FEW BITS, ONE LAW: TOWARD W2A4KV2

Anonymous authors
Paper under double-blind review

ABSTRACT

Extreme low-bit LLM compression is most challenging when weights, activations, and KV caches are quantized together: their distributions differ, and quantization errors interact throughout the network. We introduce CanonQ, a unified quantization-aware training framework that addresses these challenges by separating source canonicalization from task-aware adaptation. Fixed rotations and energy normalization map heterogeneous tensor sources to canonical coordinates, enabling frozen Gaussian-reference codebooks to be reused across layers and models. Joint training then adapts the network to the coupled errors of weight, activation, and cache quantization within a common scalar/vector interface. We bound frozen-codebook transfer error and local task loss, and derive an exact normalization-aware straight-through Jacobian that links quantization distortion to gradient bias. The strongest gains arise under joint `W2A4KV2` compression: across `LLaMA3-1B/3B/8B`, CanonQ-Omni achieves up to $14.28\times$ lower `WikiText-2` perplexity and up to 57.9% higher mean zero-shot accuracy than prior state-of-the-art and representative quantization baselines. The benefits extend to `Qwen3-1.7B`, code generation, and mathematical reasoning: on instruction-tuned `MobileLLM-Pro-1B` at `W2A16KV16`, CanonQ achieves relative improvements of 41.7% in `HumanEval` pass@1 and 39.1% in `GSM8K` exact match over the strongest evaluated quantization baseline.

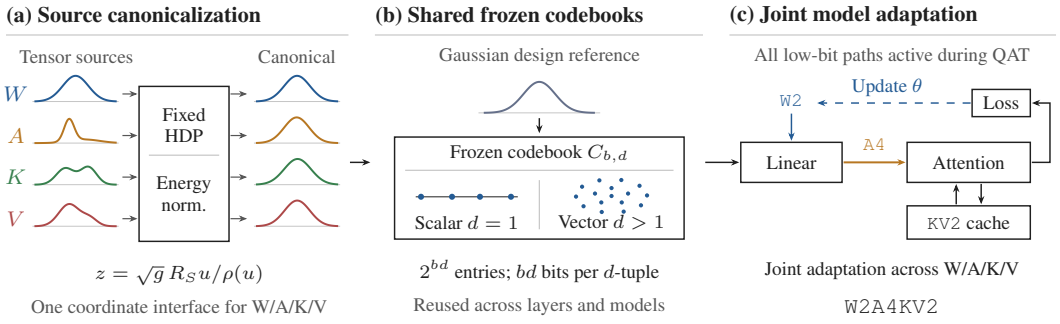


Figure 1: **Schematic overview of CanonQ.** Fixed source-specific transforms and normalization map tensor blocks to canonical coordinates. Frozen Gaussian-reference codebooks are reused across layers and models for matching (b, d). Joint QAT adapts the model to coupled quantization errors at `W2A4KV2`.

1 INTRODUCTION

LLM inference is increasingly limited by memory movement: weights dominate model storage and matrix-multiplication bandwidth, activations affect prefill, and the KV cache grows with batch size and context length. Quantization is therefore central to deployment, yet the most attractive regimes, which include `W1/W2` weights, `A4` activations, and `KV2` caches, remain difficult. Post-Training Quantization (PTQ) is effective at moderate precision but often collapses when several components are compressed simultaneously (Dettmers et al., 2022; Xiao et al., 2023; Frantar et al., 2023; Lin et al., 2024; Shao et al., 2024; Egiazarian et al., 2024; Malinovskii et al., 2024). Quantization-Aware Training (QAT) can recover quality by adapting the model to the quantized computation (Liu et al.,

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

2024a; 2025b; Chen et al., 2025; Li et al., 2026; Yi et al., 2026), but the result still depends on the discrete representation to which the model is trained.

At such low precision, that representation is not a minor implementation choice. Uniform quantization uses equally spaced reconstruction levels, even though weights, activations, and KV states have different distributional shapes. A data-fitted codebook can accommodate those differences, but ties the representation to the tensors or calibration data used to construct it. Joint compression makes the choice harder: weights shape the activation distribution, and quantized activations generate the keys and values that must themselves survive a two-bit cache. This motivates a common quantization interface that supports adaptation across the coupled computation.

We ask: *can a small quantization interface be designed once, then reused while an LLM learns to operate through it?* Our approach separates two decisions. Source-specific transforms and normalization determine the coordinates presented to the quantizer; a source-independent reference determines the reconstruction alphabet. QAT then adapts the network to the resulting fixed codec. We introduce CanonQ around this principle: canonicalize the source, then train the model to use the codec. A fixed source-specific Hadamard–diagonal–permutation (HDP) transform mixes each block, and energy normalization removes its magnitude. The codebooks and HDP maps stay fixed throughout QAT. CanonQ-W applies this interface to weights; CanonQ-Omni applies it to the complete W/A/K/V computation.

Our central finding is that the benefit of canonical quantization becomes most pronounced when weights, activations, and KV caches are compressed together.

Our contributions are: **(1) A unified framework for joint W/A/KV quantization.** We introduce CanonQ, which unifies scalar and vector quantization of weights, activations, and KV caches through a common canonical interface. Frozen reference codebooks are reusable across layers and models, while joint QAT adapts the network to interacting quantization errors across the full computation. **(2) A unified theory linking source alignment to optimization.** We establish a quantitative chain from source mismatch to codebook distortion, local task loss, and optimization error. Wasserstein bounds characterize when frozen reference codebooks transfer to actual tensor distributions, while an exact normalization-aware straight-through estimator (STE) Jacobian identity reveals how assignment distortion controls backward sensitivity and gradient bias. We further derive distortion-dependent stationarity guarantees for a smooth reference objective, providing a theoretical foundation for the canonical training interface. **(3) Large quality gains under extreme joint compression.** At `W2A4KV2`, CanonQ-Omni achieves up to $14.28\times$ lower `WikiText-2` perplexity and up to 57.9% higher mean zero-shot accuracy than strong quantization baselines across `LLaMA3-1B/3B/8B`. Further evaluations demonstrate transfer to `Qwen3-1.7B`, substantial coding and `GSM8K` improvements under `W2A16KV16`, and stable relative perplexity within the `8K` continued-QAT training range.

2 Related Work

Low-bit optimization and codebooks. PTQ methods including GPTQ, AWQ, and OmniQuant improve low-bit reconstruction (Frantar et al., 2023; Lin et al., 2024; Shao et al., 2024); LLM-QAT, EfficientQAT, ParetoQ, WinQ, and JacQuant adapt models through quantized training (Liu et al., 2024a; Chen et al., 2025; Liu et al., 2025b; Li et al., 2026; Yi et al., 2026). Binary and ternary methods likewise make representation a first-order design choice (Wang et al., 2023; Ma et al., 2024; Shang et al., 2023; Huang et al., 2024; Xu et al., 2024). Lloyd–Max and Linde–Buzo–Gray provide classical scalar and vector constructions (Max, 1960; Lloyd, 1982; Linde et al., 1980); modern non-uniform, additive, and lattice approaches include NF4, SqueezeLLM, AQLM, PV-Tuning, and QuIP# (Dettmers et al., 2023; Kim et al., 2024; Egiazarian et al., 2024; Malinovskii et al., 2024; Tseng et al., 2024). LC-QAT is a close weight-only QAT comparison: it represents two-bit quantized weights through a learned affine map over discrete vectors and avoids explicit codebook lookup in the training forward pass (Wang et al., 2026). CanonQ instead freezes the reference table, retains nearest-codeword assignment, and trains all enabled W/A/K/V paths.

Rotations, canonicalization, and caches. Randomized rotations support distributed compression and LLM incoherence processing (Vargaftik et al., 2021; 2022; Chee et al., 2023; Tseng et al., 2024). QuaRot and SpinQuant suppress outliers or learn rotations (Ashkboos et al., 2024; Liu et al., 2025a);

KIVI and KVQuant exploit cache-specific structure (Liu et al., 2024b; Hooper et al., 2024). TurboQuant separates norm from direction and constructs scalar quantizers for rotation-induced marginals; its inner-product variant additionally quantizes a residual with QJL (Zandieh et al., 2025). This is the closest source-modeling connection. We study task-aware adaptation to a fixed canonical codec, allow joint tuple assignment, and quantify the implemented HDP map's mismatch from its design reference. Building on joint four-bit compression exemplified by QuaRot, CanonQ targets the more aggressive `W1/W2-A4/A8-KV2` regime through a reusable canonical interface and joint task-aware adaptation.

3 CanonQ: Unified W/A/KV Quantization

As illustrated in Fig. 1, CanonQ separates source-specific coordinates from source-independent reconstruction choices. Let $u \in \mathbb{R}^{g_S}$ be a block from $S \in \{W, A, K, V\}$. Draw and fix the orthogonal map $R_S = P_S H D_S$, where H is a normalized Hadamard matrix, D_S contains random signs, and P_S is a permutation. Each source shares its map across layers; keys and values share one KV map. With $\rho_S(u) = \|u\|_2 + \epsilon$, define

$$z = \sqrt{g_S} R_S u / \rho_S(u). \tag{1}$$

Rotation redistributes energy; normalization removes magnitude. For $\epsilon = 0$, the representation is invariant to positive rescaling of nonzero u. The Gaussian law $Q_d = \mathcal{N}(0, I_d)$ provides a reusable codebook design reference, motivated by the large-g_S behavior of Haar-rotated sphere coordinates (Diaconis & Freedman, 1984). The canonicalization defect in Def. 1 quantifies deviations under the implemented structured HDP map.

3.1 Reusable Codebooks in Canonical Coordinates

For b bits per scalar and tuple dimension d, the population target is the best $K = 2^{bd}$-point quantizer of Q_d,

$$C^\star_{b,d} \in \arg\min_{1 \le |C| \le 2^{bd}} \frac{1}{d} \mathbb{E}_{V \sim Q_d} \operatorname{dist}(V, C)^2. \tag{2}$$

Generalized Lloyd/LBG iterations approximate this target offline; we denote the deployed numerical table by $C^{\text{can}}_{b,d} := \widehat{C}_{b,d}$ and retain its finite fitting gap ε_{vq} in Eq. (6). For $d = 1$ the construction is Lloyd–Max; for $d > 1$ assignment is genuinely joint rather than a Cartesian product of scalar decisions. For example, `W2` with $d = 4$ uses 256 four-dimensional codewords, so one 8-bit index represents four weights. A table depends only on (b, d) and stays frozen across groups, layers, models, and QAT updates.

Group energies and source maps handle scale and coordinates, while QAT adapts the model without refitting each tensor's alphabet. The shared construction therefore accommodates different raw W/A/K/V distributions.

Partition z into consecutive tuples $v_j \in \mathbb{R}^{d_S}$, with $d_S \mid g_S$. The codec is

$$\widehat{v}_j = \Pi_{C^{\text{can}}_{b_S, d_S}}(v_j), \quad \widehat{u} = \gamma_{S,\ell} \frac{\rho_S(u)}{\sqrt{g_S}} R_S^\top \operatorname{concat}_j(\widehat{v}_j). \tag{3}$$

Let $q = \operatorname{concat}_j(\widehat{v}_j)$ denote the assigned canonical block. The nonnegative gain $\gamma_{S,\ell}$ adjusts its scale after assignment; clipping and decoding metadata remain source-specific. Here g_S is the normalization-block size and d_S is the assignment dimension. Assignment error $\|q - z\|_2$ and reconstruction error $\|\gamma_{S,\ell} q - z\|_2$ therefore describe different quantities. With identity assignment and unit gain, the codec recovers u exactly.

3.2 Joint Adaptation to Quantization Errors

Nearest-codeword assignment uses the identity STE,

$$\widetilde{v}_j = v_j + \operatorname{sg}\Big(\Pi_{C^{\text{can}}_{b_S, d_S}}(v_j) - v_j\Big), \quad \min_{\theta,\phi} \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell\big(f_{Q_C(\theta;\phi)}(x), y\big),$$

where sg stops gradients, θ denotes shadow-model parameters, and ϕ contains enabled trainable quantizer parameters. The reported canonical activation and KV paths use energy normalization and frozen tables without an additional learned, absmax, or percentile clipping parameter. All low-bit paths in CanonQ-Omni are active during QAT, so training sees the coupled errors that the evaluation operator will introduce. Evaluation re-quantizes the shadow weights and uses the labeled activation/cache path. The norm is recomputed from the current block. Differentiating it in both encoding and decoding makes the block STE differ from the identity, as derived in Sec. 4.3.

3.3 Scalar/Vector Variants and Computational Cost

CanonQ-W canonicalizes weights and retains uniform quantizers for enabled A/KV paths; CanonQ-Omni canonicalizes every enabled source. We also implement VQLLM-QAT, a raw-space control with group-fitted weight codebooks, the same tuple dimension and index rate, nearest-codeword assignment, and STE. This control changes transformation and table fitting together. Sources literally share a table only when (b, d) matches; otherwise they share the construction. The suffix d denotes d_W; reported canonical A/K/V paths use $d_A = d_K = d_V = 1$. Each tuple uses bd index bits. A shared table occupies $2^{bd} d q_C$ bits for q_C-bit components. Exhaustive assignment costs $O(g_S 2^{bd})$ arithmetic per block and the Hadamard transform costs $O(g_S \log g_S)$. These arithmetic costs motivate small d. Group metadata, table bytes, and unquantized modules are accounted for in Sec. A.4.

4 Theory: From Source Alignment to Optimization

The fixed interface raises three questions: when do reference codebooks transfer, how do quantization residuals affect local task loss, and how does distortion influence training?

4.1 Bounding Frozen-Codebook Transfer Error

Let $Q_d = \mathcal{N}(0, I_d)$. Haar-rotated normalized blocks converge to this reference in W_2 for fixed d (Prop. 1). The implemented $R = PHD$ is a structured orthogonal transform. For blocks independent of its random signs, Prop. 2 establishes sub-Gaussian coordinates, a marginal Gaussian-approximation bound controlled by $\|u\|_3^3/\|u\|_2^3$, and an exact mean tuple-covariance defect controlled by $g\|u\|_4^4/\|u\|_2^4$. Our transfer analysis explicitly accounts for deviations from the Gaussian reference through the checkpoint-specific joint tuple law.

Write $\omega = (\theta, \phi)$ and let $P_{S,d}^{\omega}$ denote the post-HDP tuple law induced by checkpoint ω under a fixed sampling rule over layers, groups, tokens, heads, and tuple positions.

Definition 1 (Canonicalization defect)**.** *For any tuple law P with finite second moment, define*

$$\Delta_d(P) := \frac{1}{\sqrt{d}} W_2(P, Q_d), \quad \Delta_{S,d}(\omega) := \Delta_d(P_{S,d}^{\omega}). \tag{4}$$

This checkpoint-specific defect measures mismatch in the full joint tuple law, including coordinate dependence, structured-transform effects, and $\epsilon > 0$ normalization. It connects the reference-codebook design to the tensor distributions encountered during QAT.

