QUANTIZATION-AWARE PRECONDITIONING FOR PHYSICS-INFORMED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Ill-conditioned physics-informed neural networks (PINNs) residual objectives can slow or destabilize optimization; residual preconditioning reshapes this geometry, while quantization reduces memory and arithmetic cost. These two design choices are usually treated independently, even though the same residual transform acts on both the physical error and the perturbations induced by low-bit weights and activations. We introduce a quantization-aware framework for residual preconditioning in physics-informed neural networks, in which the preconditioner is designed jointly from the physical tangent geometry and measured quantization-induced residual statistics. The resulting quantization-aware preconditioning (QAP) framework combines isotropic and covariance-aware regularization with scalable low-rank estimates of the quantization geometry. Our analysis separates optimization contraction from quantization injection and shows why a preconditioner that is effective in full precision can become ineffective after quantization. Across ten-seed end-to-end quantization-aware training (QAT) experiments, QAP reduces deployed solution error by up to $2.07\times$ on heterogeneous one-dimensional diffusion and by $1.45\times$ on a $1000{:}1$ two-dimensional checkerboard-diffusion problem relative to no preconditioning; in the two-dimensional setting, the stabilized directional variant also outperforms isotropic QAP. These results identify quantization-induced residual geometry as a preconditioner-design variable. On the evaluated diffusion problems, QAP improves deployed solution accuracy over quantized PINNs trained without preconditioning, at the same weight and activation bitwidths.

1 INTRODUCTION

Physics-informed neural networks (PINNs) approximate PDE solutions by minimizing residuals of governing equations together with boundary or initial conditions (Raissi et al., 2019; Karniadakis et al., 2021). Their appeal is strongest when differentiable surrogates, inverse problems, irregular observations, or repeated downstream evaluations matter. Their optimization can nevertheless be stiff and spectrally imbalanced, producing long plateaus, oscillatory training, and large solution error even when the residual loss is small (Wang et al., 2021; Krishnapriyan et al., 2021; Wang et al., 2022).

Residual preconditioning changes the geometry of this optimization problem. If $r_\theta \in \mathbb{R}^m$ is the vector of PDE residuals evaluated at collocation points and $J = \partial r_\theta / \partial \theta$ is its parameter Jacobian, the unpreconditioned local curvature is governed by $J^\top J$. A residual transform P replaces it by $J^\top P^\top P J$, allowing weak and stiff residual directions to be rebalanced. Operator-preconditioning analyses make this connection explicit (De Ryck et al., 2024), and condition-number-based PINN preconditioning has been linked to both error control and convergence (Liu et al., 2024). In parallel, quantization-aware training (QAT) simulates low-bit weights and activations during optimization to obtain efficient deployed networks (Jacob et al., 2018; van den Dool et al., 2023).

These mechanisms interact because the same residual transform acts on physical error and low-bit error. Let $r_{\theta,q} = r_\theta + \xi_q$ be the residual of the quantized network on the same collocation points, where $\xi_q \in \mathbb{R}^m$ is the quantization-induced residual perturbation. For $P \in \mathbb{R}^{m\times m}$,

$$Pr_{\theta,q} = Pr_\theta + P\xi_q. \tag{1}$$

Aggressive equalization of a weak physical mode can therefore amplify the low-bit perturbation carried by that mode. The relevant design problem is to improve physical conditioning while

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

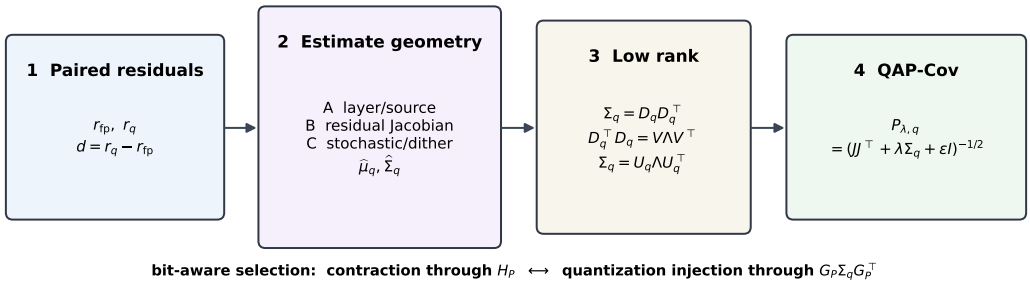


Figure 1: **QAP mechanism and covariance construction.** Full-precision and low-bit residuals define perturbation samples. Their geometry can be estimated by layer/source decomposition, first-order Jacobian propagation, or repeated stochastic quantization, then compressed as a low-rank factor. QAP uses this geometry inside a noise-aware inverse so that the same preconditioner jointly controls tangent contraction and quantization injection.

controlling this additional injection. Throughout this paper, *precision* means arithmetic bitwidth of weights, activations, or a stored preconditioner. Numerical approximation parameters such as an ILU drop tolerance instead control how accurately a structured preconditioner approximates an operator (Liu et al., 2024); these are separate design variables.

The resulting quantization-aware preconditioning (QAP) framework models two effects of P: the preconditioned tangent geometry that governs local contraction and the residual-perturbation geometry that governs low-bit injection. Learned and hybrid preconditioners can also be interpreted through error-propagation operators and complementary spectral corrections (Kopaničáková & Karniadakis, 2024); QAP retains this mode-aware viewpoint while making the quantization-induced residual statistics explicit.

Contributions. We make five contributions.

1. We formulate **quantization-aware preconditioning** for PINNs and derive local QAT dynamics that separate contraction from the injection of quantization-induced residual perturbations.
2. We introduce a constructive **covariance-aware regularization** principle and prove a mode-wise conditioning–noise tradeoff together with a bit-dependent crossover between aggressive and conservative preconditioning.
3. We develop three ways to construct the low-bit residual geometry: **layer/source decomposition**, a **first-order Jacobian construction**, and repeated stochastic/dithered quantization; we also give a matrix-free **low-rank eigendecomposition** that avoids dense residual covariances.
4. We separate centered quantization variance from persistent bias, analyze quantization of the preconditioner itself, and give residual-space and tangent-space criteria for selecting the regularization without ground-truth PDE solutions.
5. We validate QAP in controlled, PDE-calibrated, and **end-to-end QAT** experiments spanning heterogeneous one-dimensional diffusion and a 1000:1 two-dimensional checkerboard problem. A reliability-gated low-rank realization stabilizes directional covariance information in the larger residual space and improves over no preconditioning, full-precision-derived preconditioning, and isotropic QAP.

2 Related Work

Training pathologies and residual preconditioning in PINNs. PINN optimization failures have been connected to loss imbalance, gradient-flow pathologies, frequency bias, and tangent-spectrum imbalance (Wang et al., 2021; 2022; Krishnapriyan et al., 2021). De Ryck et al. (2024) develop an operator-preconditioning perspective. Liu et al. (2024) relate condition number to PINN error control and convergence and apply incomplete-LU (ILU)-type transforms to discretized residual systems;

their drop-tolerance studies vary the approximation fidelity of the numerical preconditioner. QAP adds arithmetic bitwidth and the resulting residual-perturbation statistics to the preconditioner design criterion.

Learned and hybrid preconditioners. Kopaničáková & Karniadakis (2024) construct DeepONet-based hybrid preconditioners that combine learned low-frequency correction with classical smoothing and analyze convergence through an error-propagation operator. QAP permits analogous structured or learned residual transforms, but evaluates them under the tangent geometry of the quantized PINN and its measured low-bit residual perturbations.

Quantization of neural PDE solvers. QAT simulates low-bit arithmetic during training, commonly using straight-through estimators (STEs) (Jacob et al., 2018; Bengio et al., 2013), and neural PDE solvers can retain useful accuracy under QAT (van den Dool et al., 2023). In diffusion models, Q-Sched adapts scheduling coefficients to quantization (Frumkin & Marculescu, 2025); SmoothQuant, QuaRot, and SpinQuant likewise show that transformations adapted to the quantized regime can improve low-bit behavior (Xiao et al., 2023; Ashkboos et al., 2024; Liu et al., 2025). In statistical inverse problems, noise covariance can enter whitening or preconditioning transforms (Calvetti, 2007). QAP uses quantization-generated *PDE residual* geometry to regularize a physics preconditioner jointly for contraction and low-bit injection.

Notation used in the analysis. The residual tangent Gram matrix is $K_r = JJ^\top \in \mathbb{R}^{m\times m}$. For a fixed low-bit regime, $\mu_q = \mathbb{E}[\xi_q] \in \mathbb{R}^m$ is the mean residual perturbation, $\Sigma_q = \mathbb{E}[(\xi_q - \mu_q)(\xi_q - \mu_q)^\top] \in \mathbb{R}^{m\times m}$ is its centered covariance, and $M_q = \Sigma_q + \mu_q\mu_q^\top$ is its second moment. For a residual preconditioner P, H_P denotes the preconditioned tangent operator and G_P the residual-to-parameter injection map; both are defined explicitly in section 4.1.

3 Problem Setup

3.1 Physics-informed residual minimization

Let Ω be the PDE domain, μ an optional vector of physical parameters, $\mathcal{D}_\mu$ the interior differential operator, and $\mathcal{B}_\mu$ the boundary (or initial-condition) operator. The target field u satisfies

$$\mathcal{D}_\mu u = f_\mu \quad \text{in } \Omega, \qquad \mathcal{B}_\mu u = g_\mu \quad \text{on } \partial\Omega, \tag{2}$$

where f_μ and g_μ are known source and boundary data. A PINN u_θ, with trainable parameters $\theta \in \mathbb{R}^{d_\theta}$, is evaluated at interior collocation points $X_r = \{x_j\}_{j=1}^m$. Its physics-residual vector $r_\theta(\mu) \in \mathbb{R}^m$ is defined componentwise by $[r_\theta(\mu)]_j = \mathcal{D}_\mu u_\theta(x_j) - f_\mu(x_j)$. For boundary points $X_b = \{y_k\}_{k=1}^{n_b}$, $\mathcal{L}_{\mathrm{bc}}(\theta;\mu)$ denotes the boundary/initial-condition mismatch, e.g., $\frac{\lambda_{\mathrm{bc}}}{2n_b}\sum_k |\mathcal{B}_\mu u_\theta(y_k) - g_\mu(y_k)|^2$, where $\lambda_{\mathrm{bc}} \geq 0$ is the boundary-loss weight; it is zero when the condition is imposed exactly by the network ansatz. When μ is fixed, we suppress it from r_θ, P_μ, and related notation.

A residual preconditioner $P_\mu \in \mathbb{R}^{m\times m}$ acts on the interior residual before its squared norm is formed:

$$\mathcal{L}_P(\theta;\mu) = \frac{1}{2}\|P_\mu r_\theta(\mu)\|_2^2 + \mathcal{L}_{\mathrm{bc}}(\theta;\mu). \tag{3}$$

This is the meaning of *preconditioning* in this work. Around a reference parameter $\theta^\star$, let $e = \theta - \theta^\star$ and $J = \partial r_\theta/\partial\theta|_{\theta^\star} \in \mathbb{R}^{m\times d_\theta}$. The local residual is $r_\theta \approx Je$. With $P = I$ and learning rate $\eta > 0$, gradient descent gives $e^+ \approx (I - \eta J^\top J)e$: the largest singular value of J restricts the stable step size, while small-singular-value modes contract slowly. Preconditioning replaces $J^\top J$ by $J^\top P^\top P J$ and thereby reshapes the active spectrum (De Ryck et al., 2024; Liu et al., 2024). If P is injective, $Pr_\theta = 0$ if and only if $r_\theta = 0$, so the zero-physics-residual solution set is unchanged. Spectral, diagonal, block, ILU-derived, multilevel, convolutional, and learned transforms all fit this definition; when a numerical method writes $M^{-1}r$, our notation sets $P = M^{-1}$.

