FEATURE CORRELATION PRESERVING QUANTIZATION FOR ANALOG COMPUTE-IN-MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Analog compute-in-memory (ACIM) enables efficient neural network inference by performing matrix operations within memory arrays, but its limited numerical precision requires low-bit quantization. Existing post-training quantization (PTQ) methods mainly minimize reconstruction error between full-precision and quantized representations. This objective does not explicitly preserve feature relations, which are important when multiple features share the same quantization configuration within an ACIM mapping group. We propose Feature Correlation Preserving Quantization (FCPQ), a groupwise PTQ framework that preserves feature relations while optimizing reconstruction. FCPQ consists of two components. First, Feature Correlation Calibration (FCC) standardizes features and aligns the correlation matrices of full-precision and quantized representations within each mapping group. Second, Correlation Aware Quantization Optimization (CQO) estimates the contribution and distortion of feature relations and uses them to assign adaptive reconstruction weights. Together, these components guide quantization toward preserving both numerical values and feature relations. Experiments across multiple datasets and settings show that FCPQ improves accuracy and maintains the feature structure of full-precision models under low-bit quantization.

1 INTRODUCTION

Analog compute in memory (ACIM) performs computation inside memory arrays instead of moving all data between memory and separate computing units. It is an analog realization of compute-in-memory (CIM), where matrix-vector multiplication is carried out through physical operations in memory cells and peripheral circuits. By combining computation and storage, CIM reduces data movement during neural network inference. ACIM further exploits the parallel computation of analog arrays, which can provide high throughput with low data movement and energy cost (Shafiee et al., 2016; Ankit et al., 2019; Verma et al., 2019).

ACIM performs computation with analog signals inside memory arrays, which limits the numerical precision of the computation. Practical ACIM arrays support only finite cell states and limited input and output precision. Digital-to-analog and analog-to-digital conversion also introduces additional area, energy, and quantization error (Jain et al., 2020b). Neural network weights and activations therefore need to use low-bit representations that match the precision supported by the hardware. Quantization is thus necessary for ACIM inference.

Quantization in ACIM is also constrained between the hardware mapping of neural network layers and physical arrays. A layer can be larger than a single physical array, so its weights and output channels are partitioned across multiple arrays. We call each partition an ACIM mapping group. Quantization parameters are shared within a group, and the corresponding features are processed by the same hardware resources. Their quantization errors are therefore coupled rather than independent. Features within the same group also play different roles in the representation. This means that ACIM quantization should consider not only the magnitude of numerical errors but also the relations among jointly quantized features.

Since full model retraining is costly, post-training quantization (PTQ) offers a practical way to obtain low-bit models. Reconstruction-based PTQ methods optimize quantization parameters by matching the floating-point model. AdaRound learns data-dependent rounding through reconstruction, while BRECQ extends reconstruction to network blocks (Li et al., 2021). Yet, reconstruction mainly

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.

measures numerical differences and does not explicitly preserve relations among features within an ACIM mapping group. Ignoring these relations disrupt the feature structure learned by the full-precision model and lead to accuracy degradation. For ACIM quantization, PTQ should therefore preserve both numerical values and feature relations.

In this work, we propose Feature Correlation Preserving Quantization (**FCPQ**), a groupwise PTQ framework that extends reconstruction-based quantization from numerical matching to structural preservation. FCPQ jointly preserves feature responses and relational structures during quantization to maintain the representations learned by full-precision models, satisfying the dual requirements of numerical accuracy and structural consistency in low-bit grouped ACIM computation. Specifically, we introduce Feature Correlation Calibration (**FCC**) to preserve feature dependencies by reducing correlation distortion between floating-point and quantized representations. Furthermore, we propose Correlation-aware Quantization Optimization (**CQO**), which incorporates correlation importance into reconstruction optimization and adaptively focuses on features that contribute more to representation structures. Overall, the contributions of this paper are three-fold:

- We propose FCPQ, a groupwise PTQ framework that adds feature relation constraints to reconstruction while keeping reconstruction as the main optimization objective.
- We design Feature Correlation Calibration (FCC), a relation-preserving objective that captures the structural dependency among features within ACIM mapping groups and reduces correlation distortion after quantization.
- We design Correlation-aware Quantization Optimization (CQO), a relation-guided reconstruction strategy that adaptively emphasizes features with larger structural impacts, improving the robustness of quantized representations.

2 Related Work

ACIM Inference and Quantization. ACIM architectures execute matrix vector multiplication in resistive crossbars or SRAM arrays to reduce data movement during neural network inference (Shafiee et al., 2016; Ankit et al., 2019; Verma et al., 2019). Practical deployment is constrained by finite array capacity, limited cell states, data conversion, and device and circuit variation . Prior algorithm and hardware studies address these constraints through mapping, noise aware optimization, mixed precision, and hardware cost aware quantization (Wang et al., 2019). These studies mainly optimize mapping efficiency, numerical precision, or robustness to physical errors. Recent ACIM designs further study routing, shuffle-based analog processing, heterogeneous arrays, and mixed-precision co-design (Tu et al., 2022; Chen et al., 2024; Han et al., 2024). They do not directly preserve the feature relations altered by groupwise low precision quantization.

