BPDQ: Bit-Plane Decomposition Quantization on a Variable Grid for Large Language Models

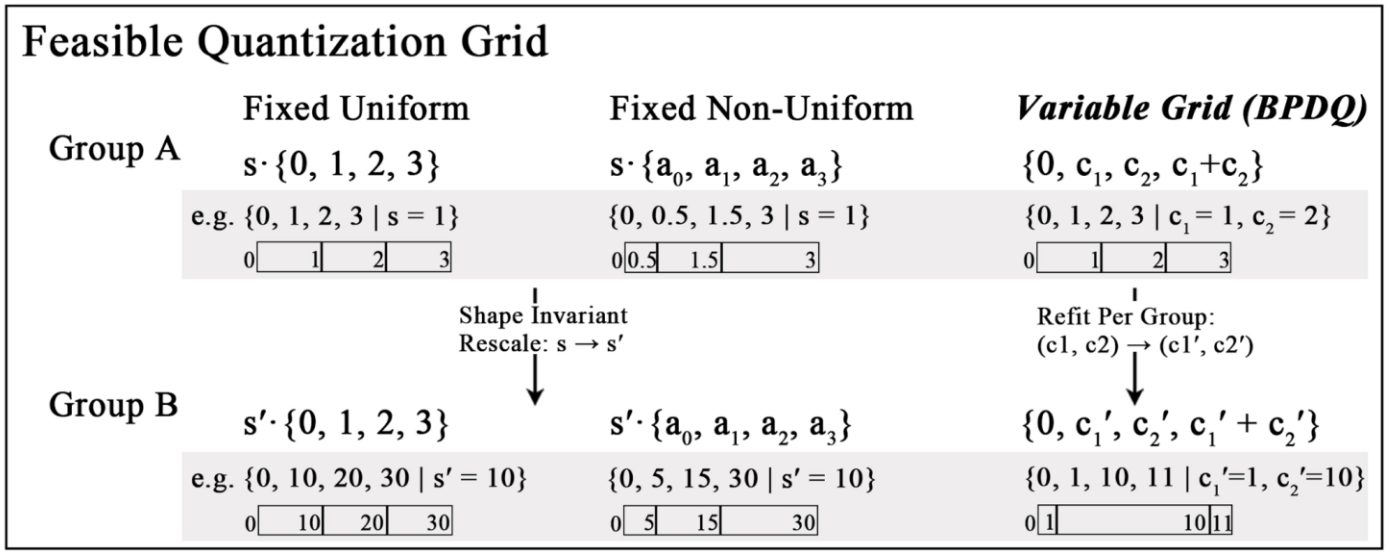
Junyu CHEN

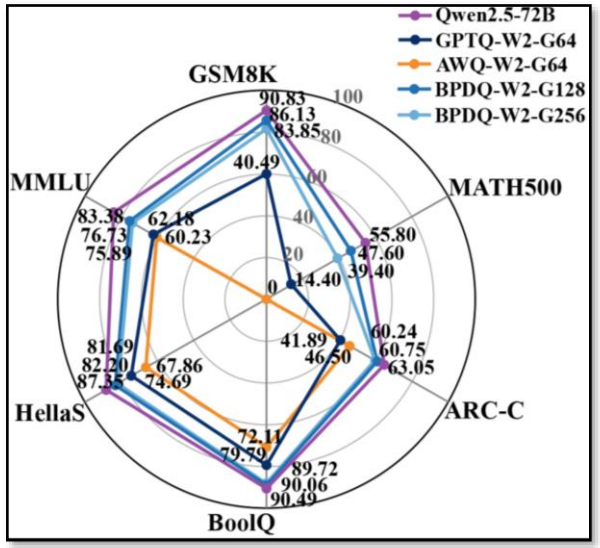
30/05/2026

ICML
International Conference
On Machine Learning

Code: github.com/KingdalfGoodman/BPDQ
Model: huggingface.co/goodman20241017/Qwen2.5-72B-BPD2-G256
Email: KingdalfGoodman@foxmail.com

Core Insight - Beyond the Fixed Grid





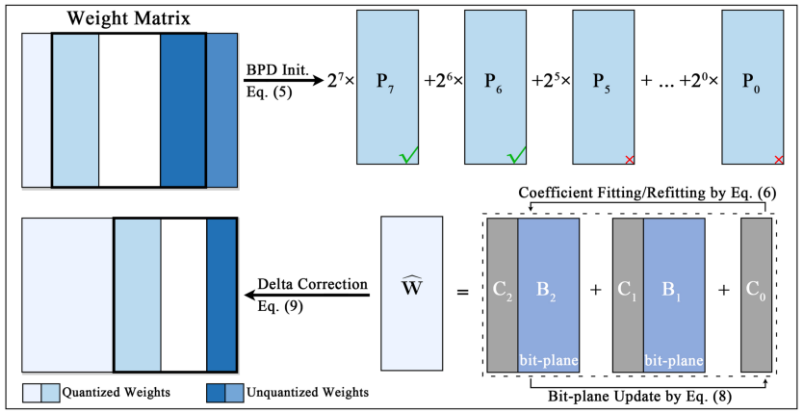
Fixed grids (Uniform/Non-Uniform) enforce shape invariance, where the relative spacing of quantization levels is shared across groups (scaled by s).

VS

BPDQ breaks this limitation by constructing a variable grid per group using bit-plane coefficients (c1, c2), expanding the feasible set.

Essentially, the problem is not a failure of the optimization objective, but the rigidity of the quantizer's feasible set under that objective.

