GLOBAL FP4 QUANTIZATION VIA LOW-DIMENSIONAL CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Native FP4 quantization encodes small weight groups with shared scales, but correlated inputs couple their reconstruction errors. Independent groupwise selection neglects these interactions, while exhaustive joint search grows exponentially with the number of groups. For reconstruction metrics with full within-group curvature and a shared rank-r component, we formulate a convex relaxation with an r-dimensional dual. We prove that any feasible candidate mixture can be reduced to at most r mixed groups without changing its aggregate residual or increasing the relaxed objective. LDC-FP4 uses this construction to coordinate codes and scales and recover native encodings with a residual-variance bound. Across Llama and Qwen models, LDC-FP4 improves language modeling and downstream performance over the evaluated quantization baselines. Matched-candidate comparisons show that global coordination improves reconstruction beyond independent selection and outperforms a wide-beam search baseline at significant lower calibration cost. All optimization is confined to calibration, leaving the native FP4 representation and inference operators unchanged.

1 INTRODUCTION

Native FP4 arithmetic makes language-model inference more economical by storing weights in small groups that share a scale. NVFP4 uses sixteen E2M1 values and one E4M3 scale per group, together with a tensor-level scale (NVIDIA et al., 2025; NVIDIA, 2026). Fine-grained scaling accommodates substantial variation in weight magnitudes without changing the dense matrix-multiplication operator (Rouhani et al., 2023). Encoding quality is determined offline, when calibration selects the codes and scales that the hardware will subsequently consume (Nagel et al., 2020; Lin et al., 2024; Shao et al., 2024).

For correlated inputs, the accuracy of these decisions cannot be assessed independently (Li et al., 2021; Frantar et al., 2022; Chee et al., 2023). If e denotes a row's quantization error and H the input second moment, its squared output error is $e^\top He$. Terms coupling different groups remain even when each group is scored with its complete local curvature. Minimizing the individual terms can select errors that reinforce one another at the output. A replacement rejected on its own can also become favorable when another group changes with it. Improving local reconstruction does not account for either interaction.

Existing quantizers retain this dependence in several ways. GPTQ and GPTAQ compensate for committed errors in weights that remain unquantized (Frantar et al., 2023; Li et al., 2025). H-Scale and PiSO improve scale selection, and WUSH constructs data-aware local coordinates (Yu et al., 2026; Amboage et al., 2026; Chen et al., 2026). Multiple-path search and global weight updates permit more extensive joint optimization (Egiazarian et al., 2024; Cha et al., 2026; Lucas et al., 2026). Their cost reflects the number of candidate assignments considered or the size of the continuous optimization state. Even sixteen candidates for each of 512 groups define 16^{512} assignments. This motivates a formulation in which global dependence is represented explicitly without retaining that combinatorial state.

We consider metrics $H = D + UU^\top$, where D retains every within-group interaction and the shared factor U has r columns. Each candidate contributes a local scalar cost and an r-dimensional residual vector. The coupled loss is the sum of those costs plus the squared norm of the sum of the residual vectors. Consequently, the information passed between groups has dimension r, although the weight

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

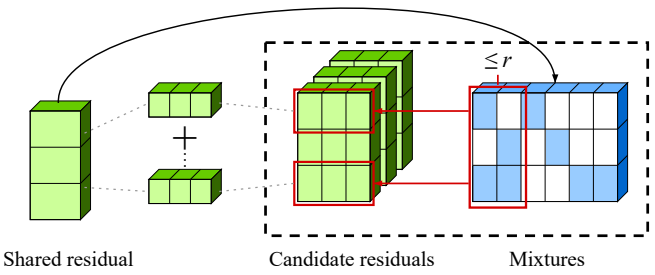


Figure 1: Support reduction preserves the shared residual and leaves at most r mixed groups (red). Blue columns are groups, with filled cells marking active candidates.

row and its reconstruction metric can both be high dimensional. Global selection can be expressed through a dual variable in this residual space, with all groups responding to the same variable.

Coordinating a continuous relaxation still leaves a discrete recovery problem. Independently taking the largest candidate probability generally changes the aggregate residual and can discard the benefit of joint optimization. Our main result gives a constructive alternative. At any feasible relaxed solution, support can be reduced at a fixed aggregate residual without increasing the objective until at most r groups retain multiple candidates. For 512 groups and $r = 16$, at least 496 complete native encodings are fixed before rounding. A residual-variance bound controls the cost of resolving the remaining support. This result follows from extreme-point geometry and applies to incomplete numerical solves as well as relaxed optima.

LDC-FP4 implements this construction for asymmetric layer reconstruction. It fits local and shared curvature using the inputs received by the quantized layer, generates legal code-and-scale candidates, and solves the low-dimensional coordination problem. Feasible support reduction connects the optimized residual to the final discrete representation. Neither the factor nor the candidate mixtures are retained in the deployed model. Across six Llama-3.1 and Qwen3 models, LDC-FP4 improves perplexity over the competing quantized encoders on both evaluation corpora and attains the highest reported downstream accuracy aggregates among quantized methods. Comparisons under identical candidate lists measure the quality and calibration cost of coordination. Rank and recovery ablations examine how shared curvature and residual preservation contribute to these gains.

Our contributions are threefold. First, we formulate globally coupled native FP4 encoding as a convex relaxation with a dual dimension determined by the shared curvature rank. Second, we establish a constructive support bound and a residual-variance rounding guarantee, including an explicit correction for curvature approximation. Third, we develop an encoder that applies these results to quantized-prefix calibration and improves model quality without changing the deployed format. Matched-candidate experiments quantify its reconstruction gains and calibration cost relative to local, pairwise, and beam-based selection.

2 Problem Formulation

2.1 Asymmetric layer reconstruction

Consider one output row of a linear layer. Let $w \in \mathbb{R}^d$ denote the reference weights and $q \in \mathbb{R}^d$ their quantized representation. Inputs $Z \in \mathbb{R}^{n \times d}$ are collected from the quantized prefix. In W4A4, they also include the current activation quantizer. Corresponding teacher outputs are $y \in \mathbb{R}^n$. We minimize

$$\mathcal{L}(q) = \frac{1}{n} \|Zq - y\|_2^2 + \gamma (q - w)^\top S (q - w), \qquad S \succ 0, \tag{1}$$

with a block-diagonal regularizer S. For $e = q - w$, define

$$H = \frac{1}{n} Z^\top Z + \gamma S, \qquad a = \frac{1}{n} Z^\top (Zw - y), \qquad \mathcal{L}_0 = \frac{1}{n} \|Zw - y\|_2^2 . \tag{2}$$

Expanding the objective gives $\mathcal{L}(w + e) = \mathcal{L}_0 + e^\top H e + 2a^\top e$. Using the actual quantized inputs incorporates accumulated prefix error, as in GPTAQ (Li et al., 2025). Curvature H is shared by the output rows of a layer. Both a and $\mathcal{L}_0$ are row dependent.

Partition the input coordinates into B groups of size $g = 16$. Each group receives a finite list $\mathcal{Q}_b = \{q_{b1}, \ldots, q_{bK_b}\}$ of complete native encodings,

$$q_{bk} = s_{\mathrm{g}} \ell_{bk} c_{bk}, \qquad \ell_{bk} \in \mathcal{S}_{\mathrm{E4M3}}, \qquad c_{bk} \in \mathcal{E}^{16}_{\mathrm{E2M1}}. \tag{3}$$

Here s_{g} is a fixed tensor-level scale, $\mathcal{S}_{\mathrm{E4M3}}$ contains positive finite scales, and $\mathcal{E}_{\mathrm{E2M1}} = \{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 3, \pm 4, \pm 6\}$. A candidate fixes the stored scale and all sixteen codes. Our optimization domain is the product of these supplied lists.

2.2 Structured reconstruction metric

We represent the metric by

$$H = D + UU^\top + R, \qquad D = \mathrm{BlkDiag}(D_1, \ldots, D_B) \succeq 0, \qquad U \in \mathbb{R}^{d \times r}. \tag{4}$$

Each D_b may be dense and full rank. The factor U captures shared directions, and R records the approximation residual. We first set $R = 0$ and address its effect in Section 3.5. For candidate error $e_{bk} = q_{bk} - w_b$, write

$$\alpha_{bk} = e_{bk}^\top D_b e_{bk} + 2 a_b^\top e_{bk}, \qquad v_{bk} = U_b^\top e_{bk} \in \mathbb{R}^r. \tag{5}$$

Selecting one candidate per group gives

$$F(k_1, \ldots, k_B) = \mathcal{L}_0 + \sum_{b=1}^{B} \alpha_{b,k_b} + \left\| \sum_{b=1}^{B} v_{b,k_b} \right\|_2^2. \tag{6}$$

The scalar α_{bk} includes the asymmetric linear term. The vector v_{bk} describes how the candidate changes the shared residual. These two quantities suffice for every comparison under the structured metric.