For a K-point codebook $C \subset \mathbb{R}^d$, let

$$D_{P,d}(C) := \frac{1}{d} \mathbb{E}_{V \sim P} \operatorname{dist}(V, C)^2, \quad \mathcal{R}_{P,d}(C) := \sqrt{D_{P,d}(C)}, \tag{5}$$

and let $\mathcal{R}^{\star}_{P,b,d} := \inf_{1 \le |C| \le 2^{bd}} \mathcal{R}_{P,d}(C)$. The numerical canonical table $\widehat{C}_{b,d}$ is allowed a squared-error fitting gap

$$D_{Q_d,d}(\widehat{C}_{b,d}) \le (\mathcal{R}^{\star}_{Q_d,b,d})^2 + \varepsilon_{\mathrm{vq}}. \tag{6}$$

Theorem 1 (Frozen-codebook transfer)**.** *Let A_b be any 2^b-level scalar alphabet and A_b^d its Cartesian product. Then*

$$\mathcal{R}^{\star}_{Q_d,b,d} \le \mathcal{R}_{Q_d,d}(A_b^d), \quad \mathcal{R}_{Q_d,d}(\widehat{C}_{b,d}) \le \mathcal{R}^{\star}_{Q_d,b,d} + \sqrt{\varepsilon_{\mathrm{vq}}}. \tag{7}$$

For every actual tuple law P with finite second moment,

$$\mathcal{R}_{P,d}(\widehat{C}_{b,d}) - \mathcal{R}^{\star}_{P,b,d} \le 2\Delta_d(P) + \sqrt{\varepsilon_{\mathrm{vq}}}. \tag{8}$$

More directly, define the reference margin $m_Q(A_b^d, \widehat{C}_{b,d}) := \mathcal{R}_{Q_d,d}(A_b^d) - \mathcal{R}_{Q_d,d}(\widehat{C}_{b,d})$. *Then*

$$\mathcal{R}_{P,d}(\widehat{C}_{b,d}) - \mathcal{R}_{P,d}(A_b^d) \le 2\Delta_d(P) - m_Q(A_b^d, \widehat{C}_{b,d}). \tag{9}$$

Hence a validated lower bound on the reference shaping gain transfers whenever it exceeds twice a validated upper bound on the canonicalization defect.

The bound separates two routes to reliable transfer: better reference-table fitting reduces $\varepsilon_{\mathrm{vq}}$, while better canonical coordinates reduce $\Delta_d(P)$. The margin condition in Eq. (9) quantifies when a reference reconstruction advantage survives deployment mismatch. These assignment results use $\gamma = 1$; Lem. 3 extends the accounting to assign-then-gain decoding. Tuple dimension affects reference distortion, fitting error, and joint source mismatch together. We next incorporate downstream sensitivity to connect these representation effects to local task loss.

4.2 Bounding Local Task Loss from Quantization

Let $\Pi_C^{(d)}$ apply C tuple by tuple to $z = \sqrt{g}Ru/\rho(u)$. The implementation first assigns against C and then applies a nonnegative source-specific gain γ. Its actual canonical residual and block perturbation are therefore

$$\eta_{C,\gamma}(z) := \gamma\Pi_C^{(d)}(z) - z, \quad e_{C,\gamma}(u) := \widehat{u} - u = \frac{\rho(u)}{\sqrt{g}} R^\top \eta_{C,\gamma}(z). \tag{10}$$

Quantization errors also interact across sources. For a linear block, write $\widehat{W} = W + E_W$ and $\widehat{a} = a + e_A$. Then

$$\widehat{W}\widehat{a} - Wa = E_W a + W e_A + E_W e_A.$$

The cross term, and the analogous query/key/value terms in attention (Lem. 4), motivate joint adaptation.

Theorem 2 (Gain-aware local task-loss certificate)**.** *Consider a single-site intervention that replaces block* u *by* $u + e_{C,\gamma}(u)$ *while holding the other sites fixed. Suppose* $\mathcal{L}$ *is* C^3 *on a neighborhood of the joining segment,* $\|\nabla^2\mathcal{L}(u)\|_{\mathrm{op}} \le \Lambda_u$, *and* $\|\nabla^3\mathcal{L}(w)\|_{\mathrm{op}} \le M_3$ *on that segment. Then*

$$\begin{aligned}|\mathcal{L}(u + e_{C,\gamma}) - \mathcal{L}(u)| \le{}& \frac{\rho(u)}{\sqrt{g}}\|\nabla\mathcal{L}(u)\|_2\|\eta_{C,\gamma}(z)\|_2 \\ &+ \frac{\rho(u)^2\Lambda_u}{2g}\|\eta_{C,\gamma}(z)\|_2^2 + \frac{M_3\rho(u)^3}{6g^{3/2}}\|\eta_{C,\gamma}(z)\|_2^3.\end{aligned} \tag{11}$$

The certificate applies to a smooth downstream continuation or a hard-quantized branch whose downstream assignments remain fixed throughout a neighborhood of the intervention segment. It retains the local gradient term explicitly, covering activation and KV interventions away from coordinate-wise stationarity, and separates first-order sensitivity, curvature, and higher-order effects. For random blocks, set $Z = \sqrt{g}RU/\rho(U)$ and choose $\Lambda_U \ge \|\nabla^2\mathcal{L}(U)\|_{\mathrm{op}}$. The quadratic contribution is bounded by one half of the sensitivity-weighted distortion

$$D_{S,\gamma}^{\mathrm{curv}}(C) := \frac{1}{g}\mathbb{E}\left[\rho(U)^2\Lambda_U\|\eta_{C,\gamma}(Z)\|_2^2\right], \tag{12}$$

without residual-isotropy assumptions. Expected certificates are finite when the displayed sensitivity-weighted moments are finite. Common-cut interventions retain cross-site Hessian terms; causally ordered sites require telescoping with the smoothness condition checked at each step (Cor. 1).

4.3 Quantization Distortion Controls STE Error

The hard nearest-codeword objective is generally discontinuous at cell boundaries. We therefore compare its STE with a smooth gain-preserving reference. Set $s(u) = \rho(u)/\sqrt{g}$, $q = \Pi_C^{(d)}(z)$, and $\delta_a = \|q - z\|_2/\sqrt{g}$. If the backward pass uses $Dq/Dz = I$ and treats γ as independent of u, for $u \ne 0$, its surrogate Jacobian is

$$J_{\mathrm{STE}}(u) = \gamma\left[I + R^\top(q - z)\nabla s(u)^\top\right], \quad \|J_{\mathrm{STE}} - \gamma I\|_{\mathrm{op}} = \gamma\delta_a. \tag{13}$$

Thus, an identity STE at assignment induces a distortion-dependent block correction of rank at most one. If the group energy is detached in both encoding and decoding, the local surrogate Jacobian becomes γI instead; additional clipping introduces its own derivative.

For a downstream loss f with L_f-Lipschitz gradient, this identity yields the explicit bias certificate

$$\|g_{\mathrm{STE}} - \gamma \nabla f(\gamma u)\|_2 \le \gamma \delta_a \|\nabla f(\gamma u)\|_2 + \gamma^2 (1+\delta_a) L_f \rho(u) \delta_a =: B_C(u). \tag{14}$$

For the uniform tuple law P_u of a fixed group, $\delta_a(u)^2 = D_{P_u,d}(C)$. Under matching random-group/tuple sampling, $\mathbb{E}\delta_a(U)^2 = D_{P,d}(C)$. If gains, energies, downstream gradients, and assignment residuals are bounded on the analyzed region, then $B_C \le K\delta_a$ and $\mathbb{E}B_C(U)^2 \le K^2 D_{P,d}(C)$ (Cor. 3). A trained gain contributes an additional Jacobian-column defect $\rho(u)\delta_a$. These give codec-dependent sufficient conditions, rather than assuming alignment independently of the quantizer (Yin et al., 2019).

For the full network, let F be the L-smooth gain-preserving reference obtained by replacing all enabled hard quantizers (including any uniform control paths) by the identity. Let G_t be the stochastic STE direction, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ condition on the pre-minibatch history, and $b_t = \mathbb{E}_t G_t - \nabla F(\omega_t)$. Under backward-map stability, integrability, and valid differentiation under expectation, Lem. 5 lifts the codec-dependent errors to a full-network bias certificate $\|b_t\|_2 \le B_t$.

Theorem 3 (Distortion-controlled STE stationarity)**.** *Assume $F_\star > -\infty$, $\|b_t\|_2 \le B_t$, and the conditional STE variance satisfies $\mathbb{E}_t\|G_t - \mathbb{E}_t G_t\|_2^2 \le \sigma^2$. For $\omega_{t+1} = \omega_t - \alpha_t G_t$ with deterministic steps $0 < \alpha_t \le 1/(4L)$ and $S_T = \sum_{t<T} \alpha_t$,*

$$\frac{1}{S_T}\sum_{t<T} \alpha_t \mathbb{E}\|\nabla F(\omega_t)\|_2^2 \le \frac{2(F(\omega_0) - F_\star)}{S_T} + \frac{5}{2S_T}\sum_{t<T}\alpha_t \mathbb{E}B_t^2 + \frac{L\sigma^2}{S_T}\sum_{t<T}\alpha_t^2. \tag{15}$$

With $\alpha_t = c/\sqrt{T} \le 1/(4L)$, the bound separates $O(T^{-1/2})$ optimization and sampling terms from the codec-dependent gradient-bias contribution. This establishes distortion-dependent stationarity bounds for the smooth gain-preserving reference under the stated SGD update rule. The appendix further develops a conditional PL bound and Voronoi-margin stability.

Connection to evidence. The analysis separates source mismatch, sensitivity-weighted reconstruction error, and assignment distortion. Scalar W_1/KS diagnostics assess marginal alignment; zero-update/QAT comparisons test adaptation to the fixed codec; precision and dimension sweeps establish how these choices affect model quality.

5 Experiments

The evaluation first locates the benefit across compression settings, then examines source alignment and adaptation, and finally tests longer-context likelihood, code generation, and mathematical reasoning.

5.1 Evaluation Setup

We evaluate `LLaMA3-1B/3B/8B` and `Qwen3-1.7B` (Dubey et al., 2024; Yang et al., 2025), using `1B` for the broad component study. Standard QAT uses `FineWeb-Edu`, sequence length `2048`, effective batch size `64`, AdamW, and up to `240K` updates ($\approx$ 31.5B tokens). Learning rates are selected from $\{1, 2, 4\} \times 10^{-5}$ on held-out validation data. The `8B` CanonQ-Omni configuration reuses the selected `3B` setting. The context study separately continues QAT at `8192` for `40K` steps. We report `WikiText-2` perplexity (PPL) (Merity et al., 2017) and the unweighted mean zero-shot accuracy over eight benchmarks. Accuracy is reported in percent; task definitions and evaluation details appear in Sec. A.1.

The instruction-tuned extension reports greedy coding pass@1 and `GSM8K` final-answer exact match under a separate protocol (Sec. A.5) (Austin et al., 2021; Chen et al., 2021; Cobbe et al., 2021).

ParetoQ denotes its weight quantizer combined with the uniform A/KV paths specified in Sec. A.2. Uniform `A4/KV2` uses 15/3 active levels; canonical scalar tables use 16/4 at the same index rate.

Comparisons evaluate complete trained configurations, including these differences. Full settings and per-task results appear in Secs. A and B.3. Each reported configuration uses one training seed.

Table 1: Joint-compression and cross-model results. Columns report PPL ↓ and eight-task accuracy ↑. The suffix denotes d_W; canonical A/K/V paths use $d = 1$. CanonQ-Omni uses canonical KV quantization in the headline KV2 results. Full task scores appear in Tabs. 11 to 13.

LLaMA3-1B, quantized KV

	W2A8KV2		W2A4KV2	
Method	PPL ↓	Acc. ↑	PPL ↓	Acc. ↑
FP Model	9.75	58.54	9.75	58.54
RTN (PTQ)	5.8e5	37.31	5.5e5	36.72
GPTQ (PTQ)	5.3e4	38.23	9.5e4	37.83
QuaRot (PTQ)	3.4e4	37.39	4.1e4	38.11
KIVI	1.1e5	34.85	3.2e5	35.16
ParetoQ	47.58	45.13	90.57	40.47
VQLLM-QAT-d4	34.35	45.35	76.51	42.32
CanonQ-W-d1	30.07	46.41	64.40	42.49
CanonQ-W-d2	30.19	47.56	64.48	42.95
CanonQ-W-d4	31.27	46.91	76.19	41.97
CanonQ-Omni-d1	**15.70**	**53.45**	**15.42**	**53.56**
CanonQ-Omni-d2	16.43	52.67	15.90	53.16

LLaMA3-3B

	W2A4		W2A4KV2	
Method	PPL ↓	Acc. ↑	PPL ↓	Acc. ↑
FP Model	7.73	65.11	7.73	65.11
RTN (PTQ)	4.2e5	35.74	1.1e6	36.78
GPTQ (PTQ)	7.9e4	38.12	8.9e4	37.88
QuaRot (PTQ)	6.7e2	37.23	1.5e4	37.95
ParetoQ	17.88	53.15	92.64	39.39
VQLLM-QAT-d4	30.71	46.30	52.82	42.07
CanonQ-Omni-d1	10.66	**61.16**	**10.93**	58.37
CanonQ-Omni-d2	**10.50**	61.14	11.16	**60.04**

Qwen3-1.7B and LLaMA3-8B

	Q3-1.7B (W2A4)		L3-8B (W2A4KV2)	
Method	PPL ↓	Acc. ↑	PPL ↓	Acc. ↑
FP Model	16.33	58.10	6.14	70.14
ParetoQ	43.56	42.35	134.36	37.87
CanonQ-Omni-d1	**17.80**	**50.96**	**9.41**	**59.81**

5.2 Large Quality Gains at W2A4KV2

The central result is the large separation under W2A4KV2 (Tab. 1). On LLaMA3-1B, CanonQ-Omni-d1 reaches 15.42 PPL and 53.56% mean accuracy, versus 90.57 and 40.47% for ParetoQ: 5.87× lower PPL and a 13.09-point accuracy gain. Relative to weight-only canonicalization with CanonQ-W-d1, it lowers PPL by 4.18× and raises accuracy by 11.07 points. This direct comparison locates the larger benefit in extending the canonical interface to activations and caches.

The gains persist at larger scales. On 3B, CanonQ-Omni-d2 reaches 11.16/60.04 (PPL/accuracy), versus 92.64/39.39 for ParetoQ: 8.30× lower PPL and a 20.65-point gain. At 8B, CanonQ-Omni-d1 reaches 9.41/59.81 versus 134.36/37.87, giving 14.28× lower PPL and 57.9% higher accuracy. To reduce tuning costs, we reuse the best training hyperparameters identified on 3B for 8B. We expect a dedicated 8B hyperparameter search to further improve accuracy and narrow the gap to FP.

On Qwen3-1.7B W2A4, CanonQ-Omni-d1 obtains 17.80/50.96 versus 43.56/42.35, improving PPL by 2.45× and accuracy by 8.61 points. Together, these results show that the same frozen-codebook construction supports joint quantization across model scales and transfers to a second model family.