Preliminaries



3.1. Preliminaries

Optimization Objective. Consider a linear layer with weight $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and input activations $\mathbf{X} \in \mathbb{R}^{d_{\text{in}} \times N}$ computed from N calibration samples. The quantized weight $\widehat{\mathbf{W}} \in \mathcal{Q}$ is obtained by minimizing the output reconstruction error:

$$\begin{aligned} \widehat{\mathbf{W}} &= \underset{\widetilde{\mathbf{W}} \in \mathcal{Q}}{\operatorname{argmin}} \left\| (\mathbf{W} - \widetilde{\mathbf{W}})\mathbf{X} \right\|_F^2 \\ &= \underset{\widetilde{\mathbf{W}} \in \mathcal{Q}}{\operatorname{argmin}} \operatorname{tr}\big((\mathbf{W} - \widetilde{\mathbf{W}})\,\mathbf{H}\,(\mathbf{W} - \widetilde{\mathbf{W}})^\top\big), \end{aligned} \tag{2}$$

where $\mathbf{H} = \mathbf{X}\mathbf{X}^\top$ is an approximate Hessian metric induced by calibration data, and $\mathcal{Q}$ denotes the set of admissible low-bit weight matrices.

Quantization Error Compensation. Due to the enormous parameter space of LLMs, repeatedly updating the inverse Hessian is prohibitively expensive. To address this, GPTQ (Frantar et al., 2022) operates within the Hessian-induced geometry by using the upper-triangular Cholesky factorization $\mathbf{U} = \operatorname{chol}(\mathbf{H}^{-1})$ (i.e., $\mathbf{H}^{-1} = \mathbf{U}^\top\mathbf{U}$), and performs error propagation via triangular updates to compensate the induced quantization error on the remaining free coordinates. When quantizing the l-th column, let $\mathbf{W}_{:,l}$ and $\widehat{\mathbf{W}}_{:,l}$ denote the current working and quantized column vectors, respectively. Then define the error coordinate:

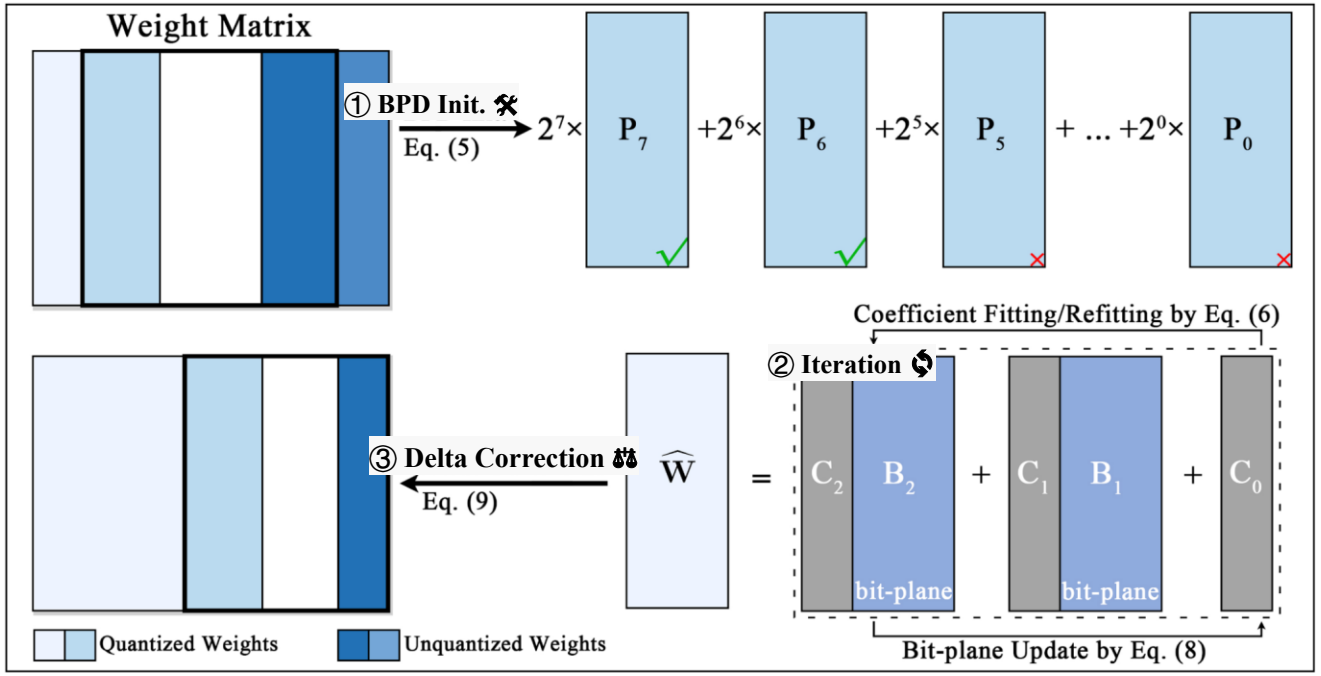
$$\mathbf{E}_{:,l} = \frac{\mathbf{W}_{:,l} - \widehat{\mathbf{W}}_{:,l}}{\mathbf{U}_{l,l}}. \tag{3}$$

The Hessian-aware compensation update is:

$$\mathbf{W}_{:,\,l:} \leftarrow \mathbf{W}_{:,\,l:} - \mathbf{E}_{:,l}\,\mathbf{U}_{l,\,l:}, \tag{4}$$

which maintains the optimization-based PTQ error-propagation state under the Hessian-induced geometry.