Post Training Quantization. PTQ converts a pretrained model to low precision using a small calibration data set and no full retraining. Reconstruction based methods optimize clipping ranges, scales, or rounding variables by matching floating point weights, layer outputs, or block outputs. AdaRound learns discrete rounding decisions through layer reconstruction (Nagel et al., 2020), BRECQ introduces block reconstruction and sensitivity (Li et al., 2021), and BOF4 exposes blocks to partially quantized inputs during calibration (Blumenberg et al., 2026). AnyBCQ further incorporates prediction differences into the Optimization objective (Park et al., 2026). Although these methods reduce numerical discrepancy, a small reconstruction error does not ensure that the relations among quantized features remain consistent with the floating point representation.

Relation Based Representation Transfer. Relation based learning transfers structural information through sample similarity, Gram matrices, or feature flow statistics (Tung & Mori, 2019; Yim et al., 2017). These objectives preserve dependencies that are not directly represented by elementwise feature matching. Most relation based methods are designed for supervised retraining or knowledge distillation between networks with different capacities. The broader quantization literature also includes LSQ, PACT, QDrop, and prediction-difference objectives (Esser et al., 2020; Choi et al., 2018; Wei et al., 2022; Liu et al., 2023). They do not address calibration only optimization, quantization parameters, or feature relation distortion within ACIM mapping groups.

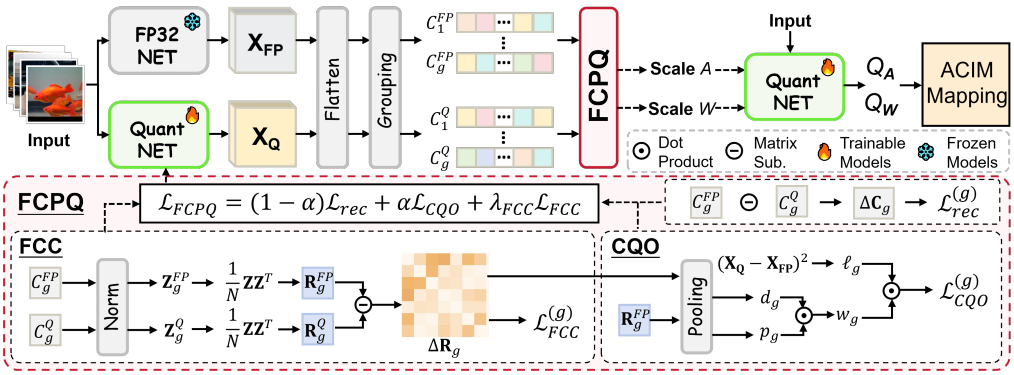


Figure 1: Overview of FCPQ. FCPQ preserves groupwise feature correlations and assigns adaptive reconstruction weights during offline PTQ, producing quantization parameters for low-bit ACIM inference.

3 METHOD

In this section, we first formulate the groupwise ACIM quantization problem and introduce the reconstruction objective. We then present FCC to preserve feature relation structures within each mapping group and CQO to incorporate relation-aware priorities into reconstruction optimization. Finally, we integrate these components into a unified relation-preserving PTQ framework for accurate and robust low-bit ACIM deployment.

3.1 MOTIVATION

ACIM reduces data movement by performing matrix vector multiplication inside memory arrays. To translate this advantage into low precision inference, the quantized operands must be supported by the cell states and interfaces of the target array. When a layer is wider than one physical array, its weights and output channels are partitioned across arrays; we denote each partition by an ACIM mapping group. The group is the unit at which quantization parameters are shared and several output features are perturbed together. Reconstruction based PTQ calibrates these parameters by matching floating point weights or outputs, but it measures each output element independently. It therefore lacks an explicit constraint on the relation structure generated by a group and a rule for allocating updates among features with different structural roles.

We describe the target mapping by a set of groups $\mathcal{G}$. For every (l, g), the weights and output features share the quantization parameters $\boldsymbol{\theta}_{l,g}$. FCPQ uses this same group boundary for all relation measurements and reconstruction weights, so the Optimization objective follows the execution granularity of ACIM. Figure 1 presents the complete calibration process.

Let $f(\cdot; \mathbf{W})$ denote a pretrained floating point network with weights $\mathbf{W}$. For layer l and ACIM mapping group g, the floating point weight tensor $\mathbf{W}_{l,g}$ is quantized as

$$\widehat{\mathbf{W}}_{l,g} = \mathcal{Q}_b(\mathbf{W}_{l,g}; \boldsymbol{\theta}_{l,g}), \tag{1}$$

where b is the target precision and $\boldsymbol{\theta}_{l,g} = (s_{l,g}, \mathbf{r}_{l,g})$ contains the group scale and weight-rounding variables. The integer range and $s_{l,g}$ jointly determine the effective clipping boundary; there is no separate clipping parameter in our implementation. The floating point weights remain fixed, and calibration updates only $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{l,g}\}_{l,g}$.

For a calibration batch, we flatten the batch and spatial dimensions of the group output. The floating point and quantized feature matrices are $\mathbf{X}_{l,g}^{\mathrm{f}}, \mathbf{X}_{l,g}^{\mathrm{q}} \in \mathbb{R}^{N \times C_g}$, where N is the number of observations and C_g is the number of features in group g. The floating point model is frozen throughout calibration, and only the quantization parameters are updated. Conventional reconstruction minimizes

$$\mathcal{L}_{\mathrm{rec}} = \sum_{l,g} \frac{1}{NC_g} \left\| \mathbf{X}_{l,g}^{\mathrm{f}} - \mathbf{X}_{l,g}^{\mathrm{q}} \right\|_F^2. \tag{2}$$

This objective constrains elementwise discrepancy but does not explicitly preserve feature relations. It remains in FCPQ as the numerical anchor for the calibrated model.

