

# From Distribution to Decision: A Diagnostic for Composable PTQ Primitives

Dorsa Zeinali

Northeastern University  
zeinali.d@northeastern.edu

Yun Fu

Northeastern University  
y.fu@northeastern.edu

## Abstract

Post-training quantization (PTQ) methods are typically architecture-specific. Yet two distributional pathologies they target, dense cores with sparse outliers and input-conditioned activation clusters, recur across model families. We argue that for these pathologies the unit of selection should be the tensor distribution, not the architecture. We propose two composable primitives, dispatched from one calibration pass by a two-statistic diagnostic: Dual-Band Affine Folding (DBAF) where a per-tensor outlier gate fires, and Prompt-Conditioned Scale Anchoring (PCSA) where a per-prompt compactness ratio does. The diagnostic is a selection heuristic, not a calibrated predictor. Four prediction-then-outcome checks are consistent with the dispatch: a rotation cliff on a six-host LLaMA-3-8B W4A4 matrix, cross-family transfer to Qwen-2.5-7B, per-site gating on the KV-cache and SAM-B (+5pp NIAH at 8k, +4.8 mAP), and a negative SwinIR- $\times 3$  prediction at training-free W4A16. Both primitives are training-free and calibrate in seconds. The gate’s boundary is learnable, not merely hand-set: a logistic dispatcher trained with each model family held out reproduces it at mean AUC 0.962. Calibrated once on LLaMA-3-8B, the same thresholds dispatch across seven model families in six domains; off distribution the gate errs toward skipping.

## 1 Introduction

Post-training quantization (PTQ) research is typically organized by model family. SR CNNs (Wang et al., 2025a), SR transformers (Liu et al., 2024a; Zeinali et al., 2026), and large language models (Xiao et al., 2023; Tseng et al., 2024; Sun et al., 2025; Ashkboos et al., 2024; Liu et al., 2025b; Liang et al., 2025) each get their own specialized machinery. Yet the same distributional pathologies recur across all of them: dense cores with sparse outliers, and input-conditioned activation

shifts. Similar mechanisms address them in each family, including per-region scaling, magnitude migration, and orthogonal rotations. Table 1 maps representative prior work onto these two axes. Each row is well established within its family, but the techniques rarely transfer between architectures.

For these pathologies the unit of selection should therefore be the tensor distribution, not the architecture that produced it: low-bit error is dominated by how weights and activations are distributed, not by what task they encode. We do not claim this holds for PTQ in general — it is a claim about two specific, recurring pathologies, and we report where it breaks down (Limitations; Appendix I). We design two composable primitives, one per pathology: Dual-Band Affine Folding (DBAF) for outliers and Prompt-Conditioned Scale Anchoring (PCSA) for input-conditioned clusters. A two-statistic diagnostic indicates whether each primitive should apply.

**Observation.** The diagnostic reads off both features in one calibration forward pass: a per-tensor outlier gate and a per-prompt compactness ratio. At each (model, host quantizer, tensor) it identifies which primitives the gate selects, where they saturate, and where they should be skipped. The outlier gate also predicts saturation in other outlier-handling methods. On rotation hosts the gate is disabled because rotation has already absorbed the outlier mass, so further outlier-targeting machinery contributes little. This is the rotation cliff of §4.2: a discrete boundary between hosts where DBAF helps and hosts where it saturates.

### Contributions.

- Two composable primitives (DBAF, PCSA) and a two-statistic diagnostic that dispatches them from one calibration pass (§3.1). The diagnostic is a selection heuristic, not a calibrated predictor.

- A rotation cliff on a six-host LLaMA-3-8B (Grattafiori et al., 2024) W4A4 matrix (§4.2). Non-rotation hosts cross a discrete boundary where DBAF helps (PPL drops  $\sim 60\times$ ); on rotation hosts the outlier mass is already absorbed, so DBAF saturates within noise. The outlier gate alone predicts which side each host falls on. TesseraQ (Li and Panda, 2024) trains without rotation and lands on the non-rotation side, as predicted ( $-82\%$  PPL).
- The compactness gate as a per-site, not per-model, signal (§4.4). One model selects PCSA at one tensor site and skips it at another:  $+5\text{pp}$  NIAH at 8k on the LLaMA-3-8B KV-cache, self-disable on LLaMA-3-8B Q ( $c=0.79$ ), and  $+4.8$  mAP on SAM-B Q ( $c=0.19$ ,  $\approx 0.4$  on real COCO) when the same test is applied outside NLP.
- A negative prediction computed from gate counts alone, with no task-level signal (§4.5). SwinIR (Liang et al., 2021)- $\times 3$  regresses in the direction the gate indicates ( $-0.31$  dB Set5 vs.  $-0.32$  dB expected), while  $\times 2$  and  $\times 4$  improve as the diagnostic anticipates.
- The gate boundary is learnable and architecture-transferable, not merely hand-set. A logistic dispatcher over the same three statistics, trained with each model family fully held out, recovers our hand-drawn region at mean AUC 0.962 (0.93–0.99 per family, 1010 layers; Appendix E), and the thresholds survive  $\pm 20\%$  perturbation with 93.9% of decisions unchanged. Calibrated once on LLaMA-3-8B, the same statistics dispatch across seven model families in six domains (Appendix I).
- A characterization of where the dispatch fails. Outside the calibrated regime the gate is conservative rather than wrong: on DiT-XL (Peebles and Xie, 2023) it declines to fire on 30% of tensors that the primitive would have improved (Appendix I), and the per-layer *gain* regression does not transfer across architectures even though the binary gate does (Appendix E).

Code, configurations, and the prediction file are available at <https://github.com/dorsa-zeinali/distributional-ptq>.

**Scope.** The diagnostic is calibrated on the LLM family and held fixed everywhere else. Its opera-

tional scope is W4 weights with W4A4 or W4A16 activations, single-turn prompts for the compactness measurement, and the two distributional patterns of Table 1; on distributions outside this scope the gate self-disables, which Appendix I shows is conservative rather than always correct. The binary gate boundary is validated on held-out architectures (Appendix E); calibrated probabilities, transfer of the per-layer *gain* regression, and scale generalization beyond 8B remain out of scope (Limitations).

## 2 Related Work

Table 1 maps prior PTQ work onto two recurring distributional pathologies: dense regions with sparse outliers and input-conditioned distribution shifts. The same pathology is solved separately in each model family; no existing technique transfers unmodified across families.

**Dense regions with sparse outliers.** The dense-core-with-outliers pattern appears across vision and language models, but solutions are bespoke. Dettmers et al. (2022) first characterized the emergent outlier phenomenon at scale (Sun et al., 2024). For SR CNNs, Wang et al. (2025a) use dual-region scales; Hong et al. (2022) characterize channel-wise differences. For SR transformers, Zeinali et al. (2026) apply Hadamard rotations to SwinIR; Qin et al. (2023) is a Swin-family baseline. For LLMs, SmoothQuant (Xiao et al., 2023) and Outlier Suppression+ (Wei et al., 2023) migrate magnitudes via per-channel shift-and-scale; rotation-based methods (Chee et al., 2023; Tseng et al., 2024; Sun et al., 2025; Ashkboos et al., 2024; Liu et al., 2025b; Lin et al., 2024a) suppress outliers via orthogonal transforms; Liang et al. (2025) target extreme tails. For SAM, AHCPTQ (Zhang et al., 2025), PTQ4SAM (Lv et al., 2024), PQ-SAM (Liu et al., 2024b), and Mix-QSAM (Ranjan and Savakis, 2025) address bimodality, asymmetric transforms, and mixed precision. Distribution-aware allocation (Zhao et al., 2021) and outlier channel splitting (Zhao et al., 2019) provide complementary primitives. DBAF (§3) addresses this pattern with a single primitive that composes on any host quantizer, while improving performance.

**Input-conditioned distribution shifts.** LLM activation statistics vary with input: Liu et al. (2024c) document up to 70% performance variation from calibration-data choice; Chang et al. (2025) and Williams and Aletras (2024) study correlated er-

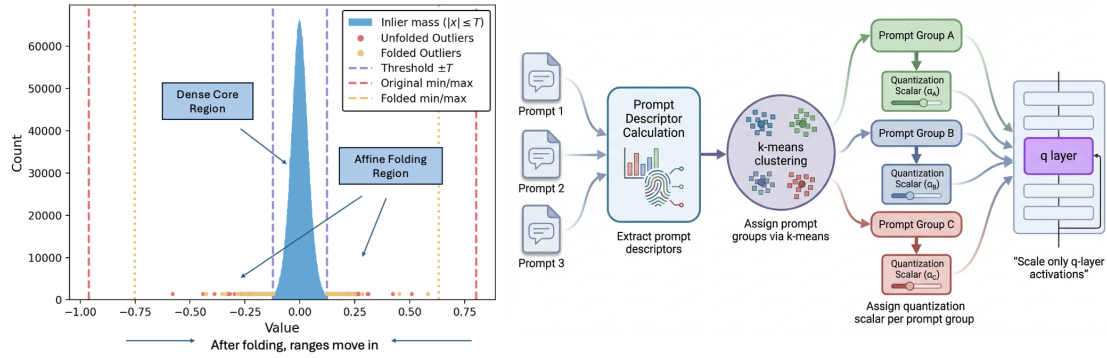


Figure 1: Overview of the distribution-centric perspective. **Left:** DBAF compresses dense-core-with-outliers distributions via a piecewise-affine fold; the fold preserves ordering and is invertible up to quantization error. **Right:** PCSA clusters input descriptors and routes each input to a per-cluster activation or  $K/V$ -cache scale. Both primitives target recurring distributions rather than architecture-specific features, and admit training-free realizations (closed-form  $\alpha^*$  for DBAF;  $k$ -means anchors for PCSA-tf). AI partially generated this figure.

| Family          | Repr.           | Problematic distribution | Quantization technique                          |
|-----------------|-----------------|--------------------------|-------------------------------------------------|
| SR CNNs         | RDN, EDSR       | Dense + outliers         | Dual-region (Wang et al., 2025a)                |
| SR Transformers | SwinIR, MambaIR | Dense + outliers         | Hadamard (Zeinali et al., 2026)                 |
| LLMs            | LLaMA, Qwen     | Dense + outliers         | Rotation (Tseng et al., 2024; Sun et al., 2025) |
| LLMs            | LLaMA           | Input-conditioned        | (Liu et al., 2024c; Chang et al., 2025)         |
| SAM             | SAM-B/L/H       | Dense + outliers         | DBAF (ours)                                     |
| SAM             | SAM-B/L/H       | Input-conditioned        | PCSA (ours)                                     |

Table 1: Quantization approaches across model families. The recurring distributional pathologies (dense core + outliers, input-conditioned shifts) appear across diverse architectures, motivating distribution-centric primitives rather than architecture-specific quantizers.

ror patterns. OmniQuant (Shao et al., 2024) absorbs variation via learnable clip parameters; PrefixQuant (Chen et al., 2026) isolates attention-sink contributions. SAM (Kirillov et al., 2023; Wang et al., 2025b) exhibits analogous per-image shifts, but existing SAM quantizers treat them within a single family. PCSA (§3) generalizes per-prompt anchor-routing as an architecture-neutral primitive.

**Composability and training-free calibration.** Strong modern quantizers require gradient-based calibration: learnable rotations (Liu et al., 2025b; Sun et al., 2025; Hu et al., 2025) or learned clip parameters (Shao et al., 2024). Our primitives compose on top of any such host without retraining: DBAF’s closed-form scale and PCSA’s  $k$ -means anchor variant reduce calibration cost on training-free hosts (RTN, GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024b)) to seconds. Patterns outside our scope (post-ReLU asymmetric, bimodal post-softmax, heavy-tailed unimodal) cause DBAF’s gate to self-disable; AHCPTQ’s bimodal handler (Zhang et al., 2025) hints at the form future distribution-specific primitives could take, which

the composable API of §3 should accommodate.

### 3 Method

#### 3.1 The Diagnostic

The diagnostic identifies the two distributional pathologies the primitives can address. Across SAM-B/L/H, LLaMA-2/3, Qwen-2.5 (Qwen Team, 2024), and SwinIR, weight and activation tensors share the same signature: a dense Gaussian-like core near zero with sparse extreme values past  $3\sigma$  (Figure 2). These outliers inflate the quantization scale, reduce effective precision in the dense core, and dominate low-bit task error. An outlier gate detects this pattern.

Activation distributions also vary across input prompts at some tensor sites but not others. Per-prompt scales recover information only at the sites where prompts induce tight clusters. A compactness ratio measures whether such clustering is present at a given site. Together, the two statistics determine the dispatch.

The per-tensor outlier-gate indicator  $g(M, T)$  is one if tensor  $T$  under model  $M$  has skewness at

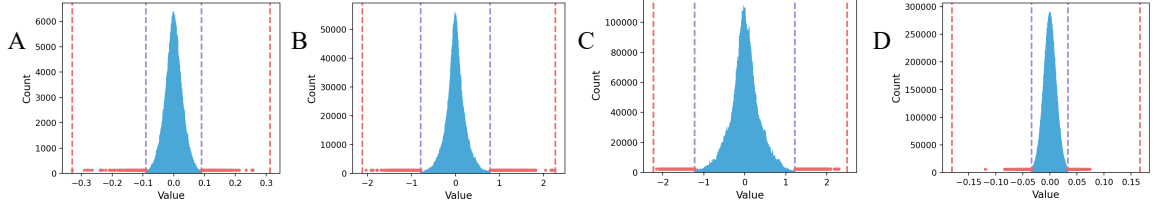


Figure 2: Recurring distributional pattern across architectures. Each panel shows a representative weight or activation distribution with a dense Gaussian-like core (blue) and sparse extreme outliers (red). Dashed purple lines:  $\pm 3\sigma$  folding threshold; dashed red lines: original min / max. (A) SAM-B activations (H-DETR (Jia et al., 2023) prompts) (B) SAM-B activations (YOLOX (Ge et al., 2021) prompts) (C) SwinIR  $\times 2$  activation (D) LLaMA-3-8B activation. The same dense-core-with-outliers distributional signature recurs across architecture families, scales, and downstream detectors, motivating transfer.

most 0.7, kurtosis in  $[3, 30]$ , and fraction above three sigma in  $[10^{-4}, 2 \times 10^{-2}]$ , and zero otherwise. Because DBAF folds both paths, the rule is evaluated twice per layer: on the weight tensor offline and on the input activation at runtime. The  $3\sigma$  boundary follows the 68–95–99.7 rule for an approximately Gaussian core; the remaining bounds are calibrated on LLaMA-3-8B once and held fixed for every subsequent model. Across the eight configurations of Table 25 the weight gate passes almost everywhere (94–100%) while the activation gate is selective (36.5% of 1010 layers) and carries the dispatch; transfer outside the dense-core regime is not claimed (Limitations).

The per-site compactness ratio  $c(M, s)$  is computed at a specific PCSA application site  $s$  of model  $M$  (a particular projection, attention head set, or KV-cache slice). It is the mean within-cluster distance of  $K=4$   $K$ -means clusters fit on  $\ell_2$ -normalized mean-pooled hidden states at that site, normalized by the mean distance under an isotropic baseline;  $c=1$  means no structure and lower values indicate tighter clustering. A small  $c(M, s)$  identifies the regime where per-cluster scale routing has value. The same model can have low  $c$  at one site (e.g. KV-cache at long context) and high  $c$  at another (self-attention  $q$ -projection on short prompts), so PCSA dispatches per-site rather than per-model.

The diagnostic assigns each (primitive, model-tensor-host) triple to one of four outcomes, pairing the gate’s decision with whether the primitive would have helped. A primitive **fires** when the gate triggers and performance improves. It **saturates** when the gate triggers but the host quantizer has already addressed the pathology (e.g. rotation hosts for DBAF), so there is no marginal gain. It **fails the gate** when the statistic rejects a tensor the primitive would not have helped. It is a **false negative**

when the statistic rejects one it would have helped. The fourth outcome follows by construction: the skew and kurtosis bounds test whether a tensor *is* the dense-core pattern, while the fold’s benefit depends only on  $T$ ,  $M$  and  $p_{\text{out}}$  (Eq. 2), so the shape tests are sufficient but not necessary. Only its cost varies: inside the calibrated family the forgone gain is marginal (0.4–4.6% further MSE reduction), while off-distribution it dominates (Appendix I).

The dispatch rule follows directly. DBAF fires on tensor  $T$  iff  $g(M, T)=1$ . PCSA fires at site  $s$  iff  $c(M, s) \leq 0.4$ . The compactness threshold separates SAM-B mask-decoder  $q$ -projection ( $c \approx 0.19$ , fires) from LLaMA-3-8B standard  $q$ -projection ( $c \approx 0.79$ , skipped); both endpoints are reported in §4.4. Elsewhere, both primitives are no-ops and the host quantizer runs alone.

The compactness ratio measures cluster structure, not raw per-input variation: LLM activations vary more per input than SAM in absolute terms (max-abs CV 0.112 vs. 0.037, Appendix M) yet sit at  $c \approx 0.79$  against SAM-B’s 0.19. The gate fires where clusters tighten below an isotropic baseline, not where per-input magnitude is large. This is why per-channel scaling (e.g. SmoothQuant) does not transfer to the sites where PCSA helps — it targets magnitude, not cluster structure.

### 3.2 Dual-Band Affine Folding

DBAF is applied elementwise to weight tensors (offline) and activation tensors (at runtime, with unfolding after dequantization). With  $T=3\sigma$ , the fold is identity inside the bulk and a sign-preserving affine compression outside. Letting  $\sigma_i = \text{sgn}(x_i)$ :

$$\tilde{x}_i = \begin{cases} \sigma_i [T + \alpha(|x_i| - T)] & |x_i| > T \\ x_i & |x_i| \leq T \end{cases} \quad (1)$$

where  $\alpha \in (0, 1)$  inversely controls outlier compression and preserves outlier ordering, a property we show is essential in §4.6. Under a uniform quantization-noise model the MSE proxy  $\tilde{\mathcal{L}}(\alpha) = (T + \alpha(M - T))^2(p_{\text{in}} + p_{\text{out}}/\alpha^2)$  (with  $M = \text{percentile}_{0.999}(|x|)$  and  $p_{\text{out}} = \Pr(|x| > T)$ ) is strictly convex on  $(0, \infty)$  and minimized at

$$\alpha^* = \sqrt[3]{T p_{\text{out}} / ((M - T) p_{\text{in}})} \quad (2)$$

(MSE-proxy derivation and proof in Appendix A). The closed form is evaluated separately on the two tensor paths DBAF folds. On LLaMA-3-8B (Qwen-2.5-7B) activations it gives  $\alpha_A^* \approx 0.20$  (0.18), below the empirical operating point  $\alpha = 0.25$  (Table 7); on the weight path it gives  $\alpha_W^* \approx 0.28$  on both models, slightly above the same operating point. Across five architectures spanning language, vision, speech and diffusion the weight-path closed form lands within 26% of the sweep-selected  $\alpha$  (Appendix I), but selecting at  $\alpha^*$  directly is worse than sweeping on the task metric.