5.3 Gains Grow Beyond Weight-Only Quantization

Weight-only quantization gives modest differences. At 1B W2A16, ParetoQ and CanonQ-W-d2 achieve similar quality: 12.66/55.53 and 12.86/55.45 (PPL/accuracy), respectively (Tab. 2). At W1A16, CanonQ-W-d1 lowers PPL from 18.13 to 16.93, while $d_W = 2$ gives the highest accuracy, 51.81%. At matched $d_W = 4$, canonical coding improves raw-space VQLLM-QAT from 22.36/47.90 to 17.47/51.17, showing a clearer benefit from canonical coordinates and reference codebooks at one-bit weight precision.

Activation compression produces the larger separation. At W2A8, CanonQ-Omni-d1 improves ParetoQ's 18.22/51.09 to 13.25/55.60. At W2A4, weight-only canonicalization remains near 35 PPL and 44% accuracy, whereas CanonQ-Omni-d1 reaches 13.84/54.71: 2.35× lower PPL and a 10.59-point gain over ParetoQ. All eight tasks improve, with gains from 1.51 points on BoolQ to 23.62 on HellaSwag (Tab. 10 and Fig. 3). Moving from A8 to A4 raises canonical-path PPL by only 0.59 and lowers accuracy by 0.89 points, versus 14.35 PPL and 6.97 points for ParetoQ. Extending the interface to activations thus preserves quality as activation precision tightens.

Vector dimension changes the tradeoff. At 1B W2A4KV2, scalar CanonQ-Omni-d1 leads $d_W = 2$ on both metrics. At 3B, $d_W = 2$ raises accuracy by 1.67 points while PPL increases from 10.93 to

Table 2: LLaMA3-1B results without KV-cache quantization. Best and second-best non-FP results in each column are shaded and underlined, respectively. Full task scores appear in Tabs. 9 and 10.

(a) Weight-only quantization

	W1A16		W2A16	
Method	PPL ↓	Acc. ↑	PPL ↓	Acc. ↑
FP Model	9.75	58.54	9.75	58.54
RTN (PTQ)	4.2e8	33.7	1.2e6	36.40
GPTQ (PTQ)	7.9e4	38.03	1.2e4	38.21
QuaRot (PTQ)	1.3e4	38.34	3.2e3	38.25
ParetoQ	18.13	51.55	**12.66**	**55.53**
VQLLM-QAT-d4	22.36	47.90	13.77	55.12
VQLLM-QAT-d8	20.10	49.85	N/A	N/A
CanonQ-W-d1	**16.93**	51.42	12.94	55.11
CanonQ-W-d2	17.41	**51.81**	12.86	55.45
CanonQ-W-d4	17.47	51.17	12.96	55.45
CanonQ-W-d8	18.79	50.56	N/A	N/A

(b) Weight–activation quantization

	W1A8		W2A8		W2A4	
Method	PPL ↓	Acc. ↑	PPL ↓	Acc. ↑	PPL ↓	Acc. ↑
FP Model	9.75	58.54	9.75	58.54	9.75	58.54
RTN (PTQ)	1.1e6	34.94	1.2e6	36.14	1.1e6	36.78
GPTQ (PTQ)	1.0e5	38.09	1.3e4	38.16	6.9e4	36.86
QuaRot (PTQ)	1.3e4	38.36	3.2e3	38.42	4.6e3	38.02
ParetoQ	21.27	48.65	18.22	51.09	32.57	44.12
VQLLM-QAT-d4	25.98	46.74	16.99	52.73	42.65	43.45
VQLLM-QAT-d8	25.37	47.47	N/A	N/A	N/A	N/A
CanonQ-W-d1	23.54	48.37	14.38	54.76	35.16	44.14
CanonQ-W-d2	21.23	49.45	13.49	53.82	35.20	43.66
CanonQ-W-d4	22.10	48.75	16.01	52.95	35.53	44.53
CanonQ-W-d8	20.54	49.59	N/A	N/A	N/A	N/A
CanonQ-Omni-d1	17.76	**51.25**	**13.25**	**55.60**	**13.84**	**54.71**
CanonQ-Omni-d2	**17.47**	49.96	13.42	54.63	14.51	54.53

11.16. The reference-law comparison in Thm. 1 explains the reconstruction choices available at a fixed index rate; task-level results determine which choice is useful after adaptation. Together, these results support selecting dimension by validation quality and lookup cost within the shared interface.

5.4 Canonicalization and QAT Play Complementary Roles

We next examine the complementary roles of source canonicalization and model adaptation through distribution diagnostics, zero-update comparisons, and cache-path ablations.

HDP aligns dynamic-source marginals. Layer-mean W_1 distances to the Gaussian reference fall from 0.101 to 0.010 for activations, 0.060 to 0.026 for keys, and 0.030 to 0.018 for values; weights change from 0.015 to 0.017 (Fig. 2). The largest distributional change is in activations, which also show the strongest separation between weight-only and full-stack canonicalization at A4. Keys and values move closer to the same reference, while weights need little marginal correction. These marginal measurements support the use of a shared reference construction for heterogeneous W/A/K/V sources.

The fixed codec benefits from task-aware adaptation. At 1B W2A4, applying the scalar canonical W/A codec with zero updates gives 7043.43/38.83, versus 13.84/54.71 after QAT (Tab. 7). PPL falls approximately 509× while the table stays fixed. The codec supplies reconstruction choices, but the pretrained network must still adapt to using them throughout the computation. Marginal alignment and task-aware training thus address different parts of the problem. The KV2 zero-update comparison also changes the cache path and measures their combined effect.

The cache path also matters. The uniform-KV CanonQ-Omni-d1 ablation reaches 22.82/47.72 at W2A8KV2 and 21.65/48.30 at W2A4KV2, compared with 15.70/53.45 and 15.42/53.56 for canonical KV. The corresponding accuracy gains are 5.73 and 5.26 points (Tab. 8), so the cache contribution persists at both activation precisions. At 1B, moving from W2A4 to W2A4KV2 costs the complete canonical path 1.58 PPL and 1.15 accuracy points, compared with 58.00 PPL and 3.65 points for ParetoQ. The ablation evaluates the complete canonical cache path, combining transformation, active levels, normalization, and backward behavior.

5.5 Stable Relative Perplexity Within the 8K Training Range

After 8K continued QAT for 40K updates, we evaluate LLaMA3-1B W2A4KV2 at 2K/4K/8K. At 8K, CanonQ-Omni achieves 14.86 PPL, versus 39.50 for CanonQ-W and 45.63 for ParetoQ. Its PPL relative to the same-length FP reference stays within 1.69–1.74×, showing a stable relative quality gap as the evaluation window grows within the trained range. Full results and protocol appear in Sec. C.

Table 3: Code generation and mathematical reasoning on `MobileLLM-Pro-Instruct-1B` `W2A16KV16`. Scores are percentages, with correct-task counts in parentheses. `MBPP/HumanEval` report greedy pass@1; `GSM8K` reports final-answer exact match (EM). Quantized settings share the teacher and evaluation step. `GSM8K` also guides learning-rate selection; see Sec. A.5.

Benchmark	Metric	Tasks	FP base	ParetoQ	VQLLM-QAT-d4	CanonQ-d1
`MBPP`	pass@1	500	46.4 (232)	28.0 (140)	27.8 (139)	**32.2** (161)
`HumanEval`	pass@1	164	60.4 (99)	43.9 (72)	33.5 (55)	**62.2** (102)
`GSM8K`	EM	1319	53.8 (709)	34.3 (453)	32.7 (431)	**47.8** (630)

5.6 Gains Extend to Code Generation and Reasoning

We evaluate `MobileLLM-Pro-Instruct-1B` at `W2A16KV16` using the shared-teacher protocol in Sec. A.5. These experiments extend the evaluation to generated solutions under weight quantization.

Mathematical reasoning. On `GSM8K`, CanonQ-d1 solves 630 of 1,319 problems (47.8%), versus 453 for ParetoQ and 431 for VQLLM-QAT-d4. This gives a 39.1% relative improvement over the stronger quantized baseline and retains 88.9% of the FP score. `GSM8K` guides learning-rate selection as specified in Sec. A.5.

Code generation. CanonQ-d1 achieves 32.2% `MBPP` and 62.2% `HumanEval` greedy pass@1 (Tab. 3), exceeding ParetoQ by 4.2 and 18.3 points. The `HumanEval` gain is a 41.7% relative improvement. The FP reference scores 60.4% (paired comparison: $p = 0.74$). Neither coding benchmark guides selection, and CanonQ leads the evaluated quantized baselines on all three generation benchmarks.

6 Discussion and Limitations

A shared interface for coupled errors. CanonQ separates reusable reconstruction choices from task-aware adaptation. Canonicalization makes a fixed reference alphabet useful across heterogeneous sources, while joint QAT adapts the network to their interacting errors. The larger gains under joint compression, together with the alignment and zero-update studies, support this division of roles. These results favor selecting the codec under the complete target precision configuration: methods with similar weight-only quality separate sharply when activations and caches are also quantized. Evaluating the intended W/A/KV format therefore captures interactions that weight-only comparisons leave untested.

Choosing scalar or vector assignment. The interface supports both choices at a fixed index rate. Scalar tables give the strongest reported `1B W2A4KV2` result; vector assignment improves `3B` accuracy with a small PPL tradeoff. Dimension is therefore a practical choice guided by validation quality and lookup cost. Joint tuple statistics and sensitivity-weighted residuals offer concrete quantities for explaining these differences.

Deployment and further evaluation. The present study establishes model quality at extreme low bit rates. Deployment evaluation will complement these results with exported-model size, KV-storage, and prefill/decode throughput measurements for fused kernels, accounting for group metadata and unquantized modules (Sec. A.4). Evaluation beyond the `8K` training range, reasoning benchmarks with independent hyperparameter selection, and repeated training seeds are directions for extending this evidence.

7 Conclusion

We introduced CanonQ, a unified framework for weight, activation, and KV-cache quantization using canonical coordinates and frozen reference codebooks. Our theory links source mismatch to reconstruction, task loss, and normalization-aware STE bias. Experiments establish substantial quality gains under `W2A4KV2` across model scales, with further improvements in coding and reasoning and stable relative perplexity within the `8K` training range.

Reproducibility Statement

Sec. 3 specifies the canonical codec and training interface. Secs. 4 and D state the theoretical assumptions and provide complete proofs. Sec. A documents the standard QAT recipe, quantizer paths, group sizes, seeds, table construction, evaluation procedures, and storage accounting. Sec. B.3 provides per-task results. Sec. C describes context-length evaluation, and Sec. A.5 details the coding and mathematical-reasoning metrics, selection protocol, and interpretation.

AI Use Statement

We used ChatGPT to improve the clarity and readability of the manuscript. The authors take full responsibility for the final content.

Ethics Statement

This work studies quantization of pretrained language models using existing training corpora and benchmarks. More compact models can broaden access to useful capabilities while also lowering barriers to misuse. Quantized models may retain biases and harmful behaviors of their source models, and compression can introduce additional reliability failures. Our quality evaluations do not establish safety. Deployment should respect model and data licenses and include application-specific reliability and safety assessment.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.00456.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Ran Ben-Basat, William Kuszmaul, Michael Mitzenmacher, Amit Portnoy, and Shay Vargaftik. Quantizing with randomized hadamard transforms: Efficient heuristic now proven. *arXiv preprint arXiv:2605.06014*, 2026. URL https://arxiv.org/abs/2605.06014.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2307.13304.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. Efficientqat: Efficient quantization-aware training for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10081–10100, 2025.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2208.07339.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.14314.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984. doi: 10.1214/aos/1176346703.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2401.06118.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.17323.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2401.18079.

Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. BiLLM: Pushing the limit of post-training quantization for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2402.04291.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2306.07629.

Dongyue Li, Zechun Liu, Kai Yi, Zhenshuo Zhang, Changsheng Zhao, Raghuraman Krishnamoorthi, Harshit Khaitan, Hongyang R Zhang, and Steven Li. Winq: Accelerating quantization-aware training of language models around saddle points. *ICML*, 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, 2024. URL https://arxiv.org/abs/2306.00978.

Yoseph Linde, Andres Buzo, and Robert M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, January 1980. doi: 10.1109/TCOM.1980.1094577.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484. Association for Computational Linguistics, 2024a. URL https://aclanthology.org/2024.findings-acl.26/.

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

Zechun Liu, Changsheng Zhao, Hanxian Huang, Sijia Chen, Jing Zhang, Jiawei Zhao, Scott Roy, Lisa Jin, Yunyang Xiong, Yangyang Shi, Lin Xiao, Yuandong Tian, Bilge Soran, Raghuraman Krishnamoorthi, Tijmen Blankevoort, and Vikas Chandra. ParetoQ: Scaling laws in extremely low-bit LLM quantization. *arXiv preprint arXiv:2502.02631*, 2025b. URL https://arxiv.org/abs/2502.02631.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR, 2024b. URL https://proceedings.mlr.press/v235/liu24bz.html.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit LLMs: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024. URL https://arxiv.org/abs/2402.17764.

Vladimir Malinovskii, Denis Mazur, Ivan Ilin, Denis Kuznedelev, Konstantin Burlachenko, Kai Yi, Dan Alistarh, and Peter Richtarik. Pv-tuning: Beyond straight-through estimation for extreme llm compression. *Advances in Neural Information Processing Systems*, 37:5074–5121, 2024.

Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, 1960. doi: 10.1109/TIT.1960.1057548.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Yuzhang Shang, Zhihang Yuan, Qiang Wu, and Zhen Dong. PB-LLM: Partially binarized large language models. *arXiv preprint arXiv:2310.00034*, 2023. URL https://arxiv.org/abs/2310.00034.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8Wuvhh0LYW.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656. PMLR, 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. DRIVE: One-bit distributed mean estimation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 362–377. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0397758f8990c1b41b81b43ac389ab9f-Abstract.html.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. EDEN: Communication-efficient and robust distributed mean estimation for federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21984–22014. PMLR, 2022. URL https://proceedings.mlr.press/v162/vargaftik22a.html.