3.2 Adaptive Rounding Optimization

The reconstruction objective provides a stable numerical constraint for updating the quantization parameters. However, the final quantized weights are determined by discrete rounding decisions, whose optimization introduces additional difficulty. We therefore optimize the rounding variables with a relaxed continuous representation and gradually approach discrete states.

For a weight w with group scale $s_{l,g}$, the quantized code is formulated as

$$q(w) = \mathrm{clip}\left(\left\lfloor \frac{w}{s_{l,g}} \right\rfloor + h(r), q_{\min}, q_{\max}\right), \qquad \widehat{w} = s_{l,g} q(w), \tag{3}$$

where r is the learnable rounding variable. During continuous optimization, the rounding offset is relaxed as

$$h(r) = \sigma(r). \tag{4}$$

To reduce the discrepancy between optimization and deployment, the final stage adopts hard rounding in the forward pass:

$$h_{\mathrm{hard}}(r) = \mathbb{1}[\sigma(r) \geq 0.5], \tag{5}$$

while maintaining differentiable optimization through the straight-through estimator:

$$h_{\mathrm{STE}}(r) = \mathrm{sg}\left(h_{\mathrm{hard}}(r) - \sigma(r)\right) + \sigma(r). \tag{6}$$

We further regularize uncertain rounding states using

$$\mathcal{L}_{\mathrm{round}} = \frac{1}{|\mathbf{r}|} \sum_j \left[1 - |2\sigma(r_j) - 1|^{\beta_t}\right], \tag{7}$$

where β_t is gradually increased during optimization.

This formulation allows the quantization parameters to progressively approach discrete rounding states while preserving stable gradient-based updates.

3.3 Feature Correlation Calibration

Elementwise reconstruction cannot directly constrain the feature relations within an ACIM mapping group. We therefore design FCC to construct a magnitude independent representation of these relations and align it between the floating point and quantized models. The floating point features are computed once for each calibration batch and are treated as fixed targets, whereas the quantized features are recomputed after every parameter update. For a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times C_g}$, its ith feature is standardized as

$$\mathbf{z}_i = \frac{\mathbf{x}_i - \mu_i \mathbf{1}}{\sigma_i + \epsilon}, \tag{8}$$

where $\mathbf{x}_i \in \mathbb{R}^N$ is the ith column of $\mathbf{X}$, μ_i and σ_i are its mean and standard deviation, and ϵ provides numerical stability. Stacking the standardized features gives $\mathbf{Z} \in \mathbb{R}^{N \times C_g}$. The groupwise correlation matrix is defined as

$$\mathbf{R}(\mathbf{X}) = \frac{1}{N} \mathbf{Z}^\top \mathbf{Z}. \tag{9}$$

Thus, $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^{C_g \times C_g}$ and its (i, j) entry measures the normalized relation between features i and j. For group (l, g), we write $\mathbf{R}^{\mathrm{f}}_{l,g} = \mathbf{R}(\mathbf{X}^{\mathrm{f}}_{l,g})$ and $\mathbf{R}^{\mathrm{q}}_{l,g} = \mathbf{R}(\mathbf{X}^{\mathrm{q}}_{l,g})$. FCC aligns these matrices using

$$\mathcal{L}_{\mathrm{FCC}} = \sum_{l,g} \frac{1}{C_g(C_g - 1)} \left\| \mathbf{M}_g \odot \left(\mathbf{R}^{\mathrm{f}}_{l,g} - \mathbf{R}^{\mathrm{q}}_{l,g}\right) \right\|_F^2, \tag{10}$$

where $\mathbf{M}_g = \mathbf{1}\mathbf{1}^\top - \mathbf{I}_{C_g}$ masks the diagonal. The standardization removes the feature mean and scale before relation measurement. FCC therefore preserves the overall relation structure without

allowing feature magnitude to dominate the Optimization objective. Because the matrices are constructed separately for every (l, g), the loss is sensitive to relations that are coupled by the ACIM mapping. The diagonal entries of a correlation matrix are fixed by normalization; the off diagonal entries carry the pairwise information that distinguishes one group representation from another. The normalization by $C_g(C_g - 1)$ averages exactly these off-diagonal entries and keeps the FCC scale comparable across mapping groups of different sizes.

3.4 Correlation Aware Quantization Optimization

FCC preserves group-wise relation consistency, but its matrix-level objective treats all relation entries equally. We therefore introduce CQO to assign feature-wise reconstruction weights according to the relation importance in the floating-point model and the relation deviation introduced by quantization.

For the i-th feature in group g of layer l, we define the relation participation as

$$p_{l,g,i} = \frac{1}{C_g - 1} \sum_{j \neq i} \left| [\mathbf{R}^{\mathrm{f}}_{l,g}]_{ij} \right|, \tag{11}$$

which measures the contribution of feature i to the floating-point relation structure.

The relation distortion after quantization is defined as

$$d_{l,g,i} = \frac{1}{C_g - 1} \sum_{j \neq i} \left| [\mathbf{R}^{\mathrm{f}}_{l,g} - \mathbf{R}^{\mathrm{q}}_{l,g}]_{ij} \right|, \tag{12}$$

where $\mathbf{R}^{\mathrm{f}}_{l,g}$ and $\mathbf{R}^{\mathrm{q}}_{l,g}$ denote the relation matrices of floating-point and quantized features, respectively.

