilon(u) = \begin{cases} u/\epsilon, & u \le \epsilon, \\ 1 + \log(u/\epsilon), & u > \epsilon. \end{cases} \tag{41}$$

Indeed $\rho'_\epsilon(u) = 1/\max\{u, \epsilon\}$, and $\nabla\Phi_l^\top v_{l,k} = -\nabla\Phi_l^\top P_{l,k}\nabla\Phi_l \le 0$. This is an instantaneous descent property of the stated smooth, frozen model. Straight-through derivatives, hard-code transitions, and statistic refreshes define the implemented sequence of local surrogate fields.

The finite code and chart budget imposes a nonzero representation floor in general: the best achievable activation fidelity is the infimum of the teacher–student discrepancy over the legal deployed family. Task-equivalent changes can nevertheless preserve downstream decisions despite positive activation error, as described by Eq. 46. The G–C–L stages therefore enlarge and refine a restricted endpoint family; strict improvement remains an empirical question of which legal endpoint the optimizer reaches.

E G–C–L Pipeline and Configuration

This appendix isolates the operational handoff between Gauge Alignment (G), granularity calibration (C), and Lattice Pinning (L), together with the configuration used in the experiments.

E.1 G–C–L quantization pipeline

The following procedure specifies how the three stages are connected in an actual block reconstruction. We first run the floating-point teacher and the current quantized block on the calibration images and cache their boundary activations. The G stage then freezes the weight codes, normalization parameters, and all downstream blocks. It updates only the activation rotation and positive scale parameters. After each geometry update, the activation observer is refit in the new coordinates and the block endpoint is rematerialized; the restoring field and the two local pins are evaluated in these refreshed coordinates. This alternating fit prevents the activation scale from describing the old coordinate system.

Once the G endpoint is selected, stage C converts its provisional activation granularity to the granularity used by L. C refits the observer and quantization cuts in that final layout, rematerializes the endpoint, and freezes the granularity and scales. The resulting chart is passed to L, which initializes soft weight codes and the activation scale and optimizes only these code and scale variables with the same pinned reconstruction objective. A code is allowed to cross a neighboring cell only after the accumulated preconditioned gradient supports the move. The endpoint-preserving rebase rewrites the continuous code as an equivalent base plus fractional part, so the deployed integer tensor is unchanged at the instant of the transition while the optimizer receives a new local coordinate. Field statistics are refreshed at the declared checkpoints, and the final hard-code materialization is followed by one fresh endpoint pass. The resulting boundary activations, observer state, and legal code/scale states are then passed to the next block. Thus G changes the activation quantization surface, C fixes its deployment granularity, and L uses the remaining finite weight/scale freedom without reopening either choice.

The stage handoff can be written compactly as

$$(u_0, v_0) \xrightarrow{G} (u^\star, v_0(u^\star)) \xrightarrow{C} (u^\star, \bar{v}_0, \bar{a}) \xrightarrow{L} (u^\star, v^\star), \qquad v^\star \in \mathcal{C}(u^\star). \tag{42}$$

Here u contains activation geometry, v contains weight codes and activation scale, and $(\bar{v}_0, \bar{a})$ denotes the fixed L-stage granularity and its refitted observer state. This separates the continuous coordinate search from the final discrete code search.

E.2 Pipeline configuration

Table 14 records the G–C–L configuration used in the experiments described in Sec. 4. Model and bit-cell changes do not alter the stage roles or parameterization described in the main text.

Table 14. Representative configuration for Gauge Alignment (G), granularity calibration (C), and Lattice Pinning (L).

Stage	Component	Configuration snapshot
Shared	Calibration context	uint8-folded image input; ImageNet calibration; patch embedding kept at 8-bit; Q/K and value paths quantized at the target bit setting; per-head activation granularity; channel-wise value direction; fused STE linear and attention; per-block CUDA-graph compilation and replay.
G	Geometry search	Scalar activation granularity (not aligned with L); role-aware arms (QKV/FC1: local gradient-guided search, FC2: Hadamard-only); 3,000 updates; batch size 64; 4,096 calibration rows; 64 images per fold.
G	Gauge parameterization	Blockwise full-layout orthogonal search initialized at identity; Cayley optimizer with learning rate 5×10^{-3}.
C	Granularity bridge	Refit the observer and quantization cuts from the provisional G granularity to the fixed L-stage granularity; rematerialize the endpoint and freeze the resulting granularity and weight scales before L.
L	Lattice optimization	5,000 updates on the matched B96 clock; batch size 96; 4,096 optimization samples; FP32 trainable parameters and gradients; BF16 fixed modules; soft-code/rebase with log-scale optimization.
L	Adam policy	Weight codes use Adam soft-code with no scheduler, maximum code step 0.01, gradient-balance target ratio 0.5, and rounding-weight EMA factors $0.95/0.9$. Activation scales use Adam in log space with cosine decay, $\beta_1 = 0.85$, $\beta_2 = 0.995$, scale factor 14, maximum log-step 0.02, and $\epsilon = 10^{-8}$; weight decay is zero.
L	Fused execution	Triton kernels fuse STE linear/attention paths and loss reduction inside the compiled CUDA-graph loop; QDrop is executed in the same optimized path which is fused in fake quantizers.
L	Activation field	Logit-KL suffix activation with raw guided observer; pseudo-Huber reconstruction field; patch and AV cosine weight 0.25; detached direct-loss normalization.