Independent full-block selection minimizes $\alpha_{bk} + \|v_{bk}\|_2^2$ for each group. Under the structured metric it uses $D_b + U_b U_b^\top$ and discards the cross terms $2 \sum_{b<c} v_{b,k_b}^\top v_{c,k_c}$. Equation (6) retains all of these terms through one vector sum. At rank zero, independent selection minimizes the separable objective.

Local preprocessing is compatible with this structure (Xiao et al., 2023). For $R = 0$, an invertible block-diagonal transform T gives $THT^\top = TDT^\top + (TU)(TU)^\top$. For W4A4, nonlinear activation quantization changes the moments, so H is estimated after the actual transform and quantizer. The encoder applies in either the original coordinates or a fixed WUSH basis. Our experiments use the latter (Chen et al., 2026).

3 Convex Relaxation and Recovery

3.1 Convex relaxation

Associate each group with $p_b \in \Delta_{K_b}$, a probability vector over its encodings, and solve

$$\Phi^* = \min_{p_b \in \Delta_{K_b}} \Phi(p), \qquad \Phi(p) = \mathcal{L}_0 + \sum_{b,k} p_{bk} \alpha_{bk} + \left\| \sum_{b,k} p_{bk} v_{bk} \right\|_2^2. \tag{7}$$

Every discrete assignment is a vertex of the feasible set, giving $\Phi^* \le F^*$. The relaxation averages the local costs of complete encodings. Each local cost remains attached to an actual code-and-scale choice. For $\bar{e}_b = \sum_k p_{bk} e_{bk}$, the excess over evaluating the same quadratic objective at $w + \bar{e}$ is

$$\sum_b \left[\sum_k p_{bk} e_{bk}^\top D_b e_{bk} - \bar{e}_b^\top D_b \bar{e}_b \right] \ge 0. \tag{8}$$

Equation (8) retains the within-group cost associated with mixing discrete alternatives.

3.2 SUPPORT REDUCTION

Theorem 1 (Rank-controlled support)**.** *For every feasible p in (7), there exists a feasible $\tilde{p}$ with the same aggregate residual and $\Phi(\tilde{p}) \leq \Phi(p)$ such that*

$$\sum_{b=1}^{B} (|\operatorname{supp}(\tilde{p}_b)| - 1) \leq r. \tag{9}$$

At most r groups mix multiple encodings. Their joint support contains at most 2^r assignments. In particular, an optimal relaxed solution exists with these properties.

Proof. Fix $z = \sum_{b,k} p_{bk} v_{bk}$ and minimize $\sum_{b,k} \tilde{p}_{bk} \alpha_{bk}$ subject to the B simplex equalities, the r equalities specifying z, and nonnegativity. The original mixture is feasible. An optimal basic feasible solution has at most $B + r$ positive coefficients. Each group requires at least one, proving (9). If group b retains s_b coefficients, $s_b \leq 2^{s_b - 1}$ gives $\prod_b s_b \leq 2^r$. □

The argument specializes extreme-point geometry for separable nonconvex problems (Udell & Boyd, 2016; Dubois-Taine & d'Aspremont, 2026). Its consequence for quantization is constructive. The candidate mixture may involve every weight in a layer, but preserving its shared residual requires only r equalities in addition to the group sums. All surplus support beyond those equalities can be removed. The theorem applies to feasible mixtures obtained before numerical convergence as well as to a relaxed optimum.

The relevant dimension can be smaller than the width of U. Choose one reference candidate per group and let r_{eff} be the dimension of $\operatorname{span}\{v_{bk} - v_{b1}\}_{b,k}$. Components orthogonal to this span are constant across assignments. An equivalent problem can be written in $\mathbb{R}^{r_{\text{eff}}}$, replacing r by r_{eff} in Theorem 1. Appendix A gives the reduction explicitly.

3.3 NATIVE ENCODING RECOVERY

For a reduced mixture, define $\mu_b = \sum_k p_{bk} v_{bk}$ and

$$V_b(p) = \sum_k p_{bk} \|v_{bk}\|_2^2 - \|\mu_b\|_2^2 . \tag{10}$$

Proposition 2 (Rounding cost)**.** *A discrete assignment in the support of p can be constructed with*

$$F(k_1, \ldots, k_B) \leq \Phi(p) + \sum_b V_b(p). \tag{11}$$

Only groups retaining multiple candidates contribute to the variance sum.

Proof. Sample group indices independently from p_b. The expected local cost is unchanged. Independence gives $\mathbb{E} \left\| \sum_b v_{b,k_b} \right\|^2 = \left\| \sum_b \mu_b \right\|^2 + \sum_b V_b(p)$. At least one assignment is no worse than this expectation. Successive conditional-expectation rounding constructs such an assignment. □

After support reduction, at most r variances remain. For two candidates, $V_b = p_{b1}(1 - p_{b1}) \|v_{b1} - v_{b2}\|^2$, so nearby residual candidates incur a small recovery cost. If each surviving variance is at most σ^2, the additive rounding cost is at most $r\sigma^2$. This bound characterizes the cost of recovering a native encoding from the relaxation. It leaves open the possibility that a different residual outside the reduced support admits a better integer assignment.

3.4 DUAL FORMULATION

Using $\|z\|^2 = \max_\lambda (2\lambda^\top z - \|\lambda\|^2)$ gives

$$g(\lambda) = \mathcal{L}_0 - \|\lambda\|_2^2 + \sum_b \min_k \left(\alpha_{bk} + 2\lambda^\top v_{bk} \right), \qquad \max_{\lambda \in \mathbb{R}^r} g(\lambda) = \Phi^*. \tag{12}$$

Strong duality follows from convexity and compactness of the product of simplices. Given λ, every group chooses its minimum-scoring candidate independently. The term $2\lambda^\top v_{bk}$ accounts for the

candidate's contribution to the shared residual. Updating one r-dimensional variable coordinates all groups in an output row.

For any dual iterate and any native assignment,

$$g(\lambda) \le \Phi^* \le F^* \le F(k). \tag{13}$$

If the hard responses satisfy $\lambda = \sum_b v_{b,k_b(\lambda)}$, the feasible objective equals the dual bound and the candidate assignment is optimal. Otherwise, support reduction and Proposition 2 provide an explicit recovery bound. Both statements concern the coupled objective over the supplied lists.

3.5 Curvature approximation bounds

For the residual R in (4), suppose $\|R\|_2 \le \varepsilon$ and set $M = \sum_b \max_k \|e_{bk}\|^2$. Then

$$|\mathcal{L}(q) - F(q)| \le \varepsilon M, \qquad g(\lambda) - \varepsilon M \le \mathcal{L}_{\mathcal{Q}}^* \le \mathcal{L}(q). \tag{14}$$

A directly evaluated encoding supplies the upper bound for the original metric. The lower bound includes the decomposition residual. The corrected interval accounts for both incomplete discrete optimization and the error introduced by the structured metric. Constants from asymmetric reconstruction are retained in every reported objective interval.

4 LDC-FP4

4.1 Structured curvature estimation

We fit the decomposition in the coordinates used by the quantized operator. Set $S = \mathrm{BlkDiag}(Z^\top Z/n) + \delta I$ and normalize $C = S^{-1/2} H S^{-1/2}$. The fitting objective is

$$\min_{\Delta, V} \left\| C - \Delta - VV^\top \right\|_F^2, \qquad \Delta = \mathrm{BlkDiag}(\Delta_b) \succeq 0, \qquad V \in \mathbb{R}^{d \times r}. \tag{15}$$

At fixed V, each Δ_b is the positive-semidefinite projection of $C_{bb} - V_b V_b^\top$. At fixed Δ, the shared component retains the largest r positive eigenvalues of $C - \Delta$. Alternating these subproblem solutions jointly estimates the two components (Fernandez et al., 2025). We recover $D = S^{1/2}\Delta S^{1/2}$ and $U = S^{1/2}V$.

The default rank is 16, with full 16×16 local matrices at every rank. Moment estimation and decomposition operate one layer at a time. Matrix–vector products support the eigensolver, and dense moments can be released before coordination. Structured storage requires $O(dg + dr)$ entries. At $d = 8192$ and $r = 16$, the local matrices and factor together occupy 1 MiB in FP32.