The two scores are combined to obtain the feature importance:

$$w_{l,g,i} = C_g \frac{\exp(\mathrm{sg}(p_{l,g,i} d_{l,g,i})/\tau)}{\sum_{k=1}^{C_g} \exp(\mathrm{sg}(p_{l,g,k} d_{l,g,k})/\tau)}, \tag{13}$$

where τ is the temperature coefficient and $\mathrm{sg}(\cdot)$ denotes stop-gradient. The scaling factor C_g preserves the average weight within each group.

Given the channel-wise reconstruction error

$$\ell_{l,g,i} = \frac{1}{N} \left\| \mathbf{x}^{\mathrm{f}}_{l,g,i} - \mathbf{x}^{\mathrm{q}}_{l,g,i} \right\|_2^2, \tag{14}$$

CQO minimizes the weighted reconstruction objective:

$$\mathcal{L}_{\mathrm{CQO}} = \sum_{l,g} \frac{1}{C_g} \sum_{i=1}^{C_g} w_{l,g,i} \ell_{l,g,i}. \tag{15}$$

Therefore, CQO emphasizes features that are highly involved in the original relation structure and suffer larger relation changes after quantization. The weights are computed from the current calibration batch and updated with the quantization states.

3.5 Relation-guided PTQ Optimization

After introducing FCC and CQO, we integrate them into a unified relation-guided PTQ optimization framework. The objective is to jointly optimize numerical reconstruction and feature correlation preservation, enabling quantized representations to maintain both response fidelity and structural consistency under groupwise ACIM execution. Specifically, the reconstruction objective and relation-aware optimization are balanced through:

$$\mathcal{L}_{\mathrm{feat}} = (1 - \alpha)\mathcal{L}_{\mathrm{rec}} + \alpha \mathcal{L}_{\mathrm{CQO}}, \qquad 0 \leq \alpha \leq 1. \tag{16}$$

The additional relation objective is formulated as

$$\mathcal{L}_{\mathrm{rel}} = \lambda_{\mathrm{c}} \mathcal{L}_{\mathrm{FCC}} + \alpha(\mathcal{L}_{\mathrm{CQO}} - \mathcal{L}_{\mathrm{rec}}). \tag{17}$$

Since relation preservation and numerical reconstruction may dominate at different optimization stages, we introduce a scheduled relation guidance strategy. The relation constraint is activated during the middle optimization stage, allowing the quantized model to first recover basic numerical responses and then refine structural consistency. Let $u = t/T$ denote normalized optimization progress, and $u_{\mathrm{s}}, u_{\mathrm{e}}$ define the activation interval:

$$\rho(u) = \begin{cases} \sin\left(\pi \frac{u-u_{\mathrm{s}}}{u_{\mathrm{e}}-u_{\mathrm{s}}}\right), & u_{\mathrm{s}} \le u \le u_{\mathrm{e}}, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

During optimization, the contribution of relation guidance is adaptively controlled to maintain the balance between reconstruction fidelity and correlation preservation. This strategy prevents relation optimization from dominating the reconstruction objective while ensuring effective structural refinement.

The proposed relation-guided optimization is performed during the offline PTQ stage. Once optimization is completed, FCPQ retains only the optimized quantization parameters, enabling direct deployment on low-bit ACIM hardware without additional inference modules.

4 Experiments

4.1 Experimental Settings

Dataset and models. We evaluate FCPQ on CIFAR-10 (Krizhevsky, 2009) and ImageNet-1K (Deng et al., 2009). On CIFAR-10, we compare FCPQ with matched PTQ baselines. On ImageNet-1K, we evaluate the same cross-model and cross-precision performance and conduct the detailed relation-preservation, group-size, ablation, calibration-cost, and ACIM analyses. We use ResNet-18 and ResNet-50 (He et al., 2016), MobileNetV2 (Sandler et al., 2018), and RepVGG-A0 (Ding et al., 2021); these models cover residual, mobile, and structurally re-parameterized execution graphs. The full-precision checkpoints are kept fixed during calibration, and all methods use the same checkpoint and evaluation pipeline.

Training and Quantization. For every model, method, bit setting, the FP32 pretrained weights remain fixed. The first and last layers use 8-bit quantization and each ACIM mapping group contains 32 output features. Each dataset uses a calibration subset of 1,024 randomly selected training images and an independent 4,096-image training diagnostic set; the complete CIFAR-10 test set (10,000 images) and ImageNet-1K validation set (50,000 images) are used only for final evaluation. All trainable methods use 2,000 Adam iterations with batch size 64 and learning rate 10^{-2}. Min–Max, AdaRound (Nagel et al., 2020), BRECQ (Li et al., 2021), BOF4 (Blumenberg et al., 2026), and AnyBCQ (Park et al., 2026) are re-evaluated under this matched protocol.

Metrics and Comparison Protocol. We report Top-1 accuracy on the complete test or validation split, together with reconstruction error, groupwise correlation distortion, calibration time, and peak calibration memory. Reconstruction error is computed between FP32 and quantized features, while correlation distortion is the Frobenius distance between their standardized groupwise correlation matrices. We additionally report group-size robustness, component ablations, CQO weighting diagnostics, calibration-resource cost, and ACIM nonideality stress tests. All methods are compared under the same model checkpoints, bit widths, mapping-group size, calibration budget, seed set, calibration/diagnostic indices, loader order, and evaluation split; method-specific objectives and parameterizations remain unchanged. Time and memory are measured on the same GPU with data loading excluded. Results from pending configurations are omitted.