We therefore treat  $\alpha^*$  as a seed for the reconstruction sweep rather than a substitute for it. Over-aggressive folding at very low  $\alpha$  diverges, as the inverse map amplifies unfold noise on outlier positions by  $\alpha^{-1}$ ; a sensitivity-weighted account of the activation-path residual is given in the Limitations section. Per-tensor  $\alpha^*$  values for both paths are released in `results/astar_llm_Meta-Llama-3-8B.json`. We seed the one-block reconstruction sweep at  $\alpha^*$  and pick  $\alpha$  from  $\{\alpha^*, 0.25, 0.3, 0.5, 0.75, 0.95, 0.99\}$ . The selection depends on the activation bit budget: W4A16 KV4 picks SAM 0.75, SwinIR 0.95, LLaMA/Qwen 0.99; W4A4 picks  $\alpha = 0.25$  on both LLMs (§4.2). After quantization, outliers are restored by the piecewise inverse  $\hat{x}_i = \text{sgn}(\hat{\tilde{x}}_i) [T + \alpha^{-1}(|\hat{\tilde{x}}_i| - T)]$  for  $|x_i| > T$ , and identity below. DBAF wraps any host quantizer’s weight or activation path; the unfold is a lightweight elementwise floating-point op applied outside the INT4 GEMM kernel, so the quantized matmul itself is unchanged.

### 3.3 Prompt-Conditioned Scale Anchoring

PCSA’s effectiveness is governed by the per-site compactness  $c(M, s)$  of §3.1. When  $c$  is small, per-cluster scales recover information that a single global scale loses. As  $c$  approaches 1, per-cluster scales degenerate to the global one and PCSA self-disables. §4.4 confirms this per-site behavior across three sites within SAM-B and LLaMA-3-8B.

### Algorithm 1 Prompt-Conditioned Scale Anchoring (PCSA)

- 1: **Input:** prompt descriptor  $d$ , anchor bank  $\{a_1, \dots, a_K\}$ , scales  $\{s_k\}$
- 2: **Calibration:** initialize  $a_k$  randomly,  $\ell_2$ -normalize
- 3: **for** each calibration prompt **do**
- 4:   compute  $d$ , find  $k^* = \arg \min_k (1 - \langle d, a_k \rangle)$
- 5:    $a_{k^*} \leftarrow 0.9 a_{k^*} + 0.1 d$ ;  $a_{k^*} / = \|a_{k^*}\|$
- 6:   update  $s_{k^*}$  via EMA on min/max of routed activations
- 7: **end for**
- 8: freeze  $\{a_k\}$
- 9: **Inference:** for each prompt, route via  $k^*$ , quantize with  $s_{k^*}$

Given the diagnostic, the mechanism is straightforward. We partition prompts into  $K$  clusters in descriptor space and assign each cluster its own pre-calibrated activation scale. At inference, a single cosine lookup routes each prompt to the appropriate scale. The full procedure (descriptor extraction,  $K$ -means clustering, per-anchor routing) does not depend on the host architecture, though application sites are model-specific: SAM’s mask decoder cross-attention  $q$ -projection, LLM self-attention  $q$ -projection in each decoder layer, and KV-cache scales for long-context generation (KV-PCSA-tf below). The prompt descriptor  $d(P) = \text{normalize}(N^{-1} \sum_n P_{:,n,:})$  is computed from the tokens feeding the  $q$ -projection. We maintain a bank of  $K$  anchors  $\{a_k\}$ , and route each prompt by cosine distance,  $k^* = \arg \min_k (1 - \langle d, a_k \rangle)$ . During calibration, anchors update via EMA with momentum 0.9, and per-anchor scales  $s_{k^*}$  update from min/max statistics of routed activations. At inference, anchors are frozen and each prompt uses its  $s_{k^*}$ . Algorithm 1 summarizes the procedure. All experiments use symmetric uniform INT4 quantization with  $q_{\text{max}} = 7$ ; the KV-cache uses W4 with group size 128 unless otherwise specified.

### 3.4 Training-Free PCSA

To compose PCSA on training-free hosts (RTN, GPTQ, AWQ), we replace the EMA update with one offline  $k$ -means pass over  $\ell_2$ -normalized descriptors:  $\{c_j\} = \arg \min_{\{c_j\}} \sum_i \min_j \|d_i - c_j\|_2^2$ , and replace the EMA scale update with a per-anchor max-abs,  $s_j = \max_{i: \rho(d_i)=j} \|x_i\|_\infty$  where  $\rho(d) = \arg \max_j \langle d, c_j \rangle$ . At inference, prompts

route via  $\rho(d_p)$  and quantize with  $s_{j^*}$ ; both are constant-cost per prompt. The k-means step is  $\mathcal{O}(NKD)$ ; with  $N=128$  and  $K=8$ , it completes in seconds on CPU and adds no GPU time to the host’s calibration pipeline.

**KV-PCSA-tf.** The same anchor-routing primitive extends to the KV cache: the descriptor and centroid fit are unchanged, and we additionally record per-anchor max-abs over the collected key/value caches,  $s_j^K = \max_{i:\rho(d_i)=j} \|K^{(i)}\|_\infty$  and  $s_j^V$  analogously. At inference, the prompt’s K/V cache is quantized with  $(s_{j^*}^K, s_{j^*}^V)$ ; long-context RULER (Hsieh et al., 2024) evaluation is reported in Appendix P.

**Composability.** Both primitives expose narrow interfaces any host can adopt: DBAF wraps a host’s weight/activation quantizer with the gate–fold–quantize–unfold sequence (Equations (1)–(2)); PCSA wraps the activation quantizer with the descriptor–route–scale rule of §3.4. Both are orthogonal to the host’s reparameterization (rotation, smoothing, block reconstruction) and require only a hook on the FP tensor entering the quantizer.

### 3.5 Hyperparameter Calibration

DBAF has one scalar hyperparameter  $\alpha$  that controls the fold compression ratio. We pick  $\alpha$  from a seven-point single-block reconstruction sweep, seeded by the closed-form  $\alpha^*$  derived in Appendix A. PCSA has one integer hyperparameter  $K$  that controls the number of anchors. We pick  $K$  from a small mAP or PPL sweep on the calibration set; in practice  $K \in \{4, 8, 16\}$  covers the configurations reported.

For predictive use, we fit a per-layer linear regression that maps the per-tensor *weight* outlier fraction  $p_T = \Pr_{w \sim M(T)}[|w| > 3\sigma]$  to the expected DBAF MSE reduction:

$$\widehat{\Delta\text{MSE}}(M, T) = a + b \cdot p_T, \quad (3)$$

with constants  $a$  and  $b$  fit on the 224 Linear layers of LLaMA-3-8B once and frozen. Within a configuration the fit is moderate ( $r \in [0.48, 0.80]$  over the eight configurations of Table 25, all positive,  $n = 1010$  layers). We use it for sign and order of magnitude at model scale, not as a per-layer estimate on a new architecture: the gate’s boundary transfers across held-out architectures where the predicted gain does not (AUC 0.962; Appendix E). The fitted values, the input-file SHA-256, and

the per-configuration Pearson coefficients are released at predictions/regression\_constants.json and results/pearson\_r\_8configs.json.

### 3.6 Decision and Diagnostic Summary

Table 2 previews the four prediction-then-outcome checks of §4 on the studied families, and Table 3 summarizes the diagnostic’s dispatch across the regimes evaluated in this paper. Each row of the decision table pairs an empirical regime with the diagnostic state and the resulting primitive recipe.

## 4 Experiments

We test the dispatch claim of §3.1 with four prediction-then-outcome checks. Three are on language models, where the diagnostic is calibrated: the rotation cliff across six LLaMA-3-8B W4A4 hosts (§4.2), cross-family transfer to Qwen-2.5-7B (§4.3), and per-site fires and self-disables within LLaMA-3-8B itself (§4.4). The same statistics are then applied outside NLP, where nothing was calibrated: SAM-B segmentation (reported inside the compactness check) and a negative SwinIR- $\times 3$  prediction from gate counts alone (§4.5). Three further families — CLIP (Radford et al., 2021), Whisper (Radford et al., 2023) and DiT — are in Appendix I, including the regime where the dispatch stops working. §4.6 isolates the mechanism; Table 2 previews the matrix.

### 4.1 Implementation Details

All experiments run on NVIDIA A100 80GB GPUs. LLM calibration draws 128 sequences of 2048 tokens from WikiText-2 (Merity et al., 2017), with perplexity evaluated on WikiText-2 and C4 (Raffel et al., 2020) and long-context retrieval on RULER’s needle-in-a-haystack subset (§3.4). SAM uses 32 calibration images from COCO train2017 (Lin et al., 2014) (evaluated on val2017), and SwinIR is calibrated on the CompSRT (Zeinali et al., 2026) set with Set5 (Bevilacqua et al., 2012)/Urban100 (Huang et al., 2015) for evaluation. Hyperparameters: DBAF  $T_\ell = 3\sigma$  per row with  $\alpha$  selected per regime (§3.2); PCSA  $K=4$  for SAM-B/L and  $K=8$  for LLaMA-3-8B; FlatQuant baselines train 15 epochs at learning rate  $5 \times 10^{-3}$ , batch 4; symmetric INT4 throughout with  $q_{\max} = 7$ .

### 4.2 The Rotation Cliff

The diagnostic predicts the rotation cliff (defined in §1) from the outlier gate alone: rotation hosts

| Check                    | Input statistic           | Predicted outcome                                        | Observed                                                 |
|--------------------------|---------------------------|----------------------------------------------------------|----------------------------------------------------------|
| Rotation cliff (§4.2)    | host rotation status      | helps non-rotation; saturates on rotation                | 6/6 cells match                                          |
| Cross-family (§4.3)      | Qwen $\Pr[ x  > 3\sigma]$ | DBAF $\Delta$ same sign and order as LLaMA               | 2/2 cells match (wide bands)                             |
| PCSA compactness (§4.4)  | $c$ at PCSA site          | sign of effect from $c$                                  | SAM-B Q +4.8 mAP; KV +5pp NIAH @ 8k; LLM Q self-disables |
| SwinIR $\times 3$ (§4.5) | act gate-pass % per scale | regress at $\times 3$ ; improve at $\times 2 / \times 4$ | -0.31 / +0.16 / +0.43 dB                                 |

Table 2: Predictivity matrix. Each row pairs a check derived from the diagnostic of Section 3.1 with the input statistic the diagnostic reads, the predicted outcome, and the observed outcome. The subsections of Section 4 report each in full.

| Regime                                                                       | $g$      | $c$  | DBAF / PCSA outcome                           | Recipe             |
|------------------------------------------------------------------------------|----------|------|-----------------------------------------------|--------------------|
| <b>LLM regimes (all at <math>c \approx 0.79</math>, PCSA fails the gate)</b> |          |      |                                               |                    |
| Non-rotation (RTN/GPTQ/AWQ/SQ)                                               | host     | yes  | 0.79 DBAF helps (wt2 16–33 at $\alpha=0.25$ ) | host + DBAF        |
| Rotation host (QuaRot)                                                       | yes      | 0.79 | DBAF no-op (saturates)                        | host               |
| Trained, no rotation (TesseraQ)                                              | yes      | 0.79 | DBAF helps (−82%)                             | host + DBAF        |
| Gate-fail layers                                                             | disabled | 0.79 | DBAF no-op (RTN)                              | host               |
| <b>SAM and super-resolution regimes</b>                                      |          |      |                                               |                    |
| SAM-B/L/H weights                                                            | yes      | 0.19 | both help (+4.8 mAP on SAM-B)                 | host + DBAF + PCSA |
| SwinIR $\times 2 / \times 4$                                                 | yes      | n/a  | DBAF +0.16/+0.43 dB; PCSA n/a                 | host + DBAF        |
| SwinIR $\times 3$ (act 37%)                                                  | partial  | n/a  | DBAF −0.31 dB; PCSA n/a                       | host               |

Table 3: Decision table for the diagnostic. Top block: LLM regimes (LLM Q-projection compactness  $c \approx 0.79$  disables PCSA throughout; DBAF dispatch varies with host). Bottom block: SAM and super-resolution regimes. Saturated cells (rotation LLM) and the SwinIR  $\times 3$  regression are predicted no-ops or negatives, not anomalies. Numerical ranges from Section 4.

have  $\Pr[|x| > 3\sigma] \approx 0$ , so Eq. 3 predicts near-zero  $\hat{\Delta}\text{MSE}$  and DBAF saturates, whereas non-rotation hosts stay in the 0.3–1.8% peak-gain regime where DBAF fires. TesseraQ (trained, no rotation) is predicted on the non-rotation side because the gate reads activation statistics, not the host’s calibration recipe: training alone does not remove the outlier mass that the gate detects.

Table 4 reports the six-host LLaMA-3-8B W4A4 matrix. At  $\alpha = 0.25$  all four training-free non-rotation hosts (RTN, GPTQ, AWQ, SmoothQuant) collapse from catastrophic alone-baselines (wt2 970 to 9844) into the 16–33 wt2 PPL band. QuaRot (Ashkboos et al., 2024) (rotation, untrained) saturates within noise across both DBAF columns (8.41  $\rightarrow$  8.44 at  $\alpha = 0.75$ , 8.73 at  $\alpha = 0.25$ ). TesseraQ (trained, no rotation) drops from 61.84 to 10.82 wt2 at  $\alpha_A = 0.25$  (−82%), placing it with the non-rotation block, not the rotation block, as the diagnostic predicts.

The Qwen-2.5-7B W4A4 host matrix (Appendix J) shows that Qwen’s 10 $\times$  higher activation kurtosis (mean 961 vs. 95 on LLaMA) causes three

of four training-free hosts to fail at  $\alpha = 0.25$ ; only SmoothQuant’s per-channel migration (18.27 wt2) and FlatQuant’s rotation (7.31 wt2) recover usable quality, placing Qwen’s cross-host pattern in the harder outlier regime the diagnostic predicts.

**DBAF  $\alpha$  sensitivity.** Table 7 (Appendix B) sweeps  $\alpha$  on RTN+DBAF at W4A4; every host minimizes at  $\alpha = 0.25$  and degrades monotonically (the 0.25  $\rightarrow$  0.75 wt2 ratio is 11/15/34/85 $\times$  for GPTQ/RTN/AWQ/SmoothQuant, and  $\sim 700\times$  on Qwen). This gives  $\alpha = 0.25$  as the W4A4 LLM default; extended cross-host results are in Appendix N.

All six host rows fall on the predicted side of the rotation cliff: the four non-rotation hosts collapse into the 16–33 wt2 PPL band at  $\alpha = 0.25$ , QuaRot saturates within noise across both DBAF columns, and TesseraQ (trained, no rotation) collapses to 10.82 wt2 alongside the non-rotation block. For context on the rotation side, FlatQuant (Sun et al., 2025) and SpinQuant (Liu et al., 2025b) report LLaMA-3-8B W4A4 wt2 PPL in the 7–8 range in

| Host        | R | T | Wiki PPL ↓ |                    |                          | C4 PPL ↓ |                    |                          |
|-------------|---|---|------------|--------------------|--------------------------|----------|--------------------|--------------------------|
|             |   |   | alone      | +D <sub>0.75</sub> | +D <sub>0.25</sub>       | alone    | +D <sub>0.75</sub> | +D <sub>0.25</sub>       |
| RTN         | ✗ | ✗ | 969.99     | 250.60             | <b>16.31</b>             | 775.23   | 310.08             | <b>24.49</b>             |
| GPTQ        | ✗ | ✗ | 1566.59    | 348.66             | <b>32.65</b>             | 1436.94  | 363.25             | <b>53.43</b>             |
| AWQ         | ✗ | ✗ | 3502.63    | 579.96             | <b>17.02</b>             | 2993.16  | 520.03             | <b>25.84</b>             |
| SmoothQuant | ✗ | ✗ | 9844.14    | 1477.98            | <b>17.32</b>             | 6794.43  | 1005.35            | <b>27.10</b>             |
| QuaRot      | ✓ | ✗ | 8.41       | 8.44               | 8.73                     | 13.69    | 13.70              | 13.84                    |
| TesseraQ    | ✗ | ✓ | 61.84      | 60.54*             | <b>10.82<sup>†</sup></b> | 52.84    | 50.88*             | <b>16.21<sup>†</sup></b> |

Table 4: LLaMA-3-8B W4A4 host matrix. The +DBAF columns differ in fold  $\alpha$ : 0.75 (SAM default; SwinIR uses 0.95) vs. 0.25 (W4A4 LLM optimum, Table 7). At  $\alpha = 0.25$  all four non-rotation hosts collapse into the 16–33 wt2 PPL band, 2–4 $\times$  QuaRot’s 8.41 (§4.2); the gate fires correctly here, but does not close the gap to the rotation host. On QuaRot both DBAF columns regress or saturate, matching the gate-self-disable prediction. TesseraQ (trained, no rotation) drops  $-82\%$  to 10.82 wt2 at  $\alpha_A = 0.25$ , placing it with the non-rotation block as predicted. \*  $\alpha_W = 0.95$ ,  $\alpha_A = 0.75$ . <sup>†</sup>  $\alpha_A = 0.25$ ; details in Appendix T.

| Model       | Host | Alone | +DBAF | $\Delta$  |
|-------------|------|-------|-------|-----------|
| LLaMA-3-8B  | RTN  | 10.28 | 8.87  | $-13.7\%$ |
| LLaMA-3-8B  | AWQ  | 13.49 | 9.42  | $-30.2\%$ |
| Qwen-2.5-7B | RTN  | 11.62 | 8.86  | $-23.8\%$ |
| Qwen-2.5-7B | AWQ  | 53.92 | 33.73 | $-37.4\%$ |

Table 5: W4A16 DBAF margins on LLaMA-3-8B and Qwen-2.5-7B (WikiText-2 PPL). Qwen margins are within a factor of two of the LLaMA values at both hosts. The predicted recurrence range is  $[-6.9\%, -27.4\%]$  for RTN and  $[-15.1\%, -60.4\%]$  for AWQ; both Qwen cells fall inside. These intervals are wide, so this is a consistency check on the transfer rather than a sharp test; the rotation cliff (§4.2) and the SwinIR- $\times 3$  negative prediction (§4.5) carry the predictive weight.

recent benchmarks, comparable to QuaRot’s 8.41 above. The training-free non-rotation cells at  $\alpha = 0.25$  (wt2 16–33) remain 2–4 $\times$  the rotation band; the diagnostic does not claim parity, only that it correctly fires on non-rotation hosts and self-disables on rotation hosts across all six hosts of the matrix.