Methodology



$$\mathbf{Z} = \sum_{i=0}^{7} 2^i \, \mathbf{P}_i, \tag{5}$$

$$c_r = \underset{c \in \mathbb{R}^{k+1}}{\operatorname{argmin}} \left\| \mathbf{U}_{\text{loc}}^{-\top} \left(\mathbf{B}_r \, c - \mathbf{W}_{r,s:(s+g)}^{\top} \right) \right\|_2^2. \tag{6}$$

$$\mathbf{b}^* = \underset{\mathbf{b} \in \{0,1\}^k}{\operatorname{argmin}} \left(\mathbf{W}'_{r,l} - v_r(\mathbf{b}) \right)^2, \tag{8}$$

$$\Delta \mathbf{E} \, \mathbf{U}_{\text{loc}} = \widehat{\mathbf{W}}_{\text{old}} - \widehat{\mathbf{W}}_{\text{new}}, \tag{9}$$

① Bit-Plane Decomposition (BPD) Initialization 🛠

Take the MSB planes (P_7, P_6) as the initial bit-planes (B_2, B_1), since they carry the dominant magnitude of the weights. Discard the LSB planes, as the truncation error is small, giving a low-cost warm start.

③ Delta Correction — Error-Propagation Alignment ⚖

Re-align the accumulated error coordinates after each coefficient refit. The delta correction injects the change exactly into the error propagation chain, so every iterate stays consistent with the output-aligned objective. *Proof in Appendix B.3*

② Iteration under closed-form steps 🔄

Step A (Fix Bits → Refit Coeffs):Closed-form weighted least-squares for the scalar coefficients.(Yields a per-group, tailored variable grid.) *Proof in Appendix B.1*

Step B (Fix Coeffs → Update Bits):Exhaustive enumeration over bit patterns, in parallel across rows.(A nearest-neighbor projection under the Hessian-induced metric.) *Proof in Appendix B.2*

Experiments - Benefits of Variable Grid

Model	BPW	Wiki2 ↓	GSM8K ↑	MATH500 ↑	ARC-C ↑	BoolQ ↑	HellaS ↑	MMLU ↑
Qwen2.5-72B	*16*	*4.72*	*90.83%*	*55.80%*	*63.05%*	*90.49%*	*87.35%*	*83.38%*
GPTQ-W4-G64	4.31	5.01	90.52%	56.00%	**63.99%**	**90.76%**	87.04%	82.77%
AWQ-W4-G64	4.31	5.64	91.28%	**59.20%**	61.26%	90.52%	86.92%	82.14%
BPDQ-W4-G128	4.63	**4.95**	**92.65%**	55.40%	62.88%	90.70%	**87.21%**	**83.09%**
GPTQ-W3-G32	3.59	5.76	**91.74%**	51.00%	**63.05%**	90.24%	86.40%	**82.19%**
AWQ-W3-G32	3.59	5.57	90.90%	**58.60%**	62.54%	**90.52%**	86.63%	82.01%
BPDQ-W3-G64	4.00	**5.55**	91.21%	56.40%	62.71%	**90.52%**	**86.73%**	81.59%
GPTQ-W3-G64	3.30	6.04	90.07%	50.80%	59.98%	90.12%	86.22%	81.55%
AWQ-W3-G64	3.30	5.85	**90.75%**	**58.60%**	**63.74%**	**90.58%**	86.35%	81.58%
BPDQ-W3-G128	3.50	**5.73**	90.67%	56.60%	62.80%	90.52%	**86.36%**	**81.65%**
GPTQ-W2-G32	2.56	10.01	63.46%	28.40%	53.16%	86.21%	78.60%	69.59%
AWQ-W2-G32	2.56	4.0E+7	0.00%	0.00%	41.47%	68.75%	58.09%	56.94%
BPDQ-W2-G64	2.75	**8.35**	**87.72%**	**51.20%**	**59.47%**	**90.37%**	**82.71%**	**77.14%**
GPTQ-W2-G64	2.28	12.47	40.49%	14.40%	41.89%	79.79%	74.69%	62.18%
AWQ-W2-G64	2.28	1.6E+7	0.00%	0.00%	46.50%	72.11%	67.86%	60.23%
BPDQ-W2-G128	2.38	**8.66**	**86.13%**	**47.60%**	**60.75%**	**90.06%**	**82.20%**	**76.73%**
BPDQ-W2-G256	2.19	8.94	83.85%	39.40%	60.24%	89.72%	81.69%	75.89%

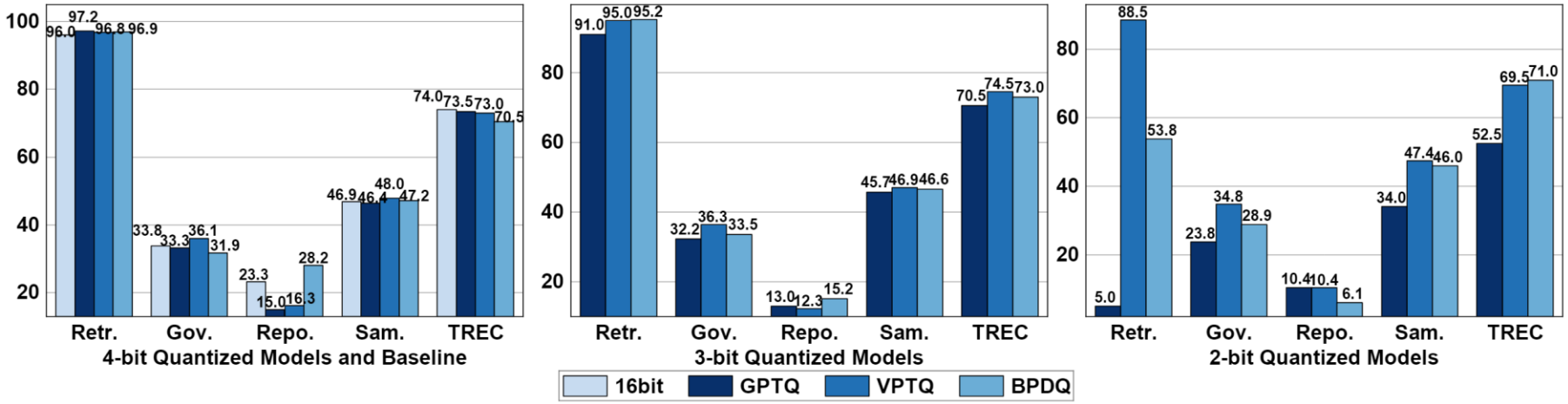
Experiments - Comparison with Bit-Plane and VQ Methods

