IDSCORE: TASK-CONDITIONED LIKELIHOOD CORRECTION FOR LOW-BIT QAT

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization-aware training (QAT) is a key approach for optimizing quantized neural networks. However, deterministic quantization introduces piecewise-constant mappings that make direct gradient-based optimization difficult, so QAT relies on surrogate gradients whose quality is critical to effective training. We propose Identification-Score (IDScore), a correction that improves surrogate gradient estimation in QAT with little additional training cost. This improvement comes from a new perspective: motivated by quantized system identification, IDScore exploits additional structure information from the quantizer itself, beyond the external signals used by existing methods. Concretely, IDScore constructs a local likelihood score from a task-selected neighboring target, the latent parameter, and quantization thresholds, and uses it to form a bounded residual correction. Our theory explains why IDScore provides a meaningful correction direction and how its correction strength should be allocated across parameters. Formally, we prove calibrated target-direction consistency and constrained-optimal likelihood allocation, and connect the resulting correction to threshold progress and a local sufficient condition for task-loss descent. Experiments on vision and language tasks show that IDScore consistently improves the tested LSQ and QuEST baselines at low additional cost. Evaluations span W1A1-W4A4 precision and models with up to 430M non-embedding parameters. It also yields further gains when combined with FOGZO or CAGE. Controlled ablations support the likelihood-based allocation mechanism characterized by our theory. Together, these results provide both an efficient approach with practical potential for low-bit training and a broader perspective on how quantization structure can inform backward optimization.

1 INTRODUCTION

As neural networks continue to scale, low-bit training has become increasingly important for reducing their computational and memory costs (Jacob et al., 2018; Esser et al., 2020; Panferov et al., 2025). In low-bit regimes, QAT provides an effective approach for preserving model performance by allowing latent full-precision parameters to adapt to quantization during training (Esser et al., 2020; Panferov et al., 2025). Its main challenge is not the forward quantization itself, but how to propagate useful optimization signals through a non-differentiable quantizer. Deterministic quantizers are piecewise constant, with zero derivatives almost everywhere (Bengio et al., 2013). Consequently, QAT relies on surrogate backward signals to update latent full-precision parameters, making the quality of gradient estimation critical to both optimization and final model performance.

This motivates the use of surrogate gradient estimators, among which the Straight-Through Estimator (STE) remains one of the most widely adopted approaches (Bengio et al., 2013; Yin et al., 2019). The STE enables efficient backpropagation through non-differentiable quantizers by replacing their derivatives with simple surrogates. Despite its empirical effectiveness, the STE is biased and can provide inaccurate update directions under aggressive quantization. Existing methods therefore introduce additional backward information: EWGS uses quantization-error-dependent gradient scaling (Lee et al., 2021), FOGZO supplements the STE with zeroth-order loss information to mitigate erroneous directions (Yang & Aamodt, 2026), CAGE introduces curvature-aware corrections (Tabesh et al., 2026), and SURGE employs an auxiliary full-precision branch to adapt surrogate gradients (Huang et al., 2026). These results suggest that information beyond the conventional STE can

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.

improve low-bit training, but typically require quantization-error heuristics, extra loss evaluations, curvature information, or auxiliary model paths.

These observations suggest a complementary direction for improving surrogate radients: exploiting structured information encoded by the quantizer itself. Its ordered levels and known thresholds define the geometry of possible discrete transitions (Nagel et al., 2022; Liu et al., 2023). This information complements the task gradient: the latter selects a locally preferred neighboring quantized state, while the latent coordinate and quantizer thresholds characterize the parameter's position relative to the corresponding transition. This leads to a natural question: *can these two sources of information be combined into a useful backward signal without additional model evaluations or changes to the forward quantizer?*

We answer this question with *Identification-Score (IDScore)*, a lightweight task-conditioned likelihood correction for low-bit QAT. The main contributions are threefold:

(i) We introduce IDScore, a lightweight correction that treats quantization structure as an additional source of information to improve surrogate gradient estimation. Its key construction, motivated by quantized system identification, is a local likelihood score built from a task-selected target, the latent parameter, and quantization thresholds, which is then used to form a bounded residual correction. IDScore readily integrates with existing QAT backward rules without requiring extra model evaluations, auxiliary full-precision branches, or additional optimizer state.

(ii) We develop a theoretical framework that explains why IDScore provides a task-consistent correction direction and how its magnitude should be allocated across parameters. Specifically, we establish calibrated target-direction consistency and constrained-optimal residual allocation for first-order likelihood decrease under a bounded budget, with the likelihood potential controlling target mismatch. We further characterize progress toward selected quantization thresholds and derive a local sufficient condition for task-loss descent under direct sign-preserving updates.

(iii) Experiments show that IDScore consistently improves low-bit QAT across vision and language, with evaluations spanning both LSQ and QuEST. They further cover W1A1–W4A4 precision, model scales up to 430M parameters, and different optimizer routes. On QuEST-30M W2A2, IDScore adds only about 2% step time and 0.4% peak allocated memory. IDScore further improves FOGZO and CAGE when combined with them. Controlled ablation studies support the role of likelihood-based residual allocation beyond target direction, residual norm, and simple boundary-distance weighting.

Together, these contributions provide a practical and low-overhead approach to low-bit QAT, while supporting a broader view of quantization as not only an obstacle to differentiation, but also a structured source of information for backward optimization.

2 IDSCORE: FROM QUANTIZATION LIKELIHOOD TO GRADIENT CORRECTION

We now formalize how the task gradient and quantizer thresholds are combined into the IDScore correction. The construction consists of three steps: selecting a neighboring target state, converting its relation to the quantizer thresholds into a likelihood score, and adding the normalized score as a bounded residual to an existing QAT backward signal form the IDScore-corrected update.

2.1 TASK-SELECTED TARGETS AND THRESHOLD INFORMATION

Consider a K-level scalar quantizer $Q(\cdot)$ with ordered levels $v_0 < \cdots < v_{K-1}$ and thresholds $c_1 < \cdots < c_{K-1}$, defined by

$$Q(u) = v_{\kappa(u)}, \qquad \kappa(u) = \sum\nolimits_{j=1}^{K-1} \mathbf{1}\{u > c_j\}, \tag{1}$$

where $c_0 = -\infty$ and $c_K = +\infty$. Equivalently, $Q(u) = v_k$ for $c_k < u \le c_{k+1}$.

For coordinate i, let u_i be the continuous quantizer input, with $k_i = \kappa(u_i)$ and $q_i = Q(u_i) = v_{k_i}$. Let $\mathbf{q} = (q_i)_i$ denote the quantized vector, and write the task loss as $L(\mathbf{q})$, suppressing other network variables for brevity. The task gradient with respect to the ith quantized coordinate is defined as

$$g_i := [\nabla_{\mathbf{q}} L(\mathbf{q})]_i = \frac{\partial L}{\partial q_i}. \tag{2}$$

Here g_i is the upstream gradient at the quantizer output, before applying any surrogate derivative through Q. Since the local descent direction is $-g_i$, $g_i > 0$ favors a smaller quantized value, while $g_i < 0$ favors a larger one. We therefore select the feasible neighboring target state as

$$k_i^{\mathrm{tar}} = \mathrm{clip}\big(k_i - \mathrm{sign}(g_i), 0, K-1\big), \qquad q_i^{\mathrm{tar}} = v_{k_i^{\mathrm{tar}}}. \tag{3}$$

The task gradient thus selects the target state. To additionally encode the latent coordinate's position relative to the quantization thresholds, we introduce the local working model $y_i = u_i + e_i, e_i \sim \mathcal{N}(0, \tau^2)$. This working model provides a smooth probabilistic representation of the deterministic threshold geometry, with closed-form threshold probabilities and gradients. It does not alter the actual training dynamics or inject noise into the forward pass, while τ controls the smoothness of the local likelihood. Together, the task-selected target, latent coordinate, and known thresholds are used to construct IDScore in the next section.

2.2 Likelihood Score and Gradient Correction

We now convert the task-selected target and the threshold-relative position of the continuous coordinate into a likelihood score for backward correction. A target quantization level k can be equivalently represented by its position relative to the $K-1$ thresholds. Specifically, define

$$s_j^{(k)} = \mathbf{1}\{k \geq j\}, \qquad p_j(u;\tau) = \Pr(Y > c_j \mid u) = \Phi\left(\frac{u - c_j}{\tau}\right), \quad j = 1, \ldots, K-1, \tag{4}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. Here, $s_j^{(k)} = 1$ indicates that target level k lies above threshold c_j, while $p_j(u;\tau)$ gives the corresponding probability under the local working model.

We define the composite negative log likelihood associated with target level k as

$$\ell_k^{\mathrm{ord}}(u;\tau) = \frac{1}{K-1} \sum_{j=1}^{K-1} \left[-s_j^{(k)} \log p_j(u;\tau) - (1 - s_j^{(k)}) \log\big(1 - p_j(u;\tau)\big) \right]. \tag{5}$$

Thus, $\ell_k^{\mathrm{ord}}(u;\tau)$ measures how well u matches the ordinal threshold pattern of target level k, capturing its relation to all $K-1$ thresholds rather than to a single boundary.