Haoyu Wang, Xingyu Yu, Haiyan Zhao, Fengxiang Wang, and Xu Han. LC-QAT: Data-efficient 2-bit QAT for LLMs via linear-constrained vector quantization. *ICML*, 2026.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. BitNet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453*, 2023. URL https://arxiv.org/abs/2310.11453.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. OneBit: Towards extremely low-bit large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.11295.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kai Yi, Vignesh Vivekraja, Harshit Khaitan, and Steven Li. Jacquant: Ste-free quantization-aware training via learned jacobian surrogates. *arXiv preprint arXiv:2605.25469*, 2026.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.05662.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. *arXiv preprint arXiv:2504.19874*, 2025. URL https://arxiv.org/abs/2504.19874.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 CanonQ: Unified W/A/KV Quantization **3**
3.1 Reusable Codebooks in Canonical Coordinates . . . 3
3.2 Joint Adaptation to Quantization Errors . . . 3
3.3 Scalar/Vector Variants and Computational Cost . . . 4

4 Theory: From Source Alignment to Optimization **4**
4.1 Bounding Frozen-Codebook Transfer Error . . . 4
4.2 Bounding Local Task Loss from Quantization . . . 5
4.3 Quantization Distortion Controls STE Error . . . 5

5 Experiments **6**
5.1 Evaluation Setup . . . 6
5.2 Large Quality Gains at W2A4KV2 . . . 7
5.3 Gains Grow Beyond Weight-Only Quantization . . . 7
5.4 Canonicalization and QAT Play Complementary Roles . . . 8
5.5 Stable Relative Perplexity Within the 8K Training Range . . . 8
5.6 Gains Extend to Code Generation and Reasoning . . . 9

6 Discussion and Limitations **9**

7 Conclusion **9**

A Experimental Details **15**
A.1 Models, Training, and Evaluation . . . 15
A.2 Quantization Paths and Baselines . . . 15
A.3 HDP Maps and Frozen Tables . . . 16
A.4 Index Rate and Stored Footprint . . . 16
A.5 Code Generation and Mathematical Reasoning . . . 17

B Supplementary Experimental Results **18**
B.1 Marginal Alignment and Task Adaptation . . . 18
B.2 Canonical versus Uniform Cache Quantization . . . 19
B.3 Per-task Results across Precisions and Models . . . 19

C Context-Length Robustness after 8K Continued QAT **24**

D Theory Details and Proofs **24**
D.1 The Haar Reference and the Implemented Hadamard Map . . . 24
D.2 Root-Distortion Stability and Frozen-Table Transfer . . . 26
D.3 Gain-Aware Task-Loss Bounds . . . 27
D.4 The STE Jacobian Induced by the Canonical Codec . . . 28
D.5 Distortion-Controlled Optimization Consequences . . . 30
D.6 Voronoi Margins, Code Switching, and Index Stability . . . 31

A Experimental Details

A.1 Models, Training, and Evaluation

We use pretrained `LLaMA-3.2-1B/3B`, `LLaMA3-8B`, and `Qwen3-1.7B` checkpoints with their native tokenizers (Dubey et al., 2024; Yang et al., 2025). The `1B` model supports the broad precision and component study; the larger models test transfer of the same frozen-table construction. Standard QAT starts from the corresponding full-precision checkpoint and samples a continuous `FineWeb-Edu` stream. The architecture, vocabulary, and positional encoding are unchanged. Training excludes the evaluation datasets.

Optimization. We train on eight `H100/A100-80G` GPUs, with eight sequences per GPU. The standard recipe uses `2048`-token sequences, effective batch size `64`, and at most `240K` updates, or approximately `31.5B` tokens. We use AdamW with $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, and global gradient-norm clipping at 1.0. A cosine schedule follows `1000` linear-warmup steps and decays to zero. Learning rates are selected from $\{1, 2, 4\} \times 10^{-5}$ on held-out validation data; enabled scale and gain parameters share the weights' learning rate. For the reported `8B` CanonQ-Omni configuration, we reuse the best training hyperparameters selected at `3B` to reduce tuning costs. The training seed is `42`. The `8K/40K` continuation and instruction-tuned coding and reasoning extension use the separate protocols in Secs. A.5 and C.

Language modeling and downstream tasks. For `WikiText-2`, we concatenate tokenized test documents without added separators, split the stream into non-overlapping `2048`-token windows, and drop the trailing partial window. PPL is the exponential of mean next-token negative log-likelihood over the scored tokens. All methods use the same tokenized corpus and evaluation implementation. Zero-shot accuracy is the unweighted mean over `ARC-Easy` and `ARC-Challenge` (Clark et al., 2018), `BoolQ` (Clark et al., 2019), `PIQA` (Bisk et al., 2020), `SIQA` (Sap et al., 2019), `HellaSwag` (Zellers et al., 2019), `OpenBookQA` (Mihaylov et al., 2018), and `WinoGrande` (Sakaguchi et al., 2020), using common prompts and answer normalization across methods. Evaluation re-quantizes shadow weights and keeps the trained activation/cache operators enabled. Each result represents one training seed unless stated otherwise.

A.2 Quantization Paths and Baselines

Tab. 4 summarizes which sources each method quantizes. `A16/KV16` denotes a floating-point path; other labels specify index bits. A canonical (b, d) table has 2^{bd} vector codewords, so a bd-bit index represents d scalars. Canonical A/K/V paths use $d = 1$ throughout; method suffixes specify d_W.

Table 4: Quantization paths. "Canonical" means HDP, energy normalization, and a frozen Gaussian-reference table. All QAT paths are active during training and evaluation.

Method	Weights	Activations	KV cache
ParetoQ	learned uniform grid	dynamic uniform	dynamic uniform
VQLLM-QAT	group-fitted raw codebook	dynamic uniform	dynamic uniform
CanonQ-W	canonical	dynamic uniform	dynamic uniform
CanonQ-Omni	canonical	canonical	canonical
HDP+LM PTQ	canonical, no updates	canonical, no updates	uniform if enabled

Baseline scope. ParetoQ supplies the weight quantizer, extended here with the uniform A/KV operators below (Liu et al., 2025b). VQLLM-QAT fits raw-space weight tables by k-means/LBG; matching its index rate and tuple dimension to CanonQ-W changes both source coordinates and table construction. RTN and GPTQ specify weight PTQ procedures; their low-A/KV rows evaluate composite configurations. KIVI specifies a cache codec with inherited W/A quantization (Frantar et al., 2023; Ashkboos et al., 2024; Liu et al., 2024b). These composite rows measure end-to-end configurations and do not isolate the original weight or cache algorithm.

Uniform weight and dynamic-source quantizers. The ParetoQ weight path uses a learned per-row scale α: `W1` reconstructs $\alpha \operatorname{sign}(w)$, while `W2` rounds to $\frac{\alpha}{4}\{-3,-1,1,3\}$. For a dynamic A/KV group x, the implemented symmetric operator is

$$M = \max_i |x_i|, \quad s = \frac{2^{b-1}-1}{M+\epsilon}, \quad \widehat{x} = \frac{\operatorname{round}(sx)}{s+\epsilon}. \tag{16}$$

Its active cardinalities are `15/255/3` at `A4/A8/KV2`, compared with `16/256/4` for canonical scalar tables. Consequently, comparisons at the same index rate can still differ in cardinality, normalization, and backward behavior.

Canonical paths. All sources use the codec in Eqs. (1) and (3), with source-specific $(g_S, b_S, d_S, R_S, \gamma_{S,\ell})$. Activations and KV states use energy normalization without an additional learned, absmax, or percentile clipping parameter. K and V are quantized independently and decoded to their original coordinates before attention. Assignment uses the identity STE, with the energy differentiated through encoding and decoding as in Eq. (13). A source with a fixed unit gain sets $\gamma_{S,\ell} = 1$.

A.3 HDP Maps and Frozen Tables

Grouping and rotations. Canonical groups contain $g_W = g_A = 128$ scalar coordinates and $g_K = g_V = 64$. Each source uses one fixed map $R_S = P_S H D_S$ across layers. Seeds are $s_S = s_0 + o_S$, with $s_0 = 0$ and offsets $o_W = 0$, $o_A = 1$, $o_K = o_V = 2$; a seeded PRNG draws signs and then the permutation. Thus K and V share a map, while W/A/KV maps differ. The normalized fast Hadamard transform uses `hammer.ops.fast_hadamard_transform` with a `PyTorch` butterfly fallback. Power-of-two groups require no padding. Quantized linear layers include attention q, k, v, o and MLP w_1, w_2, w_3 projections; token embeddings and the language-model head remain unrotated.

Offline construction and assignment. Scalar tables are fitted to $\mathcal{N}(0,1)$ using double-precision Lloyd–Max iterations; vector tables use Lloyd/LBG assignments and centroid updates. The fitting gap ε_{vq} in Eq. (6) explicitly accounts for numerical codebook optimization. The scalar `W1` levels are ± 0.7978846; the scalar `W2/KV2` levels are approximately ± 0.4528 and ± 1.5104. Exact A4/A8 levels and vector centers are serialized with the configuration. Tables remain frozen through QAT and evaluation.

For `W1`, the evaluated dimensions $d \in \{1,2,4,8\}$ give $K \in \{2,4,16,256\}$; for `W2`, $d \in \{1,2,4\}$ gives $K \in \{4,16,256\}$. Scalar `A4/A8/KV2` tables have `16/256/4` entries. Sources literally share a table only when (b,d) and the codewords match. Scalar assignment uses sorted midpoint thresholds and `torch.bucketize` with `right=False`, so midpoint ties select the lower level. Vector assignment uses `argmin` and selects the lowest index on ties. The same rule is used during training and evaluation.

A.4 Index Rate and Stored Footprint

The label `W2A4KV2` specifies an index-rate tuple of `2/4/2/2`; it excludes decoding factors, table storage, and unquantized modules. For source S with N_S scalar values, the stored rate and once-per-model allocation are

$$b_S^{\text{stored}} = b_S^{\text{idx}} + \frac{B_S^{\text{decode}} + B_S^{\text{local-table}} + B_S^{\text{aux}}}{N_S}, \tag{17}$$

$$B^{\text{resident}} = B^{\text{global-table}} + B^{\text{rotation}} + B^{\text{constant}}. \tag{18}$$

Here auxiliary bits include offsets, padding, and alignment. Globally shared objects are counted once; a group-fitted codebook contributes to local-table storage.

Decoding metadata. A group of g values with r stored q_s-bit decoding scalars adds rq_s/g bits per value. A static canonical weight group needs the folded factor $\alpha_g = \gamma_g \rho_g / \sqrt{g}$ in $\widehat{u}_g = \alpha_g R_g^\top \widehat{z}_g$. Storing this product requires one scalar; retaining norm and gain separately requires two. Table 5

Table 5: Canonical W2A4KV2 storage before padding and unquantized modules. Each decoding scalar is 16 bits; r counts independent factors after folding. A rates apply only to materialized quantized activation buffers. Global tables are separate.

Source	Index bits	Group g	Stored bits, $r = 1$	Stored bits, $r = 2$
Weights	2	128	2.125	2.250
Activations	4	128	4.125	4.250
Keys / values	2	64	2.250	2.500

gives both cases for the declared scalar group sizes. The rates are representation accounting, not measured exported bytes.

The same accounting gives 2.25 bits per value for a symmetric two-bit cache with one 16-bit scale per 64 values. An asymmetric cache with a scale and offset uses $2 + (q_s + q_z)/g$ bits per value; at $g = 32$ and $q_s = q_z = 16$, this is 3.0. Per-row ParetoQ weight scales instead contribute $16 \sum_\ell m_\ell / \sum_\ell m_\ell n_\ell$ bits per weight for matrices of shape $m_\ell \times n_\ell$. Total storage therefore accounts for both index bits and method-specific metadata.

Shared and group-fitted tables. A canonical (b, d) table with q_C-bit components occupies $2^{bd} d q_C$ bits. At 16 bits per component, W2-d1/d2/d4 tables require 8/64/2048 bytes and an A4-d1 table requires 32 bytes. Each distinct exported table is counted once. In contrast, a raw-space table shared by each block of 128 output rows contributes

$$B_W^{\text{local-table}} = \sum_\ell \lceil m_\ell / 128 \rceil \, 2^{bd} d q_C. \tag{19}$$

These table-sharing blocks differ from the 128-scalar canonical normalization groups. Any separately retained raw-codebook normalization factors add their own metadata; gains absorbed into raw codewords are not counted twice.

Deployment implications. HDP maps can be regenerated from seeds without storing dense matrices; exported signs, permutations, and replicated tables still occupy resident bytes. KV decoding factors persist with cached indices. Activation factors can remain on chip in a fused kernel, so their contribution depends on whether an activation buffer is materialized. The relevant deployment quantity is $M_{\text{total}}(B, L) = M_W + M_{\text{resident}} + M_{KV}(B, L) + M_A^{\text{peak}}$, where M_W includes unquantized modules. Actual memory and speed also depend on packing, lookup, and kernel boundaries.

A.5 Code Generation and Mathematical Reasoning

Tab. 3 evaluates MobileLLM-Pro-Instruct-1B at W2A16KV16. Token embeddings and output/language-model heads remain full precision. CanonQ-d1, VQLLM-QAT-d4, and ParetoQ use canonical scalar, raw vector, and uniform weight quantizers, respectively. All three arms start from the same instruction-tuned FP checkpoint and use it as a full-precision distillation teacher.

Checkpoint and learning-rate selection. All quantized methods are evaluated at step 279236. For each method with a learning-rate sweep, GSM8K exact match selects the setting used for all three benchmarks. The reported reasoning score therefore reflects GSM8K-guided selection, while both coding benchmarks evaluate the selected model without contributing to selection.

Coding metrics and decoding. Evaluation uses a dialog prompt: HumanEval is zero-shot with maximum prompt/generation lengths 2000/1024, while MBPP is three-shot with maximum generation length 256. Greedy decoding generates one sample per task, with seed 42 and batch size 32. The FP reference uses the same harness and decoding settings. Generated code runs in an isolated subprocess with a five-second timeout; a timeout or nonzero exit is a failure. Thus pass@1 is the fraction of solved tasks, and parentheses in Tab. 3 give integer counts.

Mathematical-reasoning metric. GSM8K exact match measures final-answer correctness on 1,319 problems. FP, ParetoQ, VQLLM-QAT-d4, and CanonQ-d1 solve 709, 453, 431, and 630 problems, respectively. We display scores to one decimal place and compute improvements from the counts

before rounding. CanonQ gives net gains of 177 and 199 correct answers over the quantized baselines. Relative to ParetoQ, this is $100(630-453)/1319 = 13.4$ accuracy points or $100(630/453-1) = 39.1\%$ relative improvement. CanonQ retains $630/709 = 88.9\%$ of the FP score and recovers $(630-453)/(709-453) = 69.1\%$ of the baseline-to-FP gap.

Uncertainty and interpretation. On `HumanEval`, CanonQ and FP solve 102 and 99 tasks, with 20 solved only by CanonQ and 17 only by FP. Exact McNemar testing gives $p = 0.74$; the paired-bootstrap 95% interval for the accuracy difference is $[-5.5, 9.1]$ points. This interval spans both signs, and no equivalence margin was tested. On `MBPP`, the discordant counts are 17 and 88 ($p < 10^{-12}$), consistent with the 14.2-point FP gap. These comparisons quantify variation across benchmark tasks under greedy decoding; training-seed variation is a separate quantity. For `GSM8K`, the available aggregate counts support accuracy and count comparisons; paired significance testing requires per-problem outcomes.

B Supplementary Experimental Results

B.1 Marginal Alignment and Task Adaptation

We test the scalar-marginal consequence of canonicalization using a trained `LLaMA3-1B W2A4KV2` checkpoint. Weights come from the checkpoint; activations and KV states are captured at quantizer inputs on a fixed calibration batch. Each comparison uses identical groups, with $g_W = g_A = 128$ and $g_K = g_V = 64$, and identical energy normalization. Thus $z_{\text{norot}} = \sqrt{g_S}x/(\|x\|_2 + \epsilon)$ and $z_{\text{HDP}} = R_S z_{\text{norot}}$ differ only by rotation. Measurements precede codeword assignment.

