QUANTUM SHARPNESS-AWARE MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum neural networks (QNNs) deployed on noisy intermediate-scale quantum (NISQ) devices often suffer from severe performance degradation due to intrinsic hardware noise. By spotlighting this challenge from a deep neural network (DNN) perspective, we raise an intriguing question: *Can we incorporate a concept of flat minima in DNNs into QNNs for quantum noise robustness?* As our answer, we propose quantum sharpness-aware minimization (Q-SAM), which integrates the flat minima concept into the Riemannian quantum parameter space to seek minima that remain *flat* or invariant under quantum state perturbations induced by hardware noise. We formally connect a bounded perturbation in the quantum state manifold to a bound in the Riemannian quantum parameter space, providing theoretical ground for Q-SAM's state noise robustness. Across image classification tasks, including MNIST-10, Fashion-MNIST, and CIFAR-10, Q-SAM with 4-, 6-, and 8-qubit QNNs consistently exhibit superior robustness against diverse noise channels compared to baselines, also validated on real IBM quantum hardware.

1 INTRODUCTION

Quantum neural networks (QNNs), instantiated by parameterized quantum circuits (PQCs), are increasingly investigated for their potential to represent stochastic model distributions that enable high expressibility in quantum machine learning tasks (Benedetti et al., 2019). However, the practical realization of this potential on noisy intermediate-scale quantum (NISQ) hardware is fundamentally bottlenecked by intrinsic noise (Preskill, 2018). In particular, PQCs are susceptible to errors such as gate infidelity, decoherence, and readout errors. These errors distort the quantum state, raising a quantum-native hardship that stifles the potential of quantum artificial intelligence.

To address this challenge, existing strategies largely fall into two categories: quantum error mitigation (QEM) (Temme et al., 2017) and noise-aware training (NAT) (Wang et al., 2022). QEM techniques, such as zero-noise extrapolation (ZNE) (Li & Benjamin, 2017) and probabilistic error cancellation (PEC) (Van Den Berg et al., 2023), typically require significant computational overhead during inference. Conversely, NAT injects simulated noise during training to enhance robustness. However, NAT has been shown to alter the loss landscape detrimentally, accelerating noise-induced barren plateaus (NIBP) and gradient collapse, thereby hindering effective training (Wang et al., 2021). Moreover, NAT relies heavily on accurate characterization of hardware-specific noise profiles. To achieve generalized robustness, it necessitates exhaustive training across a diverse spectrum of error channels, which results in prohibitive computational costs.

By shifting our attention to deep neural networks (DNNs), sharpness-aware minimization (SAM) (Foret et al., 2021) has emerged as a promising method to improve generalization and robustness by converging to flat minima, which are located on a flatter loss landscape with a lower curvature. Importantly, flatter minima learned by SAM are found to strongly foster model robustness. Specifically, SAM has been largely validated in robustness against domain shift (Cha et al., 2021), adversarial perturbations (Zhang et al., 2024) in supervised learning, action perturbations of agents in deep reinforcement learning (Lee & Yoon, 2025), and exposure bias in generative diffusion models (Lee et al., 2025), supported both in empirical and theoretical evidence.

From the perspective of DNN's flat minima, we spotlight the QNN's challenge of intrinsic hardware noise: *Can SAM in DNNs be a remedy for the robustness of QNNs against hardware noise?* Our answer is 'Yes', but a simple extension does not work; a quantum-geometry-aware manipulation of the SAM objective is required. SAM assumes the parameter space is Euclidean, whereas the

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.

mapping from parameters to quantum states is highly non-Euclidean. A small Euclidean shift in parameters can result in a drastic change in the quantum state depending on local geometry, making Euclidean flatness an inaccurate proxy for robustness in the quantum non-Euclidean Hilbert space.

We bridge the gap between quantum parameter geometry and hardware noise resilience. Built on SAM, we propose **quantum sharpness-aware minimization (Q-SAM)**, which promotes flatter loss behavior on quantum non-Euclidean Hilbert space, i.e., Riemannian quantum parameter space, by formulating a Min-Max problem for worst-case loss against quantum state perturbations.

Our key contributions are summarized as follows.

1. **Geometry-Based Noise Reinterpretation:** We theoretically link physical hardware noise to bounded quantum-state perturbations via the Stinespring dilation and Fubini-Study metric. This analysis allows us to translate the quantum state perturbations into the QNN's parameter displacements, motivating sharpness-aware optimization of QNNs.
2. **Q-SAM Algorithm:** We propose an optimizer framework that minimizes the worst-case loss within a region defined by the quantum geometric tensor (QGT), effectively finding minima of QNNs that are flat with respect to the intrinsic quantum geometry.
3. **Robustness Validation:** We reveal that a naive SAM with Euclidean perturbation does not lead to QNN's quantum noise robustness. In contrast, noise-free training with Q-SAM enhances robustness to noisy inference across multiple channels (e.g., depolarizing, amplitude/phase damping, bit/phase flip). As a proof-of-concept of Q-SAM on real IBM-Quantum hardware, we confirm that the simulation result translates to practical robustness under physical noise conditions.