The raw IDScore is the gradient of the target-conditioned likelihood loss at the current coordinate:

$$\psi_i(u;\tau) = \left. \frac{\partial \ell_{k_i^{\mathrm{tar}}}^{\mathrm{ord}}(u;\tau)}{\partial u} \right|_{u=u_i} = \sum_{j=1}^{K-1} \frac{\phi(z_{ij})}{\tau(K-1)} \frac{p_j(u_i;\tau) - s_j^{(k_i^{\mathrm{tar}})}}{p_j(u_i;\tau)\big(1 - p_j(u_i;\tau)\big)}, \quad z_{ij} = \frac{u_i - c_j}{\tau}, \tag{6}$$

where $\phi(\cdot)$ is the standard normal density. Thus, $\phi(\cdot)$ converts the current coordinate u_i, the task-selected target k_i^{tar}, and the quantizer thresholds $\{c_j\}$ into a likelihood-based correction signal.

To combine the likelihood score with an existing QAT backward rule, we normalize and bound its magnitude within each quantized tensor. A fixed route-specific adapter specifies a normalization set $\mathcal{J}$ for estimating the typical score magnitude and a residual-support mask $m_i \in \{0, 1\}$ indicating where the IDScore correction is applied. For nonempty $\mathcal{J}$, define

$$\mu_{\psi,\mathcal{J}} = \mathrm{stopgrad}\left(\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} |\psi_j| \right), \qquad \bar{\psi}_i = \mathrm{sign}(\psi_i)\, \mathrm{clip}\left(\frac{|\psi_i|}{\mu_{\psi,\mathcal{J}} + \epsilon}, c_{\min}, c_{\max} \right), \tag{7}$$

where $\epsilon > 0$ is a small constant for numerical stability, where $c_{\min}$ and $c_{\max}$ bound the normalized score magnitude and therefore control the range of the coordinate-wise residual strength.

Let $g_{\mathrm{base},i}$ denote the backward signal produced by the chosen base QAT rule for updating the latent parameter. IDScore adds the normalized likelihood score as a residual correction:

$$g_{\mathrm{ID},i} = g_{\mathrm{base},i} + \lambda m_i |g_i| \bar{\psi}_i. \tag{8}$$

Here g_i selects and scales the task-conditioned correction, while $g_{\mathrm{base},i}$ remains the primary backward signal. If $\mathcal{J}$ is empty, the residual is set to zero. Otherwise, $|g_{\mathrm{ID},i} - g_{\mathrm{base},i}| \leq \lambda c_{\max} |g_i|$ and setting $\lambda = 0$ exactly recovers the base backward rule. The likelihood construction is shared across QAT routes, while the fixed adapter $(\mathcal{J}, m_i)$ follows the native quantization coordinate and backward support. The complete IDScore training step is given in Algorithm 1.

Algorithm 1 One IDScore Training Step

Require: Quantizer Q, base QAT backward rule, route adapter, optimizer, and τ, λ, $c_{\min}$, $c_{\max}$

1: Run the quantized forward and backward passes and obtain $\mathbf{g} = \nabla_{\mathbf{q}} L$, together with $\mathbf{g}_{\text{base}}$
2: **for** each quantized tensor **do**
3: Select k_i^{tar} according to Eq. (3)
4: Compute the target-conditioned likelihood score ψ_i from Eqs. (4)–(6)
5: Obtain $(\mathcal{J}, \mathbf{m})$ from the route adapter
6: Normalize ψ_i according to Eq. (7)
7: Compute $g_{\text{ID},i}$ via Eq. (8)
8: **end for**
9: Update the latent parameters using $\mathbf{g}_{\text{ID}}$

We use $\tau = 1$, $\lambda = 0.005$, $c_{\min} = 0$, and $c_{\max} = 3$ as the default configuration. For the larger $200M$ and $430M$ settings, we set $\tau = 0.5$ for improved training stability while keeping all other hyperparameters unchanged. IDScore adds only a likelihood-based residual to the base QAT backward signal without changing the forward quantizer, requiring only elementwise operations and no extra model query or optimizer state. Framework-specific realizations are given in Appendix B.

3 Theoretical Analysis

We analyze three local properties of IDScore corresponding to the main steps of the method. First, does the likelihood score produce a descent direction toward the task-selected quantized state? Second, given this direction, does its magnitude provide a principled allocation of a bounded correction budget across coordinates? Third, does the resulting latent update move parameters toward the selected quantization transitions, and when does the induced discrete transition decrease the task loss?

We distinguish the continuous quantization input $\mathbf{u}$, the quantized output $\mathbf{q} = Q(\mathbf{u})$, and the task gradient $\mathbf{g} = \nabla_{\mathbf{q}} L(\mathbf{q})$. Since $F(\mathbf{u}) = L(Q(\mathbf{u}))$ is piecewise constant, we do not treat IDScore as gradient descent on a smooth latent objective; instead, our analysis focuses on its local likelihood geometry and induced quantized transitions.

3.1 Likelihood Identification and Target Direction

Let $Y = u_0 + E, E \sim \mathcal{N}(0, \tau^2)$ and define the random threshold indicators $S_j = \mathbf{1}\{Y > c_j\}, j = 1, \ldots, K-1$. Let $\mathbf{S} = (S_1, \ldots, S_{K-1})$, and denote by $\psi(x; \mathbf{S})$ the derivative, with respect to x, of the ordinal negative log likelihood induced by $\mathbf{S}$.

Theorem 3.1 (Likelihood identification and calibrated target direction)**.** *Define the population score*

$$\Psi(x; u_0) := \mathbb{E}[\psi(x; \mathbf{S}) \mid u_0].$$

Then, $(x - u_0)\Psi(x; u_0) > 0$ *for* $x \neq u_0$*, so* u_0 *is the unique zero of* $\Psi(\cdot; u_0)$*.*

Moreover, each target-conditioned loss $\ell_k^{\text{ord}}(\cdot; \tau)$ *is strictly convex. Under the calibration condition*

$$\psi_k(c_k; \tau) < 0 < \psi_k(c_{k+1}; \tau), \tag{9}$$

for every internal target cell, the score satisfies, away from threshold ties,

$$\operatorname{sign}\big(\psi_{k_i^{\text{tar}}}(u_i; \tau)\big) = \operatorname{sign}(g_i), \quad i \in \mathcal{S},$$

where $\mathcal{S} := \{i : m_i = 1,\ |g_i| > 0\}$*. For the two boundary target levels, the same sign relation holds without condition (9).*

Theorem 3.1 shows that the population score identifies the latent location, while the calibrated target-conditioned score provides a task-consistent descent direction. The calibration condition (9) has a direct geometric interpretation: the target-conditioned likelihood is aligned with the quantizer geometry, so likelihood descent points toward the interior of the selected target cell. Since $\ell_k^{\text{ord}}(\cdot; \tau)$ is strictly convex, it is equivalent to requiring its unique minimizer $u_k^\star(\tau)$ to lie inside the target cell (c_k, c_{k+1}). Appendix A.2 shows that sufficiently small τ is a general sufficient condition for calibration and that the admissible range extends beyond $\tau = 1$ for all quantizer geometries used in our experiments. Hence, the deployed values lie in the calibrated regime (see Table 7).

3.2 LIKELIHOOD-BASED BUDGET ALLOCATION

We next study how the likelihood magnitude allocates the IDScore correction. At a frozen training state, with the task gradients, selected targets, support set, and normalization scale fixed, consider

$$\mathcal{C}_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} |g_i| \ell^{\mathrm{ord}}_{k_i^{\mathrm{tar}}}(x_i), \qquad \mathcal{D}_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} |g_i||\kappa(x_i) - k_i^{\mathrm{tar}}|,$$

to measure the task-weighted likelihood mismatch and the corresponding discrete level mismatch.

Theorem 3.2 (Likelihood mismatch and budget allocation)**.** *For a K-level quantizer,*

$$\mathcal{D}_{\mathcal{S}}(\boldsymbol{x}) \le \frac{K-1}{\log 2} \mathcal{C}_{\mathcal{S}}(\boldsymbol{x}).$$

Let $\boldsymbol{z} = \nabla \mathcal{C}_{\mathcal{S}}(\boldsymbol{u})$, where equivalently $z_i = m_i |g_i| \psi_i$, and define the coordinate-wise residual cap $U_i = \lambda c_{\max} m_i |g_i|$. For $c_{\min} = 0$ and $\lambda > 0$, the IDScore residual $\boldsymbol{r} = (r_i)_i$ can be written as $r_i = \mathrm{clip}\big(\frac{\lambda}{\nu} z_i, -U_i, U_i\big)$, where $\nu = \mu_{\psi,\mathcal{J}} + \epsilon$. Let $R = \|\boldsymbol{r}\|_2$ and $\mathcal{B} = \{\boldsymbol{d} : |d_i| \le U_i\}$. Then,

$$\boldsymbol{r} = \underset{\boldsymbol{d}\in\mathcal{B},\ \|\boldsymbol{d}\|_2 \le R}{\arg\max} \langle \boldsymbol{z}, \boldsymbol{d} \rangle, \tag{10}$$

and the maximizer is unique.

The first inequality links the likelihood potential to the discrete mismatch. The second establishes constrained first-order optimality for likelihood decrease: among all corrections subject to the same coordinate-wise caps and no larger ℓ_2 norm, IDScore uniquely maximizes first-order decrease of $C_{\mathcal{S}}$.