4.2 Main Results

Results on ImageNet-1K. Table 1 evaluates the same protocol on ImageNet-1K. FCPQ gives the best result on 12 model–bit settings and improves the strongest reference by up to 0.53 points. The improvement is consistent across residual, mobile, and re-parameterized architectures, which indicates that the groupwise relation objective does not depend on a single block design. At W2A4,

Table 1: ImageNet-1K Top-1 accuracy (%) under groupwise ACIM quantization with group size 32. W and A denote weight and activation bits, respectively; first and last layers use 8 bits for all methods.

Model	**FP32**	**Bit**	**Min-Max**	**AdaRound**	**BRECQ**	**BOF4**	**AnyBCQ**	**FCPQ**
Source	–	–	–	ICML'20	ICLR'21	ICLR'26	ICLR'26	Ours
ResNet-18	71.47	W4A4	67.24	68.12	69.07	69.26	69.35	**69.41**
ResNet-18	71.47	W3A3	62.64	64.73	65.91	66.07	66.21	**66.42**
ResNet-18	71.47	W2A4	52.61	55.76	58.03	58.07	58.17	**58.49**
ResNet-50	76.13	W4A4	73.14	74.52	75.13	75.20	**75.21**	75.20
ResNet-50	76.13	W3A3	69.62	71.36	72.47	72.58	72.49	**72.80**
ResNet-50	76.13	W2A4	62.01	64.92	67.04	67.42	67.81	**68.34**
MobileNetV2	72.00	W4A4	66.59	67.71	68.25	69.29	69.18	**69.33**
MobileNetV2	72.00	W3A3	59.86	61.37	63.08	64.21	64.24	**64.31**
MobileNetV2	72.00	W2A4	48.16	50.32	51.11	53.02	54.57	**54.80**
RepVGG-A0	72.41	W4A4	66.46	68.04	68.79	69.28	69.25	**69.44**
RepVGG-A0	72.41	W3A3	61.22	62.19	63.76	65.07	64.93	**65.15**
RepVGG-A0	72.41	W2A4	49.78	52.05	54.10	55.03	55.52	**55.81**

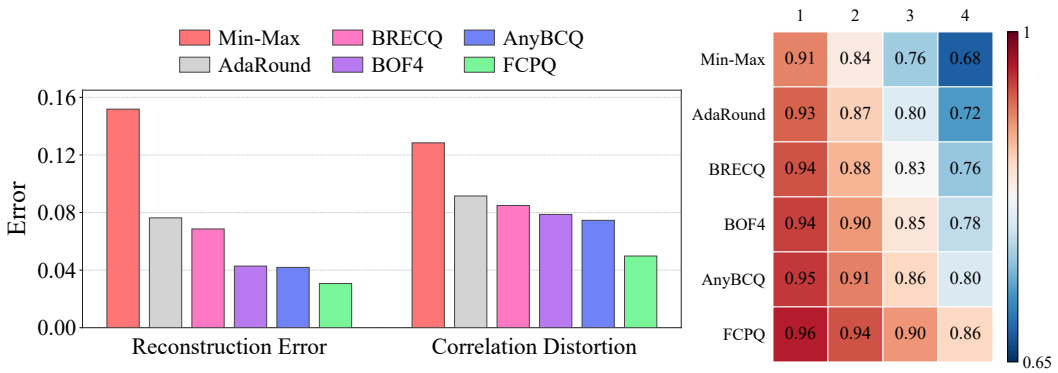
Table 2: CIFAR-10 Top-1 accuracy (%) under groupwise ACIM quantization with group size 32. W and A denote weight and activation bits, respectively; first and last layers use 8 bits for all methods.

Model	**FP32**	**Bit**	**Min-Max**	**AdaRound**	**BRECQ**	**BOF4**	**AnyBCQ**	**FCPQ**
Source	–	–	–	ICML'20	ICLR'21	ICLR'26	ICLR'26	Ours
ResNet-18	94.76	W4A4	93.21	93.64	93.78	93.85	93.92	**94.13**
ResNet-18	94.76	W3A3	90.12	91.05	91.33	91.57	91.74	**92.26**
ResNet-18	94.76	W2A4	82.47	84.22	84.86	85.31	85.73	**86.41**
ResNet-50	95.28	W4A4	93.84	94.17	94.31	94.40	**94.52**	94.48
ResNet-50	95.28	W3A3	91.03	91.86	92.14	92.37	92.55	**93.04**
ResNet-50	95.28	W2A4	84.15	85.83	86.42	86.87	87.24	**87.96**
MobileNetV2	94.12	W4A4	92.06	92.48	92.65	92.77	92.89	**93.15**
MobileNetV2	94.12	W3A3	88.24	89.37	89.72	89.96	90.21	**90.73**
MobileNetV2	94.12	W2A4	78.62	80.71	81.35	81.84	82.27	**83.02**
RepVGG-A0	94.35	W4A4	92.33	92.76	92.91	**93.08**	93.01	92.97
RepVGG-A0	94.35	W3A3	88.75	89.82	90.16	90.43	90.64	**91.21**
RepVGG-A0	94.35	W2A4	79.46	81.53	82.17	82.64	83.08	**83.85**

the gain is largest because activation quantization amplifies the coupled distortion among features assigned to one ACIM mapping group.

Results on CIFAR-10. Table 2 reports the matched W4A4 comparison on CIFAR-10. FCPQ improves over Min–Max by 0.92–5.39 points across the three bit settings and remains competitive with the strongest reconstruction baseline at W4A4. The gains are larger at W3A3 and W2A4, where the same groupwise perturbation affects a larger fraction of the representational range. These results are consistent with the motivation of relation-aware PTQ: preserving feature dependencies provides a complementary signal when elementwise reconstruction becomes less reliable.