4.2 Candidate construction

The initial encoding is GPTAQ adapted to NVFP4. Each list retains this encoding and considers sixteen E4M3 scales at index offsets $-9, \ldots, 6$ around the rounded maximum-absolute-value scale (Yu et al., 2026). At each scale we include nearest-even E2M1 rounding and a second encoding obtained by two coordinate sweeps under $e_b^\top H_{bb} e_b + 2a_b^\top e_b$. Duplicate payloads are removed, leaving at most 33 candidates.

For fixed scale and other coordinates, the continuous minimizer of the scalar subproblem is

$$q_j^{\mathrm{cont}} = w_j - \frac{a_j + \sum_{k \ne j} (H_{bb})_{jk}(q_k - w_k)}{(H_{bb})_{jj}}. \tag{16}$$

Rounding this value to the scaled E2M1 grid solves the constrained scalar update. The tensor-level scale, saturation behavior, and packing convention are identical for every candidate. Candidate generation produces the available alternatives before global coordination begins.

4.3 Coordination and recovery

We optimize the smooth dual

$$g_\tau(\lambda) = \mathcal{L}_0 - \|\lambda\|_2^2 - \tau \sum_b \log \sum_k \exp\left[-\frac{\alpha_{bk} + 2\lambda^\top v_{bk}}{\tau} \right]. \tag{17}$$

Algorithm 1 LDC-FP4 for one layer

Require: Reference weights W, calibration data (Z, Y), and rank r
1: Initialize $Q^{(0)}$ and fit $H \approx D + UU^\top$.
2: **for** each row $q^{(0)} = (Q^{(0)}_{i,:})^\top$ in output-row tiles **do**
3: Construct $\mathcal{Q}_b \ni q^{(0)}_b$ and (α_{bk}, v_{bk}) via (5).
4: Optimize g_τ numerically to obtain (λ, p) and evaluate $g(\lambda)$.
5: Reduce p to $\widetilde{p}$ at fixed $\sum_{b,k} p_{bk} v_{bk}$, leaving at most r mixed groups.
6: $\widehat{q} \leftarrow \mathrm{RecoverRefine}(\widetilde{p}, \alpha, v, \{\mathcal{Q}_b\})$ (Appendix D).
7: $Q^\top_{i,:} \leftarrow \arg\min_{q \in \{q^{(0)}, \widehat{q}\}} \mathcal{L}(q)$.
8: **end for**
9: **return** packed E2M1 codes and E4M3 scales of Q.

Softmax probabilities from (17) give the gradient $-2\lambda + 2\sum_{b,k} p_{bk} v_{bk}$. Warm-started L-BFGS follows a decreasing-temperature schedule. We compute lower bounds from the unsmoothed dual (12). Both candidate features and coordination variables are batched over output-row tiles.

The probabilities also give a stopping criterion in the original objective units. Let $z_\tau = \sum_{b,k} p_{bk} v_{bk}$ and let $\mathcal{H}(p_b)$ be the entropy of the distribution in group b. Direct substitution gives

$$\Phi(p) - g_\tau(\lambda) = \|z_\tau - \lambda\|^2 + \tau \sum_b \mathcal{H}(p_b). \tag{18}$$

Since $g(\lambda) \geq g_\tau(\lambda)$, this expression bounds the feasible-to-dual interval. Its first term is one quarter of the squared gradient norm, and the second is at most $\tau \sum_b \log K_b$. We decrease the temperature as the residual mismatch becomes small, avoiding an excessively sharp objective early in optimization. The final interval is evaluated directly under (7) and (12). Support reduction can begin from any returned feasible mixture because it preserves that mixture's attained residual.

Support reduction preserves the aggregate residual of the returned feasible mixture. Choose an active candidate in each group as an anchor. Any $r + 1$ surplus residual differences are linearly dependent. A pivot along this dependence preserves all group sums and the aggregate residual. We choose its sign to make the local cost nonincreasing, then move until one probability reaches zero. Repeating this operation leaves the support described by Theorem 1. Appendix A details the feasible update.

We enumerate supports containing at most 4,096 assignments and otherwise use conditional-expectation rounding. Two candidate-coordinate sweeps refine the recovered encoding. Finally, we evaluate this encoding and the preserved initialization under the original calibration objective (1), retaining the lower-cost result. Calibration statistics, mixtures, and shared factors are discarded after the chosen E2M1 codes and E4M3 scales have been packed.

Candidate costs and residual features require $O(BKg^2 + BKgr)$ arithmetic per output row. A dual evaluation requires $O(BKr)$. Support reduction depends on the active candidate count, and its final joint support is bounded by the shared rank. Tiling bounds the candidate workspace. Inference uses the same dense operators and numerical layout as the initialization, with no stored shared factor or additional matrix multiplication.

5 Experiments

5.1 Experimental setup

We evaluate Llama-3.1-8B/70B-Instruct and Qwen3-4B/8B/14B/32B, using non-thinking mode for Qwen3 (Grattafiori et al., 2024; Yang et al., 2025). The primary setting is native NVFP4 W4A4. Attention and feed-forward linear weights and inputs are quantized. Embeddings, normalization, output heads, softmax, and the KV cache remain in BF16. W4A16 retains BF16 inputs to the linear operators. All primary comparisons use the same fixed WUSH basis.

Calibration uses 256 FineWeb sequences of 2,048 tokens and 32 disjoint validation sequences (Penedo et al., 2024). Three document-level samples define the calibration seeds. Layers are encoded sequentially using inputs from the current quantized prefix and outputs from the BF16 teacher.

Table 1: WikiText-2 perplexity under native W4A4. HS denotes GPTAQ + H-Scale, PI PiSO + compensation, KC K-CoreQ, and AD ADMM-Q. Bold marks the best quantized result.

Model	BF16	GPTAQ	HS	PI	KC	AD	FAAR	LDC
Llama-3.1-8B	7.08	7.74	7.69	7.70	7.66	7.65	7.64	**7.53**
Llama-3.1-70B	3.63	3.87	3.84	3.85	3.83	3.82	3.82	**3.77**
Qwen3-4B	13.18	13.80	13.71	13.73	13.69	13.66	13.64	**13.55**
Qwen3-8B	9.42	9.84	9.79	9.75	9.77	9.78	9.76	**9.66**
Qwen3-14B	8.39	8.75	8.70	8.71	8.68	8.67	8.66	**8.55**
Qwen3-32B	7.36	7.66	7.62	7.59	7.61	7.61	7.60	**7.52**

Table 2: Downstream accuracy (%) and teacher–student token KL, averaged over six models. Bold marks the best quantized result in each column.

Method	Six-task ↑	MMLU ↑	GSM8K ↑	GPQA ↑	KL ↓
BF16	76.4	74.8	88.3	39.4	0
RTN	74.0	71.2	84.7	36.9	0.0831
GPTQ	74.2	71.5	85.2	37.3	0.0716
GPTAQ + H-Scale	74.7	72.5	86.4	38.0	0.0625
PiSO + compensation	74.6	72.4	86.3	37.9	0.0621
K-CoreQ	74.8	72.7	86.6	38.2	0.0614
ADMM-Q	74.8	72.9	86.5	38.1	0.0602
FAAR	74.9	73.0	86.8	38.6	0.0596
LDC-FP4	**75.6**	**73.4**	**87.1**	**38.8**	**0.0529**

No evaluation examples enter calibration or stopping decisions. Default solver settings and task configurations appear in Appendix B.

Full-pipeline baselines are GPTQ, GPTAQ, GPTAQ with H-Scale, PiSO with compensation, K-CoreQ, ADMM-Q, and FAAR. Format adaptations use the same E2M1 projection and legal E4M3 scales. FAAR uses gradient-based calibration. All costs include initialization and refinement. Matched-candidate comparisons hold the lists and objective fixed, then vary the solver among independent full-block selection, coordinate search, pairwise search, beam search, and LDC-FP4. Each receives the same final candidate-coordinate refinement, except where explicitly ablated. Converged and equal-GPU-time comparisons are reported separately.

We report WikiText-2 perplexity with 2,048-token windows and stride 1,024, and C4 perplexity on 512 fixed sequences of 2,048 tokens (Merity et al., 2017; Raffel et al., 2020). Zero-shot accuracy averages HellaSwag, PIQA, ARC-Easy, ARC-Challenge, WinoGrande, and BoolQ. MMLU, GSM8K, and GPQA Diamond use five-shot, eight-shot, and zero-shot evaluation and are reported separately (Sutawika et al., 2024). Held-out layer error and teacher–student token KL evaluate reconstruction and model-level distortion. Perplexity reductions average modelwise relative changes. Accuracy and token KL are macro-averages over models. Paired uncertainty intervals use hierarchical resampling of calibration seeds and evaluation examples.