Model	SIZE(GB)	Wiki2 ↓	GSM8K ↑	MATH500 ↑	ARC-C ↑	BoolQ ↑	HellaS ↑	MMLU ↑
Qwen2.5-7B	*14.19*	*9.42*	*75.97%*	*46.00%*	*55.29%*	*86.39%*	*80.44%*	*71.76%*
GPTQ-W4-G64	5.31	9.72	78.32%	42.20%	54.52%	**86.82%**	80.00%	71.16%
AWQ-W4-G64	5.31	10.35	78.29%	45.20%	55.12%	86.24%	79.93%	71.08%
AnyBCQ-W4-G128	6.30	11.18	29.26%	33.60%	50.68%	84.83%	79.17%	69.50%
VPTQ-W4	5.46	**9.62**	**79.45%**	**47.00%**	54.27%	86.73%	79.87%	**71.33%**
BPDQ-W4-G128	5.54	9.66	78.24%	41.60%	**55.80%**	86.42%	**80.03%**	71.19%
GPTQ-W3-G32	4.77	**10.04**	72.48%	44.40%	**54.78%**	83.91%	**78.72%**	68.97%
AWQ-W3-G32	4.76	10.70	57.16%	44.60%	51.71%	**86.36%**	78.09%	68.58%
AnyBCQ-W3-G64	5.82	12.24	26.61%	25.60%	50.34%	82.39%	77.21%	66.99%
VPTQ-W3	4.51	10.32	**78.17%**	**46.60%**	51.28%	85.96%	78.17%	69.78%
BPDQ-W3-G64	5.07	10.31	76.42%	44.00%	54.35%	85.90%	78.43%	**69.90%**
GPTQ-W3-G64	4.54	**10.27**	63.53%	39.40%	52.82%	84.68%	**78.45%**	67.53%
AWQ-W3-G64	4.54	11.28	65.58%	38.00%	50.17%	84.95%	77.27%	67.23%
AnyBCQ-W3-G128	5.06	12.44	25.63%	27.40%	52.39%	82.97%	76.77%	66.59%
BPDQ-W3-G128	4.69	10.55	**71.27%**	**40.60%**	**54.27%**	**86.21%**	77.92%	**69.53%**
GPTQ-W2-G32	3.98	21.66	0.38%	3.00%	34.04%	65.02%	66.12%	37.12%
AWQ-W2-G32	3.98	N/A	2.43%	0.00%	34.64%	45.99%	48.98%	28.30%
AnyBCQ-W2-G64	4.30	19.20	9.63%	5.80%	45.48%	69.97%	68.24%	54.26%
VPTQ-W2	4.32	**14.38**	**67.63%**	**33.40%**	**52.56%**	**86.79%**	**73.60%**	**65.81%**
BPDQ-W2-G64	4.12	15.09	44.50%	13.60%	48.29%	85.50%	69.98%	57.51%
GPTQ-W2-G64	3.77	42.59	0.00%	1.40%	29.27%	59.14%	58.51%	27.10%
AWQ-W2-G64	3.76	N/A	0.00%	0.00%	25.17%	38.81%	32.14%	26.03%
AnyBCQ-W2-G128	3.92	22.57	4.47%	5.80%	44.54%	77.52%	65.83%	48.54%
BPDQ-W2-G128	3.84	**16.85**	**35.48%**	**10.40%**	**45.90%**	**84.62%**	**68.76%**	**57.46%**

Experiments - Further Analysis of BPDQ

Model	Efficiency Profile			Outlier Statistics			
	Cost (min)	VRAM (GB)	Latency (ms)	DiagR (P95)	ΔDiagR	Cnt10	ΔCnt10
Qwen2.5-7B	N/A	14.19	14.42	3.01E4	N/A	4.32E4	N/A
GPTQ-W4-G64	16	6.63	18.74	3.28E4	+8.97%	4.28E4	-0.93%
VPTQ-W4	4×160 †	5.46	20.07	2.83E4	-5.98%	4.30E4	-0.46%
BPDQ-W4-G128	47	5.55	18.20	2.95E4	-1.99%	4.28E4	-0.93%
GPTQ-W3-G32	16	4.54	47.67	2.82E4	-6.31%	4.12E4	-4.63%
VPTQ-W3	4×160 †	4.51	17.34	2.59E4	-13.95%	4.38E4	+1.39%
BPDQ-W3-G64	40	4.69	18.21	2.96E4	-1.66%	4.29E4	-0.69%
GPTQ-W2-G32	17	3.77	33.91	2.02E4	-32.89%	3.30E4	-23.61%
VPTQ-W2	4×170 †	4.32	18.24	2.68E4	-10.96%	4.44E4	+2.78%
BPDQ-W2-G64	40	3.86	18.09	2.86E4	-4.98%	4.24E4	-1.85%

† VPTQ costs from the paper require 4 GPUs and ∼10× time relative to the single-GPU GPTQ baseline.