3.2 Quantized network and residual perturbation

For each network layer ℓ, let W_ℓ denote its latent full-precision weights and h_ℓ the activation presented to the next layer. For $b \geq 2$, a signed symmetric b-bit quantizer uses $q_{\max} = 2^{b-1} - 1$, clipping scale

$\alpha > 0$, and step $s = \alpha/q_{\max}$:

$$Q_b(x;\alpha) = s\,\mathrm{clip}(\lfloor x/s \rceil, -q_{\max}, q_{\max}). \tag{4}$$

The one-bit case uses the scaled sign quantizer described in Appendix K. With weight and activation bitwidths b_w and b_a, the QAT forward pass uses $\widehat{W}_\ell = Q_{b_w}(W_\ell)$ and $\widehat{h}_\ell = Q_{b_a}(h_\ell)$ while the STE supplies training gradients. Let $r_{\theta,q}(\mu) \in \mathbb{R}^m$ be the physics-residual vector produced by this quantized forward pass on the same collocation set X_r. We define

$$r_{\theta,q} = r_\theta + \xi_q, \qquad \xi_q := r_{\theta,q} - r_\theta, \tag{5}$$

so ξ_q collects the residual-level effect of weight quantization, activation quantization, clipping, and low-bit arithmetic. If the preconditioner is stored at bitwidth b_P, its deployed representation is written $P_q = P + E_P$, where E_P is the storage-quantization error.

4 Theory: Conditioning and Quantization Are Coupled

4.1 Local QAT dynamics

We analyze a neighborhood of $\theta^\star$ and write $e_t = \theta_t - \theta^\star$. Let

$$r_\theta \approx Je_t, \qquad J = \left.\frac{\partial r_\theta}{\partial \theta}\right|_{\theta^\star}. \tag{6}$$

Under the usual STE tangent approximation, the gradient of the quantized preconditioned loss is approximated by

$$\nabla_e \frac{1}{2}\|P(Je+\xi_t)\|_2^2 \approx J^\top P^\top P(Je+\xi_t). \tag{7}$$

Hence gradient descent obeys

$$e_{t+1} = M_P e_t - \eta G_P \xi_t, \quad M_P = I - \eta H_P, \tag{8}$$

where

$$H_P = J^\top P^\top P J, \qquad G_P = J^\top P^\top P. \tag{9}$$

H_P controls local conditioning and contraction; G_P controls how quantized residual errors enter parameter updates. Since H_P is generally singular in overparameterized networks, let $\mathcal{S} = \mathrm{range}(J^\top)$ with projector $\Pi_{\mathcal{S}}$, and let $\kappa^+(H_P)$ denote its positive-spectrum condition number. Contraction statements below are restricted to $\mathcal{S}$.

Assumption 1 (Local quantization model). *Let e_0 be square-integrable and $\mathcal{F}_t = \sigma(e_0, \zeta_0, \ldots, \zeta_{t-1})$. Write $\xi_t = \mu_q + \zeta_t$, where ζ_t is square-integrable, $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$, and $\mathbb{E}[\zeta_t\zeta_t^\top \mid \mathcal{F}_t] \preceq \Sigma_q$ a.s. Both μ_q and Σ_q may depend on bitwidth and on the local training region and are treated as fixed within the local recursion.*

With $m_t = \mathbb{E}[e_t]$, the mean follows $m_{t+1} = M_P m_t - \eta G_P \mu_q$, while the centered error $\widetilde{e}_t = e_t - m_t$ follows the same recursion as (8) with ζ_t. Hence QAP treats Σ_q as the stochastic/incoherent low-bit geometry and reports μ_q separately. For deterministic one-bit or low-bit quantization, we also use the empirical second moment $M_q = \Sigma_q + \mu_q\mu_q^\top$ as a conservative design matrix; the exact bias recursion is given in the appendix.

Theorem 1 (Active-subspace quantized error floor). *Let $\widetilde{e}_t^{\mathcal{S}} = \Pi_{\mathcal{S}}\widetilde{e}_t$ and $M_{P,\mathcal{S}} = M_P|_{\mathcal{S}}$. Suppose $\rho_P = \|M_{P,\mathcal{S}}\|_2 < 1$. Then*

$$\mathbb{E}\|\widetilde{e}_k^{\mathcal{S}}\|_2^2 \le \rho_P^{2k}\mathbb{E}\|\widetilde{e}_0^{\mathcal{S}}\|_2^2 + \eta^2\frac{1-\rho_P^{2k}}{1-\rho_P^2}\mathcal{A}_q(P), \tag{10}$$

where

$$\mathcal{A}_q(P) = \mathrm{tr}\big(G_P\Sigma_q G_P^\top\big). \tag{11}$$

Moreover, $\Pi_{\mathcal{S}^\perp}\widetilde{e}_k = \Pi_{\mathcal{S}^\perp}\widetilde{e}_0$. Consequently,

$$\limsup_{k\to\infty}\mathbb{E}\|\widetilde{e}_k^{\mathcal{S}}\|_2^2 \le \frac{\eta^2\mathcal{A}_q(P)}{1-\rho_P^2}. \tag{12}$$

If, in addition, $\mathbb{E}[\zeta_t\zeta_t^\top \mid \mathcal{F}_t] = \Sigma_q$ a.s. for every t, the active covariance obeys an exact discrete Lyapunov recursion with a unique stationary solution; see Appendix A.1.

The theorem avoids requiring contraction in parameter-null directions and exposes the central tension: improving the active spectrum of H_P can reduce ρ_P while increasing low-bit injection through G_P.

4.2 MODE-RESOLVED QAP

The interaction is especially transparent for a regularized spectral preconditioner. Let $J = U\Sigma V^\top$ with nonzero singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and define

$$P_\tau = (JJ^\top + \tau I)^{-1/2}, \qquad \tau \geq 0, \tag{13}$$

where $\tau = 0$ is understood on $\mathrm{range}(J)$ through the Moore–Penrose inverse square root. Then the nonzero eigenvalues of H_τ are

$$h_i(\tau) = \frac{\sigma_i^2}{\sigma_i^2 + \tau}. \tag{14}$$

At $\tau = 0$, the tangent directions are perfectly equalized; increasing τ makes the preconditioner more conservative.

Proposition 1 (Exact stationary error by physical/tangent mode)**.** *Under Assumption 1, suppose the conditional residual-perturbation covariance is time-invariant, $\mathbb{E}[\zeta_t \zeta_t^\top \mid \mathcal{F}_t] = \Sigma_q \succeq 0$ a.s. For the ith left singular vector u_i of J, define the corresponding quantization energy as $q_i = u_i^\top \Sigma_q u_i$. If $0 < \eta h_i(\tau) < 2$, then for $z_{i,t} = v_i^\top \widetilde{e}_t$,*

$$\lim_{t\to\infty} \mathbb{E} z_{i,t}^2 = \frac{\eta q_i}{(\sigma_i^2 + \tau)\,(2 - \eta h_i(\tau))}. \tag{15}$$

Here q_i is the residual-quantization energy along u_i, so low-σ_i modes can carry a large stationary floor when they receive appreciable perturbation energy. Appendix A.2 gives the exact total active error; Appendix A.5 gives the persistent modal-bias limit $-(u_i^\top \mu_q)/\sigma_i$, which is independent of τ.

For isotropic residual perturbations, $\Sigma_q = \nu_b I$, the instantaneous injection term becomes

$$\mathcal{A}_q(P_\tau) = \nu_b \sum_{i=1}^{r} \frac{\sigma_i^2}{(\sigma_i^2 + \tau)^2}. \tag{16}$$

Away from clipping, uniform quantization often gives the rough scaling $\nu_b \propto 2^{-2b}$, although QAP estimates ν_b empirically rather than assuming this law.

Proposition 2 (Bit-dependent crossover from the exact stationary floor)**.** *Assume $0 < \eta < 2$, $\gamma > 0$, and $\Sigma_q(b) = \nu_b \bar{\Sigma}_q$, where $\nu_b > 0$ contains the bit-dependent magnitude and $\bar{\Sigma}_q$ has fixed shape with nonzero energy on the active residual subspace, $\sum_i u_i^\top \bar{\Sigma}_q u_i > 0$. Let $\bar{q}_i = u_i^\top \bar{\Sigma}_q u_i$ and define*

$$\mathcal{J}_b(\tau) = C(\tau) + \gamma \nu_b F(\tau), \quad C(\tau) = \log \kappa^+(H_\tau), \quad F(\tau) = \sum_i \frac{\eta \bar{q}_i}{2\tau + (2-\eta)\sigma_i^2}. \tag{17}$$

For $0 \leq \tau_1 < \tau_2$, $C(\tau)$ is nondecreasing (strictly if $\sigma_1 > \sigma_r$), while $F(\tau)$ is strictly decreasing whenever some $\bar{q}_i > 0$. Hence the more conservative τ_2 is preferred over τ_1 if and only if

$$\nu_b > \frac{C(\tau_2) - C(\tau_1)}{\gamma\,(F(\tau_1) - F(\tau_2))}. \tag{18}$$

Equation (18) is an exact stationary-floor version of the paper's design principle: changing bitwidth can change the preferred preconditioner even when the PDE, architecture, and full-precision tangent geometry are unchanged.

4.3 QUANTIZING THE PRECONDITIONER ITSELF

If $P_q = P + E_P$, the local tangent operator becomes $H(P_q) = J^\top P_q^\top P_q J$.

Proposition 3 (Active-spectrum stability under low-bit preconditioner storage)**.** *Let $\delta = \|E_P J\|_2$. Then*

$$\|H(P_q) - H(P)\|_2 \leq 2\|PJ\|_2\delta + \delta^2 \leq \|J\|_2^2\big(2\|P\|_2\|E_P\|_2 + \|E_P\|_2^2\big). \tag{19}$$

If $\mathrm{rank}(PJ) = \mathrm{rank}(J) = r$ and $\delta < \sigma_r(PJ)$, then the active rank is preserved and

$$\lambda_{\min}^+(H(P_q)) \geq (\sigma_r(PJ) - \delta)^2, \quad \kappa^+(H(P_q)) \leq \left(\frac{\sigma_1(PJ) + \delta}{\sigma_r(PJ) - \delta}\right)^2. \tag{20}$$

The corresponding contraction perturbation is bounded by $\eta(2\|PJ\|_2\delta + \delta^2)$.

This tangent-aware criterion is sharper than controlling storage error through $\|J\|_2\|E_P\|_2$ alone.

4.4 CONSTRUCTING THE LOW-BIT RESIDUAL GEOMETRY

The preconditioner acts in residual space, so the relevant object is not the tensor-domain quantization error by itself but the perturbation it induces in the physics residual. We use the centered covariance Σ_q when a stochastic interpretation is justified and the second moment $M_q = \Sigma_q + \mu_q \mu_q^\top$ as a conservative geometry matrix when deterministic quantization produces a persistent offset.

A. Layer/source decomposition (default empirical construction). Freeze θ at a QAP calibration checkpoint and use the same collocation set X. Let r_{fp} be the full-precision residual. For source s (a weight tensor, activation tensor, quantized operator block, or preconditioner block), quantize only that source and compute

$$d_s = r_{q_s}(X) - r_{\mathrm{fp}}(X). \tag{21}$$

With nonnegative source weights w_s summing to one,

$$\widehat{\mu}_q^{\mathrm{src}} = \sum_s w_s d_s, \tag{22}$$

$$\widehat{\Sigma}_q^{\mathrm{src}} = \sum_s w_s (d_s - \widehat{\mu}_q^{\mathrm{src}})(d_s - \widehat{\mu}_q^{\mathrm{src}})^\top, \tag{23}$$

$$\widehat{M}_q^{\mathrm{src}} = \sum_s w_s d_s d_s^\top = \widehat{\Sigma}_q^{\mathrm{src}} + \widehat{\mu}_q^{\mathrm{src}} (\widehat{\mu}_q^{\mathrm{src}})^\top. \tag{24}$$

This construction is deterministic, architecture-aware, and particularly useful for one-bit quantizers, where a uniform additive-noise model is weak. It should be interpreted as a *source-scatter geometry*; repeated stochastic realizations within each source can be added when an actual sampling covariance is desired.