Relation Preservation Analysis. Figure 2(a) shows that reconstruction-based methods reduce feature error but retain high relation error, whereas FCPQ achieves the lowest errors in both. Table 3(a) quantifies this benefit: FCPQ reduces correlation distortion by 33.2% and reconstruction error by 26.7% relative to AnyBCQ. Figure 2(b) further evaluates representation similarity using centered kernel alignment (CKA) (Kornblith et al., 2019). FCPQ maintains the highest CKA across all four layers, with its advantage over AnyBCQ increasing from 0.013 at the first layer to 0.063 at the fourth. This trend shows that FCPQ better preserves the floating-point representations in deeper layers. Together, these results support combining FCC's direct relation constraint with CQO's relation-guided reconstruction to preserve both feature values and feature structure under groupwise quantization.



(a) Feature and relation errors (b) CKA similarity

Figure 2: Feature relation preservation on ResNet-18/ImageNet-1K under W4A4 quantization with group size 32. (a) Feature reconstruction error and correlation distortion (lower is better). (b) Centered kernel alignment (CKA) (Kornblith et al., 2019) between floating-point and quantized features at four layers (higher is better).

4.3 Group-Size Robustness

The ACIM mapping-group size determines the number of features sharing quantization parameters. We evaluate group sizes of 16, 32, and 64 on ResNet-18/ImageNet-1K under W4A4 quantization with fixed calibration settings. As shown in Table 3(b), FCPQ maintains stable performance across different group sizes, with only a 0.05% accuracy drop from group size 16 to 64, while other methods degrade by 0.39–1.61%. At group size 64, FCPQ achieves a 0.37% improvement over the strongest baseline, demonstrating better robustness under larger ACIM mapping groups.

Table 3: Relation distortion analysis and group-size robustness on ResNet-18/ImageNet-1K under W4A4 quantization. (a) Accuracy and relation distortion metrics with group size 32. (b) Accuracy comparison under different mapping-group sizes.

(a) Accuracy and distortion diagnostics

Method	Top-1 ↑	Rec. err. ↓	Corr. dist. ↓
Min-Max	67.24	0.1518	0.1284
AdaRound	68.12	0.0763	0.0915
BRECQ	69.07	0.0686	0.0849
BOF4	69.26	0.0428	0.0787
AnyBCQ	69.35	0.0419	0.0746
FCPQ	**69.41**	**0.0307**	**0.0498**

(b) Mapping-group robustness

Method	Group 16	Group 32	Group 64
Min-Max	67.52	67.24	65.91
AdaRound	68.43	68.12	67.36
BRECQ	69.21	69.07	68.52
BOF4	69.39	69.26	68.84
AnyBCQ	69.41	69.35	69.02
FCPQ	**69.44**	**69.41**	**69.39**

4.4 Ablation Study

Ablation Settings. We evaluate FCPQ variants on ResNet-18/ImageNet-1K at W4A4 with group size 32 and a fixed calibration budget. Starting from reconstruction with adaptive rounding, we add FCC and CQO separately and then combine them. To isolate the CQO weighting signals, we retain FCC and compare uniform weights, participation only, distortion only, and their product.

Component Analysis. Table 4(a) supports the complementary roles of FCC and CQO. Each component independently improves accuracy and correlation preservation over adaptive reconstruction. FCC directly constrains pairwise dependencies after removing feature mean and scale; CQO uses these relations to determine which feature reconstruction errors receive greater emphasis. Adding CQO to FCC further reduces correlation distortion by 27.5%, showing that adaptive reconstruction priorities remain useful even with explicit relation alignment. Together, the components improve accuracy by 0.66 percentage points over the baseline. These gains support coupling a groupwise structural constraint with feature-specific reconstruction weights.

CQO Weighting Analysis. Table 4(b) supports using both participation and distortion to allocate reconstruction weights. Participation measures the strength of a feature's original relations, but does not indicate whether they require correction. Distortion identifies disrupted relations, but does not account for their original strength. Their product incorporates both criteria and reduces correlation distortion by a further 11.5% relative to distortion-only weighting, the stronger individual signal. The accompanying improvement in accuracy supports prioritizing features with both strong original relations and large quantization-induced changes, as intended by CQO.

Table 4: Ablation studies of FCPQ components and CQO weighting strategies on ResNet-18/ImageNet-1K under W4A4 quantization with group size 32. (a) Component contributions of FCC and CQO. (b) Effects of different CQO weighting signals.

(a) Component contributions

Objective	Top-1 ↑	Corr. ↓
Adaptive recon.	68.75	0.0946
+ FCC	69.18	0.0687
+ CQO	69.03	0.0731
+ FCPQ	**69.41**	**0.0498**

(b) CQO weighting signals

Weighting signal	Top-1 ↑	Corr. dist. ↓
Uniform	69.18	0.0687
Participation only	69.29	0.0596
Distortion only	69.36	0.0563
Participation × distortion	**69.41**	**0.0498**

4.5 Hardware Evaluation

We evaluate FCPQ with a fabrication-ready, post-layout SRAM-based ACIM macro implemented in a TSMC 22-nm process. The macro operates at a nominal supply of 0.8 V and a clock frequency of 100 MHz. Table 5 summarizes the measured Top-1 accuracy of different quantization schemes under practical chip operating conditions. Conductance variation and drift change the effective stored weights, while limited ADC precision and accumulation noise alter the measured outputs. IR drop changes voltages within an array and introduces further error in its computation (Jain et al., 2020b;a; Kim et al., 2022).