3.3 TRANSITION PROGRESS AND TASK COMPATIBILITY

We finally examine whether the IDScore-corrected update actually moves latent coordinates toward the selected quantization transitions. Split the residual support into the movable set $\mathcal{M} = \{i \in \mathcal{S} : k_i^{\mathrm{tar}} \neq k_i\}$ and the saturated set $\mathcal{H} = \mathcal{S} \setminus \mathcal{M}$. The former contains coordinates with a feasible neighboring transition, while the latter contains those whose clipped target remains at the current extreme level. For $i \in \mathcal{M}$, let θ_i denote the threshold separating the current and target cells, and define $d_i(x) = [\mathrm{sign}(g_i)(x - \theta_i)]_+$. For a coordinate still inside its current cell, $d_i(x)$ is its remaining distance to the selected transition threshold. Then, we isolate the IDScore effect through the following direct sign-preserving update, rather than the full stateful dynamics of a specific optimizer.

Theorem 3.3 (Transition progress and task-compatible quantized updates)**.** *Assume the sign relation of Theorem 3.1 and consider $u_i^+ = u_i - \eta(\chi_i g_i + r_i)$, with the step size $\eta > 0$ and the base-gradient term $\chi_i \ge 0$. For $c_{\min} = 0$, define $\zeta_i := \min\{|\psi_i|/\nu, c_{\max}\}$. Then every $i \in \mathcal{M}$ satisfies*

$$d_i(u_i^+) = [d_i(u_i) - \eta |g_i| (\chi_i + \lambda \zeta_i)]_+ , \tag{11}$$

while every $i \in \mathcal{H}$ remains in its selected extreme cell.

Let $\mathcal{T}_{\mathrm{base}} := \{i \in \mathcal{M} : d_i(u_i) \le \eta \chi_i |g_i|\}$ denote coordinates crossing selected threshold under the base update, and $\mathcal{T}_{\mathrm{ID}} := \{i \in \mathcal{M} : d_i(u_i) \le \eta |g_i| (\chi_i + \lambda \zeta_i)\}$ those crossing it with IDScore. Then

$$\mathcal{T}_{\mathrm{base}} \subseteq \mathcal{T}_{\mathrm{ID}}. \tag{12}$$

Let $\Delta q = Q(u^+) - Q(u)$. Since Q is nondecreasing and each nonzero latent update follows $-\mathrm{sign}(g_i)$, any induced quantized transition satisfies $\Delta q_i = -\mathrm{sign}(g_i)|\Delta q_i|$. Hence, $\langle g, \Delta q \rangle = -\Gamma_{\mathrm{tr}}$, where $\Gamma_{\mathrm{tr}} := \sum_i |g_i||\Delta q_i|$. Define the finite-transition remainder $\mathcal{R}_{\mathrm{tr}} := L(q+\Delta q) - L(q) - \langle g, \Delta q \rangle$. Then $L(q + \Delta q) - L(q) = -\Gamma_{\mathrm{tr}} + \mathcal{R}_{\mathrm{tr}}$, and the task loss decreases whenever $\mathcal{R}_{\mathrm{tr}} < \Gamma_{\mathrm{tr}}$.

Moreover, if L is β-smooth along the segment from q to $q + \Delta q$, then

$$L(q + \Delta q) - L(q) \le -\Gamma_{\mathrm{tr}} + \frac{\beta}{2} \|\Delta q\|_2^2.$$

Consequently, $\Gamma_{\mathrm{tr}} > \frac{\beta}{2} \|\Delta q\|_2^2$ is a sufficient condition for $L(q + \Delta q) < L(q)$.

Eq. (11) shows that IDScore adds $\eta\lambda|g_i|\zeta_i$ of threshold-directed progress beyond the base update. Accordingly, Eq. (12) shows that IDScore preserves existing one-step task-aligned transitions and may enable additional threshold crossings. The identity $\langle g, \Delta q \rangle = -\Gamma_{\mathrm{tr}}$ further establishes first-order alignment of every induced quantized transition with the task-descent direction. The condition $\mathcal{R}_{\mathrm{tr}} < \Gamma_{\mathrm{tr}}$ gives the exact compatibility criterion, while $\Gamma_{\mathrm{tr}} > \frac{\beta}{2} \|\Delta q\|_2^2$ is a local sufficient condition under smoothness. These results are local and apply to the direct sign-preserving update $u_i^+ = u_i - \eta(\chi_i g_i + r_i)$; stateful optimizers such as AdamW and Muon are evaluated empirically.

4 EXPERIMENTS

Our experiments evaluate whether IDScore (i) improves low-bit QAT across vision and language, (ii) transfers across scales and bit widths, (iii) is compatible with gradient estimators and optimizer routes at low overhead, and (iv) benefits from likelihood-based magnitude allocation.

Experimental setup. We evaluate IDScore in two QAT families. For cross-domain evaluation, we use 2-bit LSQ on ResNet-18 and DeiT-Tiny (He et al., 2016; Touvron et al., 2021) with ImageNet (Deng et al., 2009), and on LLaMA models (Touvron et al., 2023) with C4 (Raffel et al., 2020). For scale, bit-width, composability, efficiency, and mechanism studies, we use QuEST-style weight and activation quantization on C4. Within each comparison, we match the model, data, precision, training budget, and random seed whenever paired runs are available. We use $\lambda = 0.005$ throughout, with $\tau = 1$ by default and $\tau = 0.5$ for the 200M and 430M settings to improve large-scale training stability. Unless otherwise noted, results are averaged over three matched seeds; exceptions are explicitly marked in the corresponding tables. We report top-1 accuracy for vision and validation perplexity (PPL) for language. Higher vision accuracy and lower language PPL are better.

4.1 CROSS-DOMAIN GENERALITY

We first test IDScore under 2-bit LSQ across vision and language, spanning convolutional networks, vision transformers, language models, and varied dataset sizes and training budgets. IDScore improves over STE baseline in all eight settings and over FOGZO in seven, with DeiT-Tiny/ImageNet-100 as the only exception. The gains span convolutional networks, vision transformers, and language models. For the partially unpaired settings, we interpret the differences as supporting evidence rather than paired statistical comparisons. The fully local matched settings exhibit the same overall trend.

Table 1: Cross-domain results with 2-bit LSQ.

Model and data	LSQ-STE	LSQ-FOGZO	IDScore
Vision: top-1 accuracy (↑)			
ResNet-18 / ImageNet-100	80.300 ± 0.059	80.273 ± 0.307	$\mathbf{80.507} \pm 0.109$
ResNet-18 / ImageNet-1K[†]	66.425 ± 0.044	66.015 ± 0.279	$\mathbf{67.774} \pm 0.099$
DeiT-Tiny / ImageNet-100	70.840 ± 0.263	$\mathbf{72.580} \pm 0.351$	71.587 ± 0.377
DeiT-Tiny / ImageNet-1K[†]	63.173 ± 0.165	63.493 ± 0.160	$\mathbf{64.006} \pm 0.386$
Language: validation perplexity (↓)			
LLaMA-9M / 328M C4 tokens	101.981 ± 2.583	99.842 ± 2.242	$\mathbf{93.900} \pm 1.419$
LLaMA-9M / 3.28B C4 tokens	81.569 ± 0.871	79.140 ± 0.706	$\mathbf{72.836} \pm 0.256$
LLaMA-20M / 1.3B C4 tokens	58.902 ± 0.546	58.555 ± 0.259	$\mathbf{55.198} \pm 0.371$
LLaMA-20M / 13B C4 tokens[‡]	50.853 ± 0.336	50.614 ± 0.287	$\mathbf{48.018} \pm 0.201$

Unmarked rows use three matched local seeds (five for LLaMA-9M). [†]ImageNet-1K: IDScore uses three local seeds; ResNet-18 STE is fully local, while ResNet-18 FOGZO and DeiT-Tiny baselines use local seed 0 plus author logs for seeds 3–4. [‡]LLaMA-20M/13B: IDScore uses three local seeds, while baseline statistics use five author-provided seeds, with a local matched check at seed 3.

4.2 TRANSFER ACROSS SCALE AND PRECISION

We next examine whether IDScore transfers across model scales and quantization precisions while keeping the forward quantization and base training setup unchanged.

Table 2: Transfer of the IDScore correction across model scale and bit width on C4.

(a) Model and training scale at W2A2

Scale	QuEST	+FOGZO	+IDScore
30M / 3B	37.262	36.975	**36.502**
100M / 10B	26.302	26.130	**25.700**
200M / 20B	22.583	22.434	**21.644**
430M / 43B[†]	17.029	-	**15.985**

(b) Quantization precision on 30M / 3B

Precision	QuEST	+IDScore
W1A1	54.256	**49.752**
W2A2	37.262	**36.502**
W3A3	33.745	**33.294**
W4A4	30.975	**30.731**

[†]430M uses one matched seed; the other scale settings use three.

Table 2(a) shows that IDScore consistently reduces validation PPL over QuEST from 30M/3B to 430M/43B, and over QuEST+FOGZO through 200M/20B; FOGZO is not reported at 430M because its original evaluation extends only to 200M Yang & Aamodt (2026). Table 2(b) further shows consistent improvements over QuEST from W1A1 to W4A4, with the largest gain at the most aggressive W1A1 setting. Table 2 only reports means with standard deviations in Appendix C.

4.3 Composability, Portability and Efficiency

IDScore exploits information not explicitly used by FOGZO or CAGE (Yang & Aamodt, 2026; Tabesh et al., 2026). We therefore examine its composability with their zeroth-order and curvature-aware corrections, its optimizer portability, and its computational overhead.

Composability. Under matched QuEST settings, adding IDScore to FOGZO or CAGE improves the partner method across all three matched seeds (Table 3). IDScore further improves CAGE even when standalone CAGE is stronger, supporting the complementarity of threshold-derived information with zeroth-order and curvature-aware corrections.