5.2 Quantization quality

LDC-FP4 achieves the lowest perplexity among the compared quantized encoders for all six models on both corpora (Tables 1 and 7). Relative to GPTAQ with H-Scale, mean modelwise reductions are 1.6% on WikiText-2 and 1.5% on C4. Selecting the best competing quantized encoder separately for each model and corpus gives reductions of 1.1% and 1.0%, respectively.

Across the four accuracy aggregates in Table 2, LDC-FP4 has the highest scores among the quantized methods. Its six-task average exceeds FAAR, the strongest competing quantized baseline on this metric, by 0.7 percentage points, with a paired uncertainty interval of [+0.31, +1.09] points. Relative to GPTAQ with H-Scale, MMLU, GSM8K, and GPQA Diamond improve by 0.9, 0.7, and 0.8 percentage points. Mean token KL decreases by 15.4% and is the lowest among the quantized methods.

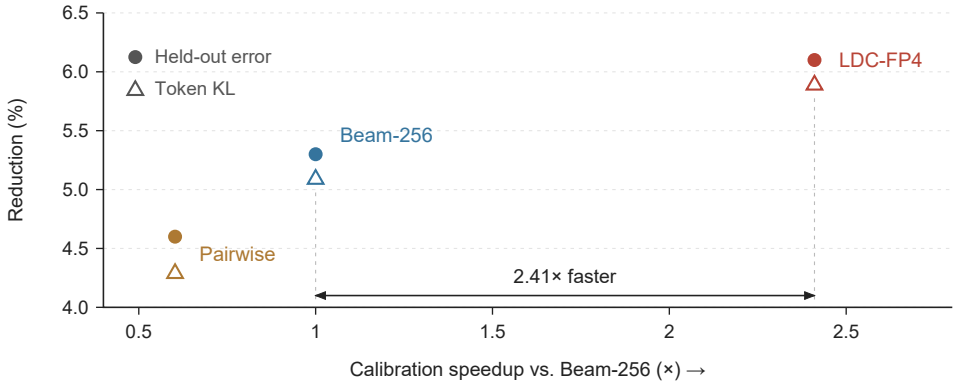


Figure 2: Held-out reconstruction error and token KL reductions under identical candidate lists, measured relative to independent full-block selection. Calibration speedup is relative to Beam-256.

Held-out layer error decreases by 16.1% in W4A16 and 12.8% in W4A4 relative to the preserved initialization. The asymmetric objective uses the actual input quantizer in each setting.

5.3 Effect of Global Coordination

With identical candidate lists, LDC-FP4 lowers held-out error by 6.1% relative to independent full-block selection, by 1.6% relative to pairwise search, and by 0.8% relative to Beam-256 (Figure 2). It uses 41.5% of Beam-256's calibration time, a 2.41× speedup measured as $T_{\mathrm{Beam-256}}/T_{\mathrm{solver}}$. In the separate comparison with equal GPU-time budgets, LDC-FP4 achieves 1.4% lower held-out error than Beam-256. Calibration costs include initialization, candidate construction, support reduction, and recovery.

Adding refined code assignments to scale-only candidates reduces independent-selection error by 6.6%. Coordinating the expanded lists reduces the remaining error by 6.1%. In this sequence of comparisons, coordination accounts for 46.3% of the total absolute improvement. These measurements use the same fixed-prefix data and error normalization.

Argmax recovery without support reduction adds 2.1% of the initialization objective. Support reduction followed by recovery and final refinement reduces this net increment to 1.1%, a 47.6% reduction. Omitting final refinement gives an increment of 1.6%. These increments measure the additional objective cost of discrete recovery. Table 6 reports the corresponding full held-out reconstruction errors.

5.4 Rank and Recovery Analysis

We test whether a small shared factor captures the curvature relevant to candidate selection and whether discrete recovery preserves the coordination gains. For candidate-error directions, define

$$\eta(e) = \frac{|e^\top (H - D - UU^\top)e|}{e^\top He + \epsilon}. \tag{19}$$

At rank 16, median and 90th-percentile η are 2.4% and 7.7%. Removing shared curvature increases held-out error by 6.5% and perplexity by 0.8%. Increasing the rank to 32 changes these metrics by -0.8% and -0.1%. Removing local curvature increases held-out error by 20.4%. Full local matrices account for within-group structure, and the shared factor captures interactions that improve candidate selection across groups.

Across layers, the smallest tested rank attaining 90th-percentile $\eta \leq 10\%$ has median 16 and 90th percentile 64. It correlates more strongly with the initial shared-error fraction (Spearman 0.61) than with input width (0.22), suggesting that rank requirements track shared error more closely than width alone.

Support reduction leaves 14.3 mixed groups per row on average and at most 16, matching the rank-16 bound. Its maximum relative residual drift is 2.94×10^{-6}. Recovery and final refinement add

1.1% of the initialization objective, with a variance bound 2.4 times the recovery cost. Median feasible-to-dual intervals are 1.1% for the relaxed structured objective and 78.2% for the original metric after curvature-residual correction. Both use the initialization objective as the denominator. Appendix C gives the interval definitions and additional ablations.

Coordination should help most when the initial shared-error fraction is large. We stratify rows by

$$\chi(e) = \frac{\left\|U^\top e\right\|^2}{e^\top De + \left\|U^\top e\right\|^2 + \epsilon}. \tag{20}$$

Under the same candidate-generation rule, held-out reconstruction improves by 20.1% in the highest quartile and 3.2% in the lowest. This concentration of gains agrees with the cross-group term in Equation (6). The shared statistics also depend on calibration data. Reducing calibration to 32 sequences changes held-out error by +3.6%, and replacing FineWeb with C4 changes six-task accuracy by -0.1 points.

5.5 Calibration and inference efficiency

Moments and factorization account for 12.1% of calibration time, candidate construction 23.9%, coordination and recovery 42.7%, and initialization and final refinement 21.3%. On B200 with matched kernels, prefill and decode throughput ratios to native GPTAQ are both 1.00 at the reported precision. Appendix B provides benchmark settings, per-model calibration costs, and BF16 comparisons.

6 Related Work

Quantization objectives and search. GPTQ uses sequential second-order compensation, and GPTAQ models quantized-prefix reconstruction mismatch (Frantar et al., 2023; Li et al., 2025). H-Scale and PiSO optimize scales (Yu et al., 2026; Amboage et al., 2026). FAAR learns FP4-aware rounding (Li et al., 2026). CoreQ keeps multiple rounding paths (Cha et al., 2026). ADMM-Q combines global weight optimization, discrete constraints, and local refinement (Lucas et al., 2026). LDC-FP4 couples complete encodings through shared residuals with bounded remaining discrete support.

Representation and curvature. Alternatives use sparse outlier storage (Dettmers et al., 2024; Kim et al., 2024) or vector and trellis codes (Liu et al., 2024; Tseng et al., 2024b). Rotations and affine transforms facilitate quantization (Ashkboos et al., 2024; Tseng et al., 2024a; Liu et al., 2025; Sun et al., 2025). WUSH constructs local transforms from full weight and activation second moments (Chen et al., 2026). Following Fernandez et al. (2025), we jointly estimate full local matrices and a low-rank shared component.

Convexification and rounding. Theorem 1 specializes extreme-point arguments for separable nonconvex problems (Starr, 1969; Aubin & Ekeland, 1976; Udell & Boyd, 2016; Bi & Tang, 2020; Kerdreux et al., 2023; Dubois-Taine & d'Aspremont, 2026). DiscQuant uses discrepancy theory and low-dimensional gradients for nearly discrete weights (Chee et al., 2025). MixedDimKV convexifies discrete cache-budget allocation (Miao et al., 2026). Our complete FP4 codes and scales couple through a quadratically penalized residual vector that determines the dual and rounding bound.

7 Conclusion

Globally coupled FP4 encoding admits a convex relaxation with an r-dimensional dual under block-local plus rank-r curvature. Support reduction preserves the shared residual and leaves at most r groups with multiple encodings. Conditional expectation resolves the remaining choices with a residual-variance bound. LDC-FP4 applies this construction to asymmetric reconstruction and retains native FP4 storage and inference. Across six models, it achieves the lowest perplexity among the compared quantized encoders on both corpora and improves the six-task accuracy average over FAAR by 0.7 percentage points. Under identical candidates, held-out error is 0.8% lower than Beam-256 at 41.5% of its calibration time. Recovery guarantees apply to the supplied candidate lists under the structured metric, with curvature-residual correction extending the objective bounds to the original metric.

AI USE STATEMENT

We used large language models (LLMs) to improve grammar and correct spelling errors in the manuscript. The authors take full responsibility for the final content of this work.