### 4.3 Cross-Family Recurrence

The diagnostic is calibrated on LLaMA; if the dispatch claim holds, a second LLM family with similar distributional signatures should land in the LLaMA-fit prediction band without refit. Qwen-2.5-7B shares RMSNorm, SwiGLU, and RoPE with LLaMA, and Eq. 3 on Qwen’s per-Linear outlier fractions places both in the same outlier regime. The prediction is that Qwen-2.5-7B’s W4A16 DBAF margin on RTN and AWQ falls within a factor of two of the matched LLaMA-3-8B values ( $-13.7\%$  RTN,  $-30.2\%$  AWQ). Table 5 reports both.

Both Qwen margins land in the predicted band;

the diagnostic generalizes to the second LLM family. The transfer mechanism is that the per-layer regression of §3.5 reads outlier mass alone, with no architectural prior beyond the shared norm/activation scaffold.

### 4.4 PCSA Descriptor Compactness

The compactness gate’s claim is that the same model can fire PCSA at one tensor site and self-disable at another, with per-site  $c$  predicting which:  $c \leq 0.2$  implies improvement of at least three points;  $c \geq 0.7$  implies PCSA-tf does not help and degrades the activation quantizer by at least one order of magnitude. Tight clusters let per-cluster scales recover information a global scale loses; sparse clusters cannot. We measure  $c$  at three sites that span the predicted range, and report results next.

**LLaMA-3-8B Q-projection** ( $c = 0.79$ ). The gate’s self-disable prediction at high  $c$  is confirmed empirically: across eight PCSA-tf variants (host  $\times$  DBAF prefold  $\times$  fold-calibration), the best lands at 653 WikiText-2 PPL versus the 10.3 W4A16 baseline for the same host (RTN, no activation quantization; Appendix O).

**LLaMA-3-8B KV-cache at long context.** At the KV-cache site the anchor descriptors cluster meaningfully: KV-PCSA-tf (§3.4) recovers 10/200 otherwise-missed needles at 8k context, a +5pp gain on needle-in-a-haystack retrieval (Hsieh et al., 2024), where a target sentence placed at varying positions must be retrieved (Table 6). The matching PPL shift stays within  $\pm 0.4$  across 2k–8k (Appendix P).

**SAM-B Q-projection: the same test outside NLP** ( $c = 0.189 \pm 0.003$ ,  $K = 4$ ;  $\approx 0.40$  on real

| Method                  | NIAH 4k      | NIAH 8k      |
|-------------------------|--------------|--------------|
| FQ+DBAF+PCSA (baseline) | 0.990        | 0.945        |
| + KV-PCSA-tf            | <b>0.995</b> | <b>0.995</b> |

Table 6: KV-PCSA-tf on LLaMA-3-8B W4A4 KV4 with FlatQuant (FQ) host. Needle-in-a-haystack retrieval accuracy at 4k and 8k context ( $n=200$  samples per length). KV-PCSA-tf recovers ten otherwise-missed needles at 8k, a +5pp gain on a capability metric (95% Wald CI on the unpaired difference [+1.7, +8.3] percentage points, excluding zero).

COCO, Table 19). The compactness statistic is not language-specific, so we apply it unchanged to a segmentation model. It fires, and trained PCSA via AHCPTQ is super-additive with DBAF (AHCPTQ  $13.4 \rightarrow 18.2$  mAP, +4.8 vs. +2.9+0.4 singletons; full breakdown Appendix L).

The three sites separate the gate’s outcomes both within and across modalities, with no threshold refit at any of them.

#### The gate reads the pooled site, not each layer.

The  $c$  values above are pooled over a site’s layers, and the separation does not survive disaggregation: SAM-B’s mask-decoder cross-attention layers span  $[0.69, 0.70]$ , inside the  $[0.67, 0.82]$  range of LLaMA-3-8B’s  $q_{\text{proj}}$  layers (Appendix M), so the  $4.2\times$  pooled gap is a shift in median, not layer-by-layer dominance. Pooled is nonetheless the granularity PCSA acts at: one anchor is selected per input and reused at every layer the routing controls. The gate is a claim about sites, not about individual layers.

#### 4.5 The SwinIR $\times 3$ Negative Prediction

The strongest test of a dispatch rule is whether it identifies where primitives should be *skipped*, and we run it on the family furthest from the calibration data. SwinIR- $\times 3$  (activation gate-pass 37%) is predicted to regress under training-free weight-only W4 (W4A16), while  $\times 2$  (61%) and  $\times 4$  (44%) are predicted to improve; the compounded  $\times 3$  prediction from Eq. 3 is  $-0.32$  dB. Observed Set5  $\Delta$  dB matches the sign at all three scales ( $-0.31$  at  $\times 3$ ,  $+0.16$  at  $\times 2$ ,  $+0.43$  at  $\times 4$ ); Urban100 agrees, and so does a W4A4 repeat ( $-0.41$  at  $\times 3$ ). The  $\times 3$  regression is a negative prediction from gate counts alone, not a post-hoc fit; the discriminating statistic is activation-side even though quantization is weight-only (Appendix R).

#### 4.6 Mechanism Isolation

DBAF’s value isolates to outlier-ordering preservation, not removal. Matched- $T$  clipping at  $\alpha=0$  gives PPL 79,256 on LLaMA-3-8B W4 weight-only against 8.87 with RTN+DBAF (Appendix C). The per-layer regression ( $r \in [0.48, 0.80]$  within a configuration) recovers the predicted task delta in the peak-gain regime (Appendices Q, S), and DBAF adds at most 1.7% latency on the LLM and SAM paths, up to +23.5% on narrow-width SwinIR-light (Appendix W), calibrating in seconds and beating fast Hadamard at  $d \geq 8192$  (Appendix W).

#### 5 Conclusion

For the distributional pathologies studied, PTQ method choice tracks tensor distribution more reliably than architecture. A two-statistic diagnostic indicates whether each of two composable primitives (DBAF, PCSA) should apply, saturate, or be skipped on a given triple. Four prediction-then-outcome checks across LLMs, SAM and image super-resolution are consistent with this view, and the gate’s decision boundary proves recoverable from data with each architecture family held out (mean AUC 0.962), dispatching across seven model families in six domains from a single calibration on LLaMA-3-8B. A matched- $T$  ablation isolates outlier-ordering preservation as the mechanism of DBAF’s efficacy. Both primitives are training-free and add at most 1.7% inference latency on the LLM and SAM paths — though up to +23.5% on narrow-width SwinIR-light, where DBAF’s elementwise work cannot amortize over a small GEMM (Appendix W).

Within the calibrated regime the gate’s skips cost little, and that is what the four checks test. Off-distribution the shape tests — sufficient but not necessary for the fold to help — become expensive: on DiT-XL the gate declines to fire on tensors DBAF would have improved threefold, and should be read as “outside the calibrated regime” rather than “DBAF would not help”. Even there the failure is one-sided: the price is a foregone gain rather than a damaged model, and the two statistics come from a calibration pass that is already being run. Replacing the hand-set boundary with the learned one is a natural next step. Predicting how much it will help, by contrast, does not yet transfer. Whether one learned dispatcher can carry both decisions remains open.

## Limitations

**Calibration scope.** The outlier-gate thresholds and the per-layer regression constants of Eq. 3 are calibrated on LLaMA-3-8B and used unchanged throughout Section 4. The DBAF  $\alpha$  is not: it is selected per regime by the reconstruction sweep of Section 3.5, which costs minutes and is how the approach adapts to a new activation profile. Automated refitting of the regression constants is future work.

**Distributional scope.** DBAF and PCSA address two recurring pathologies: dense regions with sparse outliers, and input-conditioned shifts. §4.3 identifies further patterns on which the gate self-disables. That self-disabling is conservative rather than always optimal — on DiT-XL it skips 30% of tensors DBAF would have improved 3 $\times$  (Appendix I) — and those patterns need distribution-specific primitives we leave to future work.

**Descriptor scope for PCSA-tf.** The mean-pooled PCSA-tf descriptor was validated on the calibration distribution used at inference and may not transfer to RAG, multi-turn dialog, or mixed-task batches, and the KV-PCSA long-context gain is W4A4 KV4 specific (Appendix X).

**Status of the diagnostic.** We position the dispatch as a selection heuristic, not a calibrated classifier, because its two components generalize very differently. The binary outlier gate transfers: a logistic dispatcher over the same statistics, trained leave-one-architecture-out, reproduces the hand-set boundary at mean AUC 0.962 (Appendix E). The per-layer gain regression of Eq. 3 does not: moderate within a configuration ( $r \in [0.48, 0.80]$ ), weak held out across architectures ( $r \approx 0.13$ ), so we use it as an aggregate trend estimator and never as a per-layer predictor on an unseen architecture. The diagnostic supports deciding *whether* a primitive should fire on a new model; predicting *how much* it will help remains out of reach.

**Calibrated components.** The  $3\sigma$  tail boundary follows the 68–95–99.7 rule rather than tuning, and the remaining gate parameters, fixed once on LLaMA-3-8B, retain 94–100% weight gate-pass across the eight configurations of Table 25; we have not stress-tested them under broader distribution shift. The closed-form  $\alpha^*$  of Eq. 2 minimizes an equal-weight MSE proxy and brackets the selected  $\alpha$  from below on the activation path and from above

on the weight path, so we use it to seed a reconstruction sweep rather than as the operating point. Both, with the sensitivity-weighted form of  $\alpha^*$ , are in Appendix X.

**Statistical and scale scope.** End-task results are reported at a single calibration seed, without confidence intervals and without seed variance on the end-task metrics; the dispatch decisions themselves are stable across redraws (Appendices D and X). The largest LLM evaluated is 8B, so whether the rotation cliff or the per-layer regression hold at 70B is untested, and we do not compare against adaptive-dispatch quantizers that learn a dispatcher over the same features.

## Ethical considerations

**Dual-use of efficiency.** Lower-bit deployment reduces the compute and memory cost of running large models, including for harmful applications such as bulk generation of misleading text or mass image segmentation in surveillance contexts. This work does not introduce new model capabilities, but it lowers the deployment barrier for capabilities that already exist; we view this as a standard dual-use property of PTQ research.

**Benchmark blind spots.** We validate on perplexity, mask AP, super-resolution PSNR, and long-context retrieval (NIAH). These metrics do not directly probe fairness, safety alignment, or robustness to adversarial inputs. Quantization can introduce small but systematic output shifts that standard task metrics may not surface, so deployments of the quantized models we study should include the same alignment and safety evaluations used for the full-precision originals.

**Calibration-data fairness.** The diagnostic and the DBAF  $\alpha$  choice are calibrated on WikiText-2, C4, COCO, and Set5. Calibration-data choice has been shown to drive substantial PPL variation in LLMs (Liu et al., 2024c), and analogous shifts may exist along demographic or linguistic axes that we do not measure. Practitioners deploying these primitives on populations not represented in the calibration set should validate on in-domain data before relying on the diagnostic’s dispatch decisions for that population.

**Use of generative AI.** The authors used a large language model for coding work and for partial generation of the overview figure.

## References

- Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefer, and James Hensman. 2024. [Quarot: Outlier-free 4-bit inference in rotated llms](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *British Machine Vision Conference*.
- Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. 2025. [Why do some inputs break low-bit llm quantization?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. 2023. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Mengzhao Chen, Yi Liu, Jiahao Wang, Yi Bin, Wenqi Shao, and Ping Luo. 2026. [PrefixQuant: Eliminating outliers by prefixed tokens for large language models quantization](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *International Conference on Learning Representations (ICLR)*.
- Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. [Yolox: Exceeding yolo series in 2021](#). *Preprint*, arXiv:2107.08430.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Cheeun Hong, Heewon Kim, Sungyong Baik, Junghun Oh, and Kyoung Mu Lee. 2022. Daq: Channel-wise distribution-aware quantization for deep image super-resolution networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2675–2684.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What’s the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*.
- Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Zhihang Yuan, Jiangyong Yu, Chen Xu, Zhe Jiang, and Sifan Zhou. 2025. [OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting](#). In *International Conference on Learning Representations (ICLR)*.
- Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206.
- Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. 2023. [Detrs with hybrid matching](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.
- Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.
- Yuhang Li and Priyadarshini Panda. 2024. [Tesseract: Ultra low-bit llm post-training quantization with block reconstruction](#). *Preprint*, arXiv:2410.19103.
- Guang Liang, Jie Shao, Ningyuan Tang, Xinyao Liu, and Jianxin Wu. 2025. [Tweo: Transformers without extreme outliers enables fp8 training and quantization for dummies](#). *Preprint*, arXiv:2511.23225.
- Jingyun Liang, Jie Zhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. [Swinir: Image restoration using swin transformer](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.
- Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. 2024a. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Weiming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024b. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755.
- Kai Liu, Haotong Qin, Yong Guo, Xin Yuan, Linghe Kong, Guihai Chen, and Yulun Zhang. 2024a. [2dquant: Low-bit post-training quantization for image super-resolution](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kai Liu, Dehui Wang, Zhiteng Li, Zheng Chen, Yong Guo, Wenbo Li, Linghe Kong, and Yulun Zhang. 2025a. Condiquant: Condition number based low-bit quantization for image super-resolution. *arXiv preprint arXiv:2502.15478*.
- Xiaoyu Liu, Xin Ding, Lei Yu, Yuanyuan Xi, Wei Li, Zhijun Tu, Jie Hu, Hanting Chen, Baoqun Yin, and Zhiwei Xiong. 2024b. PQ-SAM: Post-training quantization for segment anything model. In *European Conference on Computer Vision (ECCV)*.
- Yijiang Liu, Huanrui Yang, Zhen Dong, Kurt Keutzer, Li Du, and Shanghang Zhang. 2023. [Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers](#). *Preprint*, arXiv:2211.16056.
- Yijun Liu, Yuan Meng, Fang Wu, Shenhao Peng, Hang Yao, Chaoyu Guan, Chen Tang, Xinzhu Ma, Zhi Wang, and Wenwu Zhu. 2024c. [Evaluating the generalization ability of quantized llms: Benchmark, analysis, and toolbox](#). *Preprint*, arXiv:2406.12928.
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025b. [Spinquant: Llm quantization with learned rotations](#). In *International Conference on Learning Representations (ICLR)*.
- Chengtao Lv, Hong Chen, Jinyang Guo, Yifu Ding, and Xianglong Liu. 2024. [Ptq4sam: Post-training quantization for segment anything](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15941–15951.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. [Pointer sentinel mixture models](#). In *International Conference on Learning Representations (ICLR)*.
- Andrew Or, Apurva Jain, Daniel Vega-Myhre, Jesse Cai, Charles David Hernandez, Zhenrui Zheng, Driss Guessous, Vasilij Kuznetsov, Christian Puhersch, Mark Saroufim, Supriya Rao, Thien Tran, and Aleksandar Samardžić. 2025. [Torchao: Pytorch-native training-to-serving model optimization](#). *Preprint*, arXiv:2507.16099.
- William Peebles and Saining Xie. 2023. [Scalable diffusion models with transformers](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Haotong Qin, Yulun Zhang, Yifu Ding, Yifan Liu, Xianglong Liu, Martin Danelljan, and Fisher Yu. 2023. QuantSR: Accurate low-bit quantization for efficient image super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- 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. 2021. [Learning transferable visual models from natural language supervision](#). In *International Conference on Machine Learning (ICML)*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning (ICML)*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Navin Ranjan and Andreas Savakis. 2025. [Mix-qsam: Mixed-precision quantization of the segment anything model](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. [Omniquant: Omnidirectionally calibrated quantization for large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. [Massive activations in large language models](#). *Preprint*, arXiv:2402.17762.
- Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, and 1 others. 2025. [Flatquant: Flatness matters for llm quantization](#). In *International Conference on Machine Learning (ICML)*.
- Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. 2024. [Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks](#). In *International Conference on Machine Learning (ICML)*.
- Hailing Wang, Jianglin Lu, Yitian Zhang, and Yun Fu. 2025a. Outlier-aware post-training quantization for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16175–16184.

- Xiaoqi Wang, Clint Sebastian, Wenbin He, and Liu Ren. 2025b. [Prosam: Enhancing the robustness of sam-based visual reference segmentation with probabilistic prompts](#). *Preprint*, arXiv:2506.21835.
- Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. 2023. Outlier suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Miles Williams and Nikolaos Aletras. 2024. [On the impact of calibration data in post-training quantization and pruning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*.
- Dorsa Zeinali, Hailing Wang, Yitian Zhang, and Yun Fu. 2026. [Any orthogonal transform will do: Range reduction explains rotation-based quantization for super-resolution](#). *Preprint*, arXiv:2601.21069v4.
- Wenlun Zhang, Yunshan Zhong, Shimpei Ando, and Kentaro Yoshioka. 2025. [Ahcptq: Accurate and hardware-compatible post-training quantization for segment anything model](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Christopher De Sa, and Zhiru Zhang. 2019. Improving neural network quantization without retraining using outlier channel splitting. In *International Conference on Machine Learning (ICML)*.
- Sijie Zhao, Tao Yue, and Xuemei Hu. 2021. Distribution-aware adaptive multi-bit quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9281–9290.