Table 3: Composability of IDScore with FOGZO and CAGE under matched QuEST settings.

Partner and setting	Partner	IDScore	Combined	Paired gain
FOGZO b10 (30M / 3B)	36.975 ± 0.033	36.502 ± 0.043	$\mathbf{36.330 \pm 0.045}$	$+0.644 \pm 0.033$
CAGE (30M / 3B)	35.764 ± 0.069	36.502 ± 0.043	$\mathbf{35.586 \pm 0.047}$	$+0.178 \pm 0.032$
CAGE (50M / 5B)	30.605 ± 0.057	31.216 ± 0.067	$\mathbf{30.474 \pm 0.048}$	$+0.131 \pm 0.027$

Portability. IDScore is constructed from quantization information and does not use optimizer state. Table 4 shows that it improves QuEST under both AdamW (Loshchilov & Hutter, 2019) and a Muon+AdamW route (Liu et al., 2025) across all three matched seeds. These results support portability across the tested optimizer routes.

Table 4: Optimizer portability of IDScore on QuEST-30M W2A2.

Optimizer route	QuEST	QuEST+IDScore	Gain over QuEST
AdamW	37.262 ± 0.083	$\mathbf{36.502} \pm 0.043$	$+0.760$
Muon + AdamW (aux.)	39.830 ± 0.090	$\mathbf{39.508} \pm 0.069$	$+0.322$

Efficiency. With only elementwise operations and no extra model evaluation, IDScore is expected to incur little overhead over base QAT. Table 5 reports final quality together with a synchronized update-cost benchmark under QuEST-30M W2A2. IDScore adds only 2.0% mean step time and only 0.4% peak memory over QuEST, while improving PPL from 37.262 to 36.502. In contrast, FOGZO increases mean step time by about 72%. CAGE achieves lower standalone PPL at a higher measured update cost, and its combination with IDScore further improves quality. Overall, IDScore remains close to QuEST in training cost while complementing other gradient estimators.

Table 5: Final quality and synchronized update cost on QuEST-30M W2A2. Lower PPL, step time, and memory are better; higher throughput is better.

Method	Final PPL ↓	Mean step (ms) ↓	K tok./s ↑	Peak memory (GiB) ↓
QuEST	37.262 ± 0.083	71.16 ± 1.65	459.5 ± 10.4	5.556
QuEST+FOGZO	36.975 ± 0.033	122.47 ± 0.33	267.1 ± 0.7	6.062
QuEST+IDScore	36.502 ± 0.043	72.52 ± 0.12	450.6 ± 0.8	5.577
$\text{CAGE}^{\dagger}_{\text{active}}$	35.764 ± 0.069	81.69 ± 0.33	400.2 ± 1.6	5.695

Efficiency is measured over five synchronized trials on the same two NVIDIA A800 GPUs, with 300 warmup and 1,000 timed updates per trial. Memory is peak allocated GiB per GPU. $^{\dagger}\text{CAGE}_{\text{active}}$ is our local active-path measurement, and should not be interpreted as official CAGE runtime or a full-schedule average.

4.4 Mechanism Tests and Sensitivity

Controlled ablations test whether (i) likelihood magnitude adds information beyond the task-selected direction, (ii) simpler boundary-distance or norm-matched non-likelihood allocations reproduce the gain, and (iii) residual support, gradient scaling, and frozen hyperparameters affect performance.

Likelihood magnitude beyond target direction. We compare several target-conditioned corrections. Uniform retains only the likelihood-score direction with a fixed correction strength and removes the coordinate-wise magnitude information. Boundary-NLL uses the binary probit likelihood of the selected transition boundary. For the boundary c_i^{tar}, define

$$p_i^{\mathrm{bdry}}(u_i) = \Phi\left(\frac{u_i - c_i^{\mathrm{tar}}}{\tau}\right), \quad \psi_i^{\mathrm{bdry}} = \frac{\partial}{\partial u_i}\left[-s_i^{\mathrm{tar}} \log p_i^{\mathrm{bdry}} - (1 - s_i^{\mathrm{tar}}) \log(1 - p_i^{\mathrm{bdry}})\right].$$

IDScore-Cell scores the selected target cell, whereas IDScore-Ordinal uses cumulative threshold events. Figure 1(a) shows that retaining the target direction alone gives little benefit, whereas both likelihood-based scores improve QuEST. This indicates that the gain is associated with likelihood-shaped magnitude information.

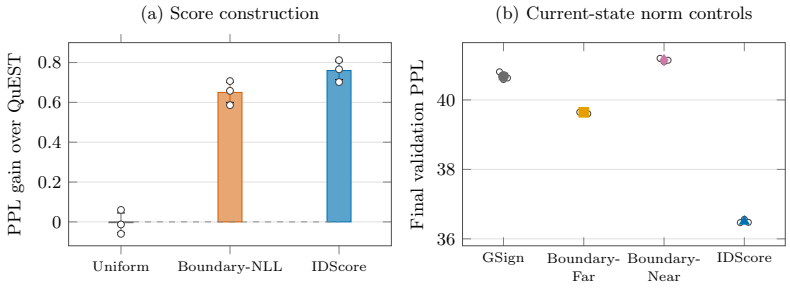


Figure 1: Mechanism tests on QuEST-30M W2A2 with 3B C4 tokens. (a) likelihood-magnitude effect across score construction. (b) final validation PPL for norm-matched non-likelihood controls.

Likelihood allocation versus norm-matched non-likelihood rules. We replace the likelihood magnitude with simpler coordinate-wise weights while matching the residual norm. At each training state, let $R_t = \left\|\mathbf{r}^{\mathrm{ID}}(\mathbf{u}_t)\right\|_2$ denote the counterfactual IDScore residual norm for that tensor. Each control uses

$$\mathbf{r}_t^{(w)} = R_t \frac{|\mathbf{g}_t| \odot \mathrm{sign}(\bar{\boldsymbol{\psi}}_t) \odot w(\boldsymbol{\delta}_t)}{\left\||\mathbf{g}_t| \odot \mathrm{sign}(\bar{\boldsymbol{\psi}}_t) \odot w(\boldsymbol{\delta}_t)\right\|_2},$$

where δ_i is the distance to the selected target boundary. GSign uses a uniform coordinate weight, $w(\delta) = 1$, while Boundary-Far and Boundary-Near favor farther or closer coordinates, respectively: $w_{\mathrm{far}}(\delta) = \frac{\delta+\epsilon}{1+\delta}(\epsilon = 10^{-6})$, $w_{\mathrm{near}}(\delta) = \frac{1}{1+\delta}$. The norm is matched along each control trajectory, so these rules preserve the residual budget while replacing IDScore's likelihood-based allocation. In particular, GSign is not the fixed-strength mask-only rule $\lambda\mathbf{g}$.

Figure 1(b) shows that IDScore outperforms GSign, Boundary-Far, and Boundary-Near on every matched seed, since these controls also match the pseudo-target, support, task-gradient magnitude, correction sign, and per-tensor residual norm, the gains support the likelihood-based allocation in Theorem 3.2, rather than a larger residual budget. Full results are reported in Appendix Table 12.

Residual support and gradient scaling. IDScore-trust restricts the residual to the native QuEST trust mask and yields a three-seed mean PPL 39.965, versus 36.502 for IDScore. EWGS-fixed, a fixed EWGS-style coefficient under the same QuEST setting, obtains 40.694. Thus, neither the trust-mask restriction nor fixed element-wise scaling reproduces the IDScore gain.

Because EWGS-fixed is only a controlled port, we further compare with native EWGS on ResNet-20/CIFAR-10 W1A1 (Lee et al., 2021). Over three matched QAT seeds, the mean final accuracies are 85.480% for STE, 85.360% for EWGS, and 85.577% for IDScore. The differences are modest, but IDScore remains competitive with native EWGS. Full results are in Appendix Tables 10 and 11.

Together, these controls support the roles of likelihood-based magnitude allocation and route-specific residual support beyond target direction, residual norm, simple boundary-distance weighting, or generic element-wise scaling.

Hyperparameter sensitivity. Figure 2 varies one hyperparameter at a time on QuEST-30M W2A2. IDScore remains effective across the tested ranges of τ and λ, with only moderate variation around the deployed settings, indicating limited sensitivity and supporting the use of fixed hyperparameters across experiments.

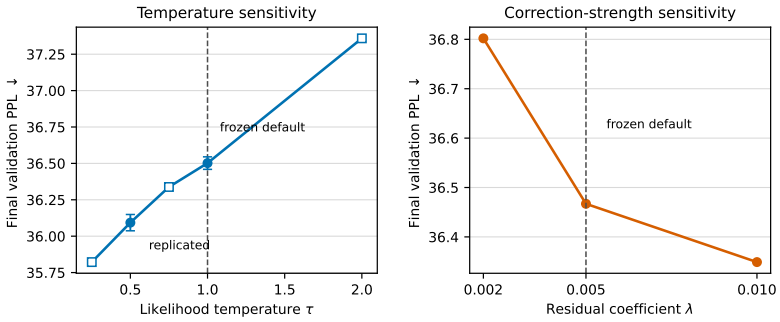


Figure 2: One-factor sensitivity to the likelihood temperature τ and residual coefficient

5 Related Work

Low-bit QAT and gradient estimation. Low-bit QAT relies on surrogate backward signals to optimize latent continuous parameters through non-differentiable quantizers (Courbariaux et al., 2015; Esser et al., 2020; Panferov et al., 2025). Existing methods differ mainly in the information used to improve these signals. STE variants use identity or smooth surrogate derivatives (Bengio et al., 2013; Yin et al., 2019; Qin et al., 2020); EWGS uses quantization-error-dependent scaling (Lee et al., 2021); FOGZO introduces zeroth-order information (Yang & Aamodt, 2026); CAGE uses curvature-aware correction (Tabesh et al., 2026); and SURGE relies on an auxiliary full-precision branch (Huang et al., 2026). In contrast, IDScore extracts information from the quantizer through its threshold structure and a task-selected target, without extra model evaluations or auxiliary paths.