REFERENCES

Juan Amboage, Pablo Monteagudo-Lago, Ian Colbert, Giuseppe Franco, and Nicholas Fraser. Optimal post-training quantization scales and where to find them. *arXiv preprint arXiv:2606.10890*, 2026.

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*, volume 37, pp. 100213–100240, 2024.

J. P. Aubin and I. Ekeland. Estimates of the duality gap in nonconvex optimization. *Mathematics of Operations Research*, 1(3):225–245, 1976.

Yingjie Bi and Ao Tang. Duality gap estimation via a refined Shapley–Folkman lemma. *SIAM Journal on Optimization*, 30(2):1094–1118, 2020.

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

Seohyeon Cha, Huancheng Chen, Dongjun Kim, Haoran Zhang, Kevin Chan, Gustavo de Veciana, and Haris Vikalo. CoreQ: Learning-free mismatch correction and successive rounding for quantization. *arXiv preprint arXiv:2602.05902*, 2026.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429, 2023.

Jerry Chee, Arturs Backurs, Rainie Heck, Li Zhang, Janardhan Kulkarni, Thomas Rothvoss, and Sivakanth Gopi. DiscQuant: A quantization method for neural networks inspired by discrepancy theory. In *Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 924–951, 2025.

Jiale Chen, Vage Egiazarian, Roberto L. Castro, Torsten Hoefler, and Dan Alistarh. WUSH: Near-optimal adaptive transforms for LLM quantization. In *International Conference on Machine Learning*, 2026.

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, 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. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024.

Benjamin Dubois-Taine and Alexandre d'Aspremont. Frank-Wolfe meets Shapley-Folkman: A systematic approach for solving nonconvex separable problems with linear constraints. *Mathematical Programming*, 219:501–551, 2026.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12284–12303, 2024.

Andres Fernandez, Felix Dangel, Philipp Hennig, and Frank Schneider. Sketching low-rank plus diagonal matrices. *arXiv preprint arXiv:2509.23587*, 2025.

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

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

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

Thomas Kerdreux, Igor Colin, and Alexandre d'Aspremont. Stable bounds on the duality gap of separable nonconvex optimization problems. *Mathematics of Operations Research*, 48(2):1044–1065, 2023.

Sehoon Kim, Coleman Richard Charles Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23901–23923, 2024.

Hanglin Li, Shuchang Tian, Chen Lin, Zhiyong Zhao, and Kun Zhan. FAAR: Format-aware adaptive rounding for NVFP4. *arXiv preprint arXiv:2603.22370*, 2026.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. GPTAQ: Efficient finetuning-free quantization for asymmetric calibration. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36690–36706, 2025.

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*, volume 6, pp. 87–100, 2024.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

Yifei Liu, Jicheng Wen, Yang Wang, Shengyu Ye, Li Lyna Zhang, Ting Cao, Cheng Li, and Mao Yang. VPTQ: Extreme low-bit vector post-training quantization for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8181–8196, 2024.

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

Ryan Lucas, Mehdi Makni, Xiang Meng, Adam Deng, and Rahul Mazumder. ADMM-Q: An improved Hessian-based weight quantizer for post-training quantization of large language models. *arXiv preprint arXiv:2605.11222*, 2026.

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

Ruijie Miao, Zhiming Wang, Wang Li, Shiwei Wu, Shufan Liu, Yanbing Jiang, and Tong Yang. Beyond token eviction: Mixed-dimension budget allocation for efficient KV cache compression. *arXiv preprint arXiv:2603.20616*, 2026.

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

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103:127–152, 2005.

NVIDIA et al. Pretraining large language models with NVFP4. *arXiv preprint arXiv:2509.25149*, 2025.

NVIDIA. Using FP8 and FP4 with Transformer Engine. *Transformer Engine documentation*, 2026.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Prabhakar Raghavan. Probabilistic construction of deterministic algorithms: Approximating packing integer programs. *Journal of Computer and System Sciences*, 37(2):130–143, 1988.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling*, 2024.

Bita Darvish Rouhani et al. Microscaling data formats for deep learning. *arXiv preprint arXiv:2310.10537*, 2023.

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.

Varun Saravanan, Gordon J. Berman, and Samuel J. Sober. Application of the hierarchical bootstrap to multi-level data in neuroscience. *Neurons, Behavior, Data Analysis, and Theory*, 3(5):1–25, 2020.

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.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Ross M. Starr. Quasi-equilibria in markets with non-convex preferences. *Econometrica*, 37(1):25–38, 1969.

Yuxuan Sun et al. FlatQuant: Flatness matters for LLM quantization. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57587–57613, 2025.

Lintang Sutawika et al. EleutherAI/lm-evaluation-harness: v0.4.3. *Zenodo*, 2024.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656, 2024a.

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with trellises and incoherence processing. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59597–59620, 2024b.

Madeleine Udell and Stephen Boyd. Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378, 2016.

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

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hao Yu, Zheng Li, Dayiheng Liu, and Jianwei Zhang. H-Scale: Hessian-guided scale refinement for NVFP4 sub-byte LLM inference. *arXiv preprint arXiv:2608.28113*, 2026.

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.

A Proofs and Recovery Guarantees

A.1 Strong duality

Let $\mathcal{P} = \prod_b \Delta_{K_b}$ and write $Vp = \sum_{b,k} p_{bk} v_{bk}$. Introducing $z = Vp$ in (7) gives the Lagrangian

$$\mathscr{L}(p, z, \lambda) = \mathcal{L}_0 + \alpha^\top p + \|z\|^2 + 2\lambda^\top (Vp - z). \tag{21}$$

Minimizing over z gives $z = \lambda$ and contribution $-\|\lambda\|^2$. Minimizing over each simplex gives the smallest candidate score, recovering (12). The squared norm is finite and continuous everywhere and $\mathcal{P}$ is compact. Fenchel duality gives equality of the primal and dual optima. The negative quadratic dominates the finite sum of affine candidate scores as $\|\lambda\| \to \infty$, so a dual maximizer exists.

For any native assignment k,

$$\begin{aligned} g(\lambda) &\le \mathcal{L}_0 - \|\lambda\|^2 + \sum_b \big(\alpha_{b,k_b} + 2\lambda^\top v_{b,k_b}\big) \\ &= F(k) - \left\| \sum_b v_{b,k_b} - \lambda \right\|^2 \le F(k). \end{aligned} \tag{22}$$

Equality holds if every k_b minimizes its score at λ and the aggregate residual equals λ. This supplies the integral optimality condition stated in the main text.

A.2 Feasible support reduction

Choose an active anchor $a(b)$ in each group. An active nonanchor coefficient is called a surplus coefficient. If more than r surplus coefficients remain, select $r + 1$ of them and form their residual differences $v_{bk} - v_{b,a(b)}$. Linear dependence gives coefficients h_{bk}, not all zero, satisfying

$$\sum_{(b,k)} h_{bk} \big(v_{bk} - v_{b,a(b)}\big) = 0. \tag{23}$$

Assign perturbation $d_{bk} = h_{bk}$ to the selected coefficients and $d_{b,a(b)} = -\sum_{k \ne a(b)} h_{bk}$ to each affected anchor. All other entries are zero. Then

$$\sum_k d_{bk} = 0 \quad \text{for every } b, \qquad \sum_{b,k} d_{bk} v_{bk} = 0. \tag{24}$$

Choose the sign of d such that $\alpha^\top d \le 0$ and set

$$\theta = \min_{d_{bk} < 0} \frac{-p_{bk}}{d_{bk}}, \qquad p \leftarrow p + \theta d. \tag{25}$$

A nonzero feasible direction has both positive and negative entries. Equation (25) preserves the group sums, aggregate residual, and nonnegativity, makes at least one coefficient zero, and cannot increase the linear cost. Updating anchors as needed and repeating this step yields (9).

Each dependence calculation involves at most $r + 1$ residual differences. The number of pivots depends on the initial active support. In floating-point arithmetic, normalization and residual errors are accumulated and checked after each batch of pivots. Entries are removed by a feasible step rather than by independently truncating small probabilities.

A.3 Conditional-expectation recovery

Suppose some groups have been fixed and the remaining groups retain distributions p_b. For an unresolved group b, let z_{-b} be the sum of fixed residuals and the means of the other unresolved residuals. The candidate-dependent part of the conditional expected objective is

$$\alpha_{bk} + \|z_{-b} + v_{bk}\|^2. \tag{26}$$

Selecting a minimum over the current support cannot exceed the previous expectation, since the latter is a convex combination of these conditional values (Raghavan, 1988). Repeating this selection proves Proposition 2 constructively.