FCPQ achieves the highest accuracy under every tested condition. Under accumulation noise, its accuracy drops by 0.64 points from the noiseless setting, compared with 2.03 points for AnyBCQ. The advantage also increases under drift and IR drop, showing that the benefit extends across different sources of ACIM error. This indicates that preserving feature relations during quantization improves tolerance to practical circuit-level variations.

Table 5: Measured Top-1 accuracy (%) on the fabricated SRAM-based ACIM macro implemented in a TSMC 22-nm process with W4A4 quantization.

Condition	Min-Max	AnyBCQ	FCPQ
Nominal (0.8V, 100MHz)	67.24	69.35	**69.41**
Low supply (0.75V)	66.71	68.79	**68.98**
Reduced ADC resolution	65.82	68.21	**68.54**
Accum. noise, $\sigma = 2\%$ FS	65.24	67.32	**68.77**
IR drop, 10%	65.91	67.65	**68.42**
Drift, 5%	65.28	67.36	**68.63**

5 Conclusion

We propose FCPQ, a relation-preserving PTQ framework for low-precision ACIM models, which jointly optimizes feature reconstruction and correlation preservation during quantization. First, we introduce FCC to preserve the feature relations between floating-point and quantized representations. Second, we design CQO to assign optimization priorities according to feature relation characteristics, further improving quantization quality. Extensive experiments demonstrate the effectiveness of FCPQ in preserving representations and improving low-bit model performance. Ablation studies further verify the complementary contributions of FCC and CQO. In the future, we will explore the extension of FCPQ to more network architectures and ACIM hardware platforms.

Reproducibility Statement

The quantization objective, rounding relaxation and relation schedule are defined in Section 3, and Section 4 specifies the datasets, models, quantization settings, calibration and diagnostic splits and evaluation metrics.

AI Use Statement

Generative AI tools assisted with language editing, figure organization, and consistency checks. The authors are responsible for the manuscript and its reported results.

References

Aayush Ankit, Izzat El Hajj, Sai Rahul Chalamalasetti, Geoffrey Ndu, Martin Foltin, R. Stanley Williams, Paolo Faraboschi, Wen-mei W. Hwu, John Paul Strachan, and Kaushik Roy. Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 715–731, 2019.

Patrick Blumenberg, Thomas Graave, and Tim Fingscheidt. Improving block-wise llm quantization by 4-bit block-wise optimal float (bof4): Analysis and variations. In *Proceedings of the International Conference on Learning Representations*, volume 2026, pp. 6796–6827, 2026.

Xuan-Jun Chen, Han-Ping Chen, and Chia-Lin Yang. Pointcim: A computing-in-memory architecture for accelerating deep point cloud analytics. In *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pp. 1309–1322, 2024.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13733–13742, 2021.

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

Lixia Han, Renjie Pan, Zheng Zhou, Hairuo Lu, Yiyang Chen, Haozhang Yang, Peng Huang, Guangyu Sun, Xiaoyan Liu, and Jinfeng Kang. Comn: algorithm-hardware co-design platform for nonvolatile memory-based convolutional neural network accelerators. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(7):2043–2056, 2024.

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

Shubham Jain, Abhronil Sengupta, Kaushik Roy, and Anand Raghunathan. Rxnn: A framework for evaluating deep neural networks on resistive crossbars. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(2):326–338, 2020a.

Shubham Jain, Abhronil Sengupta, Kaushik Roy, and Anand Raghunathan. Rxnn: A framework for evaluating deep neural networks on resistive crossbars. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(2):326–338, 2020b.

Joonhyung Kim, Kyeongho Lee, and Jongsun Park. A charge domain p-8t sram compute-in-memory with low-cost dac/adc operation for 4-bit input processing. In *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 1–6, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning*, pp. 3519–3529. PMlR, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Jiawei Liu, Lin Niu, Zhihang Yuan, Dawei Yang, Xinggang Wang, and Wenyu Liu. Pd-quant: Post-training quantization based on prediction difference metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24427–24437, 2023.

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

Gunho Park, Jeongin Bae, Beomseok Kwon, Byeongwook Kim, Se Jung Kwon, and Dongsoo Lee. Anybcq: Hardware efficient flexible binary-coded quantization for multi-precision llms. In *Proceedings of the International Conference on Learning Representations*, volume 2026, pp. 125106–125122, 2026.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R. Stanley Williams, and Vivek Srikumar. Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In *Proceedings of the International Symposium on Computer Architecture*, pp. 14–26, 2016.

Fengbin Tu, Yiqi Wang, Zihan Wu, Ling Liang, Yufei Ding, Bongjin Kim, Leibo Liu, Shaojun Wei, Yuan Xie, and Shouyi Yin. Redcim: Reconfigurable digital computing-in-memory processor with unified fp/int pipeline for cloud ai acceleration. *IEEE Journal of Solid-State Circuits*, 58(1): 243–255, 2022.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Naveen Verma, Hongyang Jia, Hossein Valavi, Yinqi Tang, Murat Ozatay, Lung-Yen Chen, Bonan Zhang, and Peter Deaville. In-memory computing: Advances and prospects. *IEEE Solid-State Circuits Magazine*, 11(3):43–55, 2019.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8612–8620, 2019.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *Proceedings of the International Conference on Learning Representations*, 2022.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

APPENDIX

This supplementary material provides additional experimental details and results that are not included in the main paper due to space limitations. The contents are organized as follows:

- Section A introduces the ACIM macro and its implementation characteristics used for hardware evaluation.
- Section B analyzes the representation transferability of quantized models through downstream evaluation.