Quantized identification and likelihood modeling. Quantized system identification and ordinal likelihood models recover latent information from thresholded observations (Ljung, 1999; McCullagh, 1980; Wang et al., 2024; Pedregosa et al., 2017). Inspired by this, IDScore treats the task-selected state as a pseudo-observation and constructs a likelihood score from the latent parameter and quantizer thresholds. Our contribution is not a new identification or ordinal model, but the task-conditioned use of quantizer-derived likelihood information for QAT backward correction.

6 Limitations and Conclusion

Limitations. Our evaluation remains below the billion-parameter regime, with models up to 430M non-embedding parameters, while the detailed efficiency and optimizer studies cover only a subset of settings. The largest QuEST experiment uses 430M non-embedding parameters and 43B C4 tokens, and runtime and memory are audited only on QuEST-30M W2A2 with two A800 GPUs. Our theory is local, characterizing target direction, constrained residual allocation, and task-compatible transitions under direct sign-preserving updates, rather than global QAT convergence, general optimizer dynamics, or optimality beyond the stated comparison class.

Conclusion. We introduced IDScore, a task-conditioned likelihood correction for low-bit QAT inspired by quantized system identification. IDScore exploits structure information encoded by the quantizer to improve backward signals while preserving the forward quantizer and incurring little additional training cost. Theoretical analysis and extensive experiments demonstrate its effectiveness and efficiency. We hope this work provides a deeper understanding of how quantization structure can inform backward optimization and inspires more effective methods for low-bit model training.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, research brainstorming and feedback on theoretical arguments, the drafting and debugging of experimental, analysis, and visualization code, and figure/table rendering and format conversion. All AI-assisted text, code, visual materials, and theoretical suggestions were reviewed by the authors, who independently ran the experiments and verified the mathematical arguments, citations, and claims. Generative AI was not used to generate experimental results or measurements. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies a training method and does not collect new human-subject or private data. Experiments use established vision and language benchmarks under their existing access conditions. Quantized models can retain biases and other risks present in their training data and downstream use. Our method does not remove those risks, although lower-cost training may reduce resource use.

REPRODUCIBILITY STATEMENT

The supplementary material provides a compact PyTorch reference implementation of Algorithm 1, including the likelihood score and bounded residual correction, together with the calibration check for the condition in Theorem 3.1. Section 2 defines the complete method, while the appendices provide the proofs, framework-specific adapter definitions, exact hyperparameter settings for each configuration, experimental protocols, statistical conventions, baseline provenance, and additional per-seed and audit results. The reference implementation is intended to illustrate the mapping from the stated algorithm to executable code rather than reproduce the full training infrastructure.

REFERENCES

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Haoyu Huang, Boyu Liu, Linlin Yang, Yanjing Li, Yuguang Yang, Xuhui Liu, Canyu Chen, Zhongqian Fu, and Baochang Zhang. SURGE: Surrogate gradient adaptation in binary neural networks. *arXiv preprint arXiv:2605.10989*, 2026.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Junghyup Lee, Dohyung Kim, and Bumsub Ham. Network quantization with element-wise gradient scaling. *arXiv preprint arXiv:2104.00903*, 2021.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Shih-Yang Liu, Zechun Liu, and Kwang-Ting Cheng. Oscillation-free quantization for low-bit vision transformers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21813–21824, 2023.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ, 2 edition, 1999.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. In *International Conference on Machine Learning*, pp. 16318–16330, 2022.

Andrei Panferov, Jiale Chen, Soroush Tabesh, Mahdi Nikdan, and Dan Alistarh. QuEST: Stable training of LLMs with 1-bit weights and activations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 47820–47836, 2025.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(55):1–35, 2017.

Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2247–2256, 2020.

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.

Soroush Tabesh, Mher Safaryan, Andrei Panferov, Alexandra Volkova, and Dan Alistarh. Cage: Curvature-aware gradient estimation for accurate quantization-aware training. *Proceedings of Machine Learning and Systems*, 8:626–638, 2026.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ying Wang, Yanlong Zhao, and Ji-Feng Zhang. Asymptotically efficient quasi-newton type identification with quantized observations under bounded persistent excitations. *Automatica*, 166: 111722, 2024.

Ningfeng Yang and Tor Aamodt. Improving the straight-through estimator with zeroth-order information. *Advances in Neural Information Processing Systems*, 38:167408–167440, 2026.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

A PROOFS FOR THE IDSCORE THEORY

This appendix provides the proofs and supporting analysis for the theoretical results in Section 3. Appendix A.1 proves Theorem 3.1, establishing likelihood identification, strict convexity, and calibrated target-direction consistency. Appendix A.2 analyzes the calibration condition, gives its geometric interpretation, establishes a general sufficient temperature regime, and verifies the calibrated ranges and implementation directions used in our experiments. Appendices A.3 and A.4 prove Theorems 3.2 and 3.3, respectively, covering likelihood-based residual allocation, threshold-transition progress, and local task compatibility. Appendix A.5 summarizes the scope and limitations of these theoretical claims. Throughout this appendix, pseudo-targets, output gradients, route-adapter sets, and normalization statistics are frozen whenever a likelihood potential is differentiated.

A.1 PROOF OF THEOREM 3.1

Write $p_j(x;\tau) = \Phi((x-c_j)/\tau)$. Since $\mathbb{E}[S_j \mid u_0] = p_j(u_0;\tau)$, taking the expectation of the ordinal negative-log-likelihood derivative gives

$$\Psi(x;u_0) = \frac{1}{K-1}\sum_{j=1}^{K-1} p_j'(x;\tau)\frac{p_j(x;\tau)-p_j(u_0;\tau)}{p_j(x;\tau)\{1-p_j(x;\tau)\}}.$$

For finite x, every denominator and every $p_j'(x;\tau)$ is positive. Because p_j is strictly increasing, every summand has the sign of $x-u_0$. Therefore $(x-u_0)\Psi(x;u_0) > 0$ for $x \neq u_0$, and u_0 is the unique zero.

Each term of ℓ_k^{ord} is either $-\log\Phi(z)$ or $-\log\Phi(-z)$, with $z = (x-c_j)/\tau$. Both functions are strictly convex. For example, with the inverse Mills ratio $\rho(z) = \phi(z)/\Phi(z)$,

$$\rho'(z) = -\rho(z)\{z+\rho(z)\} < 0,$$

where $z + \rho(z) > 0$ follows from the standard inverse-Mills-ratio inequality. Thus their positive average is strictly convex, and ψ_k is strictly increasing.

For an internal target cell k, the calibration condition

$$\psi_k(c_k;\tau) < 0 < \psi_k(c_{k+1};\tau)$$

places the unique root of $\psi_k(\cdot;\tau)$ inside that cell. Let $t = k_i^{\mathrm{tar}}$. If $g_i > 0$, then $t = k_i - 1$, so the current coordinate lies to the right of the target cell and hence to the right of its score root. Therefore,

$$\psi_t(u_i;\tau) > 0, \qquad \operatorname{sign}\big(\psi_t(u_i;\tau)\big) = \operatorname{sign}(g_i).$$

If $g_i < 0$, then $t = k_i + 1$, and the symmetric argument gives

$$\psi_t(u_i;\tau) < 0, \qquad \operatorname{sign}\big(\psi_t(u_i;\tau)\big) = \operatorname{sign}(g_i).$$

Thus, for every task-selected internal target,

$$\operatorname{sign}\big(\psi_{k_i^{\mathrm{tar}}}(u_i;\tau)\big) = \operatorname{sign}(g_i).$$

For the lowest target level $k = 0$, all target bits are zero, so every score term is positive and $\psi_0(x;\tau) > 0$. For the highest target level $k = K-1$, all target bits are one, so every score term is negative and $\psi_{K-1}(x;\tau) < 0$. These signs agree with the corresponding task-gradient directions, and hence

$$\operatorname{sign}\big(\psi_{k_i^{\mathrm{tar}}}(u_i;\tau)\big) = \operatorname{sign}(g_i)$$

also holds for the two boundary target levels.

A.2 ANALYSIS AND VERIFICATION OF THE CALIBRATION CONDITION

Geometric Interpretation of Calibration. The calibration condition has a direct geometric interpretation: it requires the target-conditioned likelihood to be aligned with the corresponding quantization cell. For an internal target level k, the loss $\ell_k^{\mathrm{ord}}(\cdot;\tau)$ is strictly convex and tends to $+\infty$ as $x \to \pm\infty$, and hence admits a unique finite minimizer

$$u_k^\star(\tau) := \arg\min_x \ell_k^{\mathrm{ord}}(x;\tau).$$

Since its score $\psi_k(x;\tau) = \partial_x \ell_k^{\mathrm{ord}}(x;\tau)$ is strictly increasing and vanishes at $u_k^\star(\tau)$, the calibration condition

$$\psi_k(c_k;\tau) < 0 < \psi_k(c_{k+1};\tau)$$

is equivalent to

$$c_k < u_k^\star(\tau) < c_{k+1}.$$

Thus, calibration means that the latent location preferred by the target-conditioned likelihood lies inside the target quantization cell, providing a geometric alignment between the likelihood and the quantizer.