F Information Visible to Few-Data Reconstruction

This appendix first characterizes the task directions visible to endpoint distillation and then the finite representation budget imposed by low-bit codes. Entropy and mutual information use base-two logarithms; KL divergences and calibration losses use natural logarithms.

F.1 What endpoint distillation observes

Proposition (local KL visibility). At a matched reference, let a boundary perturbation $e \in \mathbb{R}^d$ change logits by $z(e) = z(0) + Je + O(\|e\|^2)$ and let $p = \mathrm{softmax}(z(0)/\tau)$, with $\tau > 0$. Then

$$\begin{aligned} \mathrm{KL}\left(p \| \,\mathrm{softmax}(z(e)/\tau)\right) &= \tfrac{1}{2} e^\top G_{\mathrm{end}} e + O(\|e\|^3), \\ G_{\mathrm{end}} &= \tau^{-2} J^\top F_p J, \qquad F_p = \mathrm{Diag}(p) - pp^\top. \end{aligned} \tag{43}$$

For K classes, $\operatorname{rank} G_{\mathrm{end}} \le \min\{d, K-1, \operatorname{rank} J\}$. Also $\operatorname{tr} F_p = 1 - \|p\|_2^2$, so concentrated predictions have lower total logit-space curvature at a fixed temperature.

Proof. The gradient of KL with respect to student logits vanishes at the matched reference. Its Hessian is $\tau^{-2} F_p$. The chain rule gives the quadratic form, while $F_p \succeq 0$ and $F_p \mathbf{1} = 0$ give the rank bound. Taking the trace yields the final identity. □

For n calibration images and p_θ shared trainable coordinates, let J_i^θ map a coordinate perturbation to the logits of image i. The empirical Gauss–Newton matrix obeys

$$G_{\mathcal{C}} = \frac{1}{n\tau^2} \sum_{i=1}^{n} (J_i^\theta)^\top F_{p_i} J_i^\theta, \qquad \operatorname{rank} G_{\mathcal{C}} \le \min\{p_\theta, n(K-1)\}. \tag{44}$$

This is an upper bound on local supervision rank; the measured spectrum determines coverage and conditioning. A token-wise language objective sums over additional readout positions, whose gradients can be correlated. In a CLS-readout vision model, final non-CLS tokens lie in the exact readout

nullspace when the head depends solely on CLS. At earlier boundaries, attention can transport patch information into CLS, and its Jacobian determines which spatial directions become visible.

The teacher-matched expansion above supplies the clean visibility argument. For a quantized suffix, the baseline offset in Eq. 14 can introduce a linear costate and different local curvature.

F.2 Code budget and task-relevant collisions

Proposition (finite alphabet and retained information). Fix a deterministic quantizer and its calibrated parameters. Let $Z = q(X)$ contain m codes of b bits each, and let S be any per-input side information needed by the decoder, such as dynamic scales. Assume (Z, S) is a deterministic function of X and includes every decoder-visible per-input bypass at the considered cut. If S has at most 2^r representable states, then

$$H(Z, S) \le mb + r, \qquad I(T; Z, S) \le \min\{I(T; X), mb + r\}. \tag{45}$$

Here T denotes a task variable. Moreover,

$$I(T; X) - I(T; Z, S) = I(T; X \mid Z, S). \tag{46}$$

Thus strict information loss occurs precisely when a quantization cell retains task-relevant variation: $I(T; X \mid Z, S) > 0$. Here a cell means a fiber of the joint encoding (Z, S). A cell that merges task-equivalent inputs preserves their task information.

Proof. The joint alphabet has cardinality at most 2^{mb+r}, giving the entropy bound. Determinism gives the Markov chain $T \to X \to (Z, S)$ and $I(T; Z, S \mid X) = 0$. Expand $I(T; X, Z, S)$ in its two orders to obtain equation 46, then apply nonnegativity of conditional mutual information and $I(T; Z, S) \le H(Z, S)$. □

The proposition describes the quantized representation at a specified boundary. Applying it to an entire residual network requires a cut that includes every bypass state and its precision. Fixed learned weights and scales are conditioned on; dynamic per-input metadata contributes to r. Comparing b with a higher stored precision p gives alphabet budgets $mb + r$ and $mp + r$ at matched dimension and metadata. Their realized entropy and task information follow the input distribution and cell assignments. Weight precision describes a different budget: with fixed architecture, P weight codes of b_w bits and r_w bits of stored auxiliary parameters index at most $2^{Pb_w + r_w}$ deployed parameter settings. The realized function family can be smaller because different settings can agree as functions. Activation-rate bounds concern per-input information, whereas weight-code counts concern the available model family.

Task-loss interpretation. For a finite task alphabet and unrestricted probabilistic predictors, the optimal natural-log prediction loss from a representation $W = (Z, S)$ is $(\ln 2)H(T \mid W)$. Consequently its excess over prediction from X is

$$(\ln 2)\big[H(T \mid W) - H(T \mid X)\big] = (\ln 2) I(T; X \mid W).$$

The conditional task distribution attains the optimum by the Gibbs inequality. Thus task-relevant collisions have a direct Bayes log-loss interpretation.