For a feasible mixture p and any dual iterate λ, the recovered assignment satisfies

$$F(k_{\mathrm{round}}) - F^* \le \Phi(p) - g(\lambda) + \sum_b V_b(p). \tag{27}$$

The first term measures the solved relaxation's feasible-to-dual interval. The variance sum bounds the additional cost of discrete recovery. Further descent preserves this bound for the structured objective. The final acceptance step separately checks the original objective (1).

A.4 Effective dimension of candidate variation

Choose a reference candidate in each group and set

$$z_0 = \sum_b v_{b1}, \qquad \mathcal{V} = \mathrm{span}\{v_{bk} - v_{b1}\}_{b,k}. \tag{28}$$

Let P be an orthonormal basis of $\mathcal{V}$ with r_{eff} columns. Every attainable aggregate residual is $z = z_0 + P\xi$. Its squared norm is

$$\|z\|^2 = \left\|P^\top z_0 + \xi\right\|^2 + \left\|(I - PP^\top)z_0\right\|^2. \tag{29}$$

The second term is constant. An equivalent coordination problem has residual features $P^\top v_{bk}$ and a constant objective adjustment, with r_{eff} equalities in support reduction. Theorem 1 then applies with r_{eff}.

Support reduction preserves a particular aggregate residual of a feasible convex solution. Enumerating that support finds its best discrete assignment. The original candidate family can also contain useful assignments at other aggregate residuals. Even scalar features $v_{b,\pm} = \pm a_b$ give the sign-selection problem $\min(\sum_b \pm a_b)^2$. The native integer problem remains combinatorial even at low rank. The recovery guarantee is (27).

A.5 Bounds for the original reconstruction metric

For $\widetilde{H} = D + UU^\top$ and $\left\|H - \widetilde{H}\right\|_2 \le \varepsilon$,

$$|e^\top (H - \widetilde{H})e| \le \varepsilon \|e\|^2 \le \varepsilon \sum_b \max_k \|e_{bk}\|^2 = \varepsilon M. \tag{30}$$

Taking minima proves the lower bound in (14). Direct evaluation of the returned native encoding under H supplies the upper bound. A valid Frobenius upper bound can replace the spectral upper bound. If $H - \widetilde{H} \succeq 0$, $g(\lambda)$ is already a lower bound for the original candidate objective.

All absolute objectives include $\mathcal{L}_0$. Removing this constant preserves candidate rankings but changes any ratio normalized by the objective value. We normalize empirical intervals by the initialization objective. Relative-to-optimum bounds are computed only when their lower bound is positive.

B Encoder and Evaluation Details

B.1 Implementation settings

Table 3 gives the fixed settings. The dual is normalized by a positive row-specific factor. Dividing candidate costs by this factor and residual features by its square root preserves every native assignment ranking. Bound evaluation restores the original factor and the asymmetric constant.

The E4M3 lookup table includes finite positive normal and subnormal values. Scale indices are clipped at the representable boundaries. Signed zeros are retained during packing and merged numerically when calculating reconstruction costs. Every candidate uses the same tensor scale. All-zero groups retain their zero encoding.

Table 3: Default encoder and calibration settings.

Component	Setting
Native group size	16
Scale indices	$-9, \ldots, 6$ around the initial E4M3 scale
Encodings per scale	Nearest-even RTN and local coordinate refinement
Local refinement	Two sweeps
Retained candidate	Format-matched GPTAQ initialization
Maximum candidate count	33 before duplicate removal
Default shared rank	16
Rank sweep	$0, 4, 8, 16, 32, 64$
Calibration / validation sequences	256 / 32
Sequence length	2,048 tokens
Calibration seeds	3
Regularization coefficient	10^{-3}
Structure-fitting updates	At most 10 alternating iterations
Relative fitting tolerance	10^{-3}
Dual temperatures	$10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$
Dual optimizer	Warm-started L-BFGS
Iterations per temperature	At most 200
Support enumeration	At most 4,096 assignments
Final candidate refinement	Two coordinate sweeps
Coordination / bound precision	FP32 / FP64

B.2 Structured curvature estimation

The normalized matrix C is initialized with a rank-r positive-semidefinite spectral component. We project the residual diagonal blocks onto the positive-semidefinite cone and alternate the two updates in Section 4. The fitting objective is monitored after each update. The residual R is retained for evaluating approximation quality.

Matrix approximation is measured on held-out candidate-error directions as in (19). For rank selection statistics, a layer meets the tolerance when the 90th percentile of these errors is at most 10%. Layers not meeting this criterion at rank 64 are reported as right-censored. Correlations involving the selected rank use Spearman's coefficient (Spearman, 1904).

B.3 Candidate-coordinate refinement

Maintain $z = \sum_b v_{b,k_b}$. Replacing the current candidate k_b by k changes the structured objective by

$$\Delta F = \alpha_{bk} - \alpha_{b,k_b} + 2(z - v_{b,k_b})^\top (v_{bk} - v_{b,k_b}) + \|v_{bk}\|^2 - \|v_{b,k_b}\|^2 . \tag{31}$$

Accepted updates change z incrementally. Candidate costs and shared features are reused across all comparisons. No new scale or code is introduced in the matched-candidate refinement, which keeps the optimization domain identical across solvers.

The full-metric acceptance step evaluates both the new encoding and the initial encoding against the same Z, y, S. It chooses the smaller value of (1). This evaluation is distinct from the held-out statistics used to assess calibration transfer.

B.4 Model and task configuration

We use the original Qwen3 dense checkpoints at 4B, 8B, 14B, and 32B parameters and Llama-3.1-8B/70B-Instruct. Qwen3 runs with thinking disabled. Tokenizers, prompts, context limits, and extraction rules are identical within each model comparison. Attention projections and feed-forward linear layers use NVFP4 weights. W4A4 quantizes their inputs, and W4A16 uses BF16 inputs. KV caches remain in BF16 in both settings.

Calibration documents are split before token-window construction. The 256 calibration sequences and 32 validation sequences are disjoint. We draw three calibration document sets and reuse each set

across methods. Teacher outputs are collected on the same token positions as the actual quantized-prefix inputs.

WikiText-2 uses a 2,048-token context and stride 1,024. Each scored token contributes once to negative log-likelihood. C4 uses 512 fixed sequences under the same context setting. The six-task accuracy is the unweighted mean of HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), ARC-Easy, ARC-Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2020), and BoolQ (Clark et al., 2019). Length-normalized accuracy is used for HellaSwag and ARC. MMLU (Hendrycks et al., 2021) uses five demonstrations, GSM8K (Cobbe et al., 2021) eight, and GPQA Diamond (Rein et al., 2024) zero. Greedy generation has a 2,048-token output limit. Downstream configuration follows the Language Model Evaluation Harness (Sutawika et al., 2024).

B.5 Comparison settings

End-to-end comparisons include each method's preprocessing, initialization, candidate construction, and refinement. GPTQ, GPTAQ, PiSO, K-CoreQ, and ADMM-Q are adapted to the same native value and scale sets where required. FAAR uses gradient-based calibration. All methods use the same fixed WUSH transform in the primary comparison.

Matched-candidate comparisons hold both the candidate lists and the reconstruction target fixed. Fixed-prefix layer measurements collect inputs from the same initialization for every solver. End-to-end evaluations allow each method's quantized prefix to determine the subsequent inputs. Coordinate and pairwise methods are run to their stated convergence criteria in the converged comparison. The equal-compute comparison uses the same total GPU-time budget for each layer and records its best feasible encoding before the deadline. The two comparisons have separate labels.

B.6 Evaluation metrics and aggregation

Held-out normalized output error is

$$\mathrm{NMSE} = \frac{\left\| Z_{\mathrm{val}} \widehat{W}^{\top} - Y_{\mathrm{ref,val}} \right\|_F^2}{\left\| Y_{\mathrm{ref,val}} \right\|_F^2}. \tag{32}$$

Teacher–student token KL uses the same reference prefixes for both models. Perplexity reductions are computed for each model before averaging. Accuracy and token KL in Table 2 are macro-averages over the six models, and the reported KL reduction compares these means. Perplexity values from different tokenizers are not pooled into a single absolute score. The six-task average excludes MMLU, GSM8K, and GPQA Diamond.

Paired uncertainty intervals use hierarchical resampling of calibration seeds and evaluation examples (Saravanan et al., 2020). Layer-level summaries resample layers. Output rows sharing a layer's covariance are not treated as independent model replications. The shared-error fraction used for stratification is

$$\chi(e) = \frac{\left\| U^{\top} e \right\|^2}{e^{\top} D e + \left\| U^{\top} e \right\|^2 + \epsilon}. \tag{33}$$

The numerator and denominator use the initial encoding in the same coordinate system as the fitted metric.