2 Related Work

2.1 Noise Robustness in Quantum Machine Learning

Quantum Error Mitigation (QEM) methods are post-processing approaches to correct measurement results after circuit execution. Zero-Noise Extrapolation (ZNE) (Li & Benjamin, 2017) estimates error-free expectation values by artificially increasing noise levels and extrapolating to the zero-noise limit, while Probabilistic Error Cancellation (PEC) (Van Den Berg et al., 2023) reconstructs the ideal output by sampling from a quasi-probability distribution representing the inverse noise channel. However, these methods incur exponential sampling overhead scaling with error rates (Temme et al., 2017), rendering them computationally prohibitive for deep circuits. Moreover, they do not improve the intrinsic quality of the trained parameters themselves.

Noise-Aware Training (NAT) incorporates noise models into the training process to find robust parameters (Wang et al., 2022). However, NAT relies on precise characterization of hardware-specific noise profiles. Furthermore, injecting noise alters the loss landscape, often accelerating the onset of Noise-Induced Barren Plateaus (NIBP) and hindering convergence (Wang et al., 2021). Our method, Q-SAM, achieves intrinsic robustness without explicit noise simulation, avoiding the computational overhead and trainability issues associated with NAT.

2.2 Information Geometry in Optimization

Quantum Natural Gradient (QNG) utilizes the Fubini-Study metric to incorporate the quantum geometry in gradient directions for faster convergence, but does not explicitly optimize for geometric flatness or robustness (Stokes et al., 2020). This ensures updates follow the steepest descent direction in the Hilbert space. While QNG and Q-SAM both use Fubini-Study metric, their goals are different. QNG uses the metric to accelerate convergence, whereas Q-SAM uses it to define a robust region, ensuring the found minima are flat with respect to the intrinsic quantum geometry.

Sharpness-Aware Minimization (SAM) improves generalization in classical deep learning by searching for flat minima (Foret et al., 2021); geometry-aware variants replace its Euclidean ball with a rescaled ball (Kwon et al., 2021) or a classical-Fisher ellipsoid of the output distribution (Kim et al., 2022). Neither captures the geometry of the quantum state: Euclidean flatness is a poor proxy for quantum state robustness, and the Fisher ellipsoid measures changes in the classical output distribution rather than in the state itself. Q-SAM addresses this mismatch by defining the SAM perturbation region using the QFIM, the metric induced on parameter space by the PQC state manifold, thereby linking geometric flatness to robustness against hardware-induced state perturbations.

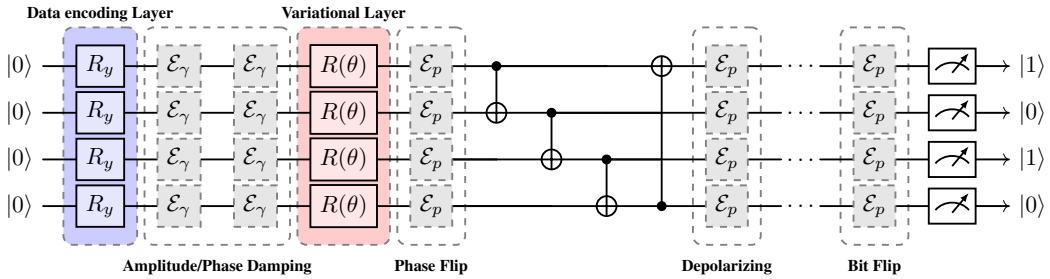


Figure 1: **PQC architecture and layer-wise noise injection.** In each circuit layer, independent quantum noise channels $\mathcal{E}_p$ (grey blocks) are injected immediately after the encoding and trainable unitary operations. This scheme simulates the accumulation of errors depth-wise.

Table 1: Single-qubit CPTP noise channels used in our experiments and their Kraus operators. p and γ denote channel strengths. I, X, Y, Z are the identity and Pauli matrices.