Theoretical Analysis - Variable Grid Expressivity

A. Analysis of the Variable Grid

A.1. Optimization-based PTQ as an H-Metric Projection

Consider a weight vector $\mathbf{w} \in \mathbb{R}^g$ within a quantization group of size g. Let $\mathbf{H} \in \mathbb{R}^{g\times g}$ denote the corresponding Hessian matrix. We define the Hessian-induced norm as $\|\mathbf{e}\|_{\mathbf{H}} = \sqrt{\mathbf{e}^\top \mathbf{H}\mathbf{e}}$. Optimization-based PTQ determines a quantized vector $\widehat{\mathbf{w}}$ from a feasible set $\mathcal{Q} \subset \mathbb{R}^g$ that minimizes the output discrepancy. This is equivalent to finding the projection of $\mathbf{w}$ onto $\mathcal{Q}$ under the $\mathbf{H}$-metric:

$$\widehat{\mathbf{w}} \;=\; \Pi_{\mathcal{Q}}^{(\mathbf{H})}(\mathbf{w}) \;=\; \operatorname*{argmin}_{\widetilde{\mathbf{w}}\in\mathcal{Q}} \|\mathbf{w}-\widetilde{\mathbf{w}}\|_{\mathbf{H}}^2 \;=\; \operatorname*{argmin}_{\widetilde{\mathbf{w}}\in\mathcal{Q}} (\mathbf{w}-\widetilde{\mathbf{w}})^\top \mathbf{H}(\mathbf{w}-\widetilde{\mathbf{w}}). \tag{10}$$

Eq. (10) establishes that optimization-based PTQ is a nearest-point projection problem. Consequently, quantization quality is restricted by the geometric richness of the feasible set $\mathcal{Q}$. In the low-bit regime (e.g., 2-3 bits), fidelity degradation stems not from a failure of the objective, but from the rigidity of the feasible set $\mathcal{Q}$ induced by a shape-invariant grid.

A.2. Feasible Set Comparison at 2-Bit Precision

The distinction between fixed and variable grids lies in the geometric degrees of freedom defining the quantization levels within each group. A fixed grid constrains the levels to a shape-invariant template governed by a scaling factor, restricting the solution to a lower-dimensional manifold. In contrast, BPDQ constructs the grid using independent scalar coefficients for each group, decoupling the quantization intervals and expanding the feasible set to a higher-dimensional geometry.

Fixed Grid (Rigid Template). Given a canonical template $\mathbf{t} = [t_0, t_1, t_2, t_3]^\top \in \mathbb{R}^4$ (e.g., $[0, 1, 2, 3]^\top$ for UINT2), a fixed grid restricts the quantization levels $\mathbf{q} \in \mathbb{R}^4$ to a scaling factor of this template. For a group scale $s \in \mathbb{R}$:

$$\mathcal{Q}_{\text{fix}}(s) \;=\; s \cdot \{t_0, t_1, t_2, t_3\}, \tag{11}$$

where s varies across groups, but the relative ratios between levels (e.g., $q_2/q_1 = (s \cdot t_2)/(s \cdot t_1) = t_2/t_1, t_1 \neq 0$) remain frozen. This rigidity restricts the feasible level vectors to a one-dimensional ray in $\mathbb{R}^4$.

Variable Grid (Adaptive Geometry). BPDQ constructs the grid via two bit-planes weighted by coefficients $c_1, c_2 \in \mathbb{R}$:

$$\mathcal{Q}_{\text{var}}(c_1, c_2) \;=\; \{0,\ c_1,\ c_2,\ c_1 + c_2\}, \tag{12}$$

where c_1 and c_2 are independent coefficients determined per group, enabling dynamic relative ratios (i.e., c_2/c_1 is flexible, $c_1 \neq 0$). This flexibility expands the feasible level vectors to a two-dimensional plane in $\mathbb{R}^4$.

Proposition 1: Strict Inclusion of Uniform Grids. *The feasible set of BPDQ strictly contains the feasible set of standard UINT2 grids.*

Proof. Without loss of generality, factoring out the group-wise bias (available to both schemes), we consider the canonical zero-based template $\mathbf{t}_{\text{uni}} = \{0, 1, 2, 3\}$. Accordingly, any uniform grid is essentially a scaled instance of this template, given by $s \cdot \{0, 1, 2, 3\} = \{0, s, 2s, 3s\}$. To show inclusion ($\mathcal{Q}_{\text{uniform}} \subset \mathcal{Q}_{\text{BPDQ}}$), we set the coefficients $c_1 = s$ and $c_2 = 2s$:

$$\mathcal{Q}_{\text{var}}(s, 2s) = \{0,\ s,\ 2s,\ s + 2s\} = \{0, s, 2s, 3s\} \equiv \mathcal{Q}_{\text{fix}}^{\text{uni}}(s). \tag{13}$$

This confirms that BPDQ can exactly reproduce any UINT2 grid. Furthermore, the inclusion is strict because BPDQ admits non-uniform spacings (e.g., when $c_2 = 10c_1$) that no linear scale s can represent. Consequently, the quantization error of BPDQ is upper-bounded by that of UINT2:

$$\min_{\mathbf{q}\in\mathcal{Q}_{\text{var}}} \|\mathbf{w}-\mathbf{q}\|_{\mathbf{H}}^2 \leq \min_{\mathbf{q}\in\mathcal{Q}_{\text{fix}}^{\text{uni}}} \|\mathbf{w}-\mathbf{q}\|_{\mathbf{H}}^2. \tag{14}$$

□

Proposition 2: Strict Error Reduction via Variable-Grid Expressivity. *Compared to shape-invariant fixed-template quantization grids parameterized only by a per-group bias c_0 and scale s, 2-bit BPDQ induces additional degrees of freedom and can realize feasible points unattainable by rigid templates. Consequently, there exists a non-empty open set of weight vectors $\mathcal{U} \subset \mathbb{R}^g$ where BPDQ achieves strictly lower quantization error.*