For each layer we compute scalar W_1 distance to $\mathcal{N}(0,1)$ and the KS distance $\sup_t |\widehat{F}(t) - \Phi(t)|$. Table 6 averages layers equally. HDP reduces activation W_1 by 90%, key W_1 by 57%, and value W_1 by 40%; weights are already close and show a small increase. The range of source-wise means narrows from 0.015–0.101 to 0.010–0.026. This explains the shared marginal reference in Fig. 2. These statistics assess marginal alignment. Estimating $\Delta_{S,d}$ requires a W_2 assessment of the corresponding tuple law, including coordinate dependence when $d > 1$.

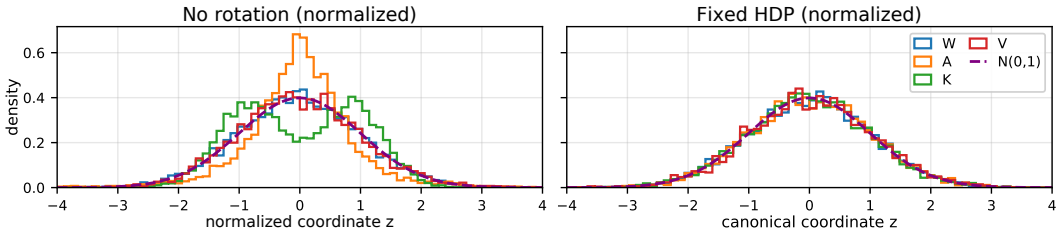


Figure 2: Paired scalar marginals before and after HDP, using identical normalized groups. Dynamic A/K/V sources approach the Gaussian reference; weights remain near it. These scalar diagnostics illustrate alignment with the shared design reference. Full statistics appear in Tab. 6.

Table 6: Paired scalar-marginal distances to $\mathcal{N}(0,1)$ for a trained `LLaMA3-1B W2A4KV2` checkpoint. Values are layer means $\pm$ standard deviations, not uncertainty across training seeds. Both conditions use the same groups and normalization.

	W_1 to $\mathcal{N}(0,1)$		KS to $\mathcal{N}(0,1)$	
Source	No rotation	HDP	No rotation	HDP
Weights W	0.015 ± 0.008	0.017 ± 0.012	0.006 ± 0.003	0.010 ± 0.006
Activations A	0.101 ± 0.055	$\mathbf{0.010 \pm 0.004}$	0.039 ± 0.019	$\mathbf{0.005 \pm 0.002}$
Keys K	0.060 ± 0.017	$\mathbf{0.026 \pm 0.013}$	0.035 ± 0.011	$\mathbf{0.014 \pm 0.006}$
Values V	0.030 ± 0.014	$\mathbf{0.018 \pm 0.008}$	0.016 ± 0.006	$\mathbf{0.009 \pm 0.003}$

Adaptation to a fixed codec. The zero-update control in Tab. 7 enables the canonical W/A codec directly on the pretrained 1B model, with scalar tables, group size 128, rotation seed 0, and no gradient updates or layer reconstruction. At W2A4, QAT reduces PPL from 7043.43 to 13.84 and raises accuracy by 15.88 points without changing table capacity. The result demonstrates the importance of model adaptation to this fixed codec. The KV2 pair additionally changes uniform KV to canonical KV, so its improvement reflects both adaptation and the cache path.

Table 7: Zero-update canonical PTQ and trained CanonQ-Omni-d1 on LLaMA3-1B. The W2A4 pair shares the W/A codec; the KV2 pair also changes the cache codec.

Format	Method / cache path	PPL ↓	Avg. ZS ↑
FP	Full precision	9.75	58.54
W2A4	HDP+LM PTQ	7043.43	38.83
W2A4	CanonQ-Omni-d1 QAT	13.84	54.71
W2A4KV2	HDP+LM PTQ, uniform KV	8868.91	36.76
W2A4KV2	CanonQ-Omni-d1 QAT, canonical KV	15.42	53.56

B.2 Canonical versus Uniform Cache Quantization

Tab. 8 compares cache codecs while keeping W/A quantization canonical. The uniform-KV variant uses the dynamic uniform operator in Sec. A.2. Canonical KV improves both PPL and eight-task accuracy at A8 and A4, extending the benefit beyond activation handling. This comparison evaluates the complete cache design, including its transform, active levels, normalization, and backward graph.

Table 8: Cache-path ablation for LLaMA3-1B CanonQ-Omni-d1. W/A paths remain canonical; the comparison changes the complete cache path.

Format	Cache path	PPL ↓	Avg. ZS ↑
W2A8KV2	uniform	22.82	47.72
	canonical	**15.70**	**53.45**
W2A4KV2	uniform	21.65	48.30
	canonical	**15.42**	**53.56**

B.3 Per-task Results across Precisions and Models

Tables 9–13 report the complete eight-task scores. ARC-e/c, HellaS., OBQA, and Wino. abbreviate ARC-Easy/Challenge, HellaSwag, OpenBookQA, and WinoGrande. The mean weights the eight tasks equally.

Weight-only and activation precision. In Tab. 9, 1B W2A16 gives closely matched results for ParetoQ and CanonQ-W-d2: 12.66/55.53 versus 12.86/55.45 (PPL/accuracy). At W1A16, CanonQ-W-d1 has the lowest PPL, while $d_W = 2$ has the highest accuracy. The matched-$d_W = 4$ comparison improves raw-space VQLLM-QAT from 22.36/47.90 to 17.47/51.17. In Tab. 10, extending canonicalization to activations preserves low PPL when the uniform activation path degrades: at W2A4, CanonQ-Omni-d1 achieves 13.84 PPL against 35.16 for CanonQ-W-d1.

Joint compression at 1B. Table 11 shows that CanonQ-Omni-d1 retains similar quality at A8 and A4 with KV2: 15.70/53.45 and 15.42/53.56. The corresponding ParetoQ results fall from 47.58/45.13 to 90.57/40.47. At W2A4KV2, CanonQ-Omni-d1 improves all eight tasks over ParetoQ; scalar canonical tables suffice for this operating point.

Weight-only and joint compression at 3B. Table 12 places all 3B precisions under one FP reference. The W2A16 difference is small: CanonQ-W-d2 reaches 9.19/63.02 versus 9.29/62.84 for ParetoQ. Under W2A4KV2, the separation grows to 11.16/60.04 versus 92.64/39.39. Dimension

Table 9: LLaMA3-1B weight-only quantization. All task scores are zero-shot accuracies (%); PPL is measured on WikiText-2.

Method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	Wino.	Avg. ZS ↑	PPL ↓
FP16 Base	64.57	42.85	65.35	75.05	44.91	64.31	50.00	61.25	58.54	9.75
W2A16										
RTN (PTQ)	27.19	24.69	54.03	50.00	35.89	25.43	23.05	50.94	36.40	1.2e6
GPTQ (PTQ)	26.95	25.31	61.66	49.12	36.67	25.47	29.69	50.78	38.21	1.2e4
QuaRot (PTQ)	28.12	24.69	62.35	50.34	36.23	26.72	28.12	49.45	38.25	3.2e3
ParetoQ	64.60	39.69	61.05	72.75	42.79	56.74	47.85	58.75	55.53	12.66
VQLLM-QAT-d4	62.89	40.62	62.65	72.61	42.63	54.43	46.09	59.06	55.12	13.77
CanonQ-W-d1	64.38	38.05	62.71	72.36	43.21	56.19	46.68	57.27	55.11	12.94
CanonQ-W-d2	62.77	41.02	63.91	71.68	43.21	56.35	45.31	59.38	55.45	12.86
CanonQ-W-d4	64.34	41.09	61.51	72.41	43.46	56.46	47.66	56.64	55.45	12.96
W1A16										
RTN (PTQ)	25.0	22.5	37.6	49.5	32.9	25.0	27.1	49.6	33.7	4.2e8
GPTQ (PTQ)	28.59	22.73	62.32	50.29	36.62	25.29	28.71	49.69	38.03	7.9e4
QuaRot (PTQ)	28.16	26.02	62.41	48.58	37.16	24.80	28.71	50.86	38.34	1.3e4
ParetoQ	59.73	35.62	61.54	68.65	42.29	46.48	43.75	54.38	51.55	18.13
VQLLM-QAT-d4	53.55	31.72	60.22	65.33	40.87	40.03	37.11	54.38	47.90	22.36
VQLLM-QAT-d8	57.77	36.80	59.83	66.65	41.60	42.93	40.62	52.58	49.85	20.10
CanonQ-W-d1	58.95	35.55	61.66	68.65	42.77	47.20	42.97	53.59	51.42	16.93
CanonQ-W-d2	60.66	36.41	61.96	68.60	41.70	46.76	42.19	56.17	51.81	17.41
CanonQ-W-d4	58.79	36.25	61.99	68.16	41.75	46.98	40.82	54.61	51.17	17.47
CanonQ-W-d8	57.77	34.63	59.68	68.10	41.63	44.85	42.38	55.47	50.56	18.79

changes the PPL–accuracy tradeoff: $d_W = 1$ yields lower PPL (10.93), while $d_W = 2$ raises accuracy from 58.37 to 60.04. Neither dimension dominates both metrics.

Transfer across model family and scale. In Tab. 13, CanonQ-Omni-d1 improves every task over ParetoQ on both Qwen3-1.7B and LLaMA3-8B. The largest gains occur on HellaSwag: 17.93 and 42.68 points, respectively. Improvements also span the remaining reasoning and commonsense tasks, so the aggregate gain reflects broad recovery with task-dependent magnitudes. The different PPL and accuracy gaps to FP reinforce the value of evaluating likelihood and task performance together.

Where the aggregate gain comes from. Figure 3 connects the precision sweep to its task-level pattern. As compression extends from W2A8 to W2A4KV2, CanonQ-Omni-d1 stays within 13.25–15.42 PPL, while the other QAT paths degrade. In W2A4, all eight tasks improve over ParetoQ; gains range from 1.51 points on BoolQ to 23.62 on HellaSwag. Thus the mean improvement is distributed across tasks, although its size depends strongly on the benchmark.

Table 10: LLaMA3-1B with quantized weights and activations, retaining the floating-point cache. The three panels vary weight and activation precision.

Method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	Wino.	Avg. ZS ↑	PPL ↓
FP base	64.57	42.85	65.35	75.05	44.91	64.31	50.00	61.25	58.54	9.75
W2A8										
RTN (PTQ)	26.91	25.16	51.47	50.29	36.43	25.72	23.05	50.08	36.14	1.2e6
GPTQ (PTQ)	26.68	25.39	62.08	48.39	37.11	25.56	29.69	50.39	38.16	1.3e4
QuaRot (PTQ)	28.71	24.38	62.44	50.15	36.08	26.74	28.91	49.92	38.42	3.2e3
ParetoQ	58.99	35.73	61.56	68.16	41.43	45.72	42.58	54.53	51.09	18.22
VQLLM-QAT-d4	61.36	37.42	61.20	70.10	42.69	49.91	43.75	55.39	52.73	16.99
CanonQ-W-d1	63.56	40.71	61.41	71.01	43.50	54.06	45.51	58.36	54.76	14.38
CanonQ-W-d2	64.15	39.95	59.92	71.07	43.04	55.67	43.16	53.59	53.82	13.49
CanonQ-W-d4	61.06	39.86	62.11	70.20	42.49	50.69	43.36	53.83	52.95	16.01
CanonQ-Omni-d1	65.84	40.20	62.35	72.14	43.85	55.99	47.85	56.56	55.60	13.25
CanonQ-Omni-d2	61.99	38.01	64.17	72.04	43.04	56.63	43.75	57.42	54.63	13.42
W1A8										
RTN (PTQ)	28.24	25.47	39.39	51.46	36.04	24.70	23.63	50.55	34.94	1.1e6
GPTQ (PTQ)	27.77	23.05	62.38	50.98	36.47	25.09	29.69	49.30	38.09	1.0e5
QuaRot (PTQ)	28.01	25.62	62.41	48.88	37.30	24.89	28.32	51.41	38.36	1.3e4
ParetoQ	55.62	32.35	61.92	66.11	41.13	40.65	40.82	50.62	48.65	21.27
VQLLM-QAT-d4	51.90	31.50	62.35	63.95	40.27	36.76	35.94	51.25	46.74	25.98
VQLLM-QAT-d8	51.94	32.26	60.77	63.85	41.08	39.13	38.87	51.88	47.47	25.37
CanonQ-W-d1	55.83	33.78	60.16	65.14	41.48	41.07	36.72	52.73	48.37	23.54
CanonQ-W-d2	55.83	36.57	61.83	65.41	41.23	42.37	39.84	52.50	49.45	21.23
CanonQ-W-d4	54.73	32.77	61.68	66.22	40.47	41.30	41.99	50.86	48.75	22.10
CanonQ-W-d8	54.65	34.88	59.83	67.40	42.19	43.75	40.43	53.59	49.59	20.54
CanonQ-Omni-d1	58.15	36.99	60.25	68.10	42.54	47.21	42.19	54.53	51.25	17.76
CanonQ-Omni-d2	57.39	34.54	57.98	68.37	42.09	46.71	39.84	52.73	49.96	17.47
W2A4										
RTN (PTQ)	26.60	26.02	54.24	50.34	36.82	25.30	23.83	51.09	36.78	1.1e6
GPTQ (PTQ)	27.54	24.06	51.20	52.15	37.06	24.79	25.98	52.11	36.86	6.9e4
QuaRot (PTQ)	26.45	25.31	60.61	51.27	37.35	25.81	27.34	50.00	38.02	4.6e3
ParetoQ	47.76	27.87	58.68	61.96	40.57	31.66	32.23	52.27	44.12	32.57
VQLLM-QAT-d4	45.99	27.62	60.13	61.31	41.03	30.71	31.05	49.77	43.45	42.65
CanonQ-W-d1	45.90	27.96	59.38	60.34	40.98	31.39	34.38	52.81	44.14	35.16
CanonQ-W-d2	47.76	27.79	56.64	61.75	39.67	31.56	33.01	51.09	43.66	35.20
CanonQ-W-d4	47.30	29.48	60.53	60.78	40.12	32.01	34.77	51.25	44.53	35.53
CanonQ-Omni-d1	63.89	39.10	60.19	71.39	43.40	55.28	46.09	58.36	54.71	13.84
CanonQ-Omni-d2	64.40	39.78	62.04	70.64	42.24	54.85	45.90	56.41	54.53	14.51

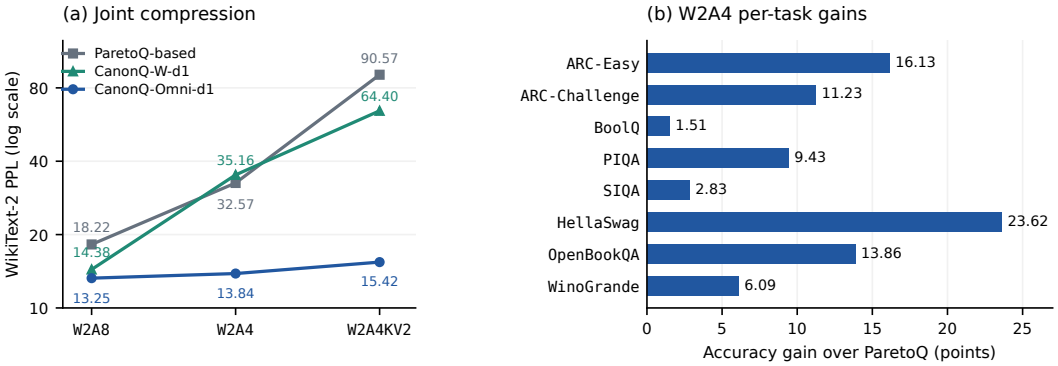


Figure 3: Quality under joint compression on LLaMA3-1B. (a) CanonQ-Omni-d1 retains low PPL as activation and cache precision decrease; the vertical axis is logarithmic. (b) Its W2A4 improvement over ParetoQ spans all eight tasks, with uneven gain magnitudes. Points summarize separately trained configurations; task scores appear in Tab. 10.