Noise Channel	**Description**	**Kraus Operators**
Bit Flip	Applies a Pauli-X error with probability p.	$K_0 = \sqrt{1-p}I, \quad K_1 = \sqrt{p}X$
Phase Flip	Applies a Pauli-Z error with probability p.	$K_0 = \sqrt{1-p}I, \quad K_1 = \sqrt{p}Z$
Depolarizing	Applies a random Pauli error (X, Y, Z) with total probability p (each with $p/3$).	$K_0 = \sqrt{1-p}I, \quad K_1 = \sqrt{p/3}X$ $K_2 = \sqrt{p/3}Y, \quad K_3 = \sqrt{p/3}Z$
Phase Damping	Pure dephasing (T2), suppresses off-diagonal terms without population loss.	$K_0 = \left(\begin{smallmatrix}1 & 0\\ 0 & \sqrt{1-\gamma}\end{smallmatrix}\right), \quad K_1 = \left(\begin{smallmatrix}0 & 0\\ 0 & \sqrt{\gamma}\end{smallmatrix}\right)$
Amplitude Damping	Energy relaxation from $\|1\rangle \to \|0\rangle$ (spontaneous emission).	$K_0 = \left(\begin{smallmatrix}1 & 0\\ 0 & \sqrt{1-\gamma}\end{smallmatrix}\right), \quad K_1 = \left(\begin{smallmatrix}0 & \sqrt{\gamma}\\ 0 & 0\end{smallmatrix}\right)$

3 Preliminaries

3.1 Parameterized Quantum Circuits (PQCs)

PQCs serve as the fundamental building blocks of QNNs (Benedetti et al., 2019), illustrated in Figure 1. Formally, a PQC consists of a parameterized unitary $U(\theta)$ acting on an initial state $|0\rangle^{\otimes n}$, producing the state: $|\psi(\theta)\rangle = U(\theta)|0\rangle$. In the density matrix formalism, adopted to accommodate noise effects, this state is represented as: $\rho(\theta) = |\psi(\theta)\rangle\langle\psi(\theta)|$. The network output is typically defined as the expectation value of an observable O, which can be written as $f(\theta) = \langle\psi(\theta)|O|\psi(\theta)\rangle = \text{Tr}[O\rho(\theta)]$.

3.2 Quantum Information Geometry

The geometry of the quantum state space is characterized by the Fubini-Study metric (Kolodrubetz et al., 2017). This metric induces a geometry on the parameter space, where the Quantum Fisher Information Matrix (QFIM), i.e., $G(\theta)$, serves as the metric tensor. For a small parameter change $d\theta$, the distance s^2_{FS} between states is approximated by: $ds^2_{FS} \approx \sum_{i,j} G_{i,j}(\theta)d\theta_i d\theta_j = d\theta^\top G(\theta)d\theta$, where (i,j)-th element of the QFIM is given by: $G_{i,j}(\theta) = \text{Re}\left[\langle\partial_i\psi|\partial_j\psi\rangle - \langle\partial_i\psi|\psi\rangle\langle\psi|\partial_j\psi\rangle\right]$.

3.3 Quantum Noise Channels

Quantum noise is modeled as a completely positive trace-preserving (CPTP) map $\mathcal{E}$ acting on the density matrix ρ. The noisy output state $\tilde{\rho}$ is given by: $\tilde{\rho} = \mathcal{E}(\rho)$. This admits complementary representations, each offering a distinct perspective on the noise process utilized in our framework.
Operator-Sum Representation (Kraus et al., 1983). First, the Kraus representation decomposes the channel into a set of operators $\{K_k\}$, which is particularly suitable for numerical simulations: $\mathcal{E}(\rho) = \sum_k K_k \rho K_k^\dagger$, subject to $\sum_k K_k^\dagger K_k = I$. For a single-qubit presentation of noise injection, bit flip, phase flip, depolarizing, phase damping, and amplitude damping, are detailed in Table 1

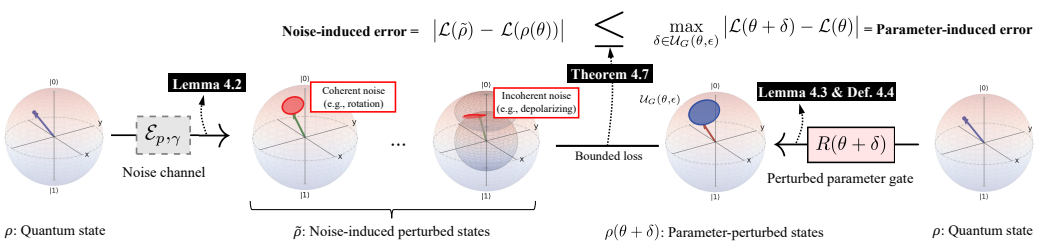


Figure 2: **A quick overview of the theoretical chain from noise to parameter flatness.**

Stinespring Dilation (Stinespring, 1955). Any CPTP map can be represented by a unitary interaction U_{SE} with an auxiliary environment: $\mathcal{E}(\rho) = \mathrm{Tr}_E\left[U_{SE}(\rho \otimes |0\rangle_E\langle 0|)U_{SE}^\dagger\right]$, where $\mathrm{Tr}_E[\cdot]$ denotes the partial trace that discards the unobserved environment and returns the reduced noisy state of the system. The unitary operator is modeled as $U_{SE}(\alpha) = e^{-i\alpha H_{SE}}$, relating the non-unitary noise on the system to a geometric rotation by an angle α in the extended Hilbert space.
Stochastic Pauli Channel (Wallman & Emerson, 2016). Finally, general noise channels can be mapped into stochastic Pauli errors via Pauli twirling. The resulting channel $\mathcal{E}_{PT}$ corresponds to an average over the n-qubit Pauli group $\mathcal{P}$: $\mathcal{E}_{PT}(\rho) = \frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}} P^\dagger\mathcal{E}(P\rho P^\dagger)P$. This simplifies arbitrary noise into a probabilistic mixture of Pauli errors while preserving the average gate fidelity.