B. First-order Jacobian construction. Let $R_{W,\ell} = \partial r / \partial \mathrm{vec}(W_\ell)$ and $R_{a,\ell} = \partial r / \partial h_\ell$. The first-order residual perturbation is

$$\xi_q \approx \sum_{\ell=1}^{L} R_{W,\ell} \varepsilon_{W,\ell} + \sum_{\ell=1}^{L} R_{a,\ell} \varepsilon_{a,\ell}. \tag{25}$$

If cross-source perturbations are approximately uncorrelated,

$$\Sigma_q \approx \sum_\ell R_{W,\ell} C_{W,\ell} R_{W,\ell}^\top + \sum_\ell R_{a,\ell} C_{a,\ell} R_{a,\ell}^\top. \tag{26}$$

For a multibit uniform quantizer away from clipping, $C_{W,\ell} \approx (\Delta_\ell^2/12) I$ is a useful baseline and exposes the approximate 2^{-2b} scaling. For binary weights, we instead estimate channel/block second moments from $Q_1(W_\ell) - W_\ell$; we do not impose the $\Delta^2/12$ approximation on one-bit errors. Appendix K gives a complementary Gaussian specialization based on the arcsine law and Bussgang cross-correlation, which makes the binary scale explicit in the predicted weight-error geometry. This construction explicitly composes tensor quantization with the PDE residual Jacobian and therefore gives a physics-aware covariance without evaluating the fully quantized model for every candidate P.

C. Paired stochastic/dithered construction. When stochastic rounding, dither, randomized clipping, or repeated QAT perturbations are available, keep (θ, X) fixed and obtain S paired residual differences $d^{(s)} = r_q^{(s)}(X) - r_{\mathrm{fp}}(X)$. Then use the ordinary empirical mean and covariance. This is the cleanest statistical estimator, while A and B remain available for deterministic deployment quantizers.

Low-rank eigendecomposition (preferred implementation). None of the above requires an explicit $m \times m$ covariance. Form centered columns

$$D_q = \frac{1}{\sqrt{S-1}} [d^{(1)} - \bar{d}, \ldots, d^{(S)} - \bar{d}], \qquad \widehat{\Sigma}_q = D_q D_q^\top. \tag{27}$$

Compute the small Gram eigendecomposition $D_q^\top D_q = V \Lambda V^\top$ and recover the nonzero residual-space eigenvectors as

$$U_q = D_q V \Lambda^{-1/2}, \qquad \widehat{\Sigma}_q = U_q \Lambda U_q^\top. \tag{28}$$

Thus the storage cost is $O(mS)$ rather than $O(m^2)$, covariance-vector products are $D_q(D_q^\top v)$, and only the residual directions actually excited by low-bit perturbations need to be regularized. In our implementation, this low-rank factorization is the default representation used by QAP-Cov.

5 Quantization-Aware Preconditioning

5.1 Constructive noise-aware inverse

Let $K_r = JJ^\top$ denote the local residual tangent Gram matrix, let $\widehat{\Sigma}_q$ be an empirical residual-quantization covariance, and let $\epsilon > 0$ be a small numerical ridge. Our default low-bit preconditioner is

$$\boxed{P_{\lambda,q} = \left(K_r + \lambda\widehat{\Sigma}_q + \epsilon I\right)^{-1/2}} \tag{29}$$

with $\lambda \geq 0$ selected from the QAP finite-horizon objective below. In implementation, K_r and $\widehat{\Sigma}_q$ are trace-normalized before combining them, making λ dimensionless. When deterministic quantization has a substantial mean perturbation, the robust variant replaces $\widehat{\Sigma}_q$ by $\widehat{M}_q$.

Equation (29) gives a concrete alternative to reusing the full-precision optimum. If $\widehat{\Sigma}_q \to 0$, it approaches the aggressive inverse-square-root preconditioner. If a residual direction carries large quantization variance, the denominator increases only in that direction and caps its inverse gain. For jointly diagonalizable $K_r = U\,\mathrm{diag}(\sigma_i^2)U^\top$ and $\Sigma_q = U\,\mathrm{diag}(\nu_i)U^\top$,

$$p_i(\lambda) = \frac{1}{\sqrt{\sigma_i^2 + \lambda\nu_i + \epsilon}}, \quad h_i(\lambda) = \frac{\sigma_i^2}{\sigma_i^2 + \lambda\nu_i + \epsilon}. \tag{30}$$

The corresponding tangent-space injection contribution is

$$a_i(\lambda) = \frac{\sigma_i^2\nu_i}{(\sigma_i^2 + \lambda\nu_i + \epsilon)^2}, \tag{31}$$

which decreases monotonically with λ, while $h_i(\lambda)$ also decreases. QAP therefore regularizes a mode in proportion to its measured quantization variance rather than applying the same damping to every mode.

Proposition 4 (Generalized selective suppression)**.** *Let $B = K_r + \epsilon I \succ 0$ and define the physically whitened quantization geometry*

$$C_q = B^{-1/2}\Sigma_q B^{-1/2} = U_\gamma\,\mathrm{diag}(\gamma_i)U_\gamma^\top, \tag{32}$$

where $\gamma_i \geq 0$ are generalized noise-to-physics eigenvalues. For $P_{\lambda,q} = (B + \lambda\Sigma_q)^{-1/2}$,

$$B^{1/2}P_{\lambda,q}^\top P_{\lambda,q}B^{1/2} = U_\gamma\,\mathrm{diag}\left(\frac{1}{1+\lambda\gamma_i}\right)U_\gamma^\top. \tag{33}$$

Thus generalized mode i has squared-gain factor $(1+\lambda\gamma_i)^{-1}$ and amplitude factor $(1+\lambda\gamma_i)^{-1/2}$, with stronger attenuation for larger γ_i. This characterization applies without a commutativity assumption on K_r and Σ_q. In the aligned case, $\gamma_i = \nu_i/(\sigma_i^2 + \epsilon)$.

5.2 A bound-driven objective

Let K denote a finite optimization horizon, E_0 the initial local error energy, $\widehat{\rho}_P$ an estimated contraction factor, and $\widehat{\mathcal{A}}_q(P;b)$ the estimated quantization-injection energy at bit configuration b. Define the finite geometric factor $g_K(\rho) = \sum_{j=0}^{K-1}\rho^{2j}$, so $g_K(1) = K$ and, for $\rho \neq 1$, $g_K(\rho) = (1-\rho^{2K})/(1-\rho^2)$. Let $\mathcal{P}$ be a structured preconditioner family, let $\mathcal{R}(P)$ enforce scale or structural constraints, and let $\beta \geq 0$ be its weight. The finite-horizon objective suggested by theorem 1 is

$$P_b^\star \in \arg\min_{P\in\mathcal{P}}\left\{\widehat{\rho}_P^{2K}E_0 + \eta^2 g_K(\widehat{\rho}_P)\widehat{\mathcal{A}}_q(P;b) + \beta\mathcal{R}(P)\right\}. \tag{34}$$

For a scale-insensitive alternative, let $\gamma \geq 0$ balance conditioning and injection. We use

$$\min_{P\in\mathcal{P}}\log\kappa^+(H_P) + \gamma\log\left(\varepsilon + \widehat{\mathcal{A}}_q(P;b)\right) + \beta\mathcal{R}(P). \tag{35}$$

The key difference from a full-precision design is the second term, which is bitwidth dependent.

5.3 TWO ESTIMATORS: RESIDUAL-SPACE AND TANGENT-SPACE

Fix the network state θ and a common collocation batch X containing the m residual coordinates. For $N \geq 2$ paired quantization realizations or probes, define the full residual-difference vectors

$$d^{(s)} = r^{(s)}_{\theta,q}(X) - r_{\theta,\mathrm{fp}}(X) \in \mathbb{R}^m, \qquad \bar{d} = \frac{1}{N}\sum_{s=1}^{N} d^{(s)}. \tag{36}$$

The sample-covariance residual-space proxy and its theory-aligned tangent-space counterpart are

$$\widehat{\mathcal{A}}_{\mathrm{res}}(P) = \frac{1}{N-1}\sum_{s=1}^{N} \|P(d^{(s)} - \bar{d})\|_2^2, \tag{37}$$

$$\widehat{\mathcal{A}}_{\tan}(P) = \frac{1}{N-1}\sum_{s=1}^{N} \left\| J^\top P^\top P(d^{(s)} - \bar{d}) \right\|_2^2. \tag{38}$$

Under independent identically distributed paired probes, (38) is unbiased for $\mathrm{tr}(G_P \Sigma_q G_P^\top)$. It is evaluated with vector–Jacobian products, and neither estimator requires explicitly forming Σ_q.

5.4 PRACTICAL REALIZATION

For end-to-end QAT, let J_q be the current STE/quantized residual Jacobian, $K_q = J_q J_q^\top$, $\bar{\kappa}_q = \mathrm{tr}(K_q)/m$, and $\widetilde{M}_q$ a trace-matched estimate of the residual second moment M_q. With isotropic strength $\tau \geq 0$ and directional strength $\lambda \geq 0$, we use the nested form

$$P_{\tau,\lambda,q} = \left(K_q + \tau \bar{\kappa}_q I + \lambda \widetilde{M}_q + \epsilon I \right)^{-1/2}. \tag{39}$$

The isotropic floor $\tau\bar{\kappa}_q I$ caps inverse gain in weak directions not reliably covered by a finite covariance estimate, while $\lambda\widetilde{M}_q$ adds directional protection; $\lambda = 0$ recovers QAP-I. We call this directional form QAP-Cov+. For larger residual spaces, QAP-RLR is a stabilized low-rank realization of the same principle: it shrinks the directional estimate toward isotropic geometry and falls back to QAP-I unless held-out perturbations support a stable directional correction. Appendix J.2 gives the gate and matrix-free implementation.

6 END-TO-END AND PDE-CALIBRATED VALIDATION

We evaluate deployed quantized PINNs on 25:1 jump diffusion, layered diffusion, and a W4A8 two-dimensional 1000:1 checkerboard problem with a 10×10 interior grid (100 residual equations) and a width-32 PINN. Pilot seeds $\{0, 1, 2\}$ are used only for selection; all settings are then frozen for held-out seeds $\{3, \ldots, 12\}$. Every method starts from the same per-seed full-precision warm state. Identity uses no preconditioning, the FP-derived baseline uses the full-precision residual tangent, QAP-I is isotropic, QAP-Cov+ uses the measured second moment, and the 2D directional result uses QAP-RLR. Full protocols are in Appendices I and J.2.

Table 1: End-to-end QAT on ten held-out seeds: final relative L_2 solution error (mean $\pm$ std). "Directional QAP" denotes QAP-Cov+ in 1D and QAP-RLR in 2D.

Method	Jump diffusion W6A8	Jump diffusion W4A8	Layered diffusion W4A8	2D checkerboard (1000:1), W4A8
Identity	0.283 ± 0.040	0.283 ± 0.040	0.172 ± 0.018	0.0716 ± 0.0033
FP-derived	0.283 ± 0.040	0.283 ± 0.040	0.174 ± 0.019	0.2623 ± 0.0254
QAP-I	0.148 ± 0.105	0.170 ± 0.081	0.195 ± 0.174	0.0813 ± 0.0162
Directional QAP	$\mathbf{0.137 \pm 0.063}$	$\mathbf{0.158 \pm 0.081}$	$\mathbf{0.141 \pm 0.045}$	$\mathbf{0.0493 \pm 0.0191}$
Gain vs. FP	$2.07\times$	$1.79\times$	$1.23\times$	$5.32\times$

On jump diffusion, QAP-Cov+ reduces mean solution error by $2.07\times$ at W6A8 and $1.79\times$ at W4A8 relative to the FP-derived baseline; layered diffusion gives a $1.23\times$ gain. On the 2D checkerboard

problem, QAP-RLR reaches 0.0493 ± 0.0191 and improves over identity, QAP-I, and the FP-derived preconditioner by $1.45\times$, $1.65\times$, and $5.32\times$, respectively, with $9/10$, $10/10$, and $10/10$ paired wins (one-sided Wilcoxon $p \leq 1.95 \times 10^{-3}$). The 2D result shows that reliability-gated shrinkage can preserve useful directional covariance information beyond the one-dimensional setting.