B.7 Hardware measurements

All methods use the same B200 backend, with matched hardware and parallelism for BF16 and FP4 comparisons. The prefill grid contains prompt lengths 512, 2,048, and 8,192 and batch sizes 1 and 8. Decode generates 256 tokens after a 2,048-token prompt. Fifty warm-up repetitions precede two hundred timed repetitions, with synchronization at measurement boundaries.

Quantization GPU-hours equal device count times elapsed hours. Peak calibration memory is the maximum over devices. Speedups are calculated for each model and workload before geometric averaging. Ratios against the native initialization isolate the encoder's deployment effect. Ratios against BF16 include the numerical format and its backend. No calibration feature, candidate table, or coordination variable is retained in the inference checkpoint.

Table 4 summarizes inference ratios relative to native GPTAQ and BF16.

Table 4: LDC-FP4 inference ratios on B200. Throughput ratios are geometric means.

	Throughput (×) ↑		Peak memory (×) ↓
Reference	Prefill	Decode	
Native GPTAQ	1.00	1.00	–
BF16	2.31	1.78	0.42

Table 5 gives calibration GPU-hours, peak calibration memory, and inference throughput relative to native GPTAQ for four of the evaluated models.

Table 5: LDC-FP4 calibration cost and inference throughput on B200. Memory is the peak per device. Throughput ratios use native GPTAQ under matched kernels and hardware.

	Calibration		Throughput / GPTAQ	
Model	GPU-hours	Peak memory (GB)	Prefill	Decode
Llama-3.1-8B	2.47	29.1	1.000	1.000
Qwen3-8B	2.91	30.9	1.000	1.000
Qwen3-32B	9.74	85.8	1.000	1.000
Llama-3.1-70B	43.86	53.1	1.000	1.000

C Additional Model and Ablation Results

C.1 C4 perplexity

Table 7 reports model-level C4 perplexity. LDC-FP4 improves on GPTAQ with H-Scale for all six models, with a mean relative reduction of 1.5%. It also improves on the best competing quantized encoder for every model, giving a mean relative reduction of 1.0%. The competing reference is selected separately for each model and corpus.

C.2 Rank and recovery analysis

Table 8 summarizes rank ablations and recovery diagnostics.

Structured intervals use $\Phi(p) - g(\lambda)$ for the feasible mixture returned by the dual solve and have median 1.1% of the initialization objective. For the returned native encoding q, the original-metric interval is $\mathcal{L}(q) - g(\lambda) + \varepsilon M$ from (14), with median 78.2% under the same normalization. This interval includes the curvature-residual correction as well as the cost of discrete recovery. Exhaustive comparisons on four-group, four-candidate subproblems extracted from model layers give a mean deviation of 0.4% from the restricted optimum.

C.3 Recovery and calibration sensitivity

After support reduction, an output row retains an average of 14.3 mixed groups, with maximum 16. Recovery and final refinement add 1.1% of the initialization objective. Removing support reduction and applying argmax recovery increases this increment by 90.9% relative to the complete encoder.

Rows in the highest quartile of the initial shared-error fraction (33) have mean held-out improvement 20.1%, compared with 3.2% in the lowest quartile. Reducing calibration from 256 to 32 sequences changes held-out error by +3.6% and the empirical feasible-to-dual interval by +20.1%. Changing the calibration corpus from FineWeb to C4 changes the six-task average by -0.1 percentage points.

Table 6: Identical-candidate comparison. Error, KL, and calibration time are normalized by independent full-block selection from the expanded lists. Final refinement is shared except where ablated. Times include initialization and candidate generation. The variant without support reduction uses argmax recovery. Bold marks the lowest error and KL.

Solver	Held-out error ↓	Token KL ↓	Calibration time
Independent full-block selection	1.000	1.000	1.00
Block-coordinate search	0.970	0.974	1.81
Pairwise search	0.954	0.957	35.48
Beam-64	0.951	0.952	6.79
Beam-256	0.947	0.949	21.38
LDC without support reduction	0.958	0.962	7.04
LDC without final refinement	0.946	0.945	8.51
LDC-FP4	**0.939**	**0.941**	8.87

Table 7: C4 perplexity under native W4A4 quantization in the fixed WUSH basis. Best baseline is the lowest perplexity among competing quantized encoders for each model. Bold marks the best quantized result.

Model	GPTAQ + H-Scale	Best baseline	LDC-FP4
Llama-3.1-8B	11.22	11.16	**11.00**
Llama-3.1-70B	7.57	7.52	**7.44**
Qwen3-4B	17.04	16.96	**16.86**
Qwen3-8B	13.72	13.66	**13.53**
Qwen3-14B	12.29	12.25	**12.11**
Qwen3-32B	11.01	10.96	**10.86**

Table 8: Rank ablations (relative to rank 16) and recovery diagnostics. Recovery cost includes final refinement. Cost and the relaxed feasible-to-dual gap are percentages of the initialization objective.

Ablation	$r = 0$	$r = 32$	No local curvature
Δ held-out error (%)	+6.5	-0.8	+20.4
Δ PPL (%)	+0.8	-0.1	–

η (%) Median / 90th pct.	Rank for $\eta_{90} \leq 10\%$ Median / 90th pct.	Mixed groups Mean / maximum
2.4 / 7.7	16 / 64	14.3 / 16

Max. relative drift	Recovery cost (%)	Structured gap (%)	Variance bound / cost
2.94×10^{-6}	1.1	1.1	2.4

C.4 Downstream task comparison

Task scores in Table 2 are macro-averages over six models. LDC-FP4 achieves the highest quantized score on each of the four reported accuracy aggregates and the lowest token KL. Its six-task advantage over FAAR is 0.7 percentage points, with a paired uncertainty interval of [+0.31, +1.09] points. Relative to GPTAQ with H-Scale, MMLU, GSM8K, and GPQA Diamond improve by 0.9, 0.7, and 0.8 percentage points, respectively. Mean token KL falls from 0.0625 to 0.0529, a 15.4% reduction.

D Detailed Algorithms

Algorithm 2 expands the layer encoder. Algorithms 3 and 4 give its support-reduction and recovery steps. Row computations are batched in the implementation. All objective comparisons include the asymmetric term from (2).

For group b, $\mathcal{S}_b$ contains the sixteen legal E4M3 scales described in Section 4, with boundary indices clipped. The initial native payload is retained even if its scale lies outside this set. The positive factors c_i denote the row normalization from Appendix B. Barred objectives use $\bar{\alpha} = \alpha/c_i$, $\bar{v} = v/\sqrt{c_i}$, and $\bar{\mathcal{L}}_0 = \mathcal{L}_0/c_i$. The returned dual point is converted to the original objective units before evaluating its bound.