4 Q-SAM and Its Theory

4.1 Theory: A Link from Noise to Parameters

Our theory follows three steps: (1) defining hardware noise as a perturbation in the quantum state space, (2) relating parameter perturbations to quantum state changes via information geometry, and (3) linking these two concepts by anchoring in the quantum state space, thus establishing an upper bound on the noise-induced loss. Figure 2 illustrates these chains; the left part illustrates state perturbation induced by noise, whereas the right one shows state perturbation due to parameter variations.

4.1.1 Hardware Noise to State Perturbation

We first define the operation of quantum noise. To rigorously bridge our experimental noise models (represented with Kraus style) with geometric distance, we utilize the Stinespring dilation.

Definition 4.1 (Quantum Noise Channel)**.** Let $\rho(\theta) = |\psi(\theta)\rangle\langle\psi(\theta)|$ be the system density matrix. A hardware noise channel is a Completely Positive Trace-Preserving (CPTP) map, denoted as $\mathcal{E}$. By Stinespring's dilation theorem (Stinespring, 1955), there exists a unitary operator U_{SE} on an extended Hilbert space $\mathcal{H}_S \otimes \mathcal{H}_E$ such that $\mathcal{E}(\rho) = \mathrm{Tr}_E\left[U_{SE}(\rho \otimes |0\rangle_E\langle 0|)U_{SE}^\dagger\right]$, where $|0\rangle_E$ is a fixed reference state of the environment. Given an orthonormal basis $\{|k\rangle_E\}$ of $\mathcal{H}_E$, the corresponding Kraus operator is defined as $K_k := \langle k|_E U_{SE}|0\rangle_E$, which yields the equivalent operator-sum representation of noise, i.e., $\mathcal{E}(\rho) = \sum_k K_k\rho K_k^\dagger$.

Lemma 4.2 (Noise to Quantum State)**.** *Let $\tilde{\rho} = \mathcal{E}(\rho(\theta))$ be the output of a CPTP channel parameterized by a physical noise strength (p or γ in Table 1). For small noise with $\lambda \ll 1$, where $\lambda \sim \sqrt{p}$ or $\sqrt{\gamma}$, the noise-induced quantum state displacement is locally bounded as: $D(\rho(\theta), \tilde{\rho}) \leq \epsilon = \mathcal{O}(\lambda)$, where $D(\cdot,\cdot)$ denotes the Fubini–Study distance, and the geometric bound scales with the square root of the physical noise strength. For a circuit with n qubits and L layers, the accumulated radius scales as $\epsilon_{n,L} = \mathcal{O}(\sqrt{nL}\lambda)$ in the perturbative local noise regime.*

□ Proof is in Appendix A.1.

4.1.2 Parameter Perturbation to State

Next, we try mapping from perturbation of parameters θ to changes in the quantum state. This mapping is governed by the quantum Fisher information matrix (QFIM).

Lemma 4.3 (Parameter to Quantum State (Provost & Vallee, 1980)). *Consider a perturbation δ in the parameter space. The coherent distance in the quantum state space is approximated by the quadratic form of the QFIM $G(\theta)$: $D^2(|\psi(\theta)\rangle, |\psi(\theta+\delta)\rangle)=\delta^\top G(\theta)\delta + \mathcal{O}(\|\delta\|^3)$, where $G(\theta)$ acts as the Riemannian metric tensor.*

The parameter perturbation (right-hand side) is mapped to a distance in the quantum state space (left-hand side). Let us define the perturbed region on the parameter space.

Definition 4.4 (Geometry-Aware Parameter Region). For a given θ and radius $\epsilon > 0$, define the quantum state geometry-aware region in the parameter space as, $\mathcal{U}_G(\theta, \epsilon) := \{\theta + \delta : \delta^\top G(\theta)\delta \leq \epsilon^2\}$. By Lemma 4.3, $\mathcal{U}_G(\theta, \epsilon)$ corresponds to a Fubini-Study distance ball of radius ϵ up to second-order approximation.