The raw quantized residual norm need not decrease with solution error because low-bit arithmetic can inject residual components that are poor optimization targets; QAP instead regularizes directions using measured quantization statistics. Local validation, a precision sweep, and a Helmholtz control are reported in Appendices F–H.

7 Discussion and Conclusion

QAP selects residual preconditioning from physical conditioning and measured low-bit perturbation geometry. Across controlled, 1D, and 2D studies, it improves deployed low-bit accuracy even when full-precision-derived preconditioning fails, identifying quantization geometry as a distinct design variable.

AI Use Statement

We used generative AI tools to assist with proof checking, drafting and revising portions of the manuscript, literature searches, experimental code development and review, and controlled small-model experiments. We also used AI assistance to improve the visual presentation of Figure 1 and to prepare experimental scripts and Colab notebooks.

Reproducibility Statement

The submission includes complete proofs of all formal claims in Appendix A, detailed quantizer and estimator definitions in Sections 3–5 and Appendix M, end-to-end training protocols and held-out seed splits in Appendices I and J.2, and the larger-grid protocol in Appendix L. The supplementary source package contains the scripts and machine-readable result files used to produce the reported experiments and figures.

References

Saleh Ashkboos et al. QuaRot: Outlier-free 4-bit inference in rotated LLMs, 2024. arXiv preprint arXiv:2404.00456.

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

Julian J. Bussgang. Crosscorrelation functions of amplitude-distorted Gaussian signals. Technical Report 216, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1952.

Daniela Calvetti. Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. *Journal of Computational and Applied Mathematics*, 198(2): 378–395, 2007.

Tim De Ryck, Florent Bonnet, Siddhartha Mishra, and Emmanuel de Bézenac. An operator preconditioning perspective on training in physics-informed machine learning. In *International Conference on Learning Representations*, 2024.

Natalia Frumkin and Diana Marculescu. Q-Sched: Pushing the boundaries of few-step diffusion models with quantization-aware scheduling, 2025. arXiv preprint arXiv:2509.01624.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Alena Kopaničáková and George Em Karniadakis. DeepONet based preconditioning strategies for solving parametric linear systems of equations, 2024. arXiv preprint arXiv:2401.02016.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Songming Liu, Chang Su, Jiachen Yao, Zhongkai Hao, Hang Su, Youjia Wu, and Jun Zhu. Preconditioning for Physics-Informed Neural Networks, 2024.

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

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Winfried van den Dool, Tijmen Blankevoort, Max Welling, and Yuki M. Asano. Efficient neural PDE-solvers using quantization aware training. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2023.

John H. Van Vleck and David Middleton. The spectrum of clipped noise. *Proceedings of the IEEE*, 54(1):2–19, 1966.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, , and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 2023.

A PROOFS

A.1 PROOF OF THEOREM 1

Because $\mathrm{range}(H_P) \subseteq \mathcal{S}$, H_P annihilates $\mathcal{S}^\perp$, and $\mathrm{range}(G_P) \subseteq \mathcal{S}$, both $\mathcal{S}$ and $\mathcal{S}^\perp$ are invariant under M_P. Projecting the centered recursion onto $\mathcal{S}$ therefore gives

$$\widetilde{e}_{t+1}^{\mathcal{S}} = M_{P,\mathcal{S}}\widetilde{e}_t^{\mathcal{S}} - \eta G_P \zeta_t, \tag{40}$$

whereas $\Pi_{\mathcal{S}^\perp}\widetilde{e}_{t+1} = \Pi_{\mathcal{S}^\perp}\widetilde{e}_t$. Unrolling the active recursion and using the martingale-difference property yields

$$\mathbb{E}\|\widetilde{e}_k^{\mathcal{S}}\|_2^2 \le \mathbb{E}\|M_{P,\mathcal{S}}^k \widetilde{e}_0^{\mathcal{S}}\|_2^2 + \eta^2 \sum_{j=0}^{k-1} \mathbb{E}\|M_{P,\mathcal{S}}^{k-1-j} G_P \zeta_j\|_2^2 \tag{41}$$

$$\le \rho_P^{2k}\mathbb{E}\|\widetilde{e}_0^{\mathcal{S}}\|_2^2 + \eta^2 \sum_{j=0}^{k-1} \rho_P^{2(k-1-j)}\,\mathrm{tr}(G_P \Sigma_q G_P^\top), \tag{42}$$

which gives (10) and (12); the martingale-difference condition in Assumption 1 removes all cross terms between the current perturbation and the active history. For the exact covariance statement, let $V_{\mathcal{S}} \in \mathbb{R}^{d_\theta \times r}$ have orthonormal columns spanning $\mathcal{S}$, and define

$$\bar{M}_P = V_{\mathcal{S}}^\top M_P V_{\mathcal{S}}, \qquad Q_{\mathcal{S}} = V_{\mathcal{S}}^\top G_P \Sigma_q G_P^\top V_{\mathcal{S}}. \tag{43}$$

If $\mathbb{E}[\zeta_t\zeta_t^\top \mid \mathcal{F}_t] = \Sigma_q$ a.s., then for $\bar{e}_t = V_{\mathcal{S}}^\top \widetilde{e}_t$ and $C_t = \mathbb{E}[\bar{e}_t \bar{e}_t^\top]$,

$$C_{t+1} = \bar{M}_P C_t \bar{M}_P^\top + \eta^2 Q_{\mathcal{S}}. \tag{44}$$

Since $\rho_P < 1$, this discrete Lyapunov recursion has the unique stationary solution

$$\mathrm{vec}(C_\infty) = \eta^2 \left[I - \bar{M}_P \otimes \bar{M}_P\right]^{-1} \mathrm{vec}(Q_{\mathcal{S}}). \tag{45}$$

A.2 PROOF OF PROPOSITION 1

For $J = U\Sigma V^\top$ and $P_\tau^\top P_\tau = (JJ^\top + \tau I)^{-1}$ on $\mathrm{range}(J)$,

$$H_\tau = V\Sigma^\top(\Sigma\Sigma^\top + \tau I)^{-1}\Sigma V^\top, \tag{46}$$

$$G_\tau = V\Sigma^\top(\Sigma\Sigma^\top + \tau I)^{-1}U^\top. \tag{47}$$

Let $z_t = V^\top \widetilde{e}_t$ and $\widehat{\zeta}_t = U^\top \zeta_t$. For a nonzero singular direction,

$$z_{i,t+1} = (1 - \eta h_i) z_{i,t} - \eta c_i \widehat{\zeta}_{i,t}, \qquad c_i = \frac{\sigma_i}{\sigma_i^2 + \tau}. \tag{48}$$

For arbitrary $\Sigma_q \succeq 0$, $\mathbb{E}\widehat{\zeta}_{i,t}^2 = u_i^\top \Sigma_q u_i = q_i$; off-diagonal entries of $U^\top \Sigma_q U$ do not enter this scalar variance. Hence the stationary scalar Lyapunov equation is $v_i = (1 - \eta h_i)^2 v_i + \eta^2 c_i^2 q_i$, which gives (15). Summing the marginal variances over the orthonormal active coordinates yields the exact total active floor

$$\lim_{t\to\infty} \mathbb{E}\|\widetilde{e}_t^{\mathcal{S}}\|_2^2 = \sum_{i=1}^{r} \frac{\eta q_i}{2\tau + (2-\eta)\sigma_i^2}. \tag{49}$$

A.3 PROOF OF PROPOSITION 2

For the positive spectrum,

$$C(\tau) = \log\frac{\sigma_1^2(\sigma_r^2 + \tau)}{\sigma_r^2(\sigma_1^2 + \tau)}, \qquad C'(\tau) = \frac{1}{\sigma_r^2 + \tau} - \frac{1}{\sigma_1^2 + \tau} \ge 0, \tag{50}$$

with strict inequality when $\sigma_1 > \sigma_r$. Moreover,

$$F'(\tau) = -2\eta \sum_i \frac{\bar{q}_i}{[2\tau + (2-\eta)\sigma_i^2]^2} < 0 \tag{51}$$

whenever some $\bar{q}_i > 0$. Therefore $\mathcal{J}_b(\tau_2) < \mathcal{J}_b(\tau_1)$ is equivalent to the inequality in (18).

A.4 Proof of Proposition 3

With $A = PJ$ and $\Delta = E_P J$,

$$H(P_q) - H(P) = A^\top \Delta + \Delta^\top A + \Delta^\top \Delta, \tag{52}$$

so submultiplicativity gives the first inequality in (19); the second follows from $\|PJ\|_2 \le \|P\|_2 \|J\|_2$ and $\delta \le \|E_P\|_2 \|J\|_2$. Singular-value perturbation gives $|\sigma_i((P + E_P)J) - \sigma_i(PJ)| \le \delta$. Because $\ker(J) \subseteq \ker((P + E_P)J)$, the perturbed rank cannot exceed r; if $\delta < \sigma_r(PJ)$, the first r singular values remain positive and $\sigma_r(P_q J) \ge \sigma_r(PJ) - \delta$, while $\sigma_1(P_q J) \le \sigma_1(PJ) + \delta$. Squaring gives the active eigenvalue bounds and (20). The triangle inequality applied to $I - \eta H$ gives the stated contraction perturbation.

A.5 Bias–variance recursion

Taking conditional expectation in (8) with $\xi_t = \mu_q + \zeta_t$ gives

$$m_{t+1} = M_P m_t - \eta G_P \mu_q. \tag{53}$$

Subtracting this recursion from (8) yields

$$\widetilde{e}_{t+1} = M_P \widetilde{e}_t - \eta G_P \zeta_t, \tag{54}$$

so Theorem 1 applies to the active centered component. Consequently, $\mathbb{E}\|e_t\|_2^2 = \|m_t\|_2^2 + \mathbb{E}\|\widetilde{e}_t\|_2^2$. For the spectral family P_τ, let $\mu_i = u_i^\top \mu_q$. The active modal mean obeys $m_{i,t+1} = (1 - \eta h_i) m_{i,t} - \eta c_i \mu_i$, and therefore, whenever the mode is stable,

$$\lim_{t \to \infty} m_{i,t} = -\frac{\mu_i}{\sigma_i}, \tag{55}$$

which is independent of τ. Thus covariance-aware preconditioning can suppress incoherent amplification, whereas persistent active bias must be controlled by quantizer calibration/scaling or represented conservatively through M_q.