Table 11: LLaMA3-1B joint W/A/KV quantization. CanonQ-Omni uses canonical KV quantization; other QAT methods use uniform KV.

Method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	Wino.	Avg. ZS ↑	PPL ↓
FP16 Base	64.57	42.85	65.35	75.05	44.91	64.31	50.00	61.25	58.54	9.75
W2A8KV2										
RTN (PTQ)	26.88	22.50	56.16	51.07	37.26	25.30	27.54	51.80	37.31	5.8e5
GPTQ (PTQ)	27.42	26.41	61.24	49.76	37.70	25.72	25.59	52.03	38.23	5.3e4
QuaRot (PTQ)	25.86	24.06	61.90	51.95	36.43	24.81	25.00	49.14	37.39	3.4e4
KIVI	27.73	23.36	37.59	52.34	38.67	26.75	23.83	48.52	34.85	1.1e5
ParetoQ	47.76	29.81	61.89	62.98	40.68	31.72	34.96	51.25	45.13	47.58
VQLLM-QAT-d4	46.33	29.81	61.98	61.75	39.67	34.56	34.96	53.75	45.35	34.35
CanonQ-W-d1	51.65	31.67	61.41	64.33	40.32	37.06	33.59	51.25	46.41	30.07
CanonQ-W-d2	53.34	33.61	57.19	65.79	40.07	38.06	39.65	52.81	47.56	30.19
CanonQ-W-d4	50.76	31.25	61.59	64.71	40.47	37.57	37.89	51.02	46.91	31.27
CanonQ-Omni-d1	60.22	38.85	60.53	70.26	43.90	53.56	45.51	54.77	53.45	15.70
CanonQ-Omni-d2	60.73	34.97	62.65	70.15	43.25	50.52	44.14	54.92	52.67	16.43
W2A4KV2										
RTN (PTQ)	28.83	25.55	54.33	50.59	36.08	25.11	23.24	50.00	36.72	5.5e5
GPTQ (PTQ)	28.12	23.91	58.77	52.64	37.30	25.23	24.41	52.27	37.83	9.5e4
QuaRot (PTQ)	27.58	25.16	61.75	51.17	36.28	24.83	25.98	52.11	38.11	4.1e4
KIVI	26.48	25.16	37.80	52.69	37.40	25.80	25.00	50.94	35.16	3.2e5
ParetoQ	40.41	25.17	54.34	56.57	39.21	28.13	30.08	49.84	40.47	90.57
VQLLM-QAT-d4	41.72	25.93	61.95	58.41	39.62	28.84	33.40	48.67	42.32	76.51
CanonQ-W-d1	41.47	28.29	58.10	59.38	39.47	28.78	32.42	52.03	42.49	64.40
CanonQ-W-d2	42.57	29.90	56.10	60.78	39.92	29.22	36.33	48.83	42.95	64.48
CanonQ-W-d4	42.02	25.25	57.68	58.14	39.31	28.28	32.62	52.42	41.97	76.19
CanonQ-Omni-d1	61.32	40.79	62.14	70.20	42.99	52.23	43.95	54.84	53.56	15.42
CanonQ-Omni-d2	63.47	36.74	59.34	70.20	43.30	52.48	44.34	55.39	53.16	15.90

Table 12: LLaMA3-3B: weight-only, weight–activation, and joint W/A/KV quantization. The same FP reference applies to all three panels. Accuracy is in percent; lower PPL is better.

Method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	Wino.	Avg. ZS ↑	PPL ↓
FP base	72.99	49.84	75.12	78.12	48.29	74.05	53.52	68.91	65.11	7.73
W2A16										
RTN (PTQ)	27.11	25.94	47.30	50.34	36.91	25.45	25.78	49.30	36.02	3.6e5
GPTQ (PTQ)	27.34	23.83	61.54	51.12	35.35	26.10	29.49	48.83	37.95	7.3e3
QuaRot (PTQ)	28.01	25.70	56.43	50.24	37.45	25.61	26.17	50.08	37.46	3.3e2
ParetoQ	73.97	50.00	68.42	75.83	46.78	68.30	55.47	63.91	62.84	9.29
CanonQ-W-d1	73.31	50.08	69.02	75.99	46.63	68.54	51.76	64.22	62.44	9.37
CanonQ-W-d2	73.56	49.38	72.48	75.78	45.95	68.42	53.32	65.23	63.02	9.19
W2A4										
RTN (PTQ)	27.23	24.84	46.30	50.54	37.60	25.34	24.80	49.30	35.74	4.2e5
GPTQ (PTQ)	28.05	25.08	61.57	51.51	37.84	24.99	27.54	48.36	38.12	7.9e4
QuaRot (PTQ)	27.93	25.47	55.05	51.71	37.84	26.29	24.80	48.75	37.23	6.7e2
ParetoQ	62.80	40.03	55.98	68.10	44.46	53.89	45.90	54.06	53.15	17.88
VQLLM-QAT-d4	49.32	30.24	61.86	63.09	40.12	34.92	37.70	53.12	46.30	30.71
CanonQ-W-d1	62.50	36.64	60.82	69.38	42.09	50.72	38.67	53.52	51.79	16.52
CanonQ-W-d2	64.10	37.58	59.53	68.80	42.44	52.99	45.31	53.98	53.09	18.12
CanonQ-Omni-d1	72.17	46.88	69.17	75.70	44.56	66.26	52.73	61.80	61.16	10.66
CanonQ-Omni-d2	72.26	47.47	67.84	76.08	46.22	66.14	50.78	62.34	61.14	10.50
W2A4KV2										
RTN (PTQ)	27.42	25.62	50.84	50.44	37.35	25.15	26.37	51.02	36.78	1.1e6
GPTQ (PTQ)	27.62	26.33	59.83	51.07	36.77	25.45	27.54	48.44	37.88	8.9e4
QuaRot (PTQ)	27.81	24.84	61.75	50.34	36.04	25.59	27.73	49.53	37.95	1.5e4
ParetoQ	37.29	23.82	49.91	55.23	39.26	26.18	32.23	51.17	39.39	92.64
VQLLM-QAT-d4	45.14	27.20	44.54	60.51	41.08	30.78	34.77	52.58	42.07	52.82
CanonQ-W-d1	39.02	26.52	44.39	55.50	39.21	26.94	28.52	52.11	39.03	76.28
CanonQ-W-d2	43.75	30.07	55.31	59.05	39.11	34.06	33.98	52.97	43.54	30.24
CanonQ-Omni-d1	68.29	44.09	67.05	74.46	44.20	63.18	45.70	60.00	58.37	10.93
CanonQ-Omni-d2	70.62	45.62	69.20	74.61	45.46	64.38	50.78	59.61	60.04	11.16

Table 13: Cross-family and scale transfer: Qwen3-1.7B W2A4 and LLaMA3-8B W2A4KV2.

Qwen3-1.7B, W2A4										
Method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaS.	OBQA	Wino.	Avg. ZS ↑	PPL ↓
FP base	68.79	41.48	79.12	71.39	44.73	59.51	38.67	61.09	58.10	16.33
ParetoQ	41.01	25.25	60.76	59.43	41.48	28.67	30.66	51.56	42.35	43.56
CanonQ-Omni-d1	58.11	34.12	62.38	67.13	43.04	46.60	43.95	52.34	50.96	17.80
LLaMA3-8B, W2A4KV2										
FP base	80.95	57.88	83.34	81.14	49.34	79.50	55.56	73.43	70.14	6.14
ParetoQ	36.36	26.71	39.98	55.33	37.91	26.15	30.36	50.16	37.87	134.36
CanonQ-Omni-d1	70.86	42.04	71.30	73.80	45.26	68.83	43.65	62.74	59.81	9.41

Table 14: LLaMA3-1B W2A4KV2: all quantized methods receive 8K continued QAT for 40K updates. WikiText-2 PPL and its ratio to the same-length FP reference.

	PPL ↓			× FP ↓		
Method	2K	4K	8K	2K	4K	8K
FP reference	9.71	9.21	8.74	1.00	1.00	1.00
ParetoQ	50.42	46.67	45.63	5.19	5.07	5.22
CanonQ-W	43.16	40.24	39.50	4.44	4.37	4.52
CanonQ-Omni	**16.86**	**15.55**	**14.86**	**1.74**	**1.69**	**1.70**

C Context-Length Robustness after 8K Continued QAT

Protocol. All context-length results use LLaMA3-1B W2A4KV2, continued QAT with 8192-token sequences for 40K updates, and evaluation at 2048/4096/8192 tokens. The FP reference is the base checkpoint evaluated at each length. Table 14 reports PPL and the same-length ratio $r_M(L) = \mathrm{PPL}_M(L)/\,\mathrm{PPL}_{\mathrm{FP}}(L)$.

Relative quality remains stable across the tested lengths. CanonQ-Omni improves from 16.86 to 15.55 to 14.86 PPL as the evaluation window grows. Its reductions relative to ParetoQ are $2.99\times$, $3.00\times$, and $3.07\times$; at 8K, it also improves over CanonQ-W by $2.66\times$. FP improves from 9.71 to 9.21 to 8.74, leaving CanonQ-Omni at 1.69–$1.74\times$ FP across lengths, versus 4.37–$4.52\times$ for CanonQ-W and 5.07–$5.22\times$ for ParetoQ. The ordering is consistent at all three lengths.

Interpretation. All evaluation windows lie within the 8K continued-QAT training range. The experiment measures language-modeling quality across these windows. Quantized models receive continued QAT, while the FP checkpoint serves as a fixed same-length reference; the resulting ratios combine training and quantization effects. Longer windows provide additional preceding tokens and can change the scored-token set. We therefore assess context-length robustness through the consistent method ordering and approximately stable same-length FP ratio.

D Theory Details and Proofs

Throughout this appendix, $Q_d = \mathcal{N}(0, I_d)$ and all tuple laws have finite second moment. Wasserstein distance is induced by the Euclidean norm. For a finite codebook $C \subset \mathbb{R}^d$, write $d_C(v) = \mathrm{dist}(v, C)$ and let $\Pi_C(v)$ denote nearest-codeword assignment with the deterministic tie rule used by the implementation. The main transfer result uses root mean-squared distortion because it is stable under W_2 without bounded support.

D.1 The Haar Reference and the Implemented Hadamard Map

Proposition 1 (Joint Haar reference law)**.** *Let R be Haar-random on $\mathsf{O}(g)$, let $u \neq 0$ be fixed, and let $1 \leq d < g$. The first d coordinates of $V = \sqrt{g}\, Ru/\|u\|_2$ have density*

$$p_{g,d}(v) = \frac{\Gamma(g/2)}{\pi^{d/2} g^{d/2} \Gamma((g-d)/2)} \left(1 - \frac{\|v\|_2^2}{g}\right)^{(g-d-2)/2}, \qquad \|v\|_2 < \sqrt{g}, \tag{20}$$

The law is independent of the direction of u, and for fixed d it converges to Q_d in W_2 as $g \to \infty$.

Proof. Set $q = u/\|u\|_2$. Haar invariance makes Rq uniform on $\mathbb{S}^{g-1}$. If $G \sim \mathcal{N}(0, I_g)$, then $Rq \overset{d}{=} G/\|G\|_2$. For $X = (Rq)_{1:d}$,

$$\|X\|_2^2 = \frac{\sum_{j \leq d} G_j^2}{\sum_{j \leq g} G_j^2} \sim \mathrm{Beta}\left(\frac{d}{2}, \frac{g-d}{2}\right).$$

Rotational invariance in the first d coordinates and normalization in polar coordinates give

$$f_X(x) = \frac{\Gamma(g/2)}{\pi^{d/2}\Gamma((g-d)/2)} (1 - \|x\|_2^2)^{(g-d-2)/2}, \qquad \|x\|_2 \leq 1.$$

The change of variables $v = \sqrt{g}\, x$ yields Eq. (20). Finally,

$$V \overset{d}{=} \frac{G_{1:d}}{\sqrt{\|G\|_2^2/g}} \Rightarrow \mathcal{N}(0, I_d)$$

by the law of large numbers and Slutsky's theorem. Moreover, $\mathbb{E}\|V\|_2^2 = d = \mathbb{E}_{Q_d}\|V\|_2^2$ for every g. Weak convergence together with convergence of second moments is equivalent to W_2 convergence, which proves the stated metric claim. □

The preceding proposition motivates the reference law but does not describe the implemented transform. The next result applies directly to one random-sign Hadamard map and exposes the input-dependent quantities that must be checked.

Proposition 2 (One-HDP coordinate and covariance diagnostics)**.** *Let $H \in \{\pm g^{-1/2}\}^{g\times g}$ be orthogonal, let $D = \mathrm{diag}(\epsilon_1, \dots, \epsilon_g)$ with independent Rademacher signs, and let P be an independent uniform permutation. For a fixed unit vector q independent of (D, P), define $\widetilde{Z} = \sqrt{g}\, HDq$. Every fixed Hadamard row satisfies*

$$\mathbb{E}_D \widetilde{Z}_i = 0, \qquad \mathbb{E}_D \widetilde{Z}_i^2 = 1, \qquad \Pr_D(|\widetilde{Z}_i| \geq t) \leq 2e^{-t^2/2}, \tag{21}$$

and, for a universal Berry–Esseen constant C_{BE},

$$\sup_x |\Pr_D(\widetilde{Z}_i \leq x) - \Phi(x)| \leq C_{\mathrm{BE}} \sum_{j=1}^{g} |q_j|^3 \leq C_{\mathrm{BE}} \|q\|_\infty. \tag{22}$$

If $I = (i_1, \dots, i_d)$ is the ordered set of distinct Hadamard rows selected by the first d positions of P, the deployed first tuple is $Z_{1:d} := \widetilde{Z}_I$, and $\Sigma_I(q) := \mathbb{E}_D[\widetilde{Z}_I \widetilde{Z}_I^\top \mid I, q]$. Then

$$\mathbb{E}_{\mathsf{P}} \|\Sigma_I(q) - I_d\|_F^2 = d(d-1)\frac{g\|q\|_4^4 - 1}{g-1}. \tag{23}$$

Proof. Writing the unnormalized Hadamard signs as $s_{ij} \in \{\pm 1\}$ gives $\widetilde{Z}_i = \sum_{j=1}^g s_{ij}\epsilon_j q_j$. The mean and variance in Eq. (21) follow from independence and $\sum_j q_j^2 = 1$. Moreover,

$$\mathbb{E}_D e^{\lambda \widetilde{Z}_i} = \prod_j \cosh(\lambda q_j) \leq \exp\left(\frac{\lambda^2}{2}\sum_j q_j^2\right) = e^{\lambda^2/2}.$$

The standard Chernoff argument gives the stated two-sided tail. Applying the Berry–Esseen inequality to the independent summands $s_{ij}\epsilon_j q_j$ proves the first inequality in Eq. (22); the second uses $\sum_j |q_j|^3 \leq \|q\|_\infty \sum_j q_j^2$.