4.1.3 Linking Noise, State, and Parameters

Our conceptual chain is as follows: We have mapped quantum hardware noise to local state perturbation (Lemma 4.2); At other side, we borrow the results in (Provost & Vallee, 1980) to link QNN's parameter perturbation to local state perturbation (Lemma 4.3). For an n-qubit circuit with L layers, the accumulated noise displaces the state by at most $\epsilon_{n,L}$. A model robust to any parameter perturbation within this state-space radius will effectively bound the worst-case error caused by hardware noise. To connect the resulting incoherent noise channel to geometry-aware parameter region, we upper-bound their loss deviation by a worst-case Pauli envelope obtained via Pauli twirling.

Lemma 4.5 (Incoherent Noise Bounded by Coherent Pauli Envelope). *Let $\tilde{\rho} = \mathcal{E}(\rho(\theta))$ be the noisy state under a stochastic Pauli channel. Then, the loss deviation by incoherent noise satisfies*

$$\big|\mathcal{L}(\tilde{\rho}) - \mathcal{L}(\rho(\theta))\big| \;\leq\; C_{\text{lip}} \cdot \max_{P\in\mathcal{P}} \big|L_{\text{meas}}(P\rho(\theta)P^\dagger) - L_{\text{meas}}(\rho(\theta))\big|, \tag{1}$$

where $\mathcal{P}$ denotes the Pauli group, $L_{\text{meas}}(\rho) := \text{Tr}[O\rho]$ is the affine measurement-level loss, and C_{lip} is the Lipschitz constant of the classical post-processing ℓ in $\mathcal{L} = \ell \circ L_{\text{meas}}$.

For stochastic Pauli channels, this bound follows from viewing the noise as a randomized mixture of coherent Pauli perturbations (Wallman & Emerson, 2016), whose worst-case impact on affine measurement observables upper-bounds that of the mixture. We refer to Equation (1) as the *Pauli-envelope condition*; the subsequent result applies to any CPTP channel satisfying this condition.

□ Proof is in Appendix A.2.

Lemma 4.6 (Pauli Perturbation Covered by $\mathcal{U}_G$). *Let $P \in \mathcal{P}$ be a Pauli operator with $D\big(\rho(\theta), P\rho(\theta)P^\dagger\big) \leq \epsilon_{n,L}$, within the accumulated state-space radius $\epsilon_{n,L}$ in Lemma 4.2. Then, Pauli-induced loss deviation is upper-bounded by quantum state geometry-aware loss envelope:*

$$\big|\mathcal{L}(P\rho(\theta)P^\dagger) - \mathcal{L}(\rho(\theta))\big| \;\leq\; \max_{\delta\in\mathcal{U}_G(\theta,\epsilon_{n,L})} \big|\mathcal{L}(\theta+\delta) - \mathcal{L}(\theta)\big|. \tag{2}$$

□ Proof is in Appendix A.3.

Now, we have connected the worst-case perturbations in the parameter space to the Pauli-induced loss. We are ready to stretch it to the noise-induced loss via Lemma 4.5.

Theorem 4.7 (Geometry–Noise Robustness Link). *Let $\mathcal{L}(\theta)$ be the loss function and $\tilde{\rho} = \mathcal{E}(\rho(\theta))$ be the noisy state under a CPTP channel satisfying the Pauli-envelope condition in Equation (1). Under Lemmas 4.2,4.3,4.5,4.6 the noise-induced loss deviation is upper-bounded as*

$$\big|\mathcal{L}(\tilde{\rho}) - \mathcal{L}(\rho(\theta))\big| \;\leq\; C_{\text{lip}} \cdot \max_{\delta\in\mathcal{U}_G(\theta,\epsilon_{n,L})} \big|\mathcal{L}(\theta+\delta) - \mathcal{L}(\theta)\big|. \tag{3}$$

This concludes the full link between the loss deviation due to the noise (left-hand side) and the loss deviation induced by parameter shifts (right-hand side). Its empirical tightness and non-vacuity under increasing circuit size is discussed in Appendix A.4.

Corollary 4.8 (Local Geometry-Aware Loss Bound). *Under Theorem 4.7, for sufficiently small $\epsilon_{n,L}$, $|\mathcal{L}(\tilde{\rho}) - \mathcal{L}(\rho(\theta))| \leq C_{\text{lip}}\left[\epsilon_{n,L}\sqrt{\nabla_\theta\mathcal{L}(\theta)^\top G(\theta)^{-1}\nabla_\theta\mathcal{L}(\theta)} + O(\epsilon_{n,L}^2)\right]$. Thus, the local noise-induced loss deviation is controlled jointly by the accumulated state-space radius and the loss sensitivity measured in the inverse-QFIM geometry.*