A.6 Proof of Proposition 4

Factor

$$B + \lambda \Sigma_q = B^{1/2} \left(I + \lambda B^{-1/2} \Sigma_q B^{-1/2} \right) B^{1/2}. \tag{56}$$

Taking the inverse and sandwiching by $B^{1/2}$ gives

$$B^{1/2}(B + \lambda \Sigma_q)^{-1} B^{1/2} = (I + \lambda C_q)^{-1} = U_\gamma \operatorname{diag}\big((1 + \lambda \gamma_i)^{-1}\big) U_\gamma^\top, \tag{57}$$

which proves (33). Since $\gamma_i \ge 0$, each squared-gain factor is non-increasing in λ and is smaller for larger generalized noise-to-physics eigenvalue; the corresponding amplitude factor is its positive square root. In the aligned case, C_q is diagonal with $\gamma_i = \nu_i / (\sigma_i^2 + \epsilon)$. In addition, (31) satisfies

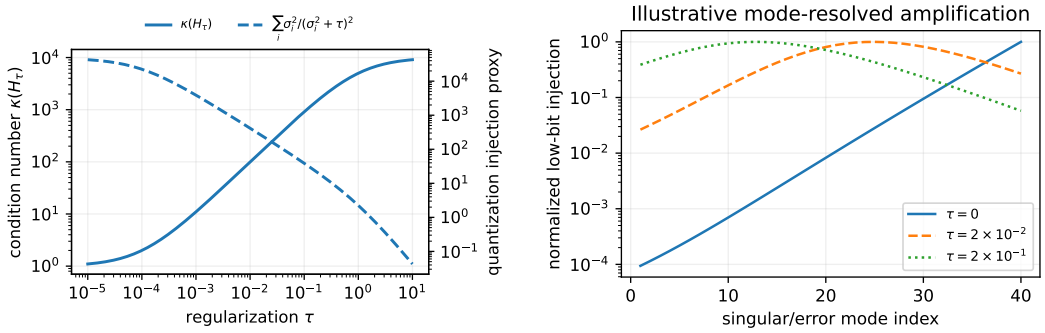
$$\frac{da_i}{d\lambda} = -\frac{2\sigma_i^2 \nu_i^2}{(\sigma_i^2 + \lambda \nu_i + \epsilon)^3} \le 0. \tag{58}$$

B A Two-Mode Example

Consider $J = \operatorname{diag}(1, \epsilon)$ with $0 < \epsilon \ll 1$. At $\tau = 0$, $H_0 = I$ gives ideal local conditioning. However, for isotropic residual quantization with variance ν_b,

$$\mathcal{A}_q(P_0) = \nu_b \left(1 + \epsilon^{-2}\right), \tag{59}$$

so the difficult mode produces arbitrarily large low-bit injection as $\epsilon \to 0$. A positive τ replaces the second contribution by $\nu_b \epsilon^2 / (\epsilon^2 + \tau)^2$. This toy model isolates the mechanism behind the crossover in proposition 2.



(a) Conditioning versus injection.

(b) Mode-resolved suppression.

Figure 2: **Theoretical illustrations, not experimental results.** For P_τ, larger τ worsens the conditioned tangent spectrum but suppresses quantization injection. The right panel illustrates that this suppression is strongest for difficult small-singular-value modes.

C QAP Optimization Loop

Algorithm 1 Quantization-Aware Preconditioning (QAP)

Require: PINN residual r_θ, preconditioner family P_ϕ, bitwidths (b_w, b_a, b_P), update period T_P

Initialize θ, ϕ
for $t = 1, 2, \ldots$ **do**
  Sample residual/boundary points and run QAT forward pass
  Update θ using $\frac{1}{2}\|P_\phi r_{\theta,q}\|_2^2 + \mathcal{L}_{\mathrm{bc}}$ with STE
  if $t \bmod T_P = 0$ **then**
    Construct $\widehat{\mu}_q, \widehat{\Sigma}_q$ by source decomposition, Jacobian propagation, or paired stochastic runs
    Compress $\widehat{\Sigma}_q$ using the low-rank Gram eigendecomposition in (28)
    Estimate contraction/conditioning of H_{P_ϕ} and $\widehat{\mathcal{A}}_{\mathrm{res}}$ or $\widehat{\mathcal{A}}_{\tan}$
    Update ϕ (or search λ, τ, δ) using (34) or (35)
    Optionally replace P_ϕ by $Q_{b_P}(P_\phi)$ during deployment-aware training
  end if
end for

D Additional Structured QAP Classes

D.1 Practical preconditioner classes

QAP-Cov / QAP-LR. Estimate $\widehat{\Sigma}_q$ by the layer/source or stochastic construction and use (29). For scalability we prefer a low-rank-plus-tail-floor approximation: retain the dominant eigenvectors from (28) and replace the discarded spectrum by its average diagonal floor. This preserves protection against weak omitted directions while requiring only matrix–vector products with the low-rank factor. When the measured noise is nearly isotropic, QAP-Spectral is preferable; we therefore select among isotropic, full-covariance, and low-rank variants using the finite-horizon objective on a held-out calibration split.

QAP-Spectral / isotropic. When only a scalar noise level is available, use P_τ from (13); this is equivalent to covariance-aware regularization under $\Sigma_q \propto I$ and remains the simplest theory/baseline. Select τ by one-dimensional search on (34) or (35).

Other structured classes. QAP also applies to ILU-derived, learned/hybrid, and parameter-conditioned preconditioners; these variants and the distinction between ILU approximation fidelity and arithmetic precision are detailed in Appendix D.

Table 2: Controlled anisotropic-noise validation. QAP-Cov protects only the modes carrying large quantization variance.

Preconditioner	Regularization	$\mathbb{E}\|e_K\|_2^2 \downarrow$	Relative to QAP-Cov
FP-style inverse	0	509.57	6237×
Best isotropic	$\tau = 2.47 \times 10^{-4}$	0.551	6.7×
QAP-Cov	$\lambda = 81.6$	**0.0817**	1×

Optimization loop. QAP is recalibrated on a slower timescale than the network: periodically construct $\widehat{\mu}_q, \widehat{\Sigma}_q$, compress the covariance, and search the structured preconditioner parameter using (34) or (35). Full pseudocode is given in Appendix C.

QAP-ILU/structured numerical. Following numerical PINN preconditioning, let M_δ be an ILU-like approximation with drop tolerance/strength parameter δ and use $P_\delta = M_\delta^{-1}$. Full-precision work typically chooses δ only for condition number and application cost. QAP additionally evaluates $\widehat{\mathcal{A}}_q(P_\delta; b)$. Importantly, δ controls *approximation fidelity*; b controls *arithmetic precision*. The two axes should be ablated separately.

QAP-Learned and hybrid. For nonlinear or parametric problems, parameterize P_ϕ as diagonal, block, convolutional, low-rank, subspace, or operator-network based. Hybrid numerical/learned preconditioners are especially natural when different components target different spectral ranges. QAP does not prescribe the architecture; it changes the training/selection objective by penalizing low-bit amplification.

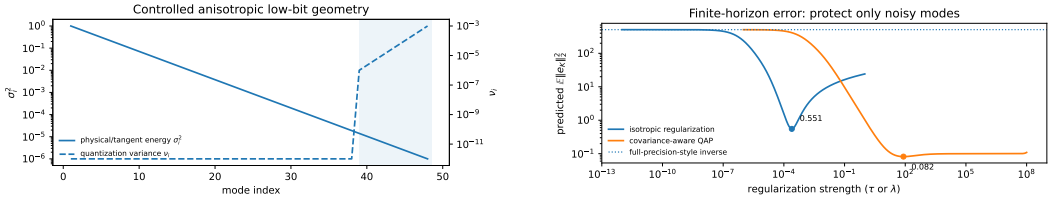
Parametric QAP. For PDE parameters $\mu \sim \nu$, a preconditioner can be conditioned on μ and optimized for average or worst-case low-bit robustness:

$$\min_\phi \ \mathbb{E}_{\mu\sim\nu}\left[\mathcal{J}_b(P_\phi(\mu);\mu)\right] \quad \text{or} \quad \min_\phi \ \sup_{\mu\in\mathcal{M}} \mathcal{J}_b(P_\phi(\mu);\mu). \tag{60}$$

This is the natural QAP analogue of parameter-robust learned preconditioning.

E Controlled Validation Details

The controlled experiment in Table 2 uses $r = 48$ modes with σ_i logarithmically spaced from 1 to 10^{-3}. The first 38 quantization variances are 10^{-12} and the final 10 are logarithmically spaced from 10^{-6} to 10^{-3}. Initial errors are 1 in the first 38 modes and 0.1 in the noisy tail. The exact finite-horizon recursion is evaluated for $K = 30$ and $\eta = 0.5$.



(a) Physical and quantization spectra.

(b) Finite-horizon regularization sweep.

Figure 3: Exact controlled validation. Covariance-aware QAP avoids globally damping the clean modes while protecting the noisy tail.

F PDE-Calibrated Local-QAT Details

The local study uses residuals obtained from finite-difference discretizations evaluated on the PINN outputs. Operator/right-hand-side pairs are scaled by the spectral norm of the discrete operator to

avoid arbitrary loss scaling. For jump diffusion, $a(x) = 1$ for $x < 1/2$ and $a(x) = 25$ otherwise, with exact solution $u(x) = \sin(\pi x)$ chosen so that the derivative vanishes at the interface.

At a common full-precision warm state, we compute the residual Jacobian J exactly. Stochastic symmetric rounding then generates paired low-bit perturbations: the first 48 samples estimate (μ_q, Σ_q), the next 48 select regularization and low-rank truncation, and the final 48 are held out for testing. QAP-LR keeps dominant covariance eigenvectors and adds an isotropic tail floor equal to the average discarded eigenvalue, preventing weak omitted directions from receiving unbounded inverse gain.

Table 3: Held-out jump-diffusion local-QAT validation. Metric is final residual energy $\mathbb{E}\|r_{40}\|_2^2$ over ten PINN seeds (mean $\pm$ std).

Method	W4A8	W4A4
FP-inverse	$(2.00 \pm 0.75) \times 10^{-7}$	$(5.33 \pm 1.10) \times 10^{-5}$
QAP-I	$(9.97 \pm 3.09) \times 10^{-8}$	$(1.85 \pm 0.47) \times 10^{-5}$
QAP-Cov	$(5.70 \pm 2.83) \times 10^{-8}$	$(4.24 \pm 2.14) \times 10^{-7}$
QAP-LR	$\mathbf{(3.92 \pm 2.20) \times 10^{-8}}$	$\mathbf{(2.34 \pm 0.51) \times 10^{-7}}$
Gain vs. FP (LR)	$5.11\times$	$227\times$

The local recursion is intentionally theory-aligned: it uses J and quantization perturbations measured from actual PINNs but holds the local geometry fixed. It therefore tests the mechanism independently of long-horizon nonlinear training. The full ten-seed precision sweep is reported next.

G Ten-Seed Jump-Diffusion Precision Sweep

To test whether covariance-aware improvement is specific to one quantization setting, we repeat the held-out local-QAT protocol for jump diffusion over ten independent PINN seeds and five precision regimes. The same 48/48/48 construction/selection/test split is used at every seed, and no test perturbation is used to choose the covariance representation or regularization strength.

Table 4: Ten-seed jump-diffusion bit sweep. Entries are held-out $\mathbb{E}\|r_{40}\|_2^2$ (mean $\pm$ std). Gains compare the listed covariance method against reuse of the FP inverse.

Method	W8A8	W6A8	W4A8	W4A6	W4A4
FP-inverse	$(1.71 \pm 0.68) \times 10^{-7}$	$(1.55 \pm 0.56) \times 10^{-7}$	$(2.00 \pm 0.75) \times 10^{-7}$	$(2.65 \pm 0.73) \times 10^{-6}$	$(5.33 \pm 1.10) \times 10^{-5}$
QAP-I	$(7.70 \pm 2.53) \times 10^{-8}$	$(7.22 \pm 2.06) \times 10^{-8}$	$(9.97 \pm 3.09) \times 10^{-8}$	$(1.08 \pm 0.26) \times 10^{-6}$	$(1.85 \pm 0.47) \times 10^{-5}$
QAP-Cov	$\mathbf{(2.00 \pm 1.16) \times 10^{-9}}$	$(8.99 \pm 3.83) \times 10^{-9}$	$(5.70 \pm 2.83) \times 10^{-8}$	$(1.64 \pm 0.86) \times 10^{-7}$	$(4.24 \pm 2.14) \times 10^{-7}$
QAP-LR	$(4.45 \pm 2.64) \times 10^{-9}$	$\mathbf{(5.91 \pm 2.24) \times 10^{-9}}$	$\mathbf{(3.92 \pm 2.20) \times 10^{-8}}$	$\mathbf{(7.63 \pm 2.78) \times 10^{-8}}$	$\mathbf{(2.34 \pm 0.51) \times 10^{-7}}$
Gain (Cov)	$85.2\times$	$17.3\times$	$3.52\times$	$16.2\times$	$125.6\times$
Gain (LR)	$38.4\times$	$26.3\times$	$5.11\times$	$34.7\times$	$227.3\times$

This sweep strengthens the claim in a deliberately limited sense: for the same PDE and PINN architecture, covariance-aware regularization repeatedly improves the *quantized local dynamics* across several precision regimes. For every regime in table 4, both QAP-Cov and QAP-LR beat the FP inverse in all $10/10$ paired seeds; an exact one-sided Wilcoxon signed-rank test gives $p = 9.77 \times 10^{-4}$ in each comparison. The gain is intentionally not monotone in nominal bitwidth: QAP responds to the measured covariance geometry, not to bit count alone.