Algorithm 2 LDC-FP4 layer encoding

Require: $W \in \mathbb{R}^{m\times d}$, $Z \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^{n\times m}$, rank r, and $\gamma, \delta > 0$
Require: Row normalizations $c_i > 0$ and an initial dual point $\bar{\lambda}_0 \in \mathbb{R}^r$
$S \leftarrow \mathrm{BlkDiag}(Z^\top Z/n) + \delta I, \quad H \leftarrow Z^\top Z/n + \gamma S$
$(D, U) \leftarrow \mathrm{FitCurvature}(H, S, r)$ ▷ Algorithm 5
$Q^{(0)} \leftarrow \mathrm{GPTAQ}_{\mathrm{FP4}}(W, Z, Y)$
for $i = 1, \ldots, m$ in output-row tiles **do**
  $w \leftarrow W_{i,:}^\top, \quad y \leftarrow Y_{:,i}, \quad q^{(0)} \leftarrow (Q_{i,:}^{(0)})^\top$
  $a \leftarrow Z^\top(Zw - y)/n, \quad \mathcal{L}_0 \leftarrow \|Zw - y\|^2/n$
  for $b = 1, \ldots, B$ **do**
    $\mathcal{Q}_b \leftarrow \{q_b^{(0)}\}$
    for $s \in \mathcal{S}_b$ **do**
      $q_b^{\mathrm{RTN}} \leftarrow$ nearest-even E2M1 rounding of w_b at scale s
      $q_b^{\mathrm{CD}} \leftarrow \mathrm{LocalRefine}(q_b^{\mathrm{RTN}}, w_b, H_{bb}, a_b, s)$ ▷ Algorithm 6
      $\mathcal{Q}_b \leftarrow \mathcal{Q}_b \cup \{q_b^{\mathrm{RTN}}, q_b^{\mathrm{CD}}\}$
    end for
    Deduplicate $\mathcal{Q}_b$ and index it as $\{q_{b1}, \ldots, q_{bK_b}\}$.
    $e_{bk} \leftarrow q_{bk} - w_b, \quad \alpha_{bk} \leftarrow e_{bk}^\top D_b e_{bk} + 2a_b^\top e_{bk}, \quad v_{bk} \leftarrow U_b^\top e_{bk}$
  end for
  $\bar{\alpha} \leftarrow \alpha/c_i, \quad \bar{v} \leftarrow v/\sqrt{c_i}, \quad \bar{\mathcal{L}}_0 \leftarrow \mathcal{L}_0/c_i, \quad \bar{\lambda} \leftarrow \bar{\lambda}_0$
  for $\tau \in (10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6})$ **do**
    $\bar{\lambda} \leftarrow \mathrm{LBFGS}(-\bar{g}_\tau, \bar{\lambda})$ ▷ at most 200 iterations
    $p_{bk} \leftarrow \mathrm{softmax}_k[-(\bar{\alpha}_{bk} + 2\bar{\lambda}^\top \bar{v}_{bk})/\tau]$
  end for
  $\lambda \leftarrow \sqrt{c_i}\,\bar{\lambda}, \quad L_{\mathrm{dual}} \leftarrow g(\lambda)$
  $\widetilde{p} \leftarrow \mathrm{SupportReduce}(p, \alpha, v, r)$ ▷ Algorithm 3
  $\widehat{q} \leftarrow \mathrm{RecoverRefine}(\widetilde{p}, \alpha, v, \{\mathcal{Q}_b\})$ ▷ Algorithm 4
  $Q_{i,:}^\top \leftarrow \arg\min_{q \in \{q^{(0)}, \widehat{q}\}} \mathcal{L}(q)$
end for
return $\mathrm{Pack}_{\mathrm{E2M1/E4M3}}(Q)$

L-BFGS minimizes the negative smooth dual numerically (Liu & Nocedal, 1989; Nesterov, 2005). Its gradient is $\nabla \bar{g}_\tau(\bar{\lambda}) = -2\bar{\lambda} + 2\sum_{b,k} p_{bk}\bar{v}_{bk}$. The previous temperature's iterate initializes the next solve. Equation (18) supplies the residual-based stopping diagnostic. The unsmoothed $g(\lambda)$ bounds the structured candidate objective. Final acceptance evaluates the original objective (1).

D.1 Support reduction

The following pivots preserve every group sum and the aggregate residual. At termination, $\sum_b(|\,\mathrm{supp}(p_b)| - 1) \le r$, so at most r groups remain mixed. The sign choice makes the objective nonincreasing. In floating-point arithmetic, group normalization and residual preservation are checked after each batch of pivots, as described in Appendix A.

Algorithm 3 SUPPORTREDUCE

Require: Feasible $p \in \prod_b \Delta_{K_b}$, costs α_{bk}, features $v_{bk} \in \mathbb{R}^r$
while $\sum_b (|\operatorname{supp}(p_b)| - 1) > r$ **do**
  Choose an anchor $a(b) \in \operatorname{supp}(p_b)$ in every group.
  Choose $r+1$ active nonanchor pairs $T \subseteq \{(b,k) : p_{bk} > 0,\ k \neq a(b)\}$.
  Find $h \neq 0$ satisfying $\sum_{(b,k)\in T} h_{bk}(v_{bk} - v_{b,a(b)}) = 0$.
  $d \leftarrow 0$
  for $(b,k) \in T$ **do**
    $d_{bk} \leftarrow h_{bk}, \quad d_{b,a(b)} \leftarrow d_{b,a(b)} - h_{bk}$
  end for
  if $\sum_{b,k} \alpha_{bk} d_{bk} > 0$ **then**
    $d \leftarrow -d$
  end if
  $\theta \leftarrow \min_{d_{bk}<0}(-p_{bk}/d_{bk})$
  $p \leftarrow p + \theta d$
end while
return p

D.2 RECOVERY AND REFINEMENT

Let $A_b = \operatorname{supp}(\widetilde{p}_b)$. Enumeration searches only the surviving support. When enumeration exceeds the budget, conditional expectation fixes each group using the means of the unresolved groups. The subsequent two sweeps search the full candidate lists. Each update minimizes the structured objective with the other groups fixed. No new scale or code is introduced.

Algorithm 4 RECOVERREFINE

Require: Reduced mixture $\widetilde{p}$, costs α, features v, and full lists $\mathcal{Q}_b$
$A_b \leftarrow \operatorname{supp}(\widetilde{p}_b)$ for every b
if $\prod_b |A_b| \le 4{,}096$ **then**
  $k \leftarrow \arg\min_{k \in \prod_b A_b} \left[\sum_b \alpha_{b,k_b} + \left\|\sum_b v_{b,k_b}\right\|^2\right]$
else
  $\mu_b \leftarrow \sum_j \widetilde{p}_{bj} v_{bj}$ for every b, $\quad z \leftarrow \sum_b \mu_b$
  for $b = 1, \dots, B$ **do**
    $z_{-b} \leftarrow z - \mu_b$
    $k_b \leftarrow \arg\min_{j \in A_b} \left[\alpha_{bj} + \|z_{-b} + v_{bj}\|^2\right]$
    $z \leftarrow z_{-b} + v_{b,k_b}$
  end for
end if
$z \leftarrow \sum_b v_{b,k_b}$
for two sweeps **do**
  for $b = 1, \dots, B$ **do**
    $z_{-b} \leftarrow z - v_{b,k_b}$
    $k_b \leftarrow \arg\min_{1 \le j \le K_b} \left[\alpha_{bj} + \|z_{-b} + v_{bj}\|^2\right]$
    $z \leftarrow z_{-b} + v_{b,k_b}$
  end for
end for
return $\widehat{q} \leftarrow (q_{1,k_1}, \dots, q_{B,k_B})$

D.3 CURVATURE FITTING

For a symmetric matrix A, let $\Pi_+(A)$ replace its negative eigenvalues by zero. Let $\mathrm{EigFactor}_r(A)$ retain its r largest positive eigenvalues and return their eigenvectors scaled by the square roots of those eigenvalues. Zero columns pad the factor if fewer than r eigenvalues are positive. These operations solve the two subproblems of (15) (Eckart & Young, 1936; Higham, 1988). The iteration limit and relative tolerance below are the defaults in Table 3.

Algorithm 5 FITCURVATURE

Require: Curvature H, block-diagonal $S \succ 0$, rank r, tolerance $\epsilon_{\text{fit}} = 10^{-3}$

$C \leftarrow S^{-1/2} H S^{-1/2}$
$V \leftarrow \text{EigFactor}_r(C)$
$\Delta \leftarrow \text{BlkDiag}\big(\Pi_+(C_{bb} - V_b V_b^\top)\big)$
$J \leftarrow \big\|C - \Delta - VV^\top\big\|_F^2$
for $t = 1, \dots, 10$ **do**
  if $J = 0$ **then**
    break
  end if
  $J_{\text{prev}} \leftarrow J$
  $\Delta \leftarrow \text{BlkDiag}\big(\Pi_+(C_{bb} - V_b V_b^\top)\big)$
  $V \leftarrow \text{EigFactor}_r(C - \Delta)$
  $J \leftarrow \big\|C - \Delta - VV^\top\big\|_F^2$
  if $|J_{\text{prev}} - J| \le \epsilon_{\text{fit}} J_{\text{prev}}$ **then**
    break
  end if
end for
$D \leftarrow S^{1/2} \Delta S^{1/2}, \quad U \leftarrow S^{1/2} V$
return (D, U)

D.4 LOCAL CANDIDATE CONSTRUCTION

At a fixed E4M3 scale s, let $\mathcal{G}_s$ be the corresponding E2M1 grid with the common tensor scale included. The projection $\Pi_{\mathcal{G}_s}$ uses nearest-even rounding and the shared saturation convention. Algorithm 6 updates coordinates within a single candidate under the full local metric H_{bb}. This differs from Algorithm 4, which selects among complete candidates using the structured objective.

Algorithm 6 LOCALREFINE

Require: Initial candidate q_b, reference w_b, local metric H_{bb}, linear term a_b, scale s

for two sweeps **do**
  for $j = 1, \dots, g$ **do**
    $u_j \leftarrow w_{b,j} - \dfrac{a_{b,j} + \sum_{\ell \neq j} (H_{bb})_{j\ell} (q_{b,\ell} - w_{b,\ell})}{(H_{bb})_{jj}}$
    $q_{b,j} \leftarrow \Pi_{\mathcal{G}_s}(u_j)$
  end for
end for
return q_b