For two distinct selected rows $i \neq k$,

$$[\Sigma_I(q)]_{ik} = \sum_{j=1}^{g} s_{ij}s_{kj}q_j^2, \qquad [\Sigma_I(q)]_{ii} = 1.$$

Fix i and average the squared off-diagonal expression over $k \neq i$. Hadamard-column orthogonality gives

$$\sum_{k=1}^{g}\left(\sum_j s_{ij}s_{kj}q_j^2\right)^2 = \sum_{j,\ell} s_{ij}s_{i\ell}q_j^2 q_\ell^2 \sum_k s_{kj}s_{k\ell} = g\sum_j q_j^4.$$

The $k = i$ term equals $(\sum_j q_j^2)^2 = 1$, so a uniformly sampled distinct pair has expected squared covariance $(g\|q\|_4^4 - 1)/(g-1)$. There are $d(d-1)$ ordered off-diagonal entries, which proves Eq. (23). □

Prop. 2 characterizes random-seed behavior for blocks independent of the sampled transform. For an adaptively trained checkpoint, Thm. 1 instead uses its actual post-HDP tuple law through $\Delta_d(P)$. The distinction is concrete: after fixing R, choosing $u = R^\top e_1$ gives $Ru = e_1$; sparse fixed blocks can also have non-Gaussian one-RHT laws. Related worst-case analyses establish sufficient multi-RHT constructions for scalar and vector guarantees (Ben-Basat et al., 2026). CanonQ uses the random-seed bounds as structural diagnostics and the checkpoint-specific defect to quantify transfer of the frozen reference codebook.

Lemma 1 (Normalization regularization defect)**.** *Sample a nonzero block $U \in \mathbb{R}^g$ by the chosen group law, then sample one of its g/d tuples uniformly. Let P_0 and P_ϵ be the tuple laws under normalization by $\|U\|_2$ and $\|U\|_2 + \epsilon$, respectively, using the same fixed orthogonal map. Then*

$$\frac{W_2(P_\epsilon, P_0)}{\sqrt{d}} \le \left[\mathbb{E}\left(\frac{\epsilon}{\|U\|_2 + \epsilon}\right)^2\right]^{1/2}, \qquad \Delta_d(P_\epsilon) \le \Delta_d(P_0) + \left[\mathbb{E}\left(\frac{\epsilon}{\|U\|_2 + \epsilon}\right)^2\right]^{1/2}. \tag{24}$$

Proof. Couple each regularized tuple to the corresponding unregularized tuple of the same block. Writing $a = \|U\|_2/(\|U\|_2 + \epsilon)$ gives $Z_\epsilon = aZ_0$. Orthogonality implies $\|Z_0\|_2^2 = g$ for the full block, so conditional mean squared tuple error is $d(1-a)^2$. This coupling bounds W_2^2; the triangle inequality gives the second claim. □

The bound isolates the regularizer from the HDP defect, but does not bound the latter. Zero blocks require the implementation's explicit convention; the differentiable-norm results below apply to nonzero blocks.

D.2 Root-Distortion Stability and Frozen-Table Transfer

Lemma 2 (Root distortion is W_2-stable)**.** *For any finite nonempty codebook C and any tuple laws P, Q with finite second moment,*

$$|\mathcal{R}_{P,d}(C) - \mathcal{R}_{Q,d}(C)| \le \frac{1}{\sqrt{d}} W_2(P, Q). \tag{25}$$

Consequently, $|\mathcal{R}^\star_{P,b,d} - \mathcal{R}^\star_{Q,b,d}| \le W_2(P, Q)/\sqrt{d}$.

Proof. For any coupling (X, Y) of (P, Q), the distance to a fixed set is one-Lipschitz. The reverse triangle inequality in L_2 therefore gives

$$\begin{aligned}\sqrt{d}\,|\mathcal{R}_{P,d}(C) - \mathcal{R}_{Q,d}(C)| &= \big|\|d_C(X)\|_{L_2} - \|d_C(Y)\|_{L_2}\big| \\ &\le \|d_C(X) - d_C(Y)\|_{L_2} \le \|X - Y\|_{L_2}.\end{aligned}$$

Taking the infimum over couplings proves Eq. (25). Applying this inequality to every feasible C, taking infima, and exchanging P, Q proves the optimal-distortion statement. □

Proof of Thm. 1. The product alphabet A_b^d has $(2^b)^d = 2^{bd}$ entries, so it is feasible in the definition of $\mathcal{R}^\star_{Q_d,b,d}$. This proves the first part of Eq. (7). From Eq. (6) and $\sqrt{a+e} \le \sqrt{a} + \sqrt{e}$,

$$\mathcal{R}_{Q_d,d}(\widehat{C}_{b,d}) \le \mathcal{R}^\star_{Q_d,b,d} + \sqrt{\varepsilon_{\mathrm{vq}}},$$

which proves its second part.

Let $\Delta = \Delta_d(P)$. By Lem. 2,

$$\mathcal{R}_{P,d}(\widehat{C}_{b,d}) \le \mathcal{R}_{Q_d,d}(\widehat{C}_{b,d}) + \Delta, \qquad \mathcal{R}^\star_{P,b,d} \ge \mathcal{R}^\star_{Q_d,b,d} - \Delta.$$

Subtracting and using the fitting gap yields Eq. (8). Applying the same two inequalities to $\widehat{C}_{b,d}$ and A_b^d gives

$$\begin{aligned}\mathcal{R}_{P,d}(\widehat{C}_{b,d}) - \mathcal{R}_{P,d}(A_b^d) &\le \mathcal{R}_{Q_d,d}(\widehat{C}_{b,d}) - \mathcal{R}_{Q_d,d}(A_b^d) + 2\Delta \\ &= 2\Delta - m_Q(A_b^d, \widehat{C}_{b,d}),\end{aligned}$$

which is Eq. (9). □

Root distortion gives a transfer bound under finite second moments. Squared distance is not globally Lipschitz on the Gaussian reference's unbounded support, so a bounded-support W_1 argument would not cover this codec.

Lemma 3 (Correction for assign-then-gain decoding)**.** *Define*

$$\mathcal{R}^{(\gamma)}_{P,d}(C) := \left(\frac{1}{d}\mathbb{E}\|V - \gamma\Pi_C(V)\|_2^2\right)^{1/2}, \qquad B_C := \max_{c\in C}\|c\|_2.$$

For the implementation that assigns with C and scales only after assignment,

$$|\mathcal{R}^{(\gamma)}_{P,d}(C) - \mathcal{R}^{(1)}_{P,d}(C)| \le \frac{|\gamma-1|B_C}{\sqrt{d}}. \tag{26}$$

Proof. Apply the reverse triangle inequality in $L_2(P)$ to $V - \gamma\Pi_C(V)$ and $V - \Pi_C(V)$. Their difference is $(1-\gamma)\Pi_C(V)$, whose norm is at most $|1-\gamma|B_C$. Division by $\sqrt{d}$ gives Eq. (26). □

Notice that $\gamma\Pi_C(V)$ is generally not the Euclidean projection onto γC: the assignment boundaries remain those of C. Thus same-rate dominance cannot simply be reused at non-unit gain. An exact projection onto the scaled table would instead assign by $\gamma\Pi_C(V/\gamma)$ for $\gamma > 0$.

D.3 Gain-Aware Task-Loss Bounds

For a hard-quantized network, downstream code-cell switches can make the loss discontinuous as the intervention moves. Theorem 2 is valid on a smooth continuation or a neighborhood in which downstream assignments remain fixed and other operations are C^3. Its smoothness hypothesis must be checked; freezing the codebook does not imply it. For activation/KV interventions, stationarity in model parameters does not imply a small gradient in the intervention coordinates.

Proof of Thm. 2. Abbreviate $e = e_{C,\gamma}(u)$ and $\eta = \eta_{C,\gamma}(z)$. Taylor's theorem gives

$$\mathcal{L}(u+e) - \mathcal{L}(u) = \nabla\mathcal{L}(u)^\top e + \frac{1}{2}e^\top\nabla^2\mathcal{L}(u)e + r_3, \qquad |r_3| \le \frac{M_3}{6}\|e\|_2^3.$$

Orthogonality and Eq. (10) imply $\|e\|_2 = \rho(u)\|\eta\|_2/\sqrt{g}$. Cauchy–Schwarz and the operator-norm Hessian bound then give exactly Eq. (11). □

For a random block U, the magnitude of the quadratic term obeys

$$\mathbb{E}\left|\frac{1}{2}e_{C,\gamma}(U)^\top\nabla^2\mathcal{L}(U)e_{C,\gamma}(U)\right| \le \frac{1}{2}D^{\mathrm{curv}}_{S,\gamma}(C).$$

This follows by applying the pointwise operator-norm inequality before taking expectations. It requires neither residual isotropy nor an almost-sure bound on $\rho(U)^2\Lambda_U$.

Corollary 1 (Simultaneous quantization at a common graph cut)**.** *Let $e = (e_1, \dots, e_m)$ concatenate intervention coordinates exposed at a common cut of the computation graph, and regard $\mathcal{L}$ as a function of their concatenation. If the full Hessian and third derivative satisfy the same operator-norm bounds as in Thm. 2, then*

$$|\mathcal{L}(x+e) - \mathcal{L}(x)| \le \|\nabla\mathcal{L}(x)\|_2\|e\|_2 + \frac{\Lambda_x}{2}\|e\|_2^2 + \frac{M_3}{6}\|e\|_2^3. \tag{27}$$

Proof. Apply the same Taylor argument to the concatenated variable. The quadratic form $e^\top\nabla^2\mathcal{L}(x)e$ contains every cross-site block $e_i^\top H_{ij}e_j$; Eq. (27) bounds the full form and does not discard these interactions. For causally dependent internal sites, apply this result sequentially and telescope the loss differences rather than treating the sites as coordinates of one cut. □

Lemma 4 (Coupled errors in linear and attention blocks)**.** *For a linear map with $\widehat{W} = W + E_W$ and $\widehat{a} = a + e_A$,*

$$\|\widehat{W}\widehat{a} - Wa\|_2 \le \|E_W\|_{\mathrm{op}}\|a\|_2 + \|W\|_{\mathrm{op}}\|e_A\|_2 + \|E_W\|_{\mathrm{op}}\|e_A\|_2.$$

For attention with head dimension h, define $S = QK^\top/\sqrt{h}$, $A = \mathrm{softmax}_{\mathrm{row}}(S)$, and $Y = AV$. Let $\widehat{Q} = Q + E_Q$, $\widehat{K} = K + E_K$, $\widehat{V} = V + E_V$, with corresponding $\widehat{S}$, $\widehat{A}$ and $\widehat{Y}$. On the same attention mask,

$$\|\widehat{S} - S\|_F \le \frac{1}{\sqrt{h}}\big(\|E_Q\|_F\|K\|_{\mathrm{op}} + \|Q\|_{\mathrm{op}}\|E_K\|_F + \|E_Q\|_F\|E_K\|_{\mathrm{op}}\big),$$
$$\|\widehat{Y} - Y\|_F \le \tfrac{1}{2}\|\widehat{S} - S\|_F\|V\|_{\mathrm{op}} + \|\widehat{A}\|_{\mathrm{op}}\|E_V\|_F.$$

Masked logits are compared only on their common active entries. These are deterministic local bounds; no independence or isotropy of the errors is assumed.

Proof. Expanding $(W+E_W)(a+e_A)-Wa$ and applying the triangle inequality gives the first result. Similarly, $\widehat{S}-S = (E_QK^\top+QE_K^\top+E_QE_K^\top)/\sqrt{h}$; the Frobenius/operator-norm product inequalities give the logit bound. For a probability vector p, the softmax Jacobian is $J = \mathrm{diag}(p) - pp^\top$. Its absolute row sum is $2p_i(1-p_i) \le 1/2$; symmetry therefore implies $\|J\|_{\mathrm{op}} \le 1/2$. Integrating along the line segment between each pair of unmasked logit rows gives $\|\widehat{A} - A\|_F \le \frac{1}{2}\|\widehat{S} - S\|_F$. Finally, $\widehat{Y} - Y = (\widehat{A} - A)V + \widehat{A}E_V$, which proves the output bound. A fixed causal mask is handled by applying the same argument on each row's active coordinates.

□

D.4 The STE Jacobian Induced by the Canonical Codec

The following exact calculation is the method-specific part of the optimization analysis. It assumes the normalization scale is differentiated, as in the stated fake-quantization graph, and that the nonnegative gain is an independent parameter when differentiating with respect to u.

Proposition 3 (Exact normalized-codebook STE Jacobian)**.** *Let $u \neq 0$, set $s(u) = \rho(u)/\sqrt{g} = (\|u\|_2 + \epsilon)/\sqrt{g}$, $z = Ru/s(u)$, $q = \Pi_C^{(d)}(z)$, and $\widehat{u} = \gamma s(u)R^\top q$ for $\gamma \ge 0$. Under the identity STE $Dq/Dz = I$, the surrogate Jacobian is*

$$J_{\mathrm{STE}}(u) = \gamma\left[I + R^\top(q - z)\nabla s(u)^\top\right]. \tag{28}$$

With the assignment defect $\delta_a(u) = \|q - z\|_2/\sqrt{g}$ and reconstruction defect $\delta_r(u) = \|\gamma q - z\|_2/\sqrt{g}$,

$$\|J_{\mathrm{STE}} - \gamma I\|_{\mathrm{op}} = \gamma\delta_a, \qquad \|J_{\mathrm{STE}}\|_{\mathrm{op}} \le \gamma(1 + \delta_a), \qquad \|\widehat{u} - u\|_2 = \rho(u)\delta_r. \tag{29}$$

The deviation from the gain-preserving reference output γu is $\|\widehat{u} - \gamma u\|_2 = \gamma\rho(u)\delta_a$. If γ is trained, the gain-column defect is

$$\|\partial_\gamma\widehat{u} - \partial_\gamma(\gamma u)\|_2 = \|sR^\top(q - z)\|_2 = \rho(u)\delta_a. \tag{30}$$

Proof. Differentiate $z = Ru/s$ using the identity STE: $dz = R\,du/s - z\,ds/s$. Since $sR^\top z = u$,

$$\begin{aligned} d\widehat{u} &= \gamma\left(R^\top q\,ds + sR^\top dz\right) \\ &= \gamma\left(du + R^\top(q - z)\,ds\right), \end{aligned}$$

which proves Eq. (28). For $u \neq 0$, $\|\nabla s(u)\|_2 = 1/\sqrt{g}$. The matrix $A(u) = R^\top(q - z)\nabla s(u)^\top$ has rank at most one and satisfies $\|A(u)\|_{\mathrm{op}} = \|q - z\|_2/\sqrt{g} = \delta_a$. Hence $J_{\mathrm{STE}} - \gamma I = \gamma A(u)$ has operator norm $\gamma\delta_a$, and the Jacobian bounds follow.