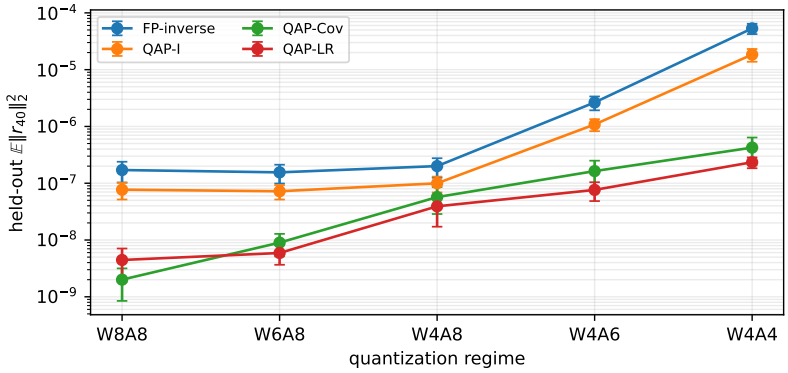


Figure 4: Ten-seed jump-diffusion precision sweep. Covariance-aware QAP consistently reduces held-out local-QAT residual energy relative to the full-precision-style inverse; the gain magnitude depends non-monotonically on both precision and measured noise geometry.

H Further Investigation: When Isotropic Regularization Is Preferable

Covariance-aware regularization is not expected to dominate when the measured perturbation geometry is weakly directional or poorly resolved. We include one control case to make this boundary explicit. For Helmholtz at W4A8, the held-out local-QAT residual energy is $(6.84 \pm 1.38) \times 10^{-6}$ for QAP-I, compared with $(8.17 \pm 2.34) \times 10^{-6}$ for QAP-Cov and $(8.67 \pm 1.88) \times 10^{-6}$ for QAP-LR. Thus the isotropic design is preferable in this case.

A plausible explanation is that the finite-sample covariance estimate does not reveal a stable set of quantization-dominated directions relative to the Helmholtz tangent modes. Under such a geometry, directional regularization can spend statistical degrees of freedom suppressing modes that still carry useful contraction, while an isotropic floor provides a lower-variance safeguard against inverse amplification. This interpretation is consistent with the validation selector reverting to the isotropic branch. A useful next step is to develop an explicit anisotropy/reliability diagnostic—for example, based on the effective rank and spectral separation of the residual covariance—that decides when to set the directional regularization strength to zero.

I End-to-End QAT Details

The end-to-end study uses two heterogeneous diffusion problems. The first is the 25:1 jump-diffusion problem used in the local analysis. The second is a layered-diffusion problem whose coefficient alternates between 1 and 50 on four subintervals. Both use a two-hidden-layer width-32 tanh PINN with hard Dirichlet boundary conditions. Pilot seeds $0, 1, 2$ are used only to select a learning rate independently for each method; seeds 3–12 form the held-out evaluation set.

All branches start from the identical per-seed full-precision warm state after 50 Adam steps at learning rate 3×10^{-3}. They then run 250 deterministic QAT+STE steps with full-batch SGD. Every 50 QAT steps, the QAP branches recompute the current quantized STE residual Jacobian and update the quantization statistics from paired stochastic-rounding probes. The FP-derived baseline is deliberately strong: at the same checkpoints it constructs its preconditioner shape from the current full-precision residual tangent, then scale-matches that transform to the quantized tangent. Hence the comparison isolates the effect of using quantization geometry rather than merely the benefit of recalibration.

For jump diffusion, the selected learning rates are 10^{-3} for identity, 10^{-4} for the FP-derived baseline, and 10^{-1} for QAP-I and QAP-Cov+ at both W6A8 and W4A8. For layered diffusion at W4A8, the selected rates are 10^{-2}, 10^{-3}, 10^{-1}, and 10^{-2}, respectively. No exact solution values are used when selecting the preconditioner itself; exact solutions are used only for final evaluation on a dense grid.

Table 5: End-to-end QAT statistics on ten held-out seeds. "Residual" is the raw deterministic quantized PDE residual divided by $\|b\|_2$. The primary deployment metric is relative solution error.

PDE	Bits	Method	Rel. L_2 error ↓	Raw rel. residual
Jump diffusion	W6A8	Identity	0.283 ± 0.040	6.49 ± 1.04
		FP-derived	0.283 ± 0.040	6.47 ± 1.01
		QAP-I	0.148 ± 0.105	12.30 ± 4.36
		QAP-Cov+	$\mathbf{0.137 \pm 0.063}$	14.34 ± 3.10
Jump diffusion	W4A8	Identity	0.283 ± 0.040	6.74 ± 1.00
		FP-derived	0.283 ± 0.040	6.81 ± 1.01
		QAP-I	0.170 ± 0.081	12.64 ± 3.97
		QAP-Cov+	$\mathbf{0.158 \pm 0.081}$	15.06 ± 2.67
Layered diffusion	W4A8	Identity	0.172 ± 0.018	0.601 ± 0.136
		FP-derived	0.174 ± 0.019	0.609 ± 0.115
		QAP-I	0.195 ± 0.174	1.221 ± 0.411
		QAP-Cov+	$\mathbf{0.141 \pm 0.045}$	0.651 ± 0.088

On jump diffusion, QAP-Cov+ beats the FP-derived baseline in all $10/10$ W6A8 seeds ($p = 9.77 \times 10^{-4}$) and in $8/10$ W4A8 seeds ($p = 4.88 \times 10^{-3}$). On layered diffusion at W4A8 it beats the FP-derived baseline in $8/10$ seeds ($p = 6.84 \times 10^{-3}$) and identity in $8/10$ seeds ($p = 9.77 \times 10^{-3}$). The covariance-aware mean is also lower than QAP-I in these three settings, but the paired difference is not statistically significant in the current end-to-end sample; we therefore do not claim universal covariance-over-isotropic superiority from these runs. The controlled anisotropic experiment in Appendix E provides the cleaner evidence for the benefit of directional covariance information over the best isotropic regularizer.

J Robust Low-Rank and Efficient QAP

This appendix studies low-rank extensions that reduce covariance-estimation variance and application cost. The generalized-eigen experiment remains exploratory; the reliability-gated low-rank covariance shrinkage in Section J.2 is used for the 2D end-to-end result in the main paper.

J.1 Reliability-gated generalized-eigen low-rank correction

Let $K_q = J_q J_q^\top$ be the current quantized residual tangent Gram matrix and let $\widehat{M}_q$ denote the estimated quantization-induced residual second moment. For an isotropic regularization level τ, define the base whitening transform

$$W_\tau = (K_q + \tau \bar{\kappa}_q I + \varepsilon I)^{-1/2}, \qquad \bar{\kappa}_q = \frac{\operatorname{tr}(K_q)}{m}. \tag{61}$$

The preconditioned quantization geometry is

$$C_{q,\tau} = W_\tau \widehat{M}_q W_\tau^\top. \tag{62}$$

This matrix measures quantization energy after normalization by the local physical/tangent geometry. Let its leading r eigenpairs be $(\gamma_i, u_i)_{i=1}^r$ and let γ_{tail} be the mean discarded eigenvalue. For $\lambda \geq 0$, define

$$a_{\text{tail}} = (1 + \lambda \gamma_{\text{tail}})^{-1/2}, \tag{63}$$

$$F_{\lambda,r} = a_{\text{tail}} I + U_r \operatorname{diag}\Big((1 + \lambda\gamma_i)^{-1/2} - a_{\text{tail}}\Big) U_r^\top, \tag{64}$$

where $U_r = [u_1, \ldots, u_r]$. The generalized low-rank preconditioner is

$$P_{\text{GE-LR}} = F_{\lambda,r} W_\tau, \tag{65}$$

followed by the same scale normalization used for the other QAP families. In contrast to a covariance eigendecomposition in raw residual coordinates, (62) ranks directions by quantization energy relative to the current preconditioned physical geometry.

Finite-sample covariance estimates can be unreliable when the perturbation spectrum is nearly isotropic or weakly separated. We therefore use a held-out reliability gate: GE-LR is activated only when its finite-horizon validation score improves on the best isotropic QAP score by at least 3%; otherwise the method uses the isotropic branch. This gate makes the isotropic design an explicit safe fallback rather than forcing directional regularization in every regime.

Table 6: Exploratory end-to-end results for reliability-gated GE-LR. Relative L_2 solution error is reported over ten held-out seeds. The QAP-I and QAP-Cov+ columns reproduce the corresponding main-study values for context.

PDE	Bits	QAP-I	QAP-Cov+	GE-LR (gated)
Jump diffusion	W6A8	0.148 ± 0.105	0.137 ± 0.063	$\mathbf{0.126 \pm 0.051}$
Jump diffusion	W4A8	0.170 ± 0.081	0.158 ± 0.081	$\mathbf{0.131 \pm 0.065}$
Layered diffusion	W4A8	0.195 ± 0.174	$\mathbf{0.141 \pm 0.045}$	0.156 ± 0.104

On jump diffusion, GE-LR lowers the mean error beyond QAP-I at both precisions. At W4A8 the gain over QAP-I is $1.30\times$ with $7/10$ paired wins (one-sided exact Wilcoxon $p = 4.20 \times 10^{-2}$). The additional improvement over QAP-Cov+ is not statistically significant in the present ten-seed sample. On layered diffusion, GE-LR remains better in mean than QAP-I and the FP-derived baseline but does not match QAP-Cov+. Accordingly, GE-LR is treated as an exploratory structured extension rather than the default directional realization.

The reliability gate also addresses the nearly isotropic Helmholtz control from Appendix H. In an end-to-end W4A8 exploratory run, isotropic QAP gives 0.06265 ± 0.01725 relative solution error, while gated GE-LR gives 0.06277 ± 0.01988; the paired difference is not significant (two-sided exact Wilcoxon $p = 0.846$). Thus the gated construction remains close to the isotropic fallback when no reproducible directional advantage is detected.

J.2 Reliability-gated low-rank covariance shrinkage in 2D

The two-dimensional experiment uses the elliptic problem $-\nabla\cdot(a(x,y)\nabla u) = f$ on $[0,1]^2$ with homogeneous Dirichlet boundary conditions and reference solution $u^\star(x,y) = \sin(\pi x)\sin(\pi y)$. The coefficient is a 2×2 checkerboard alternating between 1 and 1000. A conservative finite-volume/finite-difference discretization on a 10×10 interior grid gives $m = 100$ residual equations. A two-hidden-layer width-32 tanh PINN is trained with hard boundary conditions. We use W4A8 QAT, 100 full-precision Adam warm-start steps, 250 full-batch SGD QAT steps, and QAP recalibration every 50 steps. Pilot seeds $0, 1, 2$ are used for all structural and learning-rate selection; seeds 3–12 are held out unchanged for the reported evaluation.

Let $\widehat{M}_q = U\Lambda U^\top$ be the construction-split residual second moment, with eigenvalues in decreasing order. For rank r, let λ_{tail} be the mean discarded eigenvalue and define the low-rank-plus-tail approximation

$$\widehat{M}_{q,r} = U_r \,\mathrm{diag}((\lambda_i - \lambda_{\mathrm{tail}})_+)\, U_r^\top + \lambda_{\mathrm{tail}} I. \tag{66}$$

To reduce finite-sample variance further, shrink the directional estimate toward an isotropic target,

$$\widehat{M}_{q,r,\alpha} = \alpha \widehat{M}_{q,r} + (1-\alpha)\bar{m}_q I, \qquad \bar{m}_q = \frac{\mathrm{tr}(\widehat{M}_{q,r})}{m}. \tag{67}$$

After trace matching, form the directional candidate

$$P_c = \left(K_q + \tau\bar{\kappa}_q I + \lambda\widetilde{M}_{q,r,\alpha} + \epsilon I\right)^{-1/2} \tag{68}$$

and blend it with the isotropic fallback P_I,

$$P_{\mathrm{RLR}} = \mathcal{N}((1-\beta)P_I + \beta P_c)\,, \tag{69}$$

where $\mathcal{N}$ applies the same tangent-scale normalization as the other QAP families.