Remark 4.9 (Training Objective for Noise-Robust QNNs). Noise robustness is commonly understood as the ability of a model to preserve its nominal performance under hardware noise. For QNNs, the corresponding degradation can be quantified by the QNN-loss deviation $|\mathcal{L}(\tilde{\rho}) - \mathcal{L}(\rho)|$ for $D(\rho, \tilde{\rho}) \leq \epsilon_{n,L}$. Under Theorem 4.7, a sufficient objective for limiting this degradation is

$$\min_{\theta} \max_{D(\rho,\tilde{\rho}) \leq \epsilon_{n,L}} \mathcal{L}(\tilde{\rho}) \leq \min_{\theta} \max_{\delta \in \mathcal{U}_G(\theta, \epsilon_{n,L})} \mathcal{L}(\theta + \delta). \tag{4}$$

Thus, Q-SAM minimizes a geometry-aware upper bound on noise-induced loss degradation, promoting preservation of nominal QNN performance under hardware noise.

4.2 Q-SAM Algorithm

In brief, Q-SAM solves the right-hand side optimization in Equation (4). Building upon the inner-outer optimization of SAM, Q-SAM leverages quantum geometry. The procedure is in Algorithm 1.

4.2.1 Inner Maximization for Worst-Case Geometry-Aware Perturbation

We search for the worst-case perturbation $\delta^* \in \mathcal{U}_G$ that maximizes the loss under the geometry-aware constraint $\delta^\top G(\theta)\delta \leq \eta^2$, where $\eta \geq 0$. Using a first-order Taylor expansion of the loss around θ, the inner 'Max' problem is approximated as: $\delta^* \approx \arg\max_\delta \nabla_\theta \mathcal{L}(\theta)^\top \delta$, subject to $\delta^\top G(\theta)\delta \leq \eta^2$. Solving this with Lagrange multipliers yields a closed-form solution whose direction is aligned with quantum natural gradient (QNG) (Stokes et al., 2020):

$$\delta^* = \eta \frac{G(\theta)^{-1}\nabla_\theta \mathcal{L}(\theta)}{\sqrt{\nabla_\theta \mathcal{L}(\theta)^\top G(\theta)^{-1} \nabla_\theta \mathcal{L}(\theta)}}, \tag{5}$$

where η is a hyperparameter that controls the radius of the QFIM-defined perturbation region over which Q-SAM seeks to flatten the loss surface. Notably, Q-SAM uses this QNG update as the parameter perturbation direction, maximizing the loss.

Algorithm 1 Q-SAM Algorithm

Input: Dataset $\mathcal{D}$, Loss $\mathcal{L}$, Learning rate α, Perturbation bound η.
Initialize: Parameters θ_0.
for $t = 0$ **to** $T - 1$ **do**
  Sample mini-batch $\mathcal{B}$.
  Compute gradient $g_t = \nabla \mathcal{L}_\mathcal{B}(\theta_t)$.
  Compute QFIM $G(\theta_t)$ (or approximation).
  // Solve Inner Maximization
  Compute optimal perturbation δ^* using Equation (5).
  // Outer Minimization Step
  Compute gradient at perturbed point: $g_{adv} = \nabla \mathcal{L}_\mathcal{B}(\theta_t + \delta^*)$.
  Update parameters: $\theta_{t+1} = \theta_t - \alpha g_{adv}$.
end for

4.2.2 Outer Minimization for Update

Once δ^* is obtained, we compute the gradient at the perturbed location, i.e., $\theta + \delta^*$, and update the parameters. This update step explicitly reduces the worst-case loss $\max_{\delta \in \mathcal{U}_G} \mathcal{L}(\theta + \delta)$, moving toward the region that is low and geometrically flat in loss.

Q-SAM incurs additional cost over Adam optimizer mainly from the QFIM construction, which can be mitigated by a block-diagonal approximation (Stokes et al., 2020) and a periodic metric refresh (every k steps). A detailed cost analysis is provided in Appendix B.

5 Experiments

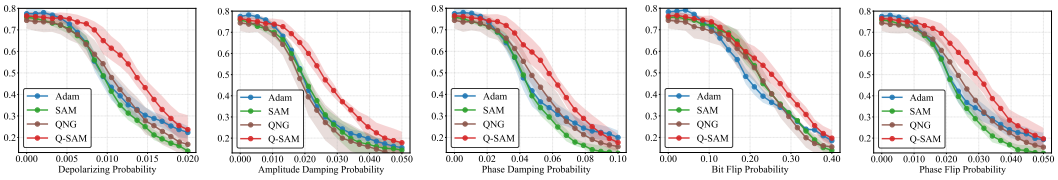
We present an empirical validation of Q-SAM under controlled noise simulations and on IBM Heron (133-qubit) hardware, assessing whether geometry-aware flatness provides robustness to diverse quantum state perturbations and transfers to realistic hardware noise. We also analyze the geometric properties of the converged minima through Riemannian loss-landscape visualization and Hessian spectrum analysis to examine how they relate to the observed robustness.