Proof. Assume group size $g \geq 3$ and Hessian $\mathbf{H} \succ 0$. Define the feasible sets for the fixed grid ($\mathcal{S}_{\text{fix}}$) and BPDQ ($\mathcal{S}_{\text{var}}$):

$$\mathcal{S}_{\text{fix}} = \{c_0\mathbf{1} + s\mathbf{z} \mid c_0, s \in \mathbb{R}, \mathbf{z} \in \{t_0, \ldots, t_3\}^g\}, \tag{15}$$

$$\mathcal{S}_{\text{var}} = \{c_0\mathbf{1} + c_1\mathbf{b}_1 + c_2\mathbf{b}_2 \mid c_0, c_1, c_2 \in \mathbb{R}, \mathbf{b}_1, \mathbf{b}_2 \in \{0, 1\}^g\}. \tag{16}$$

For any specific pattern $\mathbf{z}$ (or bit-planes $\mathbf{b}_1, \mathbf{b}_2$), the generated vectors form an affine subspace. Since the number of patterns is finite (4^g), both $\mathcal{S}_{\text{fix}}$ and $\mathcal{S}_{\text{var}}$ are finite unions of affine subspaces, and are thus closed sets.

Construction of $\mathbf{v}^*$: Define the finite set of difference ratios for $\mathbf{t} = [t_0, t_1, t_2, t_3]^\top \in \mathbb{R}^4$:

$$\mathcal{R}_\Delta(\mathbf{t}) = \left\{ \frac{t_i - t_j}{t_i - t_k} \;\middle|\; t_i, t_j, t_k \in \{t_0, \ldots, t_3\}, t_i \neq t_j, t_i \neq t_k, t_j \neq t_k \right\}. \tag{17}$$

For any $\mathbf{q} \in \mathcal{S}_{\text{fix}}$, if $\mathbf{q}$ attains three distinct values x, y, z across three coordinates (so $t_i \neq t_j$, $t_i \neq t_k$ and $t_j \neq t_k$ when $s \neq 0$), they must satisfy $x = c_0 + st_i$, $y = c_0 + st_j$, $z = c_0 + st_k$. The bias and scale cancel out in the difference ratio: $(x - y)/(x - z) = (t_i - t_j)/(t_i - t_k) \in \mathcal{R}_\Delta(\mathbf{t})$. In contrast, BPDQ can generate a vector $\mathbf{v}^*$ containing values $\{c_0, c_0 + c_1, c_0 + c_2\}$ by selecting bit-planes such that three coordinates take patterns $(0, 0), (1, 0), (0, 1)$. Let these three values be $x = c_0, y = c_0 + c_1, z = c_0 + c_2$. The difference ratio becomes:

$$\frac{x - y}{x - z} = \frac{c_0 - (c_0 + c_1)}{c_0 - (c_0 + c_2)} = \frac{c_1}{c_2}. \tag{18}$$

Since $\mathcal{R}_\Delta(\mathbf{t})$ is a finite set, we can choose $c_1, c_2 \in \mathbb{R}$ such that the ratio $c_1/c_2 \notin \mathcal{R}_\Delta(\mathbf{t})$ and the resulting values x, y, z are distinct. Suppose for contradiction that $\mathbf{v}^* \in \mathcal{S}_{\text{fix}}$. As $\mathbf{v}^*$ attains the distinct values x, y, z at the chosen coordinates, the fixed-grid constraint would necessitate that their difference ratio falls within $\mathcal{R}_\Delta(\mathbf{t})$, which contradicts our construction. Thus, $\mathbf{v}^* \in \mathcal{S}_{\text{var}}$ but $\mathbf{v}^* \notin \mathcal{S}_{\text{fix}}$.

Strict Inequality over an Open Set: Consider a weight vector $\mathbf{w} = \mathbf{v}^*$, for which the BPDQ error is zero (i.e., $F_{\text{var}}(\mathbf{v}^*) = 0$). Since $\mathbf{H} \succ 0$ induces a norm equivalent to the Euclidean norm on $\mathbb{R}^g$, and $\mathcal{S}_{\text{fix}}$ is a closed set with $\mathbf{v}^* \notin \mathcal{S}_{\text{fix}}$, the distance is strictly positive:

$$F_{\text{fix}}(\mathbf{v}^*) = \inf_{\mathbf{q}\in\mathcal{S}_{\text{fix}}} \|\mathbf{v}^* - \mathbf{q}\|_{\mathbf{H}}^2 = \delta > 0. \tag{19}$$

Define the error difference function $f(\mathbf{w}) = F_{\text{fix}}(\mathbf{w}) - F_{\text{var}}(\mathbf{w})$. Since $\mathcal{S}_{\text{fix}}$ and $\mathcal{S}_{\text{var}}$ are closed sets, their respective squared-distance functions $F_{\text{fix}}(\mathbf{w})$ and $F_{\text{var}}(\mathbf{w})$ are continuous. Specifically, the distance-to-set function $d(\mathbf{w}, \mathcal{S})$ is 1-Lipschitz under $\|\cdot\|_{\mathbf{H}}$, and squaring preserves continuity. Therefore, $f(\mathbf{w})$ is continuous, so there exists an open neighborhood $\mathcal{U}$ around $\mathbf{v}^*$ where $f(\mathbf{w}) > 0$ (i.e., $F_{\text{fix}}(\mathbf{w}) > F_{\text{var}}(\mathbf{w})$). This confirms the existence of a strict inequality over an open set. □