Directional regularization is activated only if four calibration-only conditions hold: (i) the finite-horizon score on an independent perturbation split improves over QAP-I by at least 3%; (ii) the

Frobenius alignment between construction- and validation-split second moments is at least 0.5; (iii) the leading r modes explain at least 35% of the construction-split second-moment energy; and (iv) the candidate does not worsen the positive-spectrum condition number by more than 10% or the effective contraction factor by more than 2×10^{-4}. Otherwise the method uses P_I. This gate therefore separates directional covariance use from the isotropic safety fallback.

Pilot selection chooses $r = 8$, $\alpha = 0.5$, isotropic fallback $\tau_I = 0.3$, directional $\tau = 0.1$, $\lambda = 0.01$, blend $\beta = 0.75$, and QAT learning rate 1.5. Thirty-two stochastic perturbation probes are split equally between construction and validation at each recalibration. The gate activates on 70% of held-out recalibration checkpoints on average.

Table 7: Two-dimensional $1000{:}1$ checkerboard diffusion under W4A8 QAT on ten held-out seeds. QAP-RLR uses the reliability-gated low-rank covariance shrinkage in (66)–(69).

Method	Relative L_2 error ↓	Median
Identity	0.0716 ± 0.0033	0.0718
FP-derived	0.2623 ± 0.0254	0.2603
QAP-I	0.0813 ± 0.0162	0.0805
QAP-Cov+ (ungated)	0.1217 ± 0.1439	0.0669
QAP-RLR	$\mathbf{0.0493 \pm 0.0191}$	$\mathbf{0.0515}$

QAP-RLR improves over no preconditioning by $1.45\times$ with $9/10$ paired wins (one-sided exact Wilcoxon $p = 1.95 \times 10^{-3}$), over QAP-I by $1.65\times$ with $10/10$ wins ($p = 9.77 \times 10^{-4}$), and over the FP-derived preconditioner by $5.32\times$ with $10/10$ wins ($p = 9.77 \times 10^{-4}$). The ungated QAP-Cov+ branch contains several highly accurate seeds but also a large failure (0.508 relative error), producing 0.1217 ± 0.1439. QAP-RLR retains low-error directional runs while removing this catastrophic tail; no held-out QAP-RLR seed exceeds 0.082. The paired improvement over ungated QAP-Cov+ is not statistically significant in ten seeds ($p = 0.097$), so the result is interpreted as evidence for stability and improved performance relative to the principal baselines rather than as a universal dominance claim over every covariance realization.

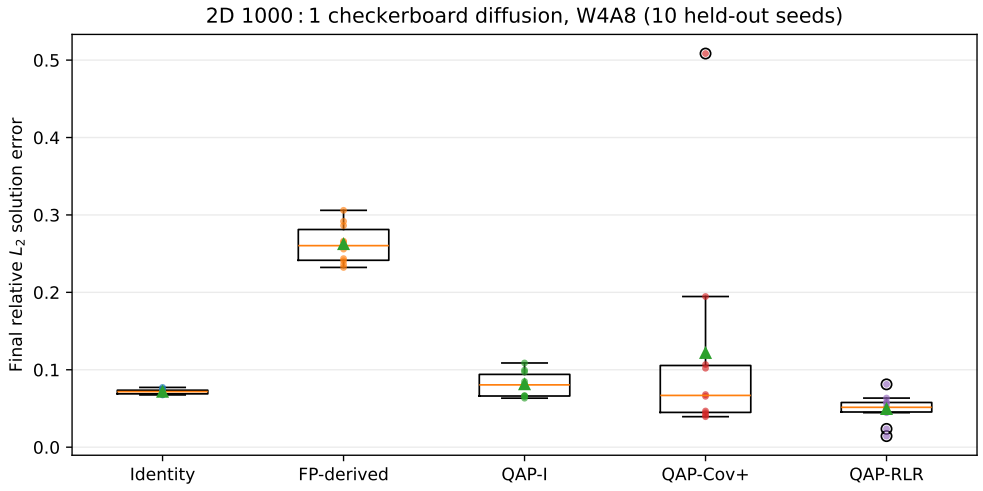


Figure 5: Held-out W4A8 solution-error distributions on the 2D $1000{:}1$ checkerboard problem. The ungated covariance design can achieve very small errors but exhibits a heavy failure tail. Reliability-gated low-rank shrinkage removes the catastrophic outlier while preserving the benefit of directional regularization.

J.3 Low-rank correction of an existing efficient preconditioner

The dense inverse-square-root form is useful for analysis but is not required for deployment. Let P_0 be any efficiently applicable base preconditioner, such as sparse ILU, multigrid, block diagonal, or

another structured numerical transform. A quantization-aware low-rank correction can be written as

$$P_{\text{eff}} = F_q P_0, \qquad F_q = aI + U_r \operatorname{diag}(c_1, \ldots, c_r) U_r^\top, \tag{70}$$

where the coefficients are obtained from the generalized noise-to-physics spectrum, as in (64). For a residual vector v, the additional QAP work after applying P_0 is

$$F_q z = az + U_r \left(c \odot (U_r^\top z) \right), \qquad z = P_0 v. \tag{71}$$

Therefore the incremental application cost is $O(mr)$ and the additional storage is $O(mr)$, compared with $O(m^2)$ storage and work for a dense residual transform. The total cost is $\mathcal{C}(P_0) + O(mr)$, so the benefit is preserved only when the base preconditioner itself is structured or efficiently applicable.

The covariance construction has a compatible matrix-free form. With S centered perturbation probes stored in $D_q \in \mathbb{R}^{m \times S}$, the Gram matrix $D_q^\top D_q$ costs $O(mS^2)$ to form and $O(S^3)$ to diagonalize, while covariance-vector products cost $O(mS)$. Because QAP is recalibrated every T_P network updates rather than at every step, this setup cost is amortized over the intervening QAT iterations.

Table 8: Illustrative single-thread CPU application-cost microbenchmark for a dense $m \times m$ residual transform versus a rank-8 correction of the form in (71). Times are median per application over seven trials on an AMD EPYC 9V74 using PyTorch 2.10; construction and application of the base preconditioner P_0 are excluded.

Residual dimension m	Dense transform (ms)	Rank-8 correction (ms)	Speedup
1024	0.317	0.0130	24.3×
2048	1.567	0.0134	117×
4096	5.988	0.0273	219×

The microbenchmark isolates the linear-algebra kernel and is not a claim about end-to-end training speed. Its purpose is to verify the expected scaling advantage of a low-rank correction. In a complete implementation, wall-clock benefit additionally depends on the cost of P_0, the number of perturbation probes, the recalibration period, automatic-differentiation cost for tangent products, and hardware efficiency. A purely diagonal base preconditioner was also tested as an aggressive cost reduction; on jump diffusion W4A8 its mean end-to-end error was 0.289, substantially worse than the structured QAP variants. This indicates that efficiency should be obtained by preserving a useful physical base preconditioner and making only the quantization-aware correction low rank, rather than discarding the conditioning structure itself.

J.4 Spectral effect of quantization on the effective tangent geometry

To visualize how quantization changes the geometry that preconditioning should target, Figure 6 reports the spectrum of the effective tangent matrix under end-to-end QAT on jump diffusion. For a quantized residual Jacobian J_q and residual-space preconditioner P, the matrix

$$H_P = J_q^\top P^\top P J_q \tag{72}$$

shares its nonzero eigenvalues with the residual-space matrix

$$B_P = P K_q P^\top, \qquad K_q = J_q J_q^\top, \tag{73}$$

because $H_P = (PJ_q)^\top (PJ_q)$ and $B_P = (PJ_q)(PJ_q)^\top$. The preconditioners used in Figure 6 are symmetric, so $B_P = PK_qP$ for the plotted experiments. Panel (a) therefore plots the positive eigenvalues of B_P, aggregated as the median over the ten held-out end-to-end QAT seeds used in the jump-diffusion experiments. Panel (b) reports two summary diagnostics from the same seeds: the active condition number $\kappa^+(H_P)$ and the effective contraction factor

$$\rho_{\text{eff}} = \max_{\lambda_i(H_P) > 0} \left| 1 - \eta \lambda_i(H_P) \right|, \qquad \eta = 0.5, \tag{74}$$

computed over the positive spectrum only.

The figure gives a spectral explanation for the empirical results. Under W6A8 and W4A8 arithmetic, the identity transform and the full-precision-derived preconditioner both produce extremely ill-conditioned effective tangent spectra with rapidly decaying tails and contraction factors arbitrarily

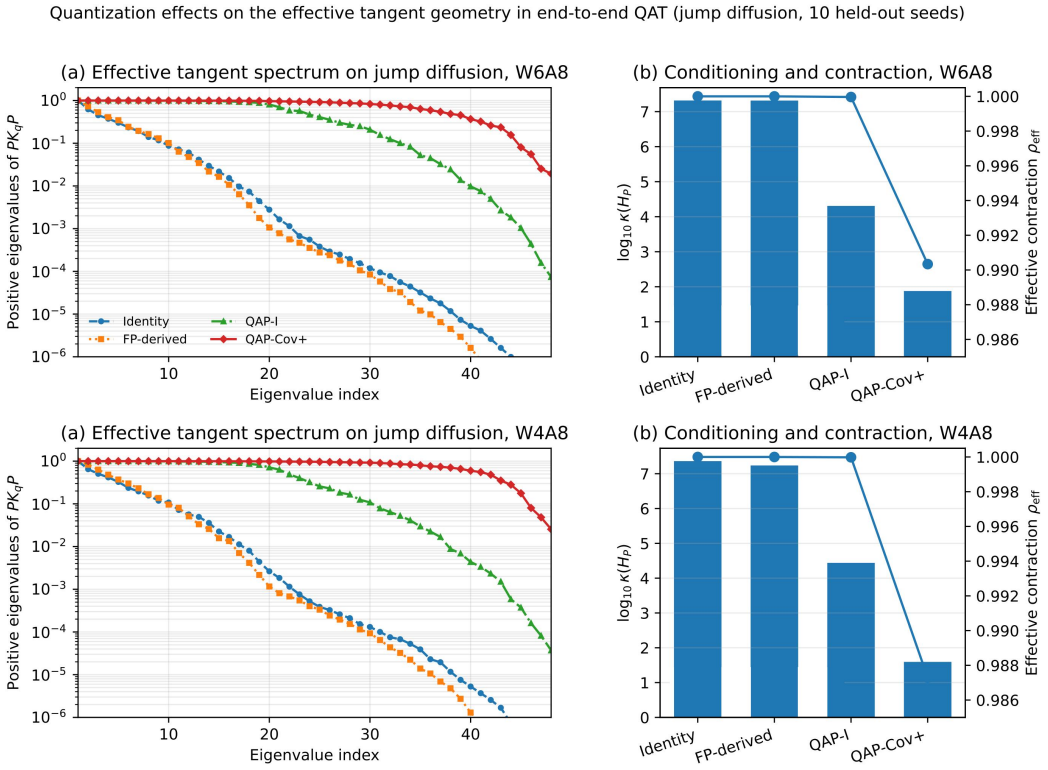


Figure 6: **Quantization modifies the effective tangent geometry in end-to-end QAT.** Panel (a) shows the median positive eigenvalues of $B_P = PK_qP^\top$, which equal the nonzero eigenvalues of $H_P = J_q^\top P^\top P J_q$. For the symmetric preconditioners plotted here, $B_P = PK_qP$. Panel (b) reports the corresponding median active condition number $\kappa^+(H_P)$ and effective contraction factor $\rho_{\text{eff}} = \max_{\lambda_i(H_P)>0} |1 - \eta\lambda_i(H_P)|$ with $\eta = 0.5$. Quantization-aware preconditioning enlarges the weakest effective tangent modes, sharply improves conditioning, and reduces the effective contraction factor relative to both no preconditioning and the full-precision-derived baseline.

close to one. Isotropic QAP raises the small-eigenvalue tail and improves the spectrum substantially, but the covariance-aware design produces the strongest spectral regularization: it lifts the weakest directions by orders of magnitude, reduces the condition number sharply, and yields the smallest effective contraction factor among the compared methods. These observations are consistent with the end-to-end gains reported in the main body and support the interpretation that quantization-aware preconditioning improves performance by modifying the effective tangent spectrum seen by low-bit training.