A General Sufficient Calibration Regime. We next establish a general existence result: for any finite-level quantizer with strictly ordered thresholds, sufficiently small τ guarantees calibration for all internal target cells.

For an internal target level k, using $\rho(z) = \phi(z)/\Phi(z)$, the score at the two cell boundaries satisfies

$$(K-1)\tau\,\psi_k(c_k;\tau) = -\rho(0) - \sum_{j<k} \rho\left(\frac{c_k - c_j}{\tau}\right) + \sum_{j>k} \rho\left(\frac{c_j - c_k}{\tau}\right),$$

$$(K-1)\tau\,\psi_k(c_{k+1};\tau) = -\sum_{j\le k} \rho\left(\frac{c_{k+1} - c_j}{\tau}\right) + \rho(0) + \sum_{j>k+1} \rho\left(\frac{c_j - c_{k+1}}{\tau}\right).$$

Because the thresholds are strictly ordered, every nonzero distance above is positive. Since $\rho(z) \to 0$ as $z \to +\infty$,

$$\lim_{\tau\to 0^+} (K-1)\tau\,\psi_k(c_k;\tau) = -\rho(0) < 0,$$

and

$$\lim_{\tau\to 0^+} (K-1)\tau\,\psi_k(c_{k+1};\tau) = \rho(0) > 0.$$

Hence, for each internal target cell, there exists $\bar\tau_k > 0$ such that $\psi_k(c_k;\tau) < 0 < \psi_k(c_{k+1};\tau), 0 < \tau < \bar\tau_k$. Since there are finitely many internal target cells, $\bar\tau := \min_k \bar\tau_k > 0$ gives a common calibrated regime $0 < \tau < \bar\tau$.

Thus, every finite quantizer with strictly ordered thresholds admits a nonempty calibrated temperature range. This small-τ result is only a sufficient condition; the actual admissible range can be substantially larger and depends on the specific quantizer geometry.

Calibration Ranges for Experimental Quantizers. We next characterize the geometry-specific calibrated ranges of the normalized quantizers used in our experiments. Unlike the general small-τ result above, these ranges are obtained directly from the endpoint calibration criterion over all internal target cells.

For example, for the signed 2-bit quantizer with levels $(-2,-1,0,1)$ and thresholds $(-3/2,-1/2,1/2)$, consider target level -1, whose ordinal code is $(1,0,0)$. At the two boundaries of the target cell,

$$\psi_{-1}(-3/2;\tau) = \frac{-\rho(0) + \rho(1/\tau) + \rho(2/\tau)}{3\tau}, \qquad \psi_{-1}(-1/2;\tau) = \frac{\rho(0)}{3\tau} > 0.$$

Hence target-cell alignment holds whenever

$$\rho(1/\tau) + \rho(2/\tau) < \rho(0).$$

The equality has the unique numerical solution

$$\bar\tau = 2.003589,$$

so the calibrated interval connected to $\tau = 0^+$ is $0 < \tau < 2.003589$. By symmetry, the same bound applies to the other internal target cell. Applying the same endpoint test to all internal cells gives the ranges in Table 6.

W1 has no internal target cell, so the calibration condition is vacuous and the boundary-target sign guarantee holds for every $\tau > 0$. For W2–W4, the calibrated upper endpoint exceeds 1 in all cases. Therefore, both deployed temperatures $\tau \in \{0.5, 1\}$ lie within the calibrated regime for all quantizer geometries used in our experiments.

Table 6: Calibrated temperature ranges for the normalized quantizers used in our experiments. The reported interval is the calibrated component connected to $\tau = 0^+$.

Precision	K	Internal cells	Calibrated range
W1	2	0	unrestricted
W2	4	2	$0 < \tau < 2.003589$
W3	8	6	$0 < \tau < 1.681188$
W4	16	14	$0 < \tau < 1.681078$

Implementation-Direction Audit. The preceding analysis establishes calibration at the likelihood level. We finally verify that the deployed implementation preserves this directional property after route-specific normalization, clipping, and support masking. Table 7 reports the audit at the default $\tau = 1$, which is used in most experiments.

Table 7: Calibration and implementation-direction audit at the default $\tau = 1$.

Configuration	M_{cal}	Failed cells	Sign mismatch
LSQ-W2	0.151679	0	0/2,778
QuEST-W1	N/A	N/A	0/30,440,903
QuEST-W2	0.151679	0	0/60,948,480
QuEST-W3	0.064352	0	0/60,948,480
QuEST-W4	0.030031	0	0/60,948,480

For internal target levels, define $M_{\mathrm{cal}} := \min_k \min\{-\psi_k(c_k; 1), \psi_k(c_{k+1}; 1)\}$. Thus, $M_{\mathrm{cal}} > 0$ certifies the endpoint calibration condition for all internal cells at $\tau = 1$. Failed cells count violations of this condition, while sign mismatch compares the deployed postprocessed score with the task gradient on its implemented support. The raw likelihood score also has zero mismatches. QuEST counts use two validation batches from completed seed-0 checkpoints, whereas LSQ-W2 uses a dense audit of both deployed LSQ backward routes. W1 has no internal target cell.

All audited internal cells have positive calibration margins, and no sign mismatch is observed for either the raw or deployed postprocessed scores relative to the task gradient. Thus, the implemented IDScore routes preserve the directional property predicted by Theorem 3.1. The larger-scale setting $\tau = 0.5$ is covered by the calibrated temperature ranges established above.

A.3 Proof of Theorem 3.2

For a scalar x, the cumulative codes of $\kappa(x)$ and target k differ at exactly $|\kappa(x) - k|$ thresholds. At each mismatched threshold, the working model assigns probability at most $1/2$ to the target event, so its negative log probability is at least $\log 2$. Since the composite loss averages over $K-1$ thresholds,

$$|\kappa(x) - k| \le \frac{K-1}{\log 2} \ell_k^{\mathrm{ord}}(x).$$

Multiplying by $|g_i|$ and summing over $\mathcal{S}$ proves the mismatch bound. Differentiating the frozen potential gives $z_i = m_i |g_i| \psi_i$.

For $c_{\min} = 0$, score normalization gives

$$r_i = \lambda m_i |g_i| \operatorname{sign}(\psi_i) \min\{|\psi_i|/\nu, c_{\max}\} = \operatorname{clip}\left(\frac{\lambda}{\nu} z_i, -U_i, U_i\right).$$

Therefore $\boldsymbol{r}$ is the Euclidean projection of $(\lambda/\nu)\boldsymbol{z}$ onto the box $\mathcal{B} = \{\boldsymbol{d} : |d_i| \le U_i\}$. Its variational inequality is

$$\left\langle \frac{\nu}{\lambda}\boldsymbol{r} - \boldsymbol{z}, \boldsymbol{d} - \boldsymbol{r} \right\rangle \ge 0 \qquad (\boldsymbol{d} \in \mathcal{B}).$$

For any $\boldsymbol{d} \in \mathcal{B}$ with $\|\boldsymbol{d}\|_2 \le \|\boldsymbol{r}\|_2$, this implies

$$\begin{aligned}\langle \boldsymbol{z}, \boldsymbol{r} - \boldsymbol{d}\rangle &\ge \frac{\nu}{\lambda}\langle \boldsymbol{r}, \boldsymbol{r} - \boldsymbol{d}\rangle \\ &= \frac{\nu}{2\lambda}\left(\|\boldsymbol{r}\|_2^2 - \|\boldsymbol{d}\|_2^2 + \|\boldsymbol{r} - \boldsymbol{d}\|_2^2\right) \\ &\ge \frac{\nu}{2\lambda}\|\boldsymbol{r} - \boldsymbol{d}\|_2^2.\end{aligned}$$

The final inequality is strict for $\boldsymbol{d} \neq \boldsymbol{r}$, proving the unique maximizer in Eq. (10). If the cap is inactive, maximizing a linear functional on a Euclidean ball gives $R\boldsymbol{z}/\|\boldsymbol{z}\|_2$.

If a numerical implementation uses $\widehat{\boldsymbol{z}}$ and its projected residual $\widehat{\boldsymbol{r}}$, the same argument yields, for every feasible $\boldsymbol{d}$ with $\|\boldsymbol{d}\|_2 \le \|\widehat{\boldsymbol{r}}\|_2$,

$$\langle \boldsymbol{z}, \widehat{\boldsymbol{r}} - \boldsymbol{d}\rangle \ge \frac{\nu}{2\lambda}\|\widehat{\boldsymbol{r}} - \boldsymbol{d}\|_2^2 - \|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|_2\|\widehat{\boldsymbol{r}} - \boldsymbol{d}\|_2.$$

Thus the analytic likelihood interpretation is stable when the implemented and analytic score vectors are close. The checkpoint probe reports this gap directly rather than assuming it is zero.