A HARDWARE EVALUATION

As shown in Table S1, the hardware experiment uses a 22-nm SRAM-based ACIM macro for W4A4 execution. The macro combines input DACs, a memory array, 4-bit SAR ADCs with programmable CDAC references, and a shuffle buffer. The physical array is 154×64: 144 rows are used for the 3×3 MAC of 16 input channels and 10 rows provide analog bias injection. Each array produces 16 output channels.

Table S1: Circuit-level implementation characteristics of the target W4A4 ACIM macro.

Metric	**Value**
Process / supply / clock	22 nm / 0.8 V / 100 MHz
Weight / activation precision	W4A4
Array organization	154×64
Mapped convolution	3×3, 16 input $\times$ 16 output channels
MAC rows / offset rows	144 / 10
Output conversion	4-bit SAR ADC with programmable CDAC
Macro area (mm^2) $\downarrow$	0.03
Macro power (mW) $\downarrow$	0.0724
Nominal DNL / INL (LSB) $\downarrow$	0.005 / 0.012
Peak energy efficiency (TOPS/W) $\uparrow$	1360

B REPRESENTATION TRANSFERABILITY ANALYSIS

Feature relations encode the structural organization of learned representations and can affect the generalization ability of quantized models. To evaluate whether the relation preservation of FCPQ translates into more transferable representations, we perform a downstream linear probing study.

We start from an ImageNet-1K pretrained ResNet-18 model and conduct W4A4 PTQ using different quantization methods. After quantization, the feature extractor and quantization parameters are fixed, while the original 1000-way classifier is replaced with a randomly initialized linear classifier for CIFAR-10. Only the linear classifier is optimized during downstream training, which evaluates the quality of the frozen quantized representations rather than additional adaptation ability.

To ensure a fair comparison, all methods use the same ImageNet calibration set and optimization settings during PTQ. The calibration data are not reused for representation evaluation, and the CIFAR-10 test set is only used after validation-based model selection.

As shown in Table S2, feature relation distortion is closely correlated with representation similarity after quantization. Methods with lower correlation distortion generally achieve higher CKA similarity, indicating better preservation of the feature structure learned by the full-precision model. Although reconstruction-based methods effectively reduce numerical errors, they do not explicitly constrain feature dependencies. For example, AnyBCQ obtains competitive reconstruction error but still exhibits larger correlation distortion than FCPQ, leading to inferior representation similarity. By introducing relation-aware optimization, FCPQ achieves the highest CKA score, demonstrating its ability to preserve the intrinsic organization of quantized features.

Table S2: Representation preservation and downstream transfer evaluation after W4A4 PTQ on ResNet-18/ImageNet-1K. Cal. and Diag. denote calibration and independent diagnostic sets, respectively. The frozen quantized backbone is evaluated by CIFAR-10 linear probing.

Method	**Cal. rec.** ↓	**Diag. rec.** ↓	**Cal. corr.** ↓	**Diag. corr.** ↓	**CKA** ↑	**CIFAR-10 Top-1** ↑
FP32	–	–	–	–	1.000	95.42
Min-Max	0.1518	0.1635	0.1284	0.1369	0.798	92.84
AdaRound	0.0763	0.0837	0.0915	0.0978	0.830	93.36
BRECQ	0.0686	0.0742	0.0849	0.0906	0.853	93.71
BOF4	0.0428	0.0476	0.0787	0.0841	0.868	93.79
AnyBCQ	0.0419	0.0468	0.0746	0.0795	0.880	93.96
FCPQ	**0.0307**	**0.0345**	**0.0498**	**0.0527**	**0.916**	**94.56**

Table S3: Downstream transfer performance and representation diagnostics after W4A4 PTQ. The backbone is frozen after ImageNet PTQ, and only a linear classifier is trained on CIFAR-10.

Method	**Cal. rec.** ↓	**Diag. rec.** ↓	**Cal. corr.** ↓	**Diag. corr.** ↓	**CKA** ↑	**CIFAR-10 Top-1** ↑
FP32	–	–	–	–	1.000	95.42
Adaptive recon.	0.0686	0.0762	0.0946	0.1028	0.785	92.84
+ FCC	0.0539	0.0607	0.0687	0.0724	0.869	94.01
+ CQO	0.0417	0.0468	0.0731	0.0792	0.843	93.76
FCPQ	**0.0307**	**0.0345**	**0.0498**	**0.0527**	**0.916**	**94.56**

To further evaluate whether the preserved representation structure benefits practical transferability, we perform a downstream linear probing experiment on CIFAR-10 with frozen ImageNet W4A4 backbones. As shown in Table S3, methods with higher representation similarity achieve better downstream performance. FCC improves transfer accuracy by preserving feature relations, while CQO further enhances the optimization by assigning larger priorities to relation-sensitive features. Combining both components, FCPQ achieves the highest CKA similarity and downstream accuracy, indicating that relation-aware quantization maintains more transferable representations beyond minimizing reconstruction error.

Overall, these results validate our motivation that feature relation degradation is an important factor affecting low-bit ACIM quantization. Preserving feature relations provides complementary benefits to conventional reconstruction-based optimization and improves the quality of quantized representations.