Remark: Geometric Degrees of Freedom. Proposition 1 establishes strictly nested feasibility for uniform grids ($\mathcal{S}_{\text{uni}} \subsetneq \mathcal{S}_{\text{BPDQ}}$), and Proposition 2 addresses general fixed templates. Geometrically, a specific choice of bit-planes $(\mathbf{b}_1, \mathbf{b}_2)$ in BPDQ yields an affine subspace of dimension up to 3 (spanning $\mathbf{1}, \mathbf{b}_1, \mathbf{b}_2$), whereas fixed templates yield subspaces of dimension at most 2 (spanning $\mathbf{1}, \mathbf{z}$). In the generic case where $(\mathbf{1}, \mathbf{b}_1, \mathbf{b}_2)$ are linearly independent, this provides an additional coefficient degree of freedom (3 parameters (c_0, c_1, c_2) vs. 2 parameters (c_0, s)). ***Crucially, the fixed grid enforces global rigidity across all groups, whereas the variable grid enables adaptability tailored for each group.***

Theoretical Analysis - Hessian-Induced Consistency

B. Consistency in Hessian-Induced Geometry

B.1. Consistency of Coefficient Fitting

Proposition. *The coefficient fitting objective in Eq. (6) is theoretically equivalent to minimizing the local contribution to the optimization objective in Eq. (2) corresponding to the current group, under the Hessian-induced geometry defined by the upper-triangular Cholesky factor.*

Proof. The optimization objective minimizes the output reconstruction error defined by the Hessian $\mathbf{H}$:

$$\mathcal{L} = \mathrm{tr}\left((\mathbf{W} - \widehat{\mathbf{W}})\mathbf{H}(\mathbf{W} - \widehat{\mathbf{W}})^\top\right). \tag{20}$$

Substituting the Cholesky factorization of the inverse Hessian $\mathbf{H}^{-1} = \mathbf{U}^\top\mathbf{U}$ (i.e., $\mathbf{H} = \mathbf{U}^{-1}\mathbf{U}^{-\top}$), we rewrite the objective using the definition of the Frobenius norm:

$$\mathcal{L} = \left\|(\mathbf{W} - \widehat{\mathbf{W}})\mathbf{U}^{-1}\right\|_F^2 = \left\|\mathbf{U}^{-\top}(\mathbf{W} - \widehat{\mathbf{W}})^\top\right\|_F^2. \tag{21}$$

This reveals that the error measures the magnitude of the weight residual projected onto the geometry defined by the inverse Cholesky factor. When determining the coefficients for a column group $\mathbf{W}_{:,s:(s+g)}$, we consider the corresponding local block $\mathbf{U}_{\text{loc}} = \mathbf{U}_{s:(s+g),s:(s+g)} \in \mathbb{R}^{g\times g}$. Accordingly, the local contribution to the error is:

$$\mathcal{L}_{\text{loc}} = \left\|\mathbf{U}_{\text{loc}}^{-\top}(\mathbf{W}_{:,s:(s+g)} - \widehat{\mathbf{W}}_{:,s:(s+g)})^\top\right\|_F^2. \tag{22}$$

Since the Frobenius norm decomposes row-wise, we minimize the error for each row r independently. In BPDQ, the quantized row vector is parameterized as $\widehat{\mathbf{W}}_{r,s:(s+g)}^\top = \mathbf{B}_r c_r$, where $c_r \in \mathbb{R}^{k+1}$ is the coefficient vector. Substituting this parameterization and the original weight vector $\mathbf{W}_{r,s:(s+g)}^\top$ into the row-wise objective yields:

$$\underset{c_r\in\mathbb{R}^{k+1}}{\mathrm{argmin}} \left\|\mathbf{U}_{\text{loc}}^{-\top}(\mathbf{B}_r c_r - \mathbf{W}_{r,s:(s+g)}^\top)\right\|_2^2. \tag{23}$$

This matches exactly the weighted least-squares problem defined in Eq. (6). Thus, solving Eq. (6) minimizes the optimization objective over the group coefficients within the Hessian-induced geometry. In implementation, a damping factor $\alpha = 10^{-4}$ is applied to the diagonal for numerical stability (omitted in the derivation for brevity). □

B.2. Consistency of Bit-Plane Update

Proposition. *During the bit-plane update (with fixed coefficients), the element-wise enumeration within column l minimizes the column-wise error term of the optimization objective in Eq. (2). Under the error propagation mechanism, this element-wise Euclidean projection ensures consistency with the Hessian-induced geometry.*

Proof. As derived in Eq. (21), the objective is equivalent to minimizing the projected residual norm $\mathcal{L} = \|(\mathbf{W} - \widehat{\mathbf{W}})\mathbf{U}^{-1}\|_F^2$. The error propagation mechanism in Eq. (4) establishes the relationship $(\mathbf{W} - \widehat{\mathbf{W}}) = \mathbf{E}\mathbf{U}$, where $\mathbf{E}$ is the matrix collecting the error coordinates $\mathbf{E}_{:,l}$ defined in Eq. (3). Substituting this relationship into the objective $\mathcal{L}$:

$$\mathcal{L} = \left\|(\mathbf{E}\mathbf{U})\mathbf{U}^{-1}\right\|_F^2 = \left\|\mathbf{E}\right\|_F^2 = \sum_{l=1}^{d_{\text{in}}} \|\mathbf{E}_{:,l}\|_2^2. \tag{24}$$

For the l-th column, the objective reduces to minimizing the error term $\|\mathbf{E}_{:,l}\|_2^2$ conditioned on the current propagation state. Based on the definition in Eq. (3), the error coordinate for the current working column $\mathbf{W}'_{:,l}$ is:

$$\mathbf{E}_{:,l} = \frac{\mathbf{W}'_{:,l} - \widehat{\mathbf{W}}_{:,l}}{\mathbf{U}_{l,l}}. \tag{25}$$