Finally,

$$\widehat{u} - u = sR^\top(\gamma q - z), \qquad \widehat{u} - \gamma u = \gamma sR^\top(q - z),$$

and orthogonality gives the two forward-error identities. Finally, $\partial_\gamma\widehat{u} = sR^\top q$ and $\partial_\gamma(\gamma u) = u = sR^\top z$, which gives Eq. (30). □

The two defects differ because the current implementation assigns in the unscaled table C and applies γ afterward. Assignment distortion controls the STE Jacobian relative to its gain-scaled identity,

whereas reconstruction distortion controls the actual error relative to u. If z is partitioned into g/d tuples V_j and P is obtained by sampling one of them uniformly, then

$$\delta_a(u)^2 = \frac{1}{g}\sum_{j=1}^{g/d} \|\Pi_C(V_j) - V_j\|_2^2 = D_{P,d}(C). \tag{31}$$

Thus Thm. 1 directly bounds the RMS assignment defect that appears in the STE Jacobian, subject only to the local sensitivity factors in the next corollary.

Corollary 2 (Codec-controlled STE bias and alignment)**.** *Let f be a downstream loss whose gradient is L_f-Lipschitz. Define $g_{\mathrm{STE}} = J_{\mathrm{STE}}^\top \nabla f(\widehat{u})$. Relative to the gain-preserving smooth continuation $u \mapsto f(\gamma u)$,*

$$\|g_{\mathrm{STE}} - \gamma \nabla f(\gamma u)\|_2 \le \gamma\delta_a \|\nabla f(\gamma u)\|_2 + \gamma^2(1+\delta_a)L_f\rho(u)\delta_a =: B_C(u). \tag{32}$$

Consequently, with $g_{\mathrm{ref}} = \gamma \nabla f(\gamma u)$,

$$\langle g_{\mathrm{STE}}, g_{\mathrm{ref}}\rangle \ge \|g_{\mathrm{ref}}\|_2\big(\|g_{\mathrm{ref}}\|_2 - B_C(u)\big). \tag{33}$$

If $g_{\mathrm{ref}} \neq 0$ and $B_C(u) \le \kappa\|g_{\mathrm{ref}}\|_2$ for some $\kappa < 1$, then

$$\cos\angle(g_{\mathrm{STE}}, g_{\mathrm{ref}}) \ge \frac{1-\kappa}{1+\kappa}. \tag{34}$$

Proof. Add and subtract $J_{\mathrm{STE}}^\top \nabla f(\gamma u)$:

$$g_{\mathrm{STE}} - \gamma\nabla f(\gamma u) = (J_{\mathrm{STE}}^\top - \gamma I)\nabla f(\gamma u) + J_{\mathrm{STE}}^\top[\nabla f(\widehat{u}) - \nabla f(\gamma u)].$$

Use Eq. (29), gradient Lipschitzness, and $\|\widehat{u} - \gamma u\|_2 = \gamma\rho(u)\delta_a$ to obtain Eq. (32). Cauchy–Schwarz yields Eq. (33). Under the relative-error condition, $\|g_{\mathrm{STE}}\|_2 \le (1+\kappa)\|g_{\mathrm{ref}}\|_2$ and the inner product is at least $(1-\kappa)\|g_{\mathrm{ref}}\|_2^2$, which proves Eq. (34). □

Corollary 3 (RMS distortion controls the blockwise STE certificate)**.** *Suppose on the analyzed region that $0 \le \gamma \le \bar{\gamma}$, $\rho(u) \le \bar{\rho}$, $\|\nabla f(\gamma u)\|_2 \le G$, and $\delta_a(u) \le \bar{\delta}$. With $K := \bar{\gamma}G + \bar{\gamma}^2(1+\bar{\delta})L_f\bar{\rho}$, one has $B_C(u) \le K\delta_a(u)$. If P is the uniformly sampled tuple law of the group, then*

$$B_C(u)^2 \le K^2 D_{P,d}(C). \tag{35}$$

For a random group U, taking expectation gives $\mathbb{E}B_C(U)^2 \le K^2\mathbb{E}D_{P_U,d}(C)$; the tuple law must use the same group-sampling distribution.

Proof. Apply the regional bounds to Eq. (32), then use Eq. (31). □

From source mismatch to STE bias. For the numerical reference table $C = \widehat{C}_{b,d}$, Lem. 2 and Eq. (6) give

$$\mathcal{R}_{P,d}(C) \le \mathcal{R}_{Q_d,d}(C) + \Delta_d(P) \le \mathcal{R}^\star_{Q_d,b,d} + \sqrt{\varepsilon_{\mathrm{vq}}} + \Delta_d(P).$$

Under the regional bounds of Cor. 3 and matching group/tuple sampling, it follows that

$$\mathbb{E}\big[B_C(U)^2\big] \le K^2\big(\mathcal{R}^\star_{Q_d,b,d} + \sqrt{\varepsilon_{\mathrm{vq}}} + \Delta_d(P)\big)^2. \tag{36}$$

This makes the roles of reference distortion, numerical fitting, source mismatch, and downstream sensitivity explicit in one blockwise bias certificate.

Comparing with the gain-scaled downstream gradient evaluated at the unquantized input u, the same decomposition and Eq. (29) give

$$\|g_{\mathrm{STE}} - \gamma\nabla f(u)\|_2 \le \gamma\delta_a\|\nabla f(u)\|_2 + \gamma(1+\delta_a)L_f\rho(u)\delta_r. \tag{37}$$

Assignment error δ_a controls the Jacobian defect, while reconstruction error δ_r controls forward error relative to u. Optimizing only the latter can leave the former large when gain is applied after assignment.

Proposition 4 (Cellwise shadow objective). *Let $f : \mathbb{R}^d \to \mathbb{R}$ have L_f-Lipschitz gradient and let V_k be the open Voronoi cell of codeword c_k. Define*

$$\mathcal{S}_C f(z) := f(c_k) + \langle \nabla f(c_k), z - c_k \rangle, \qquad z \in V_k.$$

Inside each cell, $\nabla \mathcal{S}_C f(z) = \nabla f(\Pi_C(z))$, exactly the identity-STE direction for a fixed-scale quantizer, and

$$|\mathcal{S}_C f(z) - f(z)| \le \frac{L_f}{2} \|z - \Pi_C(z)\|_2^2. \tag{38}$$

Proof. The gradient statement follows because $\mathcal{S}_C f$ is affine on V_k. The smoothness remainder for the first-order expansion of f at c_k gives Eq. (38). □

The cellwise construction relates local STE modeling error to codebook distortion. It can be discontinuous across cell boundaries, so its guarantee is local to each cell.

For a full network, let $\mathcal{B}_\xi(\omega, e)$ denote the backward map when the concatenated perturbation vector e includes forward residuals, local STE-Jacobian deviations, and trainable-quantizer Jacobian-column errors, with $\mathcal{B}_\xi(\omega, 0) = \nabla f_\xi(\omega)$. The following elementary lemma explains how the sitewise certificates aggregate.

Lemma 5 (From aggregate codec errors to global STE bias). *Assume $\mathbb{E}_\xi \nabla f_\xi(\omega) = \nabla F(\omega)$, $\mathbb{E}_\xi \|e(\xi)\|_2^2 < \infty$, and the required gradients are integrable. If, on the analyzed region,*

$$\|\mathcal{B}_\xi(\omega, e) - \mathcal{B}_\xi(\omega, 0)\|_2 \le \kappa \|e\|_2, \tag{39}$$

then for $G(\omega; \xi) = \mathcal{B}_\xi(\omega, e(\xi))$ and $F(\omega) = \mathbb{E}_\xi f_\xi(\omega)$,

$$\|\mathbb{E}_\xi G(\omega; \xi) - \nabla F(\omega)\|_2^2 \le \kappa^2 \mathbb{E}_\xi \|e(\xi)\|_2^2. \tag{40}$$

The exact blockwise bound in Eq. (32) supplies a sharper local value when the normalization graph is modeled explicitly.

Proof. Jensen's inequality followed by Eq. (39) gives

$$\|\mathbb{E}_\xi[G - \nabla f_\xi]\|_2^2 \le \mathbb{E}_\xi \|G - \nabla f_\xi\|_2^2 \le \kappa^2 \mathbb{E}_\xi \|e(\xi)\|_2^2.$$

□

D.5 Distortion-Controlled Optimization Consequences

For the SGD update in Thm. 3, the following analysis separates optimization and sampling error from the codec-dependent bias certificate. The result concerns the smooth reference objective under the stated update rule; the experimental optimizer is AdamW.

Proof of Thm. 3. Let $g_t = \nabla F(\omega_t)$, $\bar{G}_t = \mathbb{E}_t G_t$, and $b_t = \bar{G}_t - g_t$. L-smoothness gives

$$\mathbb{E}_t F(\omega_{t+1}) \le F(\omega_t) - \alpha_t \langle g_t, g_t + b_t \rangle + \frac{L\alpha_t^2}{2} \mathbb{E}_t \|G_t\|_2^2. \tag{41}$$

The variance condition implies

$$\mathbb{E}_t \|G_t\|_2^2 \le 2\|g_t\|_2^2 + 2\|b_t\|_2^2 + \sigma^2.$$

Also, $-\langle g_t, b_t \rangle \le \frac{1}{4}\|g_t\|_2^2 + \|b_t\|_2^2$. Substituting both inequalities into Eq. (41) and using $L\alpha_t \le 1/4$ yields

$$\mathbb{E}_t F(\omega_{t+1}) \le F(\omega_t) - \frac{\alpha_t}{2}\|g_t\|_2^2 + \frac{5\alpha_t}{4}\|b_t\|_2^2 + \frac{L\alpha_t^2\sigma^2}{2}. \tag{42}$$

Take total expectation, sum from $t = 0$ to $T - 1$, use $F(\omega_T) \ge F_\star$, and substitute the codec-bias bound. Multiplying the resulting inequality by two and dividing by S_T gives Eq. (15). □

For the constant schedule $\alpha_t = c/\sqrt{T} \le 1/(4L)$, $S_T = c\sqrt{T}$ and $\sum_{t<T} \alpha_t^2 = c^2$. The optimization and sampling terms in Eq. (15) therefore decay as $T^{-1/2}$, leaving the weighted codec-bias certificate. If that certificate also vanishes, the bound establishes stationarity of the smooth reference; otherwise it gives a residual upper bound determined by the codec and sensitivity conditions.

Corollary 4 (Local PL bound with codec-dependent residual)**.** *Suppose, in addition, that the iterates remain in a region satisfying the PL inequality*

$$\frac{1}{2}\|\nabla F(\omega)\|_2^2 \geq \mu(F(\omega) - F_\star), \tag{43}$$

and that $\|b_t\|_2^2 \leq \overline{B}^2$*. For a constant* $0 < \alpha \leq \min\{1/(4L), 1/\mu\}$*,*

$$\mathbb{E}[F(\omega_T) - F_\star] \leq (1 - \mu\alpha)^T (F(\omega_0) - F_\star) + \frac{5\overline{B}^2}{4\mu} + \frac{L\alpha\sigma^2}{2\mu}. \tag{44}$$

Proof. Apply Eq. (43) to Eq. (42) and take total expectations:

$$\Delta_{t+1} \leq (1 - \mu\alpha)\Delta_t + \frac{5\alpha}{4}\overline{B}^2 + \frac{L\alpha^2\sigma^2}{2}, \qquad \Delta_t := \mathbb{E}[F(\omega_t) - F_\star].$$

Unrolling the geometric recurrence and upper-bounding the finite geometric sum by $(\mu\alpha)^{-1}$ proves Eq. (44). □

D.6 Voronoi Margins, Code Switching, and Index Stability

Lemma 6 (Assignment-switch margin)**.** *For a point* v *with unique nearest codeword* $c_{a(v)}$*, define*

$$m_C(v) := \min_{j \neq a(v)} \frac{\|v - c_j\|_2^2 - \|v - c_{a(v)}\|_2^2}{2\|c_j - c_{a(v)}\|_2}. \tag{45}$$

Set $m_C(v) = 0$ *when the nearest-codeword assignment is tied. Then* $m_C(v)$ *is the distance from* v *to the boundary of its Voronoi cell. If* $\|v' - v\|_2 < m_C(v)$*, then* $a(v') = a(v)$*. Consequently, if* $\|V' - V\|_2 \leq r$ *almost surely,*

$$\Pr(a(V') \neq a(V)) \leq \Pr(m_C(V) \leq r). \tag{46}$$

Proof. For $j \neq a(v)$, the bisector between $c_{a(v)}$ and c_j is a hyperplane. The signed Euclidean distance from v to that hyperplane is precisely the corresponding ratio in Eq. (45). The minimum is therefore the distance to the boundary of the current cell. The open ball of that radius lies inside the cell, proving both claims. □

For the idealized $\epsilon = 0$ normalization $z(u) = \sqrt{g}Ru/\|u\|_2$, if $\|h\|_2 \leq \|u\|_2/2$, then

$$\|z(u + h) - z(u)\|_2 \leq \frac{2\sqrt{g}\|h\|_2}{\|u\|_2}.$$

Let $z_j(u) \in \mathbb{R}^d$ denote tuple j. A sufficient condition that one shadow update preserve every tuple index is

$$\frac{2\sqrt{g}\|h\|_2}{\|u\|_2} < \min_j m_C(z_j(u)).$$

This inequality certifies unchanged assignments when it holds. Violating it permits, but does not imply, switching or chattering.

Corollary 5 (Stability of the code-index distribution)**.** *Let* a_C *be the nearest-codeword index map, let* P, Q *be tuple laws, and let* $V \sim Q$*. Then*

$$\mathrm{TV}((a_C)_\# P, (a_C)_\# Q) \leq \inf_{r>0} \left\{ \frac{W_2(P,Q)^2}{r^2} + Q(m_C(V) \leq r) \right\}. \tag{47}$$

Proof. Take a coupling (X, Y) of (P, Q). If $\|X - Y\|_2 \leq r$ and $m_C(Y) > r$, Lem. 6 gives $a_C(X) = a_C(Y)$. Hence

$$\Pr(a_C(X) \neq a_C(Y)) \leq \Pr(\|X - Y\|_2 > r) + Q(m_C(Y) \leq r).$$

Markov's inequality bounds the first term by $\mathbb{E}\|X - Y\|_2^2/r^2$. Total variation between pushforwards is at most the mismatch probability under any coupling. Infimizing first over couplings and then over r proves Eq. (47). □

Code-index stability is governed jointly by transport distance and the probability mass near Voronoi boundaries. The bound separates these contributions: W_2 controls displacement between tuple laws, while the boundary term controls the sensitivity of discrete assignments to that displacement.