## A Derivation of $\alpha^*$

Let  $\tilde{x}_q = \tilde{x} + \epsilon$  denote the quantized folded value where  $\epsilon \sim \mathcal{U}(-s/2, s/2)$  with  $\mathbb{E}[\epsilon] = 0$ ,  $\mathbb{E}[\epsilon^2] = s^2/12$ . For inliers ( $b = 0$ , unfolding not applied),  $\mathbb{E}[(\hat{x} - x)^2 | b=0] = s^2/12$ . For outliers ( $b = 1$ , unfolding scales noise by  $1/\alpha$ ):

$$\hat{x} - x = \frac{\epsilon}{\alpha} \Rightarrow \mathbb{E}[(\hat{x} - x)^2 | b=1] = \frac{s^2}{12\alpha^2}.$$

Combining with  $p_{\text{out}} = \Pr(|x| > T_\ell)$ ,  $p_{\text{in}} = 1 - p_{\text{out}}$ :

$$\mathcal{L}(\alpha) \approx \frac{s(\alpha)^2}{12} \left( p_{\text{in}} + \frac{p_{\text{out}}}{\alpha^2} \right) \quad (4)$$

The maximum post-fold magnitude is  $M_{\text{fold}}(\alpha) = T + \alpha(M - T)$ , so for symmetric  $b$ -bit quantization with  $q_{\text{max}} = 2^{b-1} - 1$ ,  $s(\alpha) \propto T + \alpha(M - T)$ . Substituting and dropping  $\alpha$ -independent constants yields the objective  $\tilde{\mathcal{L}}(\alpha) = (T + \alpha(M - T))^2 (p_{\text{in}} + p_{\text{out}}/\alpha^2)$ .

Letting  $A = T + \alpha(M - T)$ ,  $B = p_{\text{in}} + p_{\text{out}}/\alpha^2$ :

$$\frac{d\tilde{\mathcal{L}}}{d\alpha} = 2A(M - T)B - 2A^2 \frac{p_{\text{out}}}{\alpha^3}.$$

Setting to zero and dividing by  $2A > 0$ :

$$(M - T) \left( p_{\text{in}} + \frac{p_{\text{out}}}{\alpha^2} \right) = \frac{(T + \alpha(M - T))p_{\text{out}}}{\alpha^3}.$$

Multiplying by  $\alpha^3$  and simplifying cancels the  $\alpha(M - T)p_{\text{out}}$  terms, yielding  $(M - T)p_{\text{in}}\alpha^3 = Tp_{\text{out}}$ , so:

$$\alpha^* = \sqrt[3]{\frac{Tp_{\text{out}}}{(M - T)p_{\text{in}}}} \quad (5)$$

**Convexity.** With  $c \triangleq M - T > 0$ :  $\tilde{\mathcal{L}}''(\alpha) = 2p_{\text{in}}c^2 + \frac{4p_{\text{out}}Tc}{\alpha^3} + \frac{6p_{\text{out}}T^2}{\alpha^4} > 0$  for all  $\alpha > 0$ , since each term is strictly positive. Hence  $\tilde{\mathcal{L}}$  is strictly convex on  $(0, \infty)$  and  $\alpha^*$  is the global minimum.

## B Sensitivity-Weighted $\alpha^*$ : Derivation and Scope

The Limitations section states that the operating  $\alpha$  lies above  $\alpha_A^*$  on the activation path. This appendix derives that result and states precisely what it does and does not establish.

**Setup.** DBAF folds outliers ( $|x| > T$ ) to  $\text{sgn}(x)[T + \alpha(|x| - T)]$ , quantizes in folded space with step  $s = M_{\text{fold}}/q_{\text{max}}$  where  $M_{\text{fold}} = T +$

$\alpha(M - T)$ , and unfolds outlier positions by  $\alpha^{-1}$ . The per-tensor MSE proxy of Appendix A is

$$\tilde{\mathcal{L}}(\alpha) = A^2(p_{\text{in}} + p_{\text{out}}/\alpha^2), \quad A = T + \alpha\Delta, \quad \Delta = M - T. \quad (6)$$

where the  $1/\alpha^2$  on the outlier term is the unfold noise amplification. Minimizing gives Eq. (2).

**Sensitivity weighting.** Reconstruction MSE is not the end task. Weighting the outlier-position term by its end-task sensitivity  $\lambda$  — the outlier-to-bulk Hessian ratio  $h_{\text{out}}/h_{\text{in}}$  used by GPTQ/OBQ-style objectives — gives  $\mathcal{L}(\alpha) = A^2(p_{\text{in}} + \lambda p_{\text{out}}/\alpha^2)$ . Differentiating and setting to zero,

$$2A\Delta p_{\text{in}} + 2\lambda p_{\text{out}}A(\Delta\alpha - A)/\alpha^3 = 0 \quad (7)$$

$$\Delta p_{\text{in}}\alpha^3 + \lambda p_{\text{out}}(\Delta\alpha - A) = 0. \quad (8)$$

Substituting  $A = T + \alpha\Delta$ , the  $\lambda p_{\text{out}}\Delta\alpha$  terms cancel, leaving  $\Delta p_{\text{in}}\alpha^3 = \lambda p_{\text{out}}T$ , i.e.  $\alpha^3 = \lambda(\alpha^*)^3$  and therefore

$$\alpha_{\text{sens}}^* = \lambda^{1/3} \alpha^*. \quad (9)$$

Setting  $\lambda = 1$  recovers Eq. (2) exactly, so the weighted derivation is consistent with the unweighted one. For  $\lambda > 1$  we get  $\alpha_{\text{sens}}^* > \alpha^*$ : the end-task optimum lies above the proxy optimum. This *derives* the lower-bound statement rather than asserting it.

**What this does not establish.**  $\lambda$  is not derived to a value. It is an end-task Hessian ratio, not a fold-geometry quantity, so it cannot be recovered from  $T$ ,  $M$  and  $p_{\text{out}}$  alone. We tested the natural first-principles proxy, the energy ratio  $\lambda \approx \mathbb{E}[x^2 | \text{outlier}]/\mathbb{E}[x^2 | \text{bulk}]$ , on real LLaMA-3-8B activations: it measures  $\lambda \approx 28$ , which through  $\alpha_{\text{sens}}^* = \lambda^{1/3}\alpha^*$  predicts  $\alpha \approx 0.65$  rather than the operating 0.25 (reaching 0.25 from  $\alpha_A^* = 0.20$  requires only  $\lambda \approx 2$ ). The crude energy ratio therefore overshoots substantially, and we do not claim a first-principles numerical prediction of the operating  $\alpha$ . What the derivation supports is the *form*  $\alpha_{\text{sens}}^* = \lambda^{1/3}\alpha^*$  and the *sign* result for  $\lambda > 1$ ;  $\alpha$  itself is selected by the one-block reconstruction sweep of Section 3.5.

**Weight path.** On weights the same weighting applies with a different  $\lambda$ . Outlier-position error in a weight tensor is diluted across each dot product rather than amplified into a single activation, so  $\lambda \lesssim 1$  there. Consistent with this,  $\alpha_W^* \approx 0.28$  on

| $\alpha$ | LLaMA-3-8B        |                   | Qwen-2.5-7B       |                   |
|----------|-------------------|-------------------|-------------------|-------------------|
|          | Wiki              | C4                | Wiki              | C4                |
| 0.02     | $6.6 \times 10^5$ | $7.2 \times 10^5$ | NaN*              | NaN*              |
| 0.25     | <b>16.31</b>      | <b>24.49</b>      | <b>55.53</b>      | <b>83.00</b>      |
| 0.50     | 43.08             | 62.82             | $7.3 \times 10^3$ | $1.1 \times 10^4$ |
| 0.75     | 250.6             | 310.1             | $4.0 \times 10^4$ | $4.5 \times 10^4$ |
| 0.95     | 866.8             | 701.7             | $9.0 \times 10^4$ | $1.1 \times 10^5$ |

Table 7: DBAF  $\alpha$  sensitivity on RTN+DBAF at W4A4 (WikiText-2 and C4 PPL, lower is better). The empirical minimum is  $\alpha=0.25$  on both LLMs (15 $\times$  PPL improvement vs.  $\alpha=0.75$  on LLaMA;  $\sim 700\times$  on Qwen); PPL degrades monotonically past this point. The analytic proxy optimum on the activation path,  $\alpha_A^* \approx 0.20$  (Eq. (2); 0.18 on Qwen), lies below this empirical minimum, consistent with its status as a lower bound; we use it to seed the sweep rather than to replace it. Row 1 shows the opposite regime: over-aggressive folding at  $\alpha=0.02$ , where the unfold amplification ( $\alpha^{-1}=50\times$ ) on outlier positions dominates, diverges. \* Qwen overflows at this aggressive low  $\alpha$ , consistent with its higher activation kurtosis (mean 961 vs. 95 on LLaMA).

LLaMA-3-8B and Qwen-2.5-7B lands at or slightly above the selected  $\alpha=0.25$ , the opposite side from the activation path. We report this as an observation consistent with the derivation, not as a measurement of  $\lambda$ : we have not measured  $h_{\text{out}}/h_{\text{in}}$  separately per path.

**The  $\alpha$  sweep behind the W4A4 LLM default.** Table 7 is the sweep referenced from §4.2: every training-free host minimizes at  $\alpha=0.25$ , which is the value used for the +D0.25 columns of Table 4.

## C DBAF Strictly Improves MSE Over Matched- $T$ Clipping

**Setting.** Under the noise model of Appendix A, consider a tensor  $x$  partitioned into inliers ( $|x| \leq T$ , mass  $p_{\text{in}}$ ) and outliers ( $|x| > T$ , mass  $p_{\text{out}}$ ).

**Clipping baseline.** Matched- $T$  clipping sets  $\tilde{x}_i = \text{sign}(x_i) \min(|x_i|, T)$  and then quantizes. The post-fold range is  $[-T, T]$  with step size  $s_{\text{clip}} = T/q_{\text{max}}$ . For inliers, the reconstruction error equals the quantization noise:  $\mathbb{E}[(\hat{x} - x)^2 | \text{inlier}] = s_{\text{clip}}^2/12$ . For outliers, the clipped value is exactly  $\pm T$ , so the reconstruction error is the squared clipping distance plus noise:

$$\mathbb{E}[(\hat{x} - x)^2 | \text{outlier}] = (|x| - T)^2 + \frac{s_{\text{clip}}^2}{12} \geq (|x| - T)^2.$$

The total MSE is bounded below by:

$$\mathcal{L}_{\text{clip}} \geq p_{\text{in}} \frac{T^2}{12q_{\text{max}}^2} + p_{\text{out}} \mathbb{E}[(|x| - T)^2 | \text{outlier}]. \quad (10)$$

The second term grows quadratically with outlier magnitude and cannot be reduced by clipping alone, clipping discards outlier magnitude information.

**DBAF.** DBAF maps outliers via  $\tilde{x}_i = \text{sign}(x_i)T + \alpha(x_i - \text{sign}(x_i)T)$ , preserving the ordering and magnitude. The post-fold range is  $[-T - \alpha(M - T), T + \alpha(M - T)]$ , with  $s_{\text{DBAF}}(\alpha) = (T + \alpha(M - T))/q_{\text{max}}$ . After unfolding, the outlier reconstruction error scales by a factor of  $1/\alpha$ :

$$\mathcal{L}_{\text{DBAF}}(\alpha) = \frac{s_{\text{DBAF}}(\alpha)^2}{12} \left( p_{\text{in}} + \frac{p_{\text{out}}}{\alpha^2} \right). \quad (11)$$

**Claim.** For any  $\alpha \in (0, 1)$  and outlier distribution with non-degenerate support,  $\mathcal{L}_{\text{DBAF}}(\alpha^*) < \mathcal{L}_{\text{clip}}$ .

**Proof sketch.** Clipping corresponds to  $\alpha \rightarrow 0$  in DBAF’s formulation: the post-fold range shrinks to  $T$ ,  $s_{\text{clip}}^2 = s_{\text{DBAF}}(0)^2$ , and the unfolding factor  $1/\alpha \rightarrow \infty$ . From Eq. 11,  $\mathcal{L}_{\text{DBAF}}(\alpha) \rightarrow \infty$  as  $\alpha \rightarrow 0$ .

However, by Appendix A,  $\mathcal{L}_{\text{DBAF}}$  is strictly convex on  $(0, \infty)$  with unique minimum at  $\alpha^*$ . At any  $\alpha > 0$ , the squared clipping error  $(|x| - T)^2$  that contributes to Eq. 10 is replaced in DBAF by the smaller  $\alpha^2(|x| - T)^2$  before unfolding, and after unfolding the contribution is  $(1/\alpha)^2 \cdot$  (post-quant variance), which depends only on the quantization noise, not on  $(|x| - T)^2$ . Hence:

$$\mathcal{L}_{\text{DBAF}}(\alpha^*) < \mathcal{L}_{\text{clip}}. \quad \blacksquare$$

**Empirical confirmation.** Table 35 (Appendix V) confirms this empirically: on LLaMA-3-8B W4, RTN+matched- $T$ -clipping yields PPL 79,256 (model destroyed) while RTN+DBAF yields PPL 8.87. The  $\sim 79\text{k}$  gap is the “outlier ordering loss” the bound predicts.

## D Calibration-Seed Variance of the Diagnostic

The primitives are training-free: there is no optimizer state and no random initialization, so quantization is deterministic conditional on a calibration sample. The remaining source of randomness is the calibration draw itself, which can in principle

move a tensor across a gate boundary. We therefore re-ran the diagnostic on LLaMA-3-8B across five independent WikiText-2 redraws (a fresh subset and subsample per seed, 16 calibration sequences each, 225 common Linear layers) and measured how far its decisions move.

| Quantity (5 seeds)                  | Value              |
|-------------------------------------|--------------------|
| Activation gate-pass rate           | $63.1 \pm 1.3\%$   |
| range over seeds                    | 61.3–64.4%         |
| Mean pairwise decision agreement    | 94.7%              |
| Layers identical across all 5 seeds | 89.3%              |
| PCSA site compactness $c$           | $0.856 \pm 0.045$  |
| PCSA decision at that site          | SKIP (all 5 seeds) |

Table 8: Stability of the diagnostic’s decisions under calibration redraw, LLaMA-3-8B. The gate-pass rate moves by  $\pm 1.3$  points and 89.3% of individual layer decisions are identical across every seed. The PCSA compactness at the LLaMA  $q$ -projection site stays well above the  $c \leq 0.4$  fire threshold on all five draws, so the self-disable prediction of §4.4 is not an artifact of one calibration sample.

Two points of scope. First, this measures the stability of the *dispatch*, not of the end-task metrics: PPL, mAP and PSNR in Section 4 are each reported at a single calibration seed, and we have not measured their seed variance. Second, the residual 10.7% of layers whose decision is not identical across all five seeds sit near the gate boundary by construction; the threshold-perturbation analysis reports how far the boundary can move before aggregate decisions change.

## E Threshold Robustness and Boundary Reproduction

The outlier gate of §3.1 uses three hand-set bounds (skewness  $\leq 0.7$ , kurtosis  $\in [3, 30]$ , outlier fraction  $\in [10^{-4}, 2 \times 10^{-2}]$ ) calibrated once on LLaMA-3-8B. Two questions follow: how much do the dispatch decisions move if those bounds are wrong, and is the boundary they define recoverable from data rather than hand-set? We answer both on the 1010 Linear layers of the eight configurations in Table 25 (36.5% FIRE, 63.5% SKIP). All numbers below are computed deterministically from the committed per-layer statistics in results/S4-cross-model-layer-analysis/ of the released repository.

**Threshold perturbation.** Perturbing all five gate bounds simultaneously by  $\pm 20\%$  (200 independent draws) leaves 93.9% of layer decisions unchanged on average (min 73.9%, 5th percentile

77.4%; 6.1% of decisions change). Weight gate-pass stays in the 94–100% band for every configuration (LLaMA 100.0, Qwen 99.5, SAM-B 94.1, SAM-L 97.0, SAM-H 97.7, SwinIR- $\times 2/\times 3/\times 4$  97.1/97.1/96.1). The dispatch is therefore not balanced on a knife edge: the bounds can be wrong by a fifth and nineteen decisions in twenty survive.

**Boundary reproduction, held out by architecture.** A logistic classifier over ( $\Pr[|x| > 3\sigma]$ ,  $\log_{10}$  kurtosis,  $|\text{skew}|$ ) is trained to reproduce the hand-set FIRE/SKIP boundary with one model family held out entirely, then evaluated on the held-out family. We report this for the activation-path instantiation of the gate, which is the selective one: the weight gate passes 98% of layers, whereas the activation gate fires on 36.5% and therefore carries the dispatch.

| Held-out family                   | AUC          |
|-----------------------------------|--------------|
| LLaMA                             | 0.979        |
| Qwen                              | 0.949        |
| SAM                               | 0.988        |
| SwinIR                            | 0.930        |
| Mean (leave-one-architecture-out) | <b>0.962</b> |
| Leave-one-configuration-out       | 0.957        |

Table 9: Reproduction of the hand-set gate boundary by a learned dispatcher, with each architecture family fully held out. The manual region is a stable, low-dimensional boundary that transfers to families never seen in training. This measures reproducibility of the boundary, not accuracy against downstream-benefit labels; the empirical benefit of firing is assessed separately by the forced-vs-gated comparisons of §4.6.

**The gain regression does not transfer.** The binary boundary and the gain magnitude behave very differently, and we report both. Within a configuration, the per-layer regression of Eq. 3 correlates at  $r \in [0.48, 0.80]$  at the  $\alpha = 0.95$  weight-path setting the regression is fit on (LLaMA 0.485, Qwen 0.511, SAM-B 0.595, SAM-H 0.607, SAM-L 0.650, SwinIR- $\times 2/\times 3/\times 4$  0.764/0.789/0.796; all positive, all  $p < 10^{-5}$ ). Held out across architectures the frozen single-feature regression drops to  $r \approx 0.125$ , and a learned three-feature version reaches only  $r \approx 0.262$ .

The asymmetry follows from what each half measures. The gate’s boundary is a function of distribution *shape* — skewness, kurtosis, outlier fraction — which is scale-free and therefore comparable across architectures. The predicted *gain* is a reduction relative to a particular host quantizer’s

baseline error, which depends on that host and that architecture; a frozen intercept and slope cannot follow it. One transfers because it is dimensionless; the other does not because it is not.

This asymmetry is the honest scope of the diagnostic. Deciding *whether* DBAF should fire on an unseen architecture is well supported (AUC 0.962); predicting *how much* it will help there is not. Accordingly the four checks of §4 use the regression only at model scale — the compounded SwinIR- $\times 3$  prediction, the Qwen margin band, the near-zero prediction on rotation hosts — where per-layer error averages out, and never as a per-layer point estimate on a new model.