Assuming $\mathbf{H} \succ 0$, the diagonal element $\mathbf{U}_{l,l}$ is a strictly positive scalar constant independent of the quantization choice $\widehat{\mathbf{W}}_{:,l}$. Therefore, minimizing the column-wise error coordinate $\|\mathbf{E}_{:,l}\|_2^2$ is strictly equivalent to minimizing the Euclidean distance in the weight space:

$$\widehat{\mathbf{W}}_{:,l} = \underset{\widetilde{\mathbf{W}}_{:,l}}{\mathrm{argmin}} \left\|\frac{\mathbf{W}'_{:,l} - \widetilde{\mathbf{W}}_{:,l}}{\mathbf{U}_{l,l}}\right\|_2^2 = \underset{\widetilde{\mathbf{W}}_{:,l}}{\mathrm{argmin}} \|\mathbf{W}'_{:,l} - \widetilde{\mathbf{W}}_{:,l}\|_2^2. \tag{26}$$

Since the group-wise coefficients are fixed, this problem decouples into d_{out} independent scalar nearest-neighbor projections. For each coordinate r of the column, the value $\widetilde{\mathbf{W}}_{r,l}$ must be selected from the set of values $v_r(\mathbf{b})$ generated by the bit vectors $\mathbf{b} \in \{0,1\}^k$ following Eq. (7). Thus, the problem simplifies to finding the optimal bit vector $\mathbf{b}^*$ for each coordinate:

$$\mathbf{b}^* = \underset{\mathbf{b}\in\{0,1\}^k}{\mathrm{argmin}} (\mathbf{W}'_{r,l} - v_r(\mathbf{b}))^2. \tag{27}$$

This matches Eq. (8), confirming that the element-wise Euclidean nearest-neighbor projection minimizes the column-wise error, and maintains consistency with the Hessian-induced geometry. □

B.3. Consistency of Delta Correction

Proposition. *The delta correction in Eq. (9) preserves the error-propagation consistency following coefficient refitting.*

Proof. We partition the global index space into the current group columns $\{s : (s+g)\}$ and the tail columns $\{(s+g) : d_{\text{in}}\}$. Recall the propagation invariant $(\mathbf{W} - \widetilde{\mathbf{W}}) = \mathbf{E}\mathbf{U}$ from Appendix B.2 and let $\mathbf{U}_{\text{loc}} = \mathbf{U}_{s:(s+g),s:(s+g)} \in \mathbb{R}^{g\times g}$.

To preserve the *local consistency* within the group columns, we utilize the upper triangular structure to decompose the residual:

$$\mathbf{W}_{:,s:(s+g)} - \widehat{\mathbf{W}}_{:,s:(s+g)} = (\mathbf{E}\mathbf{U})_{:,s:(s+g)} = \mathbf{E}_{:,:s}\mathbf{U}_{:s,s:(s+g)} + \mathbf{E}_{:,s:(s+g)}\mathbf{U}_{\text{loc}}. \tag{28}$$

By rearranging Eq. (28), we isolate the components that remain constant during coefficient refitting (i.e., original weights and historical errors):

$$\underbrace{\mathbf{W}_{:,s:(s+g)} - \mathbf{E}_{:,:s}\mathbf{U}_{:s,s:(s+g)}}_{\text{Constant}} = \widehat{\mathbf{W}}_{:,s:(s+g)} + \mathbf{E}_{:,s:(s+g)}\mathbf{U}_{\text{loc}}. \tag{29}$$

Since the left side of Eq. (29) is constant, the right side must be equal for the bit-plane update state (old) and the refitted state (new). Subtracting the expression for the new state from the old state:

$$(\mathbf{E}^{\text{new}}_{:,s:(s+g)} - \mathbf{E}^{\text{old}}_{:,s:(s+g)})\mathbf{U}_{\text{loc}} = \widehat{\mathbf{W}}_{\text{old}} - \widehat{\mathbf{W}}_{\text{new}}. \tag{30}$$

This strictly derives the delta correction $\Delta\mathbf{E}\,\mathbf{U}_{\text{loc}} = \widehat{\mathbf{W}}_{\text{old}} - \widehat{\mathbf{W}}_{\text{new}}$ in Eq. (9).

To preserve the *tail consistency*, we expand the tail working weights $\mathbf{W}'_{:,(s+g):}$ based on the error propagation in Eq. (4) and the decomposition analogous to Eq. (28):

$$\mathbf{W}'_{:,(s+g):} = \underbrace{\mathbf{W}_{:,(s+g):} - \sum_{j<s} \mathbf{E}_{:,j}\mathbf{U}_{j,(s+g):}}_{\text{History (Constant)}} - \underbrace{\mathbf{E}_{:,s:(s+g)}\mathbf{U}_{s:(s+g),(s+g):}}_{\text{Current Group (Varying)}}. \tag{31}$$

The first term accounts for the original weights and errors from preceding groups ($j < s$), which remain constant during the current group's coefficient refitting. The change in the working weights is derived by differencing Eq. (31) between the new and old states:

$$\mathbf{W}'^{\text{new}}_{:,(s+g):} - \mathbf{W}'^{\text{old}}_{:,(s+g):} = -(\mathbf{E}^{\text{new}}_{:,s:(s+g)} - \mathbf{E}^{\text{old}}_{:,s:(s+g)})\mathbf{U}_{s:(s+g),(s+g):} = -\Delta\mathbf{E}\,\mathbf{U}_{s:(s+g),(s+g):}. \tag{32}$$

Thus, applying the update $\Delta\mathbf{E}$ exactly synchronizes the propagation state for the tail columns. □

bpdq

THANKS!

ICML
International Conference
On Machine Learning

Code: *github.com/KingdalfGoodman/BPDQ*
Model: *huggingface.co/goodman20241017/Qwen2.5-72B-BPD2-G256*
Email: *KingdalfGoodman@foxmail.com*

Junyu CHEN
30/05/2026