A.4 Proof of Theorem 3.3

For $c_{\min} = 0$, the normalized likelihood score satisfies

$$\bar{\psi}_i = \operatorname{sign}(\psi_i)\min\left\{\frac{|\psi_i|}{\nu}, c_{\max}\right\} = \operatorname{sign}(\psi_i)\zeta_i,$$

where $\zeta_i := \min\left\{\frac{|\psi_i|}{\nu}, c_{\max}\right\}$. Hence the IDScore residual is

$$r_i = \lambda m_i |g_i| \operatorname{sign}(\psi_i)\zeta_i. \tag{13}$$

For every $i \in \mathcal{S}$, Theorem 3.1 gives $\operatorname{sign}(\psi_i) = \operatorname{sign}(g_i)$. Since $m_i = 1$ on $\mathcal{S}$, Eq. (13) becomes $r_i = \lambda|g_i|\operatorname{sign}(g_i)\zeta_i$. Therefore,

$$\chi_i g_i + r_i = |g_i|\operatorname{sign}(g_i)(\chi_i + \lambda\zeta_i),$$

and the direct update $u_i^+ = u_i - \eta(\chi_i g_i + r_i)$ can be written as

$$u_i^+ = u_i - \eta|g_i|\operatorname{sign}(g_i)(\chi_i + \lambda\zeta_i). \tag{14}$$

Consider first $i \in \mathcal{M}$. Let θ_i be the threshold separating the current and selected target cells. Since u_i is still in its current cell,

$$d_i(u_i) = \operatorname{sign}(g_i)(u_i - \theta_i) \ge 0.$$

Using Eq.(14), we have

$$\begin{aligned}\operatorname{sign}(g_i)(u_i^+ - \theta_i) &= \operatorname{sign}(g_i)(u_i - \theta_i) - \eta|g_i|(\chi_i + \lambda\zeta_i) \\ &= d_i(u_i) - \eta|g_i|(\chi_i + \lambda\zeta_i).\end{aligned}$$

Taking the positive part yields

$$d_i(u_i^+) = \left[d_i(u_i) - \eta|g_i|(\chi_i + \lambda\zeta_i)\right]_+.$$

Now consider $i \in \mathcal{H}$. By definition, the clipped target coincides with the current extreme level.

If $k_i = 0$, the outward request necessarily has $g_i > 0$. Eq. (14) then gives $u_i^+ \le u_i$, so the latent coordinate moves farther into the lowest quantization cell.

Similarly, if $k_i = K - 1$, the outward request has $g_i < 0$, and $u_i^+ \ge u_i$, so the coordinate moves farther into the highest quantization cell.

Hence every $i \in \mathcal{H}$ remains in its selected extreme cell.

We next compare the one-step transition sets.

Under the base update alone, a coordinate $i \in \mathcal{M}$ reaches or crosses the selected threshold whenever $d_i(u_i) \le \eta\chi_i|g_i|$, which is exactly the condition

$$i \in \mathcal{T}_{\text{base}}.$$

With the IDScore residual, Eq.(11) shows that the corresponding transition condition becomes $d_i(u_i) \le \eta|g_i|(\chi_i + \lambda\zeta_i)$, which is equivalent to

$$i \in \mathcal{T}_{\text{ID}}.$$

Since $\lambda \ge 0$ and $\zeta_i \ge 0$, $\eta\chi_i|g_i| \le \eta|g_i|(\chi_i + \lambda\zeta_i)$. Therefore,

$$i \in \mathcal{T}_{\text{base}} \quad \Longrightarrow \quad i \in \mathcal{T}_{\text{ID}},$$

and hence $\mathcal{T}_{\text{base}} \subseteq \mathcal{T}_{\text{ID}}$.

Let $\Delta q_i = Q(u_i^+) - Q(u_i)$. From Eq. (14), every nonzero latent displacement has direction

$$\operatorname{sign}(u_i^+ - u_i) = -\operatorname{sign}(g_i).$$

Since Q is nondecreasing, quantization cannot reverse the direction of this displacement. Therefore, whenever $\Delta q_i \neq 0$, $\operatorname{sign}(\Delta q_i) = -\operatorname{sign}(g_i)$, and equivalently

$$\Delta q_i = -\operatorname{sign}(g_i)|\Delta q_i|.$$

For $\Delta q_i = 0$, the same identity holds trivially. Consequently,

$$\langle g, \Delta q\rangle = \sum_i g_i \Delta q_i = -\sum_i |g_i||\Delta q_i| = -\Gamma_{\text{tr}}.$$

By definition of the finite-transition remainder, $\mathcal{R}_{\text{tr}} = L(q+\Delta q) - L(q) - \langle g, \Delta q\rangle$. Using $\langle g, \Delta q\rangle = -\Gamma_{\text{tr}}$, we obtain

$$L(q + \Delta q) - L(q) = -\Gamma_{\text{tr}} + \mathcal{R}_{\text{tr}}.$$

Therefore, $\mathcal{R}_{\text{tr}} < \Gamma_{\text{tr}}$ implies

$$L(q + \Delta q) < L(q).$$

Finally, since L is β-smooth along the segment from q to $q + \Delta q$, the descent lemma gives

$$L(q + \Delta q) \le L(q) + \langle \nabla_q L(q), \Delta q\rangle + \frac{\beta}{2}\|\Delta q\|_2^2. \tag{15}$$

Since $g = \nabla_q L(q)$ and $\langle g, \Delta q\rangle = -\Gamma_{\text{tr}}$, Eq. (15) gives

$$L(q + \Delta q) - L(q) \le -\Gamma_{\text{tr}} + \frac{\beta}{2}\|\Delta q\|_2^2.$$

Consequently, if $\Gamma_{\text{tr}} > \frac{\beta}{2}\|\Delta q\|_2^2$, then $-\Gamma_{\text{tr}} + \frac{\beta}{2}\|\Delta q\|_2^2 < 0$, and therefore

$$L(q + \Delta q) < L(q).$$

This proves the sufficient task-descent condition.

Remark A.1. *The smoothness condition is imposed on the task loss $L(q)$ viewed as a function of the quantized coordinates along the segment between q and $q+\Delta q$. It does not require the quantizer Q or the latent objective $L(Q(u))$ to be smooth. The two endpoints remain actual quantized states; smoothness is used only to control the nonlinear remainder of the finite transition between them.*

A.5 Scope of the Claims

Theorem 3.1 concerns the analytic Gaussian working likelihood and requires calibration for internal target cells. Appendix A.2 provides a general small-τ sufficient regime and verifies calibration for all quantizers used in our experiments. Theorem 3.2 concerns the frozen task-weighted likelihood potential. The optimality concerns first-order decrease of the frozen likelihood potential within the stated constrained class, where corrections have the same coordinate-wise caps and no larger ℓ_2 norm than the IDScore residual; it does not imply optimality for the task loss or under arbitrary correction budgets. Theorem 3.3 characterizes threshold progress, one-step transition inclusion, and first-order task alignment for the direct sign-preserving update. Its smoothness-based result is only a local sufficient condition for task-loss descent and does not model stateful optimizer dynamics. None of these results establishes global QAT convergence or unconditional task-loss improvement.

B FRAMEWORK-SPECIFIC REALIZATIONS

The target selection and likelihood-score construction in Section 2 are shared across QAT frameworks. A framework-specific realization only needs to specify (i) the continuous coordinate u in which quantization is applied and (ii) the fixed route adapter $(\mathcal{J}, \mathbf{m})$ used for score normalization and residual support in Eqs. (7)–(8). The resulting corrected signal is then mapped back to the original parameter space through the native backward transformation of the underlying QAT method. We describe these realizations for LSQ and QuEST below.

B.1 LSQ

In LSQ, quantization is applied to the normalized weight

$$u_i = \frac{w_i}{\alpha_{\ell(i)}},$$

where $\alpha_{\ell(i)}$ is the learned step size of the layer containing coordinate i. IDScore is therefore evaluated in this normalized quantization coordinate, using the corresponding normalized thresholds. For the ResNet route, the normalization set $\mathcal{J}$ is the full quantized tensor, while the residual-support mask $\mathbf{m}$ is the native LSQ in-range mask. For the LLaMA route, both normalization and residual support use the feasible-transition set

$$\mathcal{F} := \{i : g_i \neq 0,\ k_i^{\mathrm{tar}} \neq k_i\},$$

that is,

$$\mathcal{J} = \mathcal{F}, \qquad m_i = \mathbf{1}\{i \in \mathcal{F}\}.$$

These route-specific choices are fixed throughout training and follow the native backward support of their respective training pipelines rather than task-specific tuning.

After applying the IDScore correction in the normalized coordinate, the existing LSQ chain rule maps the corrected signal back to the weight coordinate w_i. IDScore does not modify the learned scale derivative $\partial L/\partial \alpha$ or its optimizer update.

B.2 QUEST

QuEST applies weight quantization in the Hadamard domain and masks its backward signal with a trust region (Panferov et al., 2025). Writing

$$\boldsymbol{w}_h = \boldsymbol{H}\boldsymbol{w}.$$

IDScore is evaluated in the same transformed coordinate in which quantization is applied. For the W2A2–W4A4 routes, the normalization set $\mathcal{J}$ is the full transformed tensor, while the residual-support mask is

$$m_i = \mathbf{1}_{\{g_{h,i} \neq 0\}}.$$

Denoting the native QuEST trust mask by $\mathbf{m}_{\mathrm{trust}}$, the corrected Hadamard-domain signal is

$$\widetilde{\boldsymbol{g}}_h = \boldsymbol{g}_h \odot \boldsymbol{m}_{\mathrm{trust}} + \lambda \boldsymbol{m}_h \odot |\boldsymbol{g}_h| \odot \bar{\boldsymbol{\psi}}_h, \tag{16}$$

and is mapped back by

$$\mathbf{g}_w = H^\top \tilde{\mathbf{g}}_h.$$

For the W1A1 route, both $\mathcal{J}$ and the residual support use the feasible-transition set

$$\mathcal{F}_h := \{i : g_{h,i} \neq 0,\ k_i^{\mathrm{tar}} \neq k_i\}.$$

In all cases, the thresholds, pseudo-targets, likelihood scores, and score normalization are computed in the Hadamard domain where quantization occurs. These route-specific choices are fixed throughout training and follow the native quantization and backward structure of QuEST rather than task-specific tuning.