## F Held-Out Distribution Shift

The gate thresholds, the DBAF fold strength  $\alpha = 0.25$ , and the regression constants are all selected on the WikiText-2 calibration split (§4.1). A natural concern is that the dispatch is tuned to that corpus. We therefore freeze every hyperparameter at its WikiText-2 value and evaluate on four held-out distributions: C4, code, multilingual text (de/fr/ru/zh) and instruction-style prompts. Nothing is refit.

**End-task gains transfer.** Table 10 reports LLaMA-3-8B W4A4 perplexity with the frozen  $\alpha = 0.25$ . RTN alone collapses on every distribution; adding DBAF prevents the collapse on every distribution, with reduction factors between  $33\times$  and  $87\times$ . The gain is not a WikiText-2 artifact.

| Distribution       | FP   | RTN    | +DBAF      | $\times$ |
|--------------------|------|--------|------------|----------|
| WikiText-2 (calib) | 6.3  | 844.2  | 17.3       | 48.9     |
| C4                 | 9.6  | 805.1  | 24.6       | 32.7     |
| Code               | 1.9  | 419.9  | <b>4.8</b> | 87.4     |
| Multilingual       | 10.5 | 1684.7 | 44.9       | 37.6     |
| Instruction        | 6.4  | 731.6  | 18.3       | 40.1     |

Table 10: LLaMA-3-8B W4A4 perplexity under held-out distribution shift, with the gate and  $\alpha = 0.25$  frozen at their WikiText-2 values (24 evaluation windows per corpus). “ $\times$ ” is the RTN $\rightarrow$ DBAF reduction factor. DBAF prevents the RTN collapse on all five distributions, including the four never used for any selection decision at calibration time.

At W3 the same pattern holds with larger margins: RTN degrades to  $1.9 \times 10^4$ – $9.9 \times 10^5$  PPL depending on corpus, and DBAF recovers it to 112–3089 (reduction  $36\times$  to  $4420\times$ , 40 windows per corpus). We report W3 only as evidence of direction; the operating regime of this paper is W4.

**Gate decisions are stable under shift.** Re-measuring the diagnostic itself on shifted calibration data moves few decisions. Against the WikiText-2 baseline (58.7% activation gate-pass over 225 layers), agreement is 95.6% on Dolly-style instructions, 91.6% on Alpaca-style instructions, and 88.0% on multilingual text, with gate-pass rates of 58.7%, 52.9% and 67.1% respectively. The PCSA site compactness stays above the  $c \leq 0.4$  fire threshold in every case ( $c = 0.854 \pm 0.026$ ,  $0.922 \pm 0.027$ ,  $0.595 \pm 0.011$ ), so the LLaMA  $q$ -projection self-disable of §4.4 is not specific to the calibration corpus.

Multilingual text is the largest mover on both axes — lowest decision agreement (88.0%) and lowest compactness (0.595, still a SKIP, but by the smallest margin). This is the expected direction: a corpus with more heterogeneous token statistics both shifts more tensors across the gate boundary and produces tighter prompt clusters.

## G Isolating DBAF from Rotation

The rotation cliff of §4.2 is measured across hosts that differ in many ways besides rotation (GPTQ’s reconstruction, AWQ’s scaling, QuaRot’s Hadamard transform, TesseraQ’s training). A reviewer could reasonably ask whether DBAF’s gains are simply a weaker substitute for rotation, rather than a mechanism that composes with it. We therefore run the controlled  $2 \times 2$ : one base quantizer (RTN), one model (LLaMA-3-8B at W4A4), rotation on/off, DBAF on/off, with everything else in the pipeline held fixed.

|             | no DBAF | +DBAF        |
|-------------|---------|--------------|
| no rotation | 854.06  | 16.93        |
| +rotation   | 13.77   | <b>10.41</b> |

Table 11: RTN on LLaMA-3-8B W4A4, WikiText-2 PPL (16 windows,  $\alpha = 0.25$ , seeded random orthogonal rotation per input dimension; FP reference 6.35). Rotation and DBAF each rescue the catastrophic RTN baseline on their own, and composing them improves on either alone.

Three readings follow. First, the two mechanisms are individually sufficient: rotation alone takes  $854 \rightarrow 13.8$ , DBAF alone takes  $854 \rightarrow 16.9$ . Second, they are not redundant — together they reach 10.4, better than either singleton. Third, and consistent with the cliff, DBAF’s *marginal* contribution shrinks sharply once rotation is present: it is worth 837 PPL without rotation and 3.4 PPL

with it. That is the same interaction the gate predicts, observed here with a single base quantizer and a single model rather than inferred across six heterogeneous hosts.

The residual 3.4 PPL gain under rotation is smaller than the saturation observed on QuaRot in Table 4 (8.41  $\rightarrow$  8.44). We attribute this to the rotation used here being a seeded random orthogonal transform rather than QuaRot’s optimized Hadamard construction: a weaker rotation absorbs less outlier mass and leaves more for DBAF to remove. The direction of the effect is what the diagnostic predicts; the exact residual depends on how completely the host’s reparameterization has already addressed the pathology.

## H Is the PCSA Descriptor Load-Bearing?

PCSA measures compactness on a specific prompt descriptor: the  $\ell_2$ -normalized mean-pooled hidden state at the application site (§3.3). That is one choice among many, and the compactness  $c$  — and therefore the FIRE/SKIP decision — is defined relative to it. If a different descriptor yielded systematically tighter clusters, the gate’s decisions would change and the diagnostic would be measuring the descriptor rather than the model. We test this at both endpoints of the paper’s compactness range.

### The SKIP endpoint (LLaMA-3-8B $q$ -projection).

We compare five descriptors at two decoder layers, under a homogeneous (WikiText) and a multi-domain (English/web/code) prompt regime,  $K = 4$ , 60 prompts per regime.

| Descriptor       | homogeneous        |       | multi-domain |       |
|------------------|--------------------|-------|--------------|-------|
|                  | L0                 | L15   | L0           | L15   |
| Mean-pool (ours) | 0.874              | 0.825 | 0.721        | 0.884 |
| Max-pool         | 0.975              | 0.946 | 0.902        | 0.914 |
| Attention-pool   | 0.885              | 0.889 | 0.642        | 0.919 |
| Sentence-BERT    | 0.826              | 0.826 | 0.859        | 0.859 |
| Last-token (CLS) | 0.000 <sup>†</sup> | 0.929 | 0.642        | 0.939 |

Table 12: PCSA compactness  $c$  at the LLaMA-3-8B  $q$ -projection across descriptors ( $c \leq 0.4$  fires). 19 of the 20 cells SKIP, agreeing with the paper’s decision. <sup>†</sup>Last-token pooling at layer 0 collapses to  $c = 0$  (zero within-cluster distance): at the first layer the last-token hidden state is a near-constant function of the start token, so all descriptors coincide. This is a degenerate artifact rather than real cluster structure, and is the reason we pool over tokens.

Four of five descriptor families preserve the SKIP decision in every cell; only the degenerate

case above flips, and it is diagnosable from its zero variance. In the multi-domain regime all five recover the three-way domain split at cluster purity 0.68–1.00 against a majority baseline of  $\approx 1/3$ , so the descriptors do capture real structure — there simply is not enough of it at this site to make per-cluster scales worthwhile.

### The FIRE endpoint (SAM-B mask decoder).

The same question at the site where PCSA does fire. Four descriptors, two protocols (synthetic prompts and real COCO images):

| Descriptor       | synthetic | real COCO |
|------------------|-----------|-----------|
| Mean-pool (ours) | 0.184     | 0.396     |
| Last-token       | 0.181     | 0.252     |
| Max-pool         | 0.273     | 0.191     |
| Attention-pool   | 0.314     | 0.286     |

Table 13: PCSA compactness at the SAM-B mask-decoder  $q$ -projection. All eight cells FIRE ( $c \leq 0.4$ ), so the positive decision is also descriptor-robust. Note that real COCO prompts give a looser  $c$  than synthetic ones under the paper’s descriptor (0.396 vs 0.184) — still firing, but close to the threshold.

Taken together, 27 of the 28 cells across both endpoints reproduce the dispatch reported in §4.4. The FIRE decision at SAM-B and the SKIP decision at LLaMA are properties of the activation geometry at those sites, not of the particular pooling rule. The honest caveat is the SAM-B real-COCO cell at  $c = 0.396$ : the firing margin there is 0.004, so on real data that decision is close to the boundary even though it lands on the correct side.

## I Beyond NLP and Vision: Three Additional Families

The paper’s hypothesis is that PTQ method choice tracks tensor distribution rather than architecture. The families in Section 4 — LLMs, SAM, super-resolution transformers — were chosen because prior work documents the dense-core-with-outliers pathology in each. A sharper test is whether the diagnostic behaves sensibly on families we did not design around. We therefore ran three: CLIP-ViT-L/14 (multimodal), Whisper-small (speech) and DiT-XL/2 (diffusion), each on a weight-only RTN host, with the gate thresholds frozen at their LLaMA-3-8B values and  $\alpha$  selected by the reconstruction sweep of §3.5.

### The gate reads distributions, not domains.

| Model         | wt. kurt. | wt. pass | act. pass |
|---------------|-----------|----------|-----------|
| CLIP-ViT-L/14 | 4.8       | 97.3%    | 27.9%     |
| Whisper-small | 6.7       | 93.8%    | 31.8%     |
| DiT-XL/2      | 956.3     | 69.8%    | 34.1%     |

Table 14: Outlier-gate statistics on three families outside the paper’s scope. CLIP and Whisper sit inside the calibrated kurtosis band  $[3, 30]$  and pass at 94–97%; DiT is two orders of magnitude outside it and the gate self-disables on 30% of weight tensors. No threshold was refit.

**CLIP and Whisper: the gate fires and the dispatch holds.** On CLIP, W4 zero-shot CIFAR-100 (Krizhevsky and Hinton, 2009) top-1 goes  $61.8 \rightarrow 71.7$  against an FP reference of 73.7; the gated recipe (71.7) slightly exceeds forcing DBAF on every layer (71.2), i.e. the gate’s skips are not costing accuracy. On Whisper, W4 WER goes  $4.43 \rightarrow 4.36$  (FP 3.74), a small gain in a regime where RTN is already close to FP; at W3, where the pathology bites, WER goes  $135.8 \rightarrow 18.8$  gated. In both families the PCSA compactness at the  $q$ -projection is high ( $c=0.794$  CLIP,  $0.514$  Whisper) and PCSA correctly self-disables, matching the LLaMA  $q$ -projection behaviour of §4.4.

**DiT: the primitive transfers, the gate is conservative.** DBAF improves DiT: at the trajectory-selected  $\alpha=0.75$ , FID-to-FP over 512 samples at 25 sampling steps improves  $242.8 \rightarrow 185.7$ , and per-step reconstruction NMSE improves  $0.297 \rightarrow 0.097$ . Two caveats come with it, and they point in different directions.

First, the selection objective must match the inference procedure. A single-pass reconstruction sweep selects  $\alpha \approx 0.3$ , which minimizes single-pass NMSE but *degrades* FID ( $242.8 \rightarrow 275.1$ ); the terminal-trajectory NMSE — still a calibration-time quantity, computed on noise/class inputs with no access to evaluation images or FID — selects  $\alpha=0.75$ , which FID then independently confirms. Iterative generation compounds outlier-position error across steps, so the one-block proxy of §3.5 is the wrong objective for this family even though the diagnostic itself is unchanged.

Second, and more importantly for the diagnostic, the gate is too conservative here. Forcing DBAF on all layers gives NMSE 0.097; the gated recipe gives 0.299, indistinguishable from RTN’s 0.297. The 30% of tensors the gate skips are ones the primitive would have helped. This is the behaviour §1 predicts for distributions outside the calibrated

regime — DiT’s weight kurtosis of 956 is far outside the  $[3, 30]$  band the gate was calibrated on, so it declines to fire — but unlike the SwinIR- $\times 3$  skip of §4.5, where forcing DBAF also degrades quality and the skip is therefore correct, here the skip forgoes a real gain.

The same conservatism appears on CLIP when the bit-width rather than the distribution moves outside the calibrated regime: at W3, forcing DBAF gives 16.8 top-1 against 3.3 gated (RTN 1.7), where at W4 the gated recipe was the better of the two. The gate was calibrated at W4 (§1), and at W3 a larger fraction of tensors falls outside the band it recognizes.

**Closed-form  $\alpha^*$  across the five weight paths.** Evaluated on weights, Eq. 2 gives  $\alpha_W^* = 0.28/0.28/0.33/0.32/0.30$  on LLaMA-3-8B, Qwen-2.5-7B, CLIP-ViT-L/14, Whisper-small and DiT-XL, against sweep-selected values of  $0.25/0.25/0.30/0.25/0.30$  — within 26% on every family, and within 13% on four of five. Selecting at  $\alpha^*$  directly is nonetheless worse than sweeping on the task metric (CLIP 70.2 vs. 71.4 top-1; Whisper 20.0 vs. 17.3 WER), which is why §3.5 uses it only to seed the sweep.

**Why the gate is conservative by construction.** This is not an implementation defect, and DiT is the case that makes the reason legible. The quantity the gate tests and the quantity that governs DBAF’s benefit are not the same. The MSE proxy of §3.2 and its minimizer  $\alpha^*$  (Eq. 2) are functions of  $(T, M, p_{\text{out}})$  alone: neither involves kurtosis or skewness. Those two statistics enter the gate as *shape* tests — evidence that a tensor is the dense-core-with-outliers pattern the fold was designed for — not as conditions on whether folding reduces error. They are therefore sufficient but not necessary.

DiT separates the two precisely. Its mean weight outlier fraction is 0.0095, comfortably inside the gate’s  $[10^{-4}, 2 \times 10^{-2}]$  band, so the criterion the benefit model actually uses is satisfied; its kurtosis of 956 fails a criterion the benefit model never consults. The gate declines, and the fold would have helped. The same reasoning predicts the direction of the error everywhere: because the shape tests can only *remove* tensors the benefit model would have accepted, the gate’s mistakes are skips rather than spurious fires.

False negatives are therefore not exclusive to off-distribution tensors — what changes is their cost. Inside the calibrated family, forcing DBAF

on the tensors the gate rejects recovers a further 0.4–4.6% MSE reduction depending on configuration (Table 25), a margin small enough that the gated recipe still matches or beats forcing on task metrics: on CLIP at W4, gating scores 71.7 top-1 against 71.2 forced. On DiT the same mechanism costs a factor of three. The gate’s conservatism is constant; only its price changes with distance from the calibrated pattern.

Closing the gap would mean widening the kurtosis bound for heavy-tailed families, or replacing the shape tests with a criterion derived from the benefit model itself — a distribution-specific extension of the kind the Limitations section anticipates, rather than a repair to the primitives.

## J Qwen-2.5-7B W4A4 Host Matrix

Table 15 reports the five-host Qwen-2.5-7B W4A4 matrix at  $\alpha = 0.25$ , using the same protocol as Table 4 in the main paper. Qwen activations are  $10\times$  more leptokurtic than LLaMA across 196 Linear layers (mean kurtosis 961 vs. 95; max  $5.8\times 10^4$  vs.  $5.5\times 10^3$ ). Despite a similar  $\Pr(|x| > 3\sigma) \approx 1.2\%$ , three of four training-free hosts fail at W4A4; only SmoothQuant’s per-channel migration and FlatQuant’s orthogonal rotation recover usable quality. FlatQuant replaces QuaRot because QuaRot’s reference implementation does not support Qwen; both serve the same outlier-flattening role in this comparison.

| Host             | Rot. | Train. | Wiki ↓      | C4 ↓         |
|------------------|------|--------|-------------|--------------|
| RTN+DBAF         | ✗    | ✗      | 55.53       | 83.00        |
| GPTQ+DBAF        | ✗    | ✗      | NaN*        | NaN*         |
| AWQ+DBAF         | ✗    | ✗      | 113.86      | 196.64       |
| SmoothQuant+DBAF | ✗    | ✗      | 18.27       | 32.58        |
| FlatQuant+DBAF   | ✓    | ✗      | <b>7.31</b> | <b>12.52</b> |

Table 15: Qwen-2.5-7B W4A4 host matrix ( $\alpha = 0.25$ ). Three of four training-free hosts fail due to Qwen’s extreme activation kurtosis; only SmoothQuant (18.27 wt2) and FlatQuant (7.31 wt2) recover usable quality. \*Hessian inversion diverged at calibration (also NaN at  $\alpha = 0.75$ ).

## K Cross-Detector AHCPTQ Transfer

To test that DBAF+PCSA do not overfit to a particular downstream detector, we evaluated AHCPTQ+DBAF+PCSA on SAM-B/L/H against the calibration detector YOLOX: SAM-B 18.22 mAP, SAM-L 33.4, SAM-H 35.3 (W4A4 segm mAP on COCO val2017). All three are

| Method                          | SAM-B W4A4 mAP |
|---------------------------------|----------------|
| AHCPTQ baseline                 | 13.4           |
| + DBAF (1% outliers)            | 15.6           |
| + DBAF (2% outliers)            | 16.3           |
| + PCSA ( $k = 2$ )              | 13.1           |
| + PCSA ( $k = 4$ )              | 13.8           |
| + PCSA ( $k = 8$ )              | 13.6           |
| + DBAF + PCSA ( $k = 4$ , ours) | <b>18.2</b>    |

Table 16: DBAF/PCSA ablation on SAM-B under W4A4 quantization with YOLOX-generated prompts (COCO segm mAP). DBAF and PCSA each contribute, but their joint application exceeds the sum of individual gains, consistent with DBAF normalizing the per-tensor heavy tail before PCSA routes per prompt.

run in this work, over the Zhang et al. (2025) Table 1 baselines. Re-running the recipe with the unseen detector H-DETR on the same hosts where the matched run is available yields SAM-B 20.6 mAP and SAM-H 35.4 mAP at W4A4 (within 0.1 of the YOLOX cell for SAM-H, and above the YOLOX cell for SAM-B), confirming the joint DBAF+PCSA recipe is robust to the downstream detector. SAM-L + H-DETR was not re-evaluated in this study.