K One-Bit Extension via Arcsine–Bussgang Geometry

This appendix specializes the quantization-geometry construction to scaled one-bit weights and derives an analytic relation between full-precision weight statistics, the binary scale, and the resulting residual perturbation geometry. The empirical second-moment estimator remains the default when the Gaussian approximation is not appropriate.

K.1 Gaussian one-bit weight-error geometry

Consider a weight vector $w \in \mathbb{R}^d$ modeled locally as zero-mean jointly Gaussian,

$$w \sim \mathcal{N}(0, C_w), \tag{75}$$

and let the scaled binary quantizer be

$$q = A\ \mathrm{sign}(w), \tag{76}$$

where A is a fixed diagonal scale matrix for the covariance calculation. In the experiments below, the standard layerwise scalar scale $A = \alpha I$ is calibrated from the current trained weight tensor using $\alpha = \operatorname{mean}|w|$ and then held fixed in the local Gaussian model. Define $D = \operatorname{Diag}(\operatorname{diag}(C_w))$ and assume its retained coordinates have strictly positive variance; zero-variance coordinates may be removed. The correlation matrix is $R = D^{-1/2} C_w D^{-1/2}$. For jointly Gaussian inputs with this fixed scale, the arcsine law and Bussgang cross-correlation give (Van Vleck & Middleton, 1966; Bussgang, 1952)

$$C_{qq} = A \left[\frac{2}{\pi} \arcsin(R) \right] A, \tag{77}$$

$$C_{qw} = A \sqrt{\frac{2}{\pi}}\, D^{-1/2} C_w, \tag{78}$$

where $\arcsin(\cdot)$ is applied entrywise. Therefore, for the one-bit weight error $e_q = q - w$,

$$\boxed{C_{e_q} = C_{qq} + C_w - C_{qw} - C_{qw}^\top.} \tag{79}$$

Equation (79) shows that the binary scale is not an unrelated tuning constant: it enters quadratically through C_{qq} and linearly through the full-precision/binary cross-covariance. In the scalar case $w \sim \mathcal{N}(0, \sigma^2)$ and $q = \alpha \operatorname{sign}(w)$,

$$\mathbb{E}(q - w)^2 = \alpha^2 + \sigma^2 - 2\alpha\sigma \sqrt{\frac{2}{\pi}}. \tag{80}$$

The minimizer is $\alpha^\star = \sigma\sqrt{2/\pi} = \mathbb{E}|w|$, yielding $\mathbb{E}(q-w)^2 = \sigma^2(1 - 2/\pi)$. Thus the familiar mean-absolute binary scale has a direct statistical interpretation inside the QAP regularization geometry.

Let $R_{W,\ell} = \partial r / \partial \operatorname{vec}(W_\ell)$ denote the residual Jacobian with respect to the weights of layer ℓ. Under a block-independent approximation across layers, the weight-induced residual covariance predicted by the one-bit model is

$$\boxed{\Sigma_{\mathrm{AB}}^{\mathrm{W1}} \approx \sum_\ell R_{W,\ell} C_{e_q,\ell} R_{W,\ell}^\top.} \tag{81}$$

For W1A8, we add the separately measured A8 activation covariance to (81). Because trained weights need not be Gaussian, this construction is used as an analytic approximation rather than as an exact replacement for the empirical second moment $M_q = \Sigma_q + \mu_q \mu_q^\top$.

K.2 Agreement with measured binary residual directions

We first test whether the analytic geometry predicts the directions excited by actual layerwise binarization. At each full-precision warm state, let

$$d_\ell = r_{Q_1(W_\ell)} - r_{\mathrm{fp}} \tag{82}$$

be the residual change obtained by binarizing only layer ℓ, and let α_ℓ be that layer's binary scale. We form the scale-aware source geometry

$$M_{\mathrm{src}}^{(\alpha)} = \sum_\ell \omega_\ell d_\ell d_\ell^\top, \qquad \omega_\ell = \frac{\alpha_\ell^2}{\sum_j \alpha_j^2}. \tag{83}$$

Across the ten held-out jump-diffusion seeds, the normalized Frobenius alignment between $\Sigma_{\mathrm{AB}}^{\mathrm{W1}}$ and $M_{\mathrm{src}}^{(\alpha)}$ is 0.963 ± 0.059. The mean overlap of their leading three-dimensional eigenspaces,

$$\frac{1}{3} \left\| U_{\mathrm{AB},3}^\top U_{\mathrm{src},3} \right\|_F^2, \tag{84}$$

is 0.966 ± 0.034. Without the α_ℓ^2 source weighting, the Frobenius alignment decreases to 0.916 ± 0.062. This provides empirical support for using the binary scales as part of the one-bit regularization geometry rather than treating all binarized layers as equally energetic.

K.3 END-TO-END W1A8 VALIDATION

We next evaluate whether one-bit-aware preconditioning remains useful in full nonlinear QAT. The experiment uses the same 25:1 jump-diffusion PINN, 50-step full-precision warm start, 250 QAT steps, and 50-step QAP recalibration period as the main end-to-end study. Learning rates are selected on pilot seeds $\{0, 1, 2\}$ and then frozen; seeds $\{3, \ldots, 12\}$ are held out. QAP-M2 uses the measured second moment M_q, while QAP-AB uses (81) for the W1 weight component together with the measured A8 activation covariance. The analytic weight geometry is magnitude-matched to the observed deterministic W1 residual energy at each calibration state, so the comparison tests its directional prediction rather than an arbitrary global scale.

Table 9: End-to-end W1A8 jump-diffusion validation on ten held-out seeds. The metric is final relative L_2 solution error of the deployed deterministic quantized network.

Method	Relative L_2 error ↓
Identity (no preconditioning)	0.254 ± 0.026
FP-derived preconditioner	0.607 ± 0.029
QAP-I	$\mathbf{0.111 \pm 0.086}$
QAP-M2 (empirical second moment)	0.113 ± 0.140
QAP-AB (arcsine–Bussgang)	0.146 ± 0.114

One-bit quantization makes reuse of the full-precision preconditioner particularly unfavorable: the FP-derived transform is worse than using no preconditioner. In contrast, QAP-I, QAP-M2, and QAP-AB all substantially reduce deployed solution error. QAP-M2 improves over identity by $2.25\times$ ($8/10$ wins, one-sided exact Wilcoxon $p = 6.84 \times 10^{-3}$) and over the FP-derived preconditioner by $5.39\times$ ($10/10$ wins, $p = 9.77 \times 10^{-4}$). The analytic QAP-AB construction improves over identity by $1.74\times$ ($8/10$ wins, $p = 1.86 \times 10^{-2}$) and over the FP-derived baseline by $4.15\times$ ($10/10$ wins, $p = 9.77 \times 10^{-4}$).

The analytic construction does not significantly outperform either isotropic QAP or the empirical second-moment design in this ten-seed study. Its value is instead structural: the arcsine–Bussgang model recovers the dominant measured binary residual directions with high alignment and provides an explicit route from the one-bit scale and full-precision weight covariance to a residual-space regularizer. For non-Gaussian trained weights or strongly biased binary errors, the empirical M_q remains the safer default.

L EXPLORATORY LARGER-GRID VALIDATION

To check whether the QAP effect persists when the residual space is enlarged, we repeated the 1000:1 two-dimensional checkerboard experiment on a 14×14 interior grid, giving $m = 196$ residual equations rather than $m = 100$. This is an exploratory scale check rather than a replacement for the ten-seed main experiment. We used W4A8 arithmetic, a width-32 PINN, 60 full-precision warm-start steps, 80 QAT steps, recalibration every 40 steps, and 10 perturbation probes. Hyperparameters were fixed before evaluating seeds $\{3, 4, 5\}$; no result from these seeds was used for tuning.

Table 10: Exploratory larger-grid 1000:1 checkerboard validation (14×14, $m = 196$), final relative L_2 solution error over three fixed evaluation seeds.

Method	Relative L_2 error ↓
Identity	0.4853 ± 0.0247
FP-derived	0.4857 ± 0.0246
QAP-I	$\mathbf{0.4662 \pm 0.0175}$
QAP-Cov+	0.4710 ± 0.0233

QAP-I improves on identity on all three evaluation seeds and lowers the mean solution error by 4.1%; QAP-Cov+ also beats identity and the FP-derived baseline on all three seeds, with a 3.0% mean reduction relative to identity. At this larger residual dimension, the isotropic realization is slightly

better in mean than the fixed directional realization, consistent with the paper's reliability-based view that finite-probe directional estimates need not dominate. Because this check has only three evaluation seeds and a shorter training budget, we treat it as supporting scale evidence and do not use it for significance claims.

M Implementation Notes

Estimating H_P. Products with $H_P v = J^\top P^\top P J v$ can be implemented with a Jacobian-vector product followed by a vector-Jacobian product. Power/Lanczos iterations can estimate extremal eigenvalues without forming J. For large PINNs, the condition-number term can be evaluated periodically rather than every optimization step.

Constructing Σ_q in practice. Our default is the layer/source construction in (23), because it remains meaningful for deterministic W1/W4 deployment quantizers. For stochastic rounding we instead use repeated paired residuals. The first-order Jacobian estimator in (26) is used as a theory-aligned ablation. In all cases we store D_q or (U_q, Λ_q) rather than a dense covariance.

Estimating tangent injection. For each full residual difference $d^{(s)}$, compute $w^{(s)} = P^\top P(d^{(s)} - \bar{d})$ and then a VJP $g^{(s)} = J^\top w^{(s)}$. The sample-covariance average $(N-1)^{-1} \sum_s \|g^{(s)}\|_2^2$ is (38). A Hutchinson sketch can be used when residual dimension is very large.

Quantizers. Start with symmetric per-channel weight QAT and per-tensor activation QAT, then ablate percentile/EMA clipping. Report clipping rates. When P is quantized, its scale and granularity must be reported separately from the network quantizers.

Nonlinear PDEs. For Burgers/Allen–Cahn, J and Σ_q are local quantities. Estimate them at a slower QAP timescale and freeze P for several QAT updates. For Newton or time-stepping formulations, QAP can be applied to each local linearized residual operator or shared over a parameter/time window.

N Scope and Limitations

The evidence combines an exact controlled study, theory-aligned local validation, nonlinear end-to-end QAT on two heterogeneous one-dimensional diffusion problems, a $1000{:}1$ two-dimensional checkerboard problem, and an exploratory W1A8 extension that tests the binary-scale and arcsine–Bussgang geometry. The end-to-end results establish that quantization-aware QAP can outperform both no preconditioning and a strong full-precision-derived preconditioner after quantization, and the 2D study shows that reliability-gated low-rank shrinkage can also outperform isotropic QAP when covariance estimation is stabilized. These results do not establish universal covariance superiority: Appendix H gives one Helmholtz control where the isotropic design is preferable. The principal 2D experiment uses a modest $m = 100$ residual equations; Appendix L adds an exploratory $m = 196$ check, but substantially larger residual spaces remain to be tested. The raw deterministic PDE residual can also increase while solution error decreases. Broader validation should therefore cover finer and irregular 2D discretizations, three-dimensional, nonlinear, and time-dependent PDEs, larger PINN architectures, alternative quantizers, low-bit preconditioner storage, wall-clock overhead, and hardware-level efficiency. The present claim is that quantization geometry is a necessary preconditioning design variable and that directional regularization is beneficial when that geometry is sufficiently reliable.