5.1 Experimental Setup

Dataset and Preprocessing. We evaluate our framework on three 10-class classification tasks: MNIST (Lecun et al., 1998), Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky et al., 2009). For MNIST-10, following Wang et al. (2022) and Schleich et al. (2024), we downsampled the

Table 2: **Robustness comparison across five noise channels on three datasets.** Train (Nominal) denotes accuracy at $p = \gamma = 0$. Test (Perturbed) denotes accuracy under five noise channels. Values in parentheses denote degradation from nominal accuracy.

Dataset	Optimizer	Train	Test (Perturbed)				
		Nominal	Depolarizing	Amp. Damping	Phase Damping	Bit Flip	Phase Flip
		$(p = 0)$	$(p^* = 0.013)$	$(\gamma^* = 0.024)$	$(\gamma^* = 0.053)$	$(p^* = 0.25)$	$(p^* = 0.026)$
MNIST	Adam	77.50	35.16 (-42.34)	34.84 (-42.66)	36.50 (-41.00)	36.32 (-41.18)	36.84 (-40.66)
4q · QNN	SAM	75.84	31.34 (-44.50)	37.68 (-38.16)	33.84 (-42.00)	40.18 (-35.66)	34.00 (-41.84)
Angle	QNG	74.40	35.86 (-38.54)	32.48 (-41.92)	42.94 (-31.46)	41.02 (-33.38)	43.20 (-31.20)
	Q-SAM	76.50	**54.18 (-22.32)**	**54.16 (-22.34)**	**55.84 (-20.66)**	**49.34 (-27.16)**	**56.00 (-20.50)**
		$(p = 0)$	$(p^* = 0.011)$	$(\gamma^* = 0.035)$	$(\gamma^* = 0.040)$	$(p^* = 0.040)$	$(p^* = 0.012)$
F-MNIST	Adam	80.46	45.46 (-35.00)	51.32 (-29.14)	56.12 (-24.34)	58.06 (-22.40)	57.26 (-23.20)
6q · QNN	SAM	79.00	47.20 (-31.80)	53.40 (-25.60)	56.46 (-22.54)	57.08 (-21.92)	57.28 (-21.72)
Amplitude	QNG	81.08	34.68 (-46.40)	45.48 (-35.60)	45.00 (-36.08)	46.60 (-34.48)	47.80 (-33.28)
	Q-SAM	76.60	**58.46 (-18.14)**	**65.86 (-10.74)**	**66.74 (-9.86)**	**61.40 (-15.20)**	**69.06 (-7.54)**
		$(p = 0)$	$(p^* = 0.025)$	$(\gamma^* = 0.12)$	$(\gamma^* = 0.36)$	$(p^* = 0.27)$	$(p^* = 0.015)$
CIFAR-10	Adam	81.06	24.66 (-56.40)	44.28 (-36.78)	47.38 (-33.68)	48.36 (-32.70)	45.94 (-35.12)
8q · QCNN	SAM	78.66	38.66 (-40.00)	44.34 (-34.32)	59.66 (-19.00)	58.62 (-20.04)	54.32 (-24.34)
Amplitude	QNG	79.68	46.72 (-32.96)	41.72 (-37.96)	57.38 (-22.30)	59.56 (-20.12)	61.28 (-18.40)
	Q-SAM	76.66	**58.78 (-17.88)**	**50.96 (-25.70)**	**65.54 (-11.12)**	**63.42 (-13.24)**	**66.04 (-10.62)**



(a) Depolarizing (b) Amp. Damping (c) Phase Damping (d) Bit Flip (e) Phase Flip

Figure 3: **Robustness evaluation curves on MNIST.** Test accuracy degradation evaluated under increasing noise probabilities. Q-SAM (Red) consistently maintains higher accuracy compared to Adam (Blue), SAM (Green), and QNG (Brown) across all noise channels.

images to 4×4 resolution ($d = 16$) and rescale the pixel values to $[0, \pi]$ for angle encoding. Fashion-MNIST images are downsampled to 8×8 resolution and normalized for amplitude encoding. For CIFAR-10, we use the Quantum Convolutional Neural Network (QCNN) architecture (Cong et al., 2019), and normalized the images for amplitude encoding. A summary of the experimental setups used in recent QNN classification studies is provided in Appendix C.

Model Architecture. We consider three model configurations to evaluate Q-SAM across different circuit architectures and encoding schemes. For MNIST, we use a 4-qubit QNN with layered data re-uploading; for Fashion-MNIST, a 6-qubit QNN with four strongly entangling layers; and for CIFAR-10, an 8-qubit QCNN with convolution and pooling stages. For the amplitude-encoding models, we apply a Hadamard layer to all measured qubits immediately before measurement to use an X-basis readout, avoiding degenerate solutions that are insensitive to phase noise under standard Z-basis readout. Detailed circuit architectures and implementation settings are provided in Appendix E.