## L DBAF/PCSA Ablation on SAM-B

We isolate the contribution of each primitive on AHCPTQ-hosted SAM-B under W4A4 with YOLOX-generated prompts (Table 16). DBAF alone closes most of the gap to the full method by addressing dense-core-with-outlier activations; PCSA alone yields a smaller improvement because it only adapts to prompt-conditioned scale shifts and cannot reshape the underlying distribution. When applied together, the two primitives combine non-additively: joint mAP (18.2) exceeds the sum of their individual improvements ( $+2.9 + 0.4 = +3.3$ ), with a +4.8 net gain. This indicates that each primitive removes a distinct source of error and that PCSA’s per-prompt routing is more effective once DBAF has tamed the heavy tail. Figure 3 visualizes the ablation alongside a sweep over PCSA anchors  $k \in \{2, 4, 8\}$  showing  $k = 4$  is optimal, prompt-conditioned activation distributions are diverse but not heterogeneous enough to require many anchors.

**SAM-L and SAM-H joint composability.** At larger SAM scales we measure only the joint DBAF+PCSA cell; the singleton ablations at those scales were not run. The joint gain narrows monotonically with model scale (Table 17), consistent

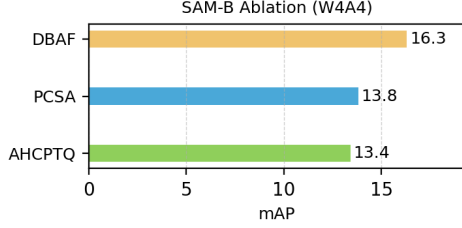


Figure 3: Bar visualization of Table 16. SAM-B W4A4 with YOLOX prompts; PCSA uses  $k = 4$  anchors.

with the AHCPTQ vanilla baseline approaching the SAM-H mAP ceiling: SAM-L gains +2.4 mAP while SAM-H, already within 2 mAP of its FP target at vanilla, gains only +0.1 mAP. The compactness gate fires on all three scales (Table 25), so the SAM-H saturation is a ceiling effect on the host metric rather than a gate self-disable.

| Method           | SAM-L mAP   | SAM-H mAP   |
|------------------|-------------|-------------|
| AHCPTQ (vanilla) | 31.0        | 35.2        |
| + DBAF + PCSA    | <b>33.4</b> | <b>35.3</b> |
| $\Delta$         | +2.4        | +0.1        |

Table 17: AHCPTQ+DBAF+PCSA on SAM-L and SAM-H at W4A4 (YOLOX segmAP, COCO val2017). Vanilla numbers from Zhang et al. (2025), Table 1.

## M Per-Input Scale Variation: SAM vs. LLM

For each Linear input we recorded the per-input maximum absolute activation value across 20 Set5/COCO images (SAM-B) and 20 WikiText-2 prompts truncated to 512 tokens (LLaMA-3-8B), then computed the coefficient of variation  $CV = \sigma_{\text{input}} / \mu_{\text{input}}$  per layer and aggregated. The aggregate CVs (Table 18) show that LLM activations actually vary more per input than SAM activations in absolute terms, which initially looks like the wrong direction for the PCSA-helps-SAM finding of §4.4.

**$K$ -means tractability.** To test whether the per-input variation is structured enough to cluster at PCSA’s resolution, we ran  $K$ -means on  $\ell_2$ -normalized mean-pooled descriptors at each model’s actual PCSA application site: SAM-B’s mask-decoder cross\_attn\_token\_to\_image.q\_proj input over 50 synthetic images (Gaussian, spanning input scales) with 2–4 random point

| Model      | Lay. | mean  | med.  | IQR          |
|------------|------|-------|-------|--------------|
| SAM-B      | 48   | 0.037 | 0.030 | [0.02, 0.05] |
| LLaMA-3-8B | 224  | 0.112 | 0.137 | [0.00, 0.17] |

Table 18: Per-input max-abs CV on 20 inputs per model (layers = number of Linears sampled; mean/median CV and 25/75 IQR over those layers).

|            | $K=2$ |       | $K=4$        |       | $K=8$        |       |
|------------|-------|-------|--------------|-------|--------------|-------|
|            | p.    | g.    | p.           | g.    | p.           | g.    |
| SAM-B      | 0.228 | 0.221 | <b>0.189</b> | 0.171 | 0.159        | 0.138 |
| LLaMA-3-8B | 0.933 | 0.932 | 0.866        | 0.860 | <b>0.790</b> | 0.781 |

Table 19:  $K$ -means compactness ratio  $d_{\text{real}}/d_{\text{baseline}}$  across  $K \in \{2, 4, 8\}$  and both baselines. Lower is more clusterable; 1.0 means no structure beyond marginals. Bold cells are PCSA’s configured  $K$  ( $K=4$  for SAM,  $K=8$  for LLM). At PCSA’s actual configurations SAM-B clusters at  $0.189 \pm 0.003$  versus LLM at  $0.790 \pm 0.001$ , a  $4.2\times$  gap. Both baselines agree directionally. Standard deviations are  $\leq 0.007$  everywhere.

prompts each, and LLaMA-3-8B’s self-attention q\_proj input across all 32 decoder layers over 50 WikiText-2 prompts. We compared mean within-cluster distance  $d_{\text{real}}$  to two baselines averaged over 20 seeds each: permute-dims (destroys cross-feature structure, keeps marginals) and isotropic Gaussian (matched mean/std). The compactness ratio  $d_{\text{real}}/d_{\text{baseline}}$  is lower when real clusters are tighter; 1.0 means no structure beyond the baseline. We sweep  $K \in \{2, 4, 8\}$  to cover PCSA’s configured  $K$  on both sides.

The asymmetry is robust to  $K$ : SAM-B is tighter than LLM at all three  $K$  values by at least a  $3\times$  margin. Variation magnitude is necessary but not sufficient for PCSA: SAM’s content-driven activation distribution provides  $K$ -means tractability under PCSA-tf’s mean-pooled descriptor while LLM’s per-prompt variation does not, consistent with the eight PCSA-tf recipe non-firings reported in §4.4 (Appendix O).

*Input protocol.* The clusterability probe above uses synthetic images to isolate the site’s structural clusterability across input scales. On real COCO images with object-center point prompts, the SAM-B site ( $K=4$ ) measures  $c \approx 0.40$  under the same mean-pooled descriptor—within, though near, the  $c \leq 0.4$  fire threshold—while the LLM site stays well above it; the PCSA fire decision, and the +4.8 mAP gain, are on real COCO data.

**Per-layer compactness control.** To rule out the concern that the cross-layer-pool gap reflects cherry-picked layer selection, we measured per-layer compactness within each layer independently across three architectures. SAM-B mask-decoder cross-attn layers:  $[0.69, 0.70]$ , median 0.69 ( $n=2$  layers). LLaMA-3-8B self-attention  $q_{\text{proj}}$  layers:  $[0.67, 0.82]$ , median 0.73 ( $n=32$  layers). Qwen-2.5-7B self-attention  $q_{\text{proj}}$  layers:  $[0.70, 0.84]$ , median 0.78 ( $n=28$  layers). The two LLMs are comparable to each other in median (0.73 vs 0.78) and both medians sit well above SAM-B’s (0.69); the per-layer ranges overlap modestly (LLaMA’s most clusterable layer reaches 0.67, slightly below SAM’s worst at 0.70), so the  $4.2\times$  aggregate gap comes from a shift in the median rather than from uniform layer-by-layer dominance. This is the property PCSA’s per-prompt routing actually needs: a descriptor identifying a cluster that remains the prompt’s cluster at every layer the routing controls. The per-layer measurements for all three architectures are in our release artifacts (`cluster_tractability_per_layer.json` and `cluster_tractability_qwen.json` in the `results/` directory).

**CV and compactness are statistically independent.** To confirm the two metrics measure different properties of a descriptor distribution, we constructed two synthetic datasets on  $\mathbb{R}^{256}$  with  $N=800$  unit-norm vectors: (A) a 4-mode Gaussian mixture with well-separated means and per-mode  $\sigma=0.1$ , and (B) isotropic Gaussian noise, with B rescaled to match A’s mean per-feature standard deviation. (A) has cluster structure that K-means should recover; (B) does not. At  $K=4$  the compactness ratios are  $0.889 \pm 0.0002$  for (A) and  $1.000 \pm 0.0004$  for (B), discriminating the two distributions even though both have comparable per-feature standard deviations. CV alone gives no signal that distinguishes structured from random data; the compactness ratio does.

**PCSA gain increases with descriptor clusterability.** We further test whether PCSA’s quantization benefit is structurally tied to clusterability by constructing synthetic activation distributions with controlled mode separation. We sample  $N=1000$  inputs from  $K=4$  modes on  $\mathbb{R}^{256}$ , vary mode separation  $S \in [0, 32]$ , and at each  $S$  measure both the compactness ratio of per-input mean-pooled descriptors and the INT4 quantization MSE reduction of per-anchor scale routing ( $K$ -means anchors with

| Mode sep. $S$ | Compactness ratio  | PCSA gain (%) |
|---------------|--------------------|---------------|
| 0.0           | $1.000 \pm 0.0002$ | <b>0.0</b>    |
| 0.5           | $0.972 \pm 0.0004$ | 7.0           |
| 1.0           | $0.889 \pm 0.0003$ | 9.1           |
| 2.0           | $0.680 \pm 0.0003$ | 10.8          |
| 4.0           | $0.419 \pm 0.0005$ | 7.6           |
| 8.0           | $0.223 \pm 0.0001$ | 11.1          |
| 16.0          | $0.119 \pm 0.0001$ | <b>11.5</b>   |
| 32.0          | $0.057 \pm 0.0001$ | 8.9           |

Table 20: Synthetic clusterability sweep: PCSA’s INT4 quantization benefit (relative MSE reduction over a global scale) tracks descriptor compactness ratio across  $S \in [0, 32]$ . At  $S=0$  (no cluster structure, compactness 1.0) PCSA gain is exactly 0%; as clusters separate, compactness drops and PCSA gain rises. This is a constructive existence proof: PCSA’s mechanism requires  $K$ -means tractable descriptors, and within that regime its quantization benefit increases with the structure the compactness ratio measures. The trend is not monotonic: the gain dips at  $S=4$  (7.6%) and  $S=32$  (8.9%) against the 11.5% peak at  $S=16$ , so the relationship is directional rather than strictly ordered.

$K=K_{\text{true}}=4$ ) over a single global max-abs scale.

**Per-token activation baseline (G1).** The sweep above compares PCSA’s per-anchor scale against a global per-tensor scale (the activation-quantization mode used by FlatQuant fused W4A4, torchao (Or et al., 2025) W4A4 INT4 GEMMs, and the RTN/GPTQ/AWQ training-free hosts in Table 35). For completeness we also evaluate a per-token activation baseline (one max-abs scale per (input, token) row, the mode used by rotation-based hosts that fuse a Hadamard into the INT4 GEMM): PCSA’s relative MSE versus per-token is  $-202.4\%$  on average across  $S \in [0.5, 32]$  (Table 21). The synthetic dataset gives per-token a trivial advantage because each input’s  $B$  tokens share a single cluster magnitude by construction, so the per-token max-abs recovers the cluster scale without descriptor routing. This matches the rotation-saturation finding in the main text (§4.4): when the host already exposes per-row scale information (whether through rotation or dynamic per-token reduction), PCSA’s static per-anchor routing offers no additional MSE benefit. PCSA’s contribution is therefore for the static-scale W4A4 deployment regime, which constitutes the FlatQuant fused, torchao, and training-free RTN/GPTQ/AWQ paths reported in Tables 35 and 37.

**Capacity matters:**  $K_{\text{calib}} \geq K_{\text{true}}$ . The sweep above fixes  $K_{\text{calib}} = K_{\text{true}} = 4$ . To test what hap-

| $S$  | Comp. | MSE <sub>ten</sub> | MSE <sub>tok</sub> | MSE <sub>anc</sub> | vs. tok. |
|------|-------|--------------------|--------------------|--------------------|----------|
| 0.5  | 0.972 | 0.0000             | 0.0000             | 0.0000             | −182%    |
| 1.0  | 0.889 | 0.0002             | 0.0001             | 0.0002             | −208%    |
| 2.0  | 0.680 | 0.0008             | 0.0002             | 0.0007             | −190%    |
| 4.0  | 0.419 | 0.0032             | 0.0010             | 0.0029             | −194%    |
| 8.0  | 0.223 | 0.0147             | 0.0040             | 0.0130             | −225%    |
| 16.0 | 0.119 | 0.0556             | 0.0159             | 0.0492             | −209%    |
| 32.0 | 0.057 | 0.2170             | 0.0639             | 0.1977             | −209%    |

Table 21: G1: synthetic clusterability sweep with per-token activation baseline added. PCSA per-anchor MSE is reported against both per-tensor (the static W4A4 INT4 GEMM mode, last column of Table 20, positive gain) and per-token (dynamic mode used by rotation-based hosts, last column here, negative gain). The synthetic construction gives per-token a structural advantage: each input’s  $B$  tokens share a single cluster magnitude, so per-token max-abs reduces to per-cluster max-abs at finer granularity. The result is consistent with the rotation-saturation finding (Appendix T): when the W4A4 INT4 GEMM exposes per-row scale information, PCSA’s static routing offers no additional MSE benefit; when it does not (FlatQuant fused, torchao, RTN/GPTQ/AWQ training-free), PCSA is the static-compatible mechanism that captures cluster structure.

pens when PCSA’s allocated capacity is below the data’s true mode count, we generate the same high-separation ( $S=8$ ) regime but with  $K_{\text{true}}=8$  modes and calibrate PCSA with  $K_{\text{calib}}=4$  anchors. The compactness ratio at  $K=4$  rises to 0.69 (PCSA can only match 4 of the 8 modes), and the PCSA quantization gain drops from 12.1% to 4.3%, a 3 $\times$  reduction. This is the empirical analogue of the SAM K-sweep (Table 16): once  $K_{\text{calib}}$  matches the natural cluster count  $K_{\text{true}}=4$  on SAM-B prompt descriptors, additional anchors stop helping ( $K=8$  on SAM-B gives 13.6 mAP versus  $K=4$ ’s 13.8); when  $K_{\text{calib}}$  is below  $K_{\text{true}}$ , capacity is the bottleneck rather than the descriptor.

## N DBAF $\alpha$ Sensitivity Across Four Training-Free Hosts

Table 7 reports the  $\alpha$ -sweep on RTN+DBAF only. To rule out a host-specific artifact, we additionally re-evaluate GPTQ, AWQ, and SmoothQuant at  $\alpha \in \{0.5, 0.75\}$  on LLaMA-3-8B W4A4 (the  $\alpha=0.25$  cells are already in Table 4). All cells use the same fold threshold  $T=3\sigma$  and the same per-tensor activation quantization wrapper.

## O LLM PCSA-tf Failure Catalog

The eight PCSA-tf variants summarized in §4.4 are listed in Table 23 with both wt2 and c4 PPL.

| Host        | $\alpha=0.25$ | $\alpha=0.5$ | $\alpha=0.75$ | Ratio         |
|-------------|---------------|--------------|---------------|---------------|
| RTN         | <b>16.31</b>  | 43.08        | 250.60        | 15.4 $\times$ |
| GPTQ        | <b>32.65</b>  | 104.10       | 348.66        | 10.7 $\times$ |
| AWQ         | <b>17.02</b>  | 73.19        | 579.96        | 34.1 $\times$ |
| SmoothQuant | <b>17.32</b>  | 121.96       | 1,477.98      | 85 $\times$   |

Table 22: Cross-host DBAF  $\alpha$ -sensitivity, LLaMA-3-8B W4A4, WikiText-2 PPL. “Ratio” is wt2 degradation from  $\alpha=0.25 \rightarrow 0.75$ . Every host minimizes at  $\alpha=0.25$  and degrades monotonically; the slope spans roughly an order of magnitude (GPTQ tolerates the loosest fold at 10.7 $\times$ ; SmoothQuant degrades fastest at 85 $\times$  because per-channel migration amplifies fold-induced noise on outlier positions). The agreement rules out an RTN-specific artifact and confirms  $\alpha=0.25$  is the right W4A4 LLM default regardless of weight-side mechanism.

| Variant                                   | wt2 $\downarrow$  | c4 $\downarrow$   |
|-------------------------------------------|-------------------|-------------------|
| RTN+DBAF+PCSA-tf ( $\alpha=0.75$ )        | <b>653</b>        | <b>700</b>        |
| GPTQ+DBAF+PCSA-tf ( $\alpha=0.75$ )       | $1.1 \times 10^3$ | $1.2 \times 10^3$ |
| AWQ+DBAF+PCSA-tf ( $\alpha=0.75$ )        | $2.6 \times 10^3$ | $2.3 \times 10^3$ |
| RTN+PCSA-tf ( $\alpha=0.75$ )             | 891               | 967               |
| GPTQ+PCSA-tf ( $\alpha=0.75$ )            | $1.8 \times 10^3$ | $2.1 \times 10^3$ |
| AWQ+PCSA-tf ( $\alpha=0.75$ )             | $6.7 \times 10^3$ | $4.8 \times 10^3$ |
| RTN+DBAF+PCSA-tf ( $\alpha=0.25$ , def.)  | $2.9 \times 10^4$ | $2.3 \times 10^4$ |
| RTN+DBAF+PCSA-tf ( $\alpha=0.25$ , fold.) | $5.5 \times 10^4$ | $5.5 \times 10^4$ |
| W4A16 baseline (RTN, no act quant)        | 10.3              | 17.1              |

Table 23: LLaMA-3-8B W4A4 PCSA-tf failure catalog (compactness  $c=0.79$  at Q-projection). All eight variants exceed 600 wt2 PPL; best 653 is 63 $\times$  worse than the W4A16 baseline.

Variants cross host  $\in \{\text{RTN}, \text{GPTQ}, \text{AWQ}\}$ , DBAF prefold  $\in \{\text{on}, \text{off}\}$ , and fold-calibration method  $\in \{\text{default}, \text{folded}\}$ . Every variant exceeds 600 wt2 PPL against a 10.28 W4A16 baseline; ordering preserves the across-host per-channel scaling intensity (RTN < GPTQ < AWQ) since per-channel host scaling amplifies the per-prompt-scale mismatch PCSA-tf chases. At  $\alpha=0.25$  both calibration modes are catastrophic: the  $1/\alpha=4\times$  unfold amplifies per-prompt quantization noise past DBAF’s outlier-recovery budget.