B.3 Frozen-Checkpoint Adapter Audit

We verify that the generic adapter formulation above reproduces the QuEST-W2 implementation used in the main experiments and examine how alternative normalization and support choices change the residual. At the final checkpoints of three training seeds, we sample four minibatches per seed. Table 8(a) compares alternative adapters at the same frozen states, while Table 8(b) reports diagnostics of the implemented route. Here "movable" denotes coordinates for which the clipped adjacent target differs from the current level, and "full" denotes all coordinates in the transformed tensor. Panel (a) reports averages over all 12 seed–minibatch pairs. In panel (b), each seed row averages its four minibatches, except that the projection error reports the maximum.

Table 8: Frozen-checkpoint audit of the implemented QuEST-W2 adapter.

(a) Alternative normalization and residual-support choices evaluated at the same frozen states; adapter names are written as normalization/support.

Adapter	Normalization set $\mathcal{J}$	Residual support	Cosine	Norm ratio	Relative ℓ_2
Implemented	Full	$g_i \neq 0$	1.000	1.000	0.000
Full/movable	Full	Movable	0.977	0.977	0.215
Movable/full	Movable	$g_i \neq 0$	1.000	0.912	0.088
Movable/movable	Movable	Movable	0.977	0.891	0.232
Full/trust	Full	QuEST trust mask	0.978	0.978	0.209

(b) Diagnostics of the implemented route.

Seed	$\mu_{\mathrm{mov}}/\mu_{\mathrm{full}}$	Sat. coord. (%)	Mov. energy (%)	Sat. energy (%)	Outside trust (%)	Proj. error	Calibration
0	1.0982	17.368	95.346	4.654	4.345	7.38×10^{-8}	4/4
1	1.0980	17.379	95.363	4.637	4.095	7.39×10^{-8}	4/4
2	1.0979	17.359	95.379	4.621	4.622	7.61×10^{-8}	4/4
Mean	1.0980	17.368	95.363	4.637	4.354	7.61×10^{-8}	12/12

The implemented residual is reproduced to numerical precision. The adapter choices are not numerically interchangeable: changing either the normalization set or residual support changes its norm or direction. Under the implemented route, movable coordinates carry 95.36% of residual energy while saturated boundary coordinates carry the remaining 4.64%; these two cases are treated separately in Theorem 3.3. Approximately 4.35% of residual energy lies outside the native QuEST trust mask. Table 8 is an implementation audit rather than a task-performance comparison. The complete-training IDScore-trust control in Table 10 restricts the residual to the native trust mask and increases mean PPL from 36.502 to 39.965. Thus, the relatively small frozen-state energy outside the trust mask should not be interpreted as implying a negligible effect on the training trajectory.

B.4 Computational Cost

For n quantized coordinates and K quantization levels, evaluating the ordinal likelihood score requires $\mathcal{O}(n(K-1))$ elementwise operations and temporary tensors. IDScore does not require an additional task-loss evaluation, forward or backward model pass, auxiliary network path, or optimizer state.

The extra computation therefore scales linearly with the number of quantized coordinates for a fixed bit width. In the QuEST-30M W2A2 benchmark of Table 5, this corresponds to only a 2.0% increase in mean step time and a 0.4% increase in peak allocated memory relative to QuEST.

C Additional Experimental Results

This section first reports complete seed statistics omitted from the main tables and then presents additional controlled comparisons. Section C.1 gives the full scale and precision statistics. Section C.2 examines likelihood-based magnitude, residual support, and fixed element-wise scaling. Section C.3 compares IDScore with the native EWGS procedure, and Section C.4 isolates coordinate-wise allocation using norm-matched non-likelihood controls.

C.1 Full Statistics for Main Results

Table 9 provides the full three-seed statistics behind the completed scale and precision results summarized in Table 2. The improvements remain positive relative to the observed seed variation. Settings without a complete matched three-seed aggregate are not mixed into this table.

Table 9: Full three-seed results for transfer across model scale and bit width on C4. Values are final validation PPL reported as mean ± population standard deviation over three matched seeds; lower is better.

(a) Model and training scale at W2A2

Scale	QuEST	QuEST+FOGZO	QuEST+IDScore
30M / 3B	37.262 ± 0.083	36.975 ± 0.033	**36.502** ± 0.043
50M / 5B	31.974 ± 0.101	31.742 ± 0.054	**31.216** ± 0.067
100M / 10B	26.302 ± 0.025	26.130 ± 0.010	**25.700** ± 0.015
200M / 20B	22.583 ± 0.043	22.434 ± 0.053	**21.644** ± 0.042

(b) Quantization precision on 30M / 3B

Precision	QuEST	QuEST+IDScore
W1A1	54.256 ± 0.453	**49.752** ± 0.109
W2A2	37.262 ± 0.083	**36.502** ± 0.043
W3A3	33.745 ± 0.085	**33.294** ± 0.050
W4A4	30.975 ± 0.074	**30.731** ± 0.056

C.2 Structural and Route Controls

Table 10 evaluates controlled variants that isolate three aspects of IDScore: likelihood-based magnitude, residual support, and element-wise gradient scaling. Uniform retains the target-consistent correction direction but removes the likelihood-derived coordinate-wise magnitude. Its performance closely matches QuEST, whereas Boundary-NLL improves it and the two likelihood-based IDScore variants achieve nearly identical mean PPL, supporting the role of likelihood-based magnitude beyond the target direction alone. IDScore-trust restricts the residual to the native QuEST trust mask and performs substantially worse than the default IDScore, indicating that the route-specific residual support also matters. EWGS-fixed replaces the likelihood-based correction with a fixed EWGS-style element-wise scaling rule ($\delta = 10^{-3}$) in the same QuEST weight-backward route. It does not use the adaptive curvature-related coefficient of native EWGS and is therefore intended as a controlled comparison rather than a full EWGS reproduction. We use IDScore as the default construction because it directly exploits the ordered threshold structure and applies to both binary and multi-bit quantization.

C.3 Native Comparison with EWGS

Because EWGS-fixed in Table 10 is only a controlled port, we further compare IDScore with native EWGS on ResNet-20/CIFAR-10 W1A1 following Lee et al. (2021). This comparison evaluates IDScore against EWGS under its native training protocol rather than through the simplified QuEST-side control. All methods use the same forward quantizer, training schedule, and pretrained full-precision checkpoint. The three runs vary only the QAT seed, and results are reported at the fixed 400-epoch endpoint rather than at each run's best epoch. IDScore is applied to weight quantizers while retaining the native STE route for activations. As shown in Table 11, the mean final accuracies

Table 10: Structural, support, and gradient-scaling controls on QuEST-30M W2A2 with 3B C4 tokens. Lower PPL is better.

Method	s0	s1	s2	Mean PPL	Std	Gain vs QuEST
QuEST	37.233	37.374	37.178	37.262	0.083	+0.000
EWGS-fixed	40.852	40.667	40.562	40.694	0.120	-3.432
Uniform	37.293	37.314	37.191	37.266	0.054	-0.004
IDScore-trust	40.062	39.959	39.873	39.965	0.077	-2.703
Boundary-NLL	36.647	36.667	36.520	36.611	0.065	+0.651
IDScore	36.467	36.562	36.476	36.502	0.043	+0.760

are 85.480% for STE, 85.360% for EWGS, and 85.577% for IDScore. The differences are modest, but IDScore remains competitive with native EWGS in this setting.

Table 11: Final accuracy (%) under the native EWGS ResNet-20/CIFAR-10 W1A1 protocol.

Method	Seed 0	Seed 1	Seed 2	Mean $\pm$ Std
STE	85.330	85.510	85.600	85.480 ± 0.112
EWGS	85.130	85.670	85.280	85.360 ± 0.228
IDScore	85.590	85.440	85.700	$\mathbf{85.577} \pm 0.107$

C.4 Matched Magnitude Controls

To test whether the IDScore gain can be explained simply by a larger residual magnitude, we construct non-likelihood controls that match the IDScore residual norm at every training state. All controls use the same pseudo-target, support, task-gradient magnitude $|g|$, correction sign, and per-tensor residual norm as IDScore; only the coordinate-wise allocation rule is changed. GSign uses uniform coordinate weights, while Boundary-Far and Boundary-Near allocate larger corrections to coordinates farther from or closer to the selected target boundary, respectively. These controls therefore preserve the scalar residual budget while replacing the likelihood-shaped allocation.

Table 12: Norm-matched controls for IDScore magnitude allocation.

Method	s0	s1	s2	Mean PPL	Std	Gap vs IDScore
IDScore-Ordinal	36.467	36.562	36.476	36.502	0.043	+0.000
GSign matched	40.810	40.588	40.633	40.677	0.096	+4.175
Boundary-Far matched	39.647	39.681	39.604	39.644	0.032	+3.142
Boundary-Near matched	41.197	41.103	41.143	41.148	0.039	+4.646

Table 12 shows that the relative norm error is below 4.1×10^{-8} and that no control reaches the clipping cap. IDScore outperforms all three alternatives on every matched seed. Since the controls match the main non-allocation factors, the performance gap supports the role of likelihood-shaped coordinate allocation rather than a larger residual budget alone.