## P Long-Context KV-PCSA

The KV-PCSA variant (§3) augments the trained FlatQuant+DBAF+PCSA baseline with per-anchor multiplicative corrections on the quantized KV-cache (per-token base scale, per-anchor multiplicative correction; the v1 variant that replaced per-token scaling entirely is dominated by the corrected version v2 used here). We evaluate transfer to long-context retrieval via RULER’s needle-in-

| Context | wt2 PPL ↓   |          | c4 PPL ↓     |              |
|---------|-------------|----------|--------------|--------------|
|         | base        | +KV-PCSA | base         | +KV-PCSA     |
| 2k      | 7.15        | 7.17     | 10.98        | <b>10.95</b> |
| 4k      | <b>6.45</b> | 6.46     | <b>10.48</b> | 10.66        |
| 8k      | <b>6.45</b> | 6.57     | <b>10.69</b> | 11.08        |

Table 24: Long-context perplexity on LLaMA-3-8B W4A4 KV4 (FlatQuant host). KV-PCSA’s effect on average-loss metrics stays inside  $\pm 0.4$  PPL across context lengths and is direction-mixed.

a-haystack (NIAH) accuracy at 4k and 8k context lengths on LLaMA-3-8B W4A4 KV4 (calibration: 128 WikiText-2 sequences, 200 NIAH evaluation prompts per context length).

The NIAH retrieval gain (+5pp at 8k context) is reported in the main paper as Table 6; this section focuses on the per-context-length PPL comparison (Table 24) above. The PPL-vs-NIAH gap mirrors the LLM-activations finding: capability metrics expose input-conditioned scale drift that average-loss metrics smear away. KV-PCSA-tf targets the per-anchor structure of attention rather than per-token or per-head scales, an axis orthogonal to rotation-based treatments.

## Q Per-Tensor Distributional Statistics

Table 25 catalogs the per-tensor distributional statistics across the eight model configurations evaluated in the main paper. Weight gate-pass is 94–100% on all hosts, confirming a shared dense-core-with-outliers substrate; mean kurtosis stays in 4–12 and  $\text{frac} > 3\sigma$  in 0.6–0.9%. Activation gate-pass varies; Qwen-2.5-7B is the low outlier (8.2%) because its per-channel asymmetric mass fails the symmetric-skew gate.

| Model              | W gate | Kurt  | frac $> 3\sigma$ | A gate |
|--------------------|--------|-------|------------------|--------|
| LLaMA-3-8B         | 100.0% | 4.47  | 0.59%            | 62.5%  |
| Qwen-2.5-7B        | 99.5%  | 5.48  | 0.68%            | 8.2%   |
| SAM-B              | 94.1%  | 12.33 | 0.83%            | 23.5%  |
| SAM-L              | 97.0%  | 10.02 | 0.75%            | 27.3%  |
| SAM-H              | 97.7%  | 5.64  | 0.76%            | 21.4%  |
| SwinIR- $\times 2$ | 97.1%  | 4.96  | 0.67%            | 61.2%  |
| SwinIR- $\times 3$ | 97.1%  | 4.20  | 0.67%            | 36.9%  |
| SwinIR- $\times 4$ | 96.1%  | 4.98  | 0.64%            | 43.7%  |

Table 25: Per-tensor distributional statistics across eight model configurations.

## R SwinIR-light Training-Free W4A16 Deltas

Table 26 reports Set5 and Urban100 PSNR deltas for RTN+DBAF vs. RTN alone on SwinIR-light at all three scales, alongside the activation-side gate-pass rate that drives the prediction of §4.5.

| Scale      | Act gate-pass | Set5 $\Delta$ dB | Urban100 $\Delta$ dB |
|------------|---------------|------------------|----------------------|
| $\times 2$ | 61.2%         | +0.16            | +0.07                |
| $\times 3$ | 36.9%         | −0.31            | −0.15                |
| $\times 4$ | 43.7%         | +0.43            | +0.25                |

Table 26: SwinIR-light training-free W4A16 (weight-only) DBAF deltas (RTN+DBAF vs. RTN alone). Act gate-pass is the fraction of activation tensors passing the sparse-outlier gate. The sign of  $\Delta$  dB agrees with the prediction at all three scales: regression at  $\times 3$  (below the 40% threshold), improvement at  $\times 2$  and  $\times 4$  (above it). The threshold is descriptive, not fitted: with three scales any value in (36.9, 43.7) separates them equally well, and the sign prediction of §4.5 comes from the compounded per-layer regression, not from this cut.

**Why the activation-side statistic drives a weight-only prediction.** The weight gate passes at 96–97% at all three SwinIR scales and so separates none of them; only the activation gate varies (61.2/36.9/43.7%), and its ordering tracks the observed outcome ordering. This is the general pattern of §3.1 — the weight gate is near-saturated while the activation gate carries the dispatch — rather than anything specific to super-resolution. Because activations are left in FP16 here, the activation gate-pass acts as a distributional proxy for the layer, not as a measurement of a tensor that is itself quantized. We do not establish a mechanism linking activation statistics to weight-only outcomes; the paragraph below shows only that the same prediction survives when activations *are* quantized.

**The same signs hold at W4A4.** The prediction of §4.5 is stated for weight-only W4A16, the regime in which the SwinIR  $\alpha = 0.95$  operating point was selected. Repeating the comparison with activations also quantized to INT4 (RTN+DBAF with both gates active, versus RTN alone) reproduces all three signs and a larger  $\times 3$  regression: Set5  $\Delta$  dB of +0.14, −0.41 and +0.48 at  $\times 2 / \times 3 / \times 4$ , with Urban100 at +0.05, −0.20 and +0.24. The negative prediction therefore holds under both quantization regimes, and is not an artifact of leaving activations in FP16.

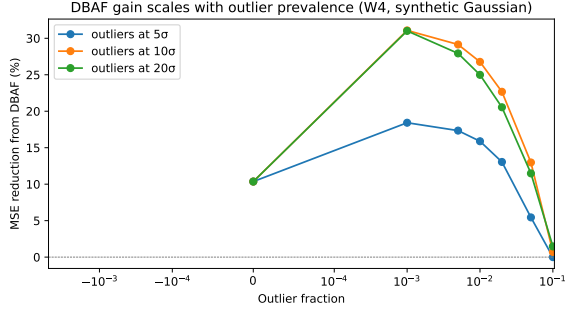


Figure 4: Synthetic study: DBAF MSE reduction vs. injected outlier fraction on Gaussian tensors. Peak gain ( $\sim 31\%$ ) occurs at fractions of  $10^{-3}$ , the regime where real LLM/SAM/SR tensors actually live.

## S Mechanism: Four Causal Experiments (Full)

The composability and saturation results raise the question of whether the observed gain is tied to the dense-with-sparse-outliers distribution, or whether any clipping-like operation would match it. Four controlled experiments isolate the mechanism.

**(1) Synthetic outlier injection (sparse).** We generate Gaussian tensors with controlled outlier injection, vary the outlier fraction from 0% to 10%, and measure DBAF MSE reduction vs. per-row RTN (Figure 4). DBAF gain peaks at outlier fractions of  $10^{-3}$ , the exact regime observed in real models (Table 25), and decays to near zero at 10% when outliers are no longer rare.

**(2) Dense-outlier control.** The synthetic study above injects sparse point outliers. To test the gate’s information-preservation claim we repeat the study with dense heavy-tailed outliers (Student- $t$ , df sweep) at matched outlier fractions. Body-MSE gain stays positive (2–7%) for moderately heavy tails ( $\text{df} \geq 3$ ) but collapses to 0.0% body gain at  $\text{df} = 2.5$ , where outliers are no longer separable from the body (Table 27). This confirms that DBAF’s fold is information-preserving only in the sparse regime; in dense regimes the gate is what prevents body-information damage.

**(3) Per-layer correlation and depth compounding.** Across all 224 nn.Linear layers in LLaMA-3-8B, per-layer outlier fraction correlates with forced-DBAF MSE reduction at Pearson  $r = 0.561$  ( $n = 224$  Linear layers,  $p < 10^{-19}$ ) at  $\alpha = 0.75$  with a per-row  $3\sigma$  threshold. The same correlation measured on the weight path at  $\alpha = 0.95$ , which is the configuration Eq. 3 is fit on, gives  $r = 0.485$

| Family                 | df / mag | frac $> 3\sigma$ | Body gain    |
|------------------------|----------|------------------|--------------|
| Sparse Gauss, 10 out.  | mag = 8  | 0.28%            | +6.04%       |
| Sparse Gauss, 100 out. | mag = 8  | 0.32%            | +6.01%       |
| Dense Student- $t$     | df = 10  | 0.72%            | +6.03%       |
| Dense Student- $t$     | df = 4   | 1.28%            | +2.15%       |
| Dense Student- $t$     | df = 3   | 1.50%            | +2.10%       |
| Dense Student- $t$     | df = 2.5 | 1.29%            | <b>0.00%</b> |

Table 27: Sparse vs. dense outlier regimes. DBAF’s body-MSE gain collapses when the gate’s frac  $> 3\sigma$  constraint would reject the tensor.

| Scale      | Mean $\Delta$ gate-pass | Max $ \Delta $ | All-layer $\Delta$ |
|------------|-------------------------|----------------|--------------------|
| $\times 2$ | +0.0008 dB              | 0.04 dB        | +0.156 dB          |
| $\times 3$ | −0.0026 dB              | 0.04 dB        | <b>−0.315 dB</b>   |
| $\times 4$ | +0.0037 dB              | 0.04 dB        | +0.432 dB          |

Table 28: Per-layer DBAF ablation on SwinIR Set5 (one layer at a time). Single-layer effects are bounded by  $\pm 0.04$  dB but compound across 103 layers.

over the same 224 layers (Appendix E); the two differ in fold strength, not in layer set, and the weaker fold yields both smaller gains (median 3.2% vs. 15.4%) and a looser fit. At  $\alpha = 0.75$ , 100% of layers show positive DBAF gain (Figure 5). The per-layer MSE reduction is itself substantial (median 15.4%, range 11.3–26.4%), but its task-level effect in isolation is tiny; what matters is that it compounds. On SwinIR-light (103 layers), per-layer DBAF deltas are bounded by  $\pm 0.04$  dB; yet applying DBAF to all layers gives +0.16 dB at  $\times 2$ , −0.32 dB at  $\times 3$ , and +0.43 dB at  $\times 4$  (Table 28). No single layer is the “culprit”, compounding depth interaction is the mechanism by which both gains and (in  $\times 3$ ) damages emerge.

**The  $\times 3$  regression was predicted, not discovered.** Before running the per-layer ablation we classified each activation tensor in SwinIR-light using the gate-aware distribution taxonomy of §3. Counts of dense-outlier activation layers came out at 3 for  $\times 2$ , 22 for  $\times 3$ , and 8 for  $\times 4$ , an  $\sim 8\times$  jump at  $\times 3$ . The framework therefore predicted  $\times 3$  as the regime where information-preserving folding should fail before any task-level numbers were measured; the measured −0.315 dB all-layer regression at  $\times 3$  (Table 28), against a compounded prediction of −0.32 dB (§4.5), confirms that prediction. The distributional framework is therefore predictive, not just descriptive.

**(4) Matched- $T$  clipping ablation.** The strongest test: replace DBAF’s affine fold with hard clipping

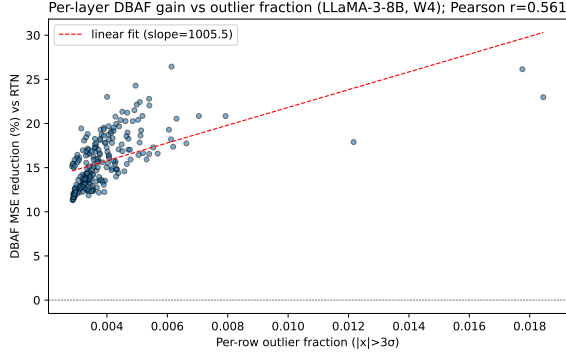


Figure 5: Per-layer DBAF gain vs. outlier fraction on LLaMA-3-8B. Real-weight evidence that DBAF gain tracks outlier prevalence.

| Method                                     | wt2 PPL ↓     |
|--------------------------------------------|---------------|
| RTN (W4, no fold)                          | 10.28         |
| RTN + matched- $T$ clipping ( $\alpha=0$ ) | <b>79,256</b> |
| RTN + DBAF ( $\alpha=0.75$ )               | <b>8.87</b>   |

Table 29: Matched- $T$  clipping vs. DBAF at the same threshold  $T = 3\sigma$  on LLaMA-3-8B W4 weight-only. Hard clipping at  $\alpha = 0$  destroys the model; DBAF at the same threshold with  $\alpha = 0.75$  recovers production-grade PPL. The  $\sim 10^4$  PPL ratio isolates the value of preserving outlier ordering from any dynamic-range-reduction benefit.

at the same threshold  $T = 3\sigma$ . Both reduce dynamic range; only DBAF preserves outlier ordering.

The  $\sim 79k$  PPL gap shows that DBAF’s value lies in folding (preserving the relative ordering of outliers), not in dynamic-range reduction. A formal proof under the MSE-proxy noise model of §3.2 is in Appendix C.

**Summary.** The four experiments establish a causal chain: the dense-with-sparse-outliers distribution recurs across architectures (§4.3); DBAF targets it (experiment 1); the gate prevents damage in the dense regime (experiment 2); per-layer effects are small but compound through depth (experiment 3); and the value is folding rather than clipping (experiment 4). The framework predicts which scale of SwinIR will regress before any task-level evaluation.

## T Saturation Evidence on Rotation-Based Hosts

The reframing of §1 predicts that, on hosts whose own machinery is already absorbing the dense-with-outliers distribution, the marginal gain from adding DBAF+PCSA should collapse, not im-

| Method (LLaMA-3-8B W4A4 KV4)      | WikiText     | C4            |
|-----------------------------------|--------------|---------------|
| FlatQuant (no DBAF, no PCSA)      | 6.964        | 11.158        |
| FlatQuant + DBAF + PCSA (gated)   | 6.966        | 11.143        |
| FlatQuant + DBAF (no-gate) + PCSA | <b>6.910</b> | <b>10.968</b> |

Table 30: Saturation on a rotation-based host. With the standard sparse-outlier gate (§3), DBAF+PCSA add  $\approx 0.002$  PPL on WikiText over pure FlatQuant, statistically indistinguishable. Disabling the gate lets DBAF fire on every layer and recovers a small but consistent improvement ( $-0.054$  WikiText,  $-0.19$  C4).

prove. Two rotation-based hosts let us test this directly: FlatQuant (Sun et al., 2025) for LLMs and CompSRT (Zeinali et al., 2026) for SwinIR.

The pattern in Table 30 is the central saturation finding: FlatQuant’s learned rotations already absorb the sparse-outlier pattern our gate targets, so on this host the gate self-disables and the gated DBAF+PCSA composition is a no-op ( $\Delta = 0.002$  PPL). Removing the gate forces DBAF onto residual outliers that survive the rotation; the resulting gain ( $-0.054$  WikiText,  $-0.19$  C4) is small but consistent across both calibration sets.

SR transformers reach the same endpoint by a different route. On the CompSRT host the gated composition changes Set5 PSNR by  $-0.01$  dB at both  $\times 2$  and  $\times 3$  (Table 32), and across all five  $\times 3$  benchmarks the deltas take both signs and all lie within  $\pm 0.01$  dB ( $-0.008$ ,  $+0.007$ ,  $+0.004$ ,  $-0.010$ ,  $-0.003$ ). Here the gate does *not* self-disable: it fires on 83.6% of decisions at  $\times 2$  and 81.8% at  $\times 3$  ( $n = 126,336$  decisions per scale), because a Hadamard rotation mixes each coefficient into a sum over its row and the resulting shape — a dense core with rare  $3\sigma$  outliers — is close to what the gate selects for. The rotation manufactures the distribution the gate looks for; this is a statement about the gate’s acceptance criterion, not about why rotation aids quantization, which we do not test here. The no-op arises downstream instead: folding at  $T = 3\sigma$  with  $\alpha = 0.95$  touches a median 0.78% of values and removes 5% of their overshoot, which on an already-Gaussian tensor is a near-identity operation. DBAF is applied to a distribution it has nothing left to correct. Bypassing the gate is nonetheless worse on four of the five  $\times 3$  benchmarks, most visibly  $-0.21$  dB on Manga109 (Urban100 is the exception, at  $+0.04$ ), so the gate is doing real work even though the composition is a wash. We read these cases not as a method failure but as a sign that the host is already absorbing the

recurring distribution implicitly, which is what the cross-architecture claim of §1 predicts.

**Rotation moves tensors into the gate’s acceptance band.** Table 26 reports gate-pass rates for the same architecture and the same scales on an *unrotated* RTN host, which makes the comparison like-for-like:

| Scale | Unrotated (RTN) |       | Rotated (CompSRT) |       |
|-------|-----------------|-------|-------------------|-------|
|       | wt.             | act.  | wt.               | act.  |
| ×2    | 96–97%          | 61.2% | 100.0%            | 73.8% |
| ×3    | 96–97%          | 36.9% | 100.0%            | 70.9% |

Rotation takes the weight gate to exactly 100% (47,376 of 47,376 weight-side decisions at each scale) and lifts activation pass by 12.6 and 34.0 percentage points. The ×3 case is the informative one: unrotated it is the only scale below the 40% threshold and the only one where DBAF regresses, and rotating the host moves it to 70.9%. This is the acceptance-band claim of this paragraph, measured rather than inferred. The claim is specifically about what the gate selects for: its criterion is a dense core with rare  $3\sigma$  outliers, and rotation supplies tensors of that shape. We are not claiming that increased Gaussianity is what makes rotation help quantization in general — that is a separate question about quantization error, and one this experiment does not test. The comparison covers the two scales for which we trained a rotated host; ×4 appears in the unrotated column of Table 26 only.

It also scopes that threshold. The rotated ×3 host passes at 70.9% — well above 40%, which would predict improvement — yet its measured delta is  $\approx 0$ . Activation gate-pass therefore predicts the sign of the DBAF delta *within* a host, not across hosts: applicability is not headroom, and on a saturated host the gate can be broadly applicable with nothing left to correct. The gate is binary and ignores magnitude, which is where the two come apart — the median 0.78% of values beyond  $3\sigma$  that folding touches on this host sits against 0.27% for an exact Gaussian, so these tensors clear the “has outliers” test while carrying almost no outlier mass to fold.

**Headline numbers on saturated hosts.** For completeness we report headline LLM and SR numbers on the saturated hosts (Tables 31, 32). The SAM (AHCPTQ) composability headline appears in Appendix U.

| Method              | LLaMA-2-7B  | LLaMA-3-8B  |
|---------------------|-------------|-------------|
| FP                  | 5.47        | 6.14        |
| SmoothQuant         | 83.12       | 210.19      |
| QuaRot              | 8.56        | 10.60       |
| SpinQuant           | 6.14        | 7.96        |
| FlatQuant           | <b>5.79</b> | 6.98        |
| FlatQuant+DBAF+PCSA | 5.80        | <b>6.96</b> |

Table 31: LLaMA W4A4 KV16 WikiText-2 PPL. Baselines: Sun et al. (2025), Table 1 (RTN); our own QuaRot run (Table 4) reaches 8.41, not 10.60. On the rotation-saturated host the gated composition matches or slightly improves the baseline.

| Method                         | PSNR (×2)    | PSNR (×3)    |
|--------------------------------|--------------|--------------|
| FP                             | 38.15        | 34.63        |
| NoisyQuant (Liu et al., 2023)  | 37.50        | 33.13        |
| 2DQuant                        | 37.87        | 34.06        |
| CondiQuant (Liu et al., 2025a) | <b>38.03</b> | <b>34.32</b> |
| CompSRT                        | 37.90        | 34.17        |
| CompSRT+DBAF                   | 37.89        | 34.16        |

Table 32: SwinIR W4 Set5 PSNR (↑); best quantized cell in bold. The two CompSRT rows are measured by us under a quantizer whose realised bitwidth is clamped to the nominal one, and are not comparable to figures reported for CompSRT elsewhere; all other rows are Liu et al. (2025a)’s 4-bit entries. A second caveat on reading this as a ranking: CompSRT rotates with the Sylvester construction, which zero-pads the embedding dimensions to the next power of two, so its rows store about 0.27 more bits per parameter than the unpadded rows despite the shared nominal W4 label. The claim it supports is the CompSRT vs. CompSRT+DBAF delta, measured on one host at one budget and unaffected by that difference; the cross-method ordering is at matched nominal bitwidth, not matched storage. The gated composition is neutral on this host (−0.01 dB at both scales): DBAF composes with a rotation-based SR quantizer at no cost rather than at a gain.

## U Composability on Less-Saturated Trained Hosts

The central methodological claim of §3, that DBAF and PCSA-tf are composable primitives that ride on top of any distribution-respecting host, is tested here on two trained hosts selected to span the architectural axes of the paper: AHCPTQ (Zhang et al., 2025) (SAM, reconstruction + Hadamard) and 2DQuant (Liu et al., 2024a) (SR transformers, dual-grid). These hosts are deliberately less saturated than the rotation-based FlatQuant/CompSRT pair examined in Appendix T, so the prediction of the framework is that composability should produce consistent moderate improvements rather than the no-op behavior of the saturated regime. Ta-

| AHCPTQ (SAM-B)   | segm mAP $\uparrow$ |
|------------------|---------------------|
| AHCPTQ (vanilla) | 13.4                |
| + DBAF           | 16.3                |
| + PCSA           | 13.8                |
| + DBAF + PCSA    | <b>18.2</b>         |

Table 33: AHCPTQ host composability on SAM-B (W4A4 YOLOX mAP  $\uparrow$ ). The vanilla row is [Zhang et al. \(2025\)](#), Table 1; the rest are run in this work. The two primitives are super-additive here: +4.8 jointly against +2.9 and +0.4 as singletons. We report the decomposition only for SAM-B, the size at which all four cells were evaluated in this work.

| 2DQuant (SR)      | $\times 2$ | $\times 3$ | $\times 4$ |
|-------------------|------------|------------|------------|
| 2DQuant (vanilla) | 37.816     | 30.113     | 31.751     |
| + DBAF            | 37.812     | 30.087     | 31.746     |
| + PCSA-tf         | 37.816     | 30.113     | 31.751     |
| + DBAF + PCSA-tf  | 37.812     | 30.087     | 31.747     |

Table 34: 2DQuant host composability on SwinIR-light SR transformers (W4A4 Set5 PSNR  $\uparrow$ ). Neither DBAF nor PCSA-tf moves the PSNR by more than  $\pm 0.03$  dB at any scale; PCSA-tf is in particular identical to vanilla to three decimal places, which matches the picture of SwinIR’s per-image content variation lacking the  $K$ -clusterable structure PCSA-tf needs (Appendix M). The 2DQuant host itself is already saturated for these primitives, similar to the rotation-based hosts in Appendix T. The SR composability gains live on the unrotated training-free host instead (Appendix R), not on CompSRT, which is rotation-saturated in the same way (Appendix T).

bles 33 and 34 report each host’s matrix.

The 2DQuant entries show a complementary saturation pattern: the host’s existing primitives already account for what DBAF and PCSA-tf would otherwise contribute, so this host mirrors the rotation-saturated regime of Appendix T rather than the moderate gains predicted above.

## V Training-Free Hosts (RTN / GPTQ / AWQ), Weight-Only

We test the same primitives on training-free hosts, no fine-tuning, no reconstruction loss, no learned rotations, to isolate the effect of the primitives from any host-side optimization. Table 35 reports the three-host  $\times$  four-arm matrix on the two LLM targets at W4 weight-only, and Table 36 reports the same four arms on the two vision targets under an RTN weight host (the W4A4 follow-up is in main Table 4).

The RTN+DBAF rows of Tables 35 and 36 already document the central training-free finding:

| Method           | LLaMA-3-8B<br>(Wiki PPL $\downarrow$ ) | Qwen-2.5-7B<br>(Wiki PPL $\downarrow$ ) |
|------------------|----------------------------------------|-----------------------------------------|
| RTN host         |                                        |                                         |
| RTN (vanilla)    | 10.279                                 | 11.623                                  |
| + DBAF           | 8.870                                  | 8.860                                   |
| + PCSA-tf        | $4.82 \times 10^{5b}$                  | $3.07 \times 10^{4b}$                   |
| + DBAF + PCSA-tf | $5.90 \times 10^{5b}$                  | $5.57 \times 10^{4b}$                   |
| GPTQ host        |                                        |                                         |
| GPTQ (vanilla)   | 11.604                                 | NaN <sup>†</sup>                        |
| + DBAF           | 9.794                                  | NaN <sup>†</sup>                        |
| + PCSA-tf        | $2.98 \times 10^{5b}$                  | NaN <sup>†</sup>                        |
| + DBAF + PCSA-tf | $4.70 \times 10^{5b}$                  | NaN <sup>†</sup>                        |
| AWQ host         |                                        |                                         |
| AWQ (vanilla)    | 13.491                                 | 53.92 <sup>‡</sup>                      |
| + DBAF           | 9.417                                  | 33.73 <sup>‡</sup>                      |
| + PCSA-tf        | $3.57 \times 10^{5b}$                  | $9.72 \times 10^{4b}$                   |
| + DBAF + PCSA-tf | $1.09 \times 10^{6b}$                  | $1.66 \times 10^{5b}$                   |

Table 35: Training-free composability matrix on the LLM targets (W4 weight-only, no fine-tuning). Both GPTQ and AWQ are broken on Qwen-2.5-7B at W4 weight-only: <sup>†</sup> GPTQ’s Hessian inversion diverges to NaN, and <sup>‡</sup> AWQ stays well outside the usable range (53.9 alone, 33.7 with DBAF; the fp16 reference is  $\approx 9$  PPL). DBAF reduces the AWQ PPL by 37% but cannot restore usability: W4 weight-only is too aggressive on Qwen for either training-free host. The numbers are reported so the failure is on the record; we omit the GPTQ/AWQ Qwen rows from comparative claims in the main text. <sup>b</sup> PCSA-tf (without the AHCPTQ-style calibration recipe) fails on both LLMs, diverging to  $10^4$ – $10^6$  perplexity as per-input descriptor routing collides with weight-only rounding under the LLM distribution (Table 19: per-layer compactness  $\geq 0.70$  on both LLMs vs.  $\leq 0.22$  on SAM). Vision targets are in Table 36.

the same primitive reduces LLaMA-3-8B PPL by  $-1.41$ , Qwen-2.5-7B PPL by  $-2.76$ , SAM-B mAP by  $+0.84$ , and SwinIR  $\times 2$  PSNR by  $+0.16$  dB. The cross-architecture consistency of this single primitive is the training-free analogue of the composability finding in Appendix U. Figure 6 visualizes the RTN+DBAF row across architectures.

## W Real Low-Bit Overhead Across Models

We complement Table 37 (LLaMA-3-8B FlatQuant W4A4) with DBAF overhead on torchao’s portable W4A8 and W4A16 paths across Qwen-2.5-7B, SAM-B/L/H, and SwinIR-light  $\times 2/ \times 3/ \times 4$ . DBAF is torch.compile-fused, applied as a forward pre-hook on each Linear input. The W4A4 path is excluded for vision models because torchao’s cutlass W4A4 kernels assume row-tile divisibility that SwinIR’s  $d = 60$  and SAM-B’s 64-dim per-head paths do not satisfy at present.

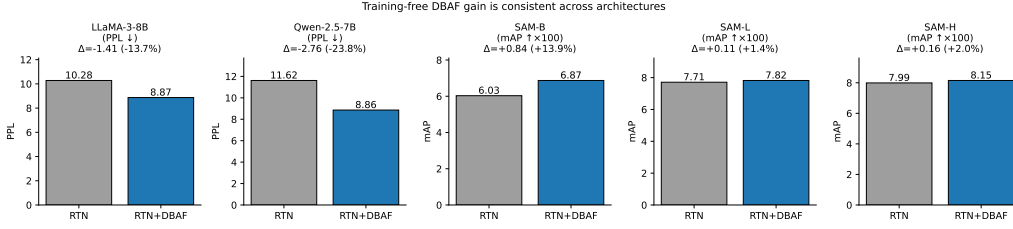


Figure 6: Training-free RTN+DBAF gain across architectures (W4 weight-only, no fine-tuning). The same primitive improves LLM perplexity (LLaMA-3-8B -13.7%, Qwen-2.5-7B -23.8%) and SAM segmentation mAP (SAM-B +13.9%, SAM-L +1.4%, SAM-H +2.0%).

| Method           | SAM-B<br>(mAP ↑)  | SwinIR ×2<br>(PSNR ↑) |
|------------------|-------------------|-----------------------|
| RTN (vanilla)    | 6.03              | 32.61                 |
| + DBAF           | 6.87              | 32.76                 |
| + PCSA-tf        | 0.03 <sup>b</sup> | 31.92 <sup>b</sup>    |
| + DBAF + PCSA-tf | 0.04 <sup>b</sup> | 32.11 <sup>b</sup>    |

Table 36: The same four training-free arms on the vision targets (W4 weight-only, no fine-tuning), under an RTN weight host. GPTQ and AWQ are evaluated on the LLM targets only (Table 35): our implementations cover the Linear layers of the decoder stack and do not extend to SwinIR’s Conv2d stages or the SAM image encoder. These rows are about the primitives under training-free weight quantization, not about the host that produced the weights. <sup>b</sup> PCSA-tf collapses SAM-B mAP (6.03 → 0.03) and drops SwinIR ×2 PSNR by -0.69 dB, because it adds INT4 activation quantization on top of W4 weights, which is too aggressive when W4-only already preserves vision quality. DBAF partially recovers vision PSNR (SwinIR ×2: 31.92 → 32.11 dB) but cannot rescue SAM mAP. PCSA-tf fits its  $K$  anchors by  $k$ -means from an unseeded initialisation, so the two PCSA rows carry a run-to-run component: over four refits of the +DBAF+PCSA-tf cell we measure a spread of 0.04 dB (s.d. 0.02). Differences of a few hundredths should not be read as effects; the +0.19 dB DBAF recovery is an order of magnitude larger.

| Stage             | Baseline INT4 | +DBAF          |
|-------------------|---------------|----------------|
| Prefill (ms)      | 185.0         | 185.2 (+0.1%)  |
| Decode (ms)       | 12,698        | 12,920 (+1.7%) |
| PPL eval (ms/seq) | 180.7         | 181.1 (+0.2%)  |

Table 37: LLaMA-3-8B real INT4 deployment overhead on FlatQuant’s fused W4A4 KV4 CUDA kernels (Sun et al., 2025). DBAF unfolding is an elementwise FP forward pre-hook outside the INT4 GEMM.

## X Extended Limitations

**Reproduction caveats on saturated hosts.** Our reproduced FlatQuant + DBAF + PCSA W4A4 KV4 number on LLaMA-3-8B (WikiText-2 PPL

| Model           | W4A16 | W4A8   |
|-----------------|-------|--------|
| Qwen-2.5-7B     | -0.0% | +0.4%  |
| SAM-B           | +0.6% | +0.8%  |
| SAM-L           | +0.6% | +0.9%  |
| SAM-H           | +0.4% | +0.8%  |
| SwinIR-light ×2 | +0.5% | +2.3%  |
| SwinIR-light ×3 | +6.6% | +2.0%  |
| SwinIR-light ×4 | +7.9% | +23.5% |

Table 38: +DBAF (torch.compile-fused) overhead on torchao’s W4A8 and W4A16 paths, A100. LLM + SAM stay below 1% on both configurations; the SwinIR-light variants ( $d=60$ ,  $\sim 140$  ms forward on ×4) show kernel-launch-dominated overhead: DBAF adds six element-wise operations per Linear that cannot amortize over a small GEMM at that width.

6.966) is within calibration noise of pure FlatQuant W4A4 KV4 (6.964); the  $\approx 0.02$  PPL gain we previously saw at W4A4 KV16 does not replicate at KV4, and Table 30 reports both honestly. The same table’s C1 ablation (FlatQuant + DBAF no-gate + PCSA: 6.910 WikiText, 10.968 C4) shows that disabling the gate lets DBAF extract a small but consistent additional gain, suggesting the gate is over-conservative when the host already absorbs the dominant outlier pattern. This is the false-negative mode of §3.1 and Appendix I, observed on a saturated rotation host rather than off-distribution. We do not recommend disabling the gate by default — the no-gate variant applies DBAF where its information-preservation guarantee no longer holds — but document it as evidence that the gate-host interaction is an open design axis.

**Calibration-descriptor proxy for PCSA-tf.** PCSA-tf uses an  $\ell_2$ -normalized mean-pooled hidden state as the prompt descriptor (§3.4), a deliberately simple choice: it needs no auxiliary model and reproduces with a single forward pass over calibration data. Where prompt structure is highly variable — retrieval-augmented generation, multi-turn

dialog, mixed-task batches — better descriptors may be needed to recover the same routing quality. Appendix H tests four alternatives — max-pooling, attention pooling, last-token pooling and Sentence-BERT embeddings — and finds the dispatch decision unchanged in 27 of 28 cells, so within the regimes we measure the choice is not load-bearing. RAG and multi-turn prompts remain untested.

**Threshold robustness.** The  $3\sigma$  tail boundary in the outlier gate follows the 68–95–99.7 rule for an approximately Gaussian core and is principled rather than tuned. The remaining gate parameters (skewness  $\leq 0.7$ , kurtosis  $\in [3, 30]$ , outlier fraction  $\in [10^{-4}, 2 \times 10^{-2}]$ ) are calibrated on LLaMA-3-8B once and verified to retain 94–100% weight gate-pass across the eight model configurations of Table 25. We have not stress-tested them under broader distribution shift (e.g., post-instruction-tuning models with substantially altered activation profiles) and we do not claim transfer beyond the dense-core regime the gate is designed to detect.

**Sensitivity-weighted  $\alpha^*$  on the activation path.** The closed-form  $\alpha^*$  of Eq. 2 minimizes an MSE proxy that weights all positions equally. On LLaMA-3-8B (Qwen-2.5-7B) activations it gives  $\alpha_A^* \approx 0.20$  (0.18), below the selected  $\alpha = 0.25$ . Weighting the outlier-position term by its end-task sensitivity  $\lambda$  (the outlier-to-bulk Hessian ratio) yields the closed form  $\alpha_{\text{sens}}^* = \lambda^{1/3} \alpha^*$ . Since  $\lambda > 1$  on this path—PPL is more sensitive to outlier-position error than to bulk error—the operating point lies above  $\alpha_A^*$ , consistent with the selected 0.25. We do not identify  $\lambda$  numerically: it is an end-task Hessian ratio rather than a fold-geometry quantity, and the natural energy-ratio proxy overshoots (Appendix B). On the weight path, where outlier-position error is diluted across each dot product,  $\alpha_W^*$  instead lands at or slightly above the selected  $\alpha$ . We therefore select  $\alpha$  by one-block reconstruction sweep seeded at  $\alpha^*$  on both paths.

**Calibration-seed stability of the diagnostic.** End-task numbers are reported at a single calibration seed, but the dispatch decisions themselves are stable across redraws: over five independent calibration redraws the activation gate-pass rate is  $63.1 \pm 1.3\%$ , 89.3% of layer decisions are identical across all five seeds, and the PCSA site compactness is  $0.856 \pm 0.045$ , SKIP on every seed (Appendix D). What we have not measured is seed variance of the end-task metrics themselves.

**Long-context KV scope.** The KV-PCSA gain at long context (+5pp NIAH at 8k, §4.4) is specific to W4A4 KV4; Table 30 reports the corresponding KV16 and KV4 results on the rotation-saturated host.

**Detail moved from the main Limitations.** Three points are stated compactly in the Limitations section and recorded in full here. First, the additional distributional patterns the gate self-disables on are post-ReLU asymmetric, bimodal post-softmax and heavy-tailed unimodal, identified by the cross-model layer analysis of Table 25; DiT-XL’s weight kurtosis of 956 places it far outside the calibrated band, which is why the gate skips there, and AHCPTQ’s bimodal handler suggests the form a distribution-specific primitive would take. Second, the leave-one-architecture-out AUC of 0.962 spans 0.93–0.99 across families, and the held-out per-layer gain correlation is  $r \approx 0.13$  with frozen features and 0.26 with learned ones. Third, we do not report probability-calibrated confidence on regimes outside the studied triples, and the per-regime  $\alpha$  values are 0.25 at W4A4 on both LLMs and 0.75–0.99 elsewhere (Tables 4 and 7).