Compared Methods. We consider four optimization strategies: (i) **Adam**, representing standard stochastic optimization, which minimizes the empirical loss directly in the Euclidean parameter space; (ii) **SAM**, which minimizes the worst-case loss within an L_2 neighborhood to seek Euclidean-flat minima; (iii) **QNG** (Stokes et al., 2020), which preconditions the gradient by the inverse QFIM to follow the quantum-state geometry, but does not enforce geometric flatness; and (iv) **Q-SAM** (Ours), which minimizes the worst-case loss within a QFIM-defined Riemannian region to explicitly optimize quantum-geometric flatness.

Training and Noise Injection Protocol. All models are trained in a *noise-free* simulation environment using matched random seeds and optimization settings across baselines. Q-SAM uses a block-diagonal approximation of the QFIM, and the perturbation bound is matched between SAM and Q-SAM within each experimental configuration. For simulated noisy evaluation, as illustrated in Figure 1, CPTP noise channels are applied after each unitary operation so that quantum errors accumulate with circuit depth. Detailed training and noise-injection settings are provided in Appendix E.

5.2 Robustness Evaluation

We evaluate the trained models under noisy inference with the noise channels defined in Section 3.3.

Evaluation on Single-Source Noise. We evaluate the robustness of Q-SAM against individual noise channels by sweeping the error probability for five noise channels. Figure 3 summarizes the full test accuracy decay on MNIST, with shaded regions indicating the 95% confidence interval over 5 seeds. Table 2 reports representative noise-level results across MNIST, Fashion-MNIST, and CIFAR-10, where **Q-SAM** consistently exhibits improved robustness over all baselines under the evaluated noise channels, despite differences in encoding schemes and circuit architectures. Figure 3 further shows that Q-SAM effectively mitigates both the sharp decay caused by cumulative errors and the linear degradation from one-shot readout errors (e.g., Bit Flip). In contrast, while SAM performs comparably to Adam in low-noise regimes, it degrades rapidly as noise intensity increases, indicating that Euclidean sharpness minimization fails to ensure robustness to severe quantum noise. QNG only marginally improves over Adam/SAM and falls far below Q-SAM, confirming that quantum geometry alone is insufficient without a flatness objective.

Evaluation on Composite Noise. We test Q-SAM under a composite noise model calibrated to the error profile of real quantum hardware (133-qubit IBM Heron processor). This model incorporates gate depolarization ($p \approx 0.01$), thermal relaxation parameters ($\gamma \approx 0.005$), and readout errors ($p \approx 0.02$), using parameters derived from actual device calibration data (McKay et al., 2023). As summarized in Table 3, **Q-SAM** achieves a test accuracy of 44.16%, maintaining a higher robustness margin than Adam (29.66%), SAM (24.84%) and QNG (26.94%). This result confirms Q-SAM effectively ensures robustness even under multi-channel hardware-emulated noise.

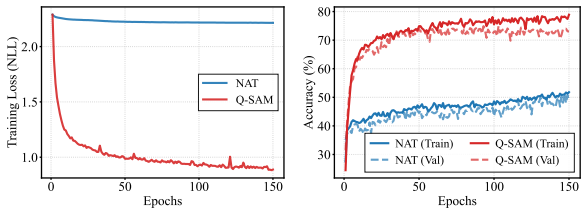
Validation on NISQ Hardware. We deploy the trained model directly on the physical IBM Heron processor. Inference is performed on a test set of 150 samples with 4096 shots per circuit. As in Table 3, **Q-SAM** achieves the highest test accuracy of 49.33%, outperforming Adam (27.33%), SAM (26.00%) and QNG (24.33%). While the standard Adam optimizer suffers a 50.17% performance degradation from its noise-free baseline, Q-SAM limits this drop to 27.17%, effectively reducing the performance loss. NISQ hardware experiments confirm the robustness of Q-SAM against real noise induced by a practical quantum computing unit, validating our theory and algorithm design.

Table 3: **Robustness validation on MNIST across simulated and physical environments.** Accuracy under noise-free simulation, calibrated composite noise simulation, and on the IBM Heron processor (NISQ Hardware, 4096 shots).

Optimizer	Train Nominal ($p = 0$)	Test (Perturbed)	
		Composite	NISQ Hardware
Adam	77.50	29.66 (-47.84)	27.33 (-50.17)
SAM	75.84	24.84 (-51.00)	26.00 (-49.84)
QNG	74.40	26.94 (-47.46)	24.33 (-50.07)
Q-SAM	**76.50**	**44.16 (-32.34)**	**49.33 (-27.17)**

5.3 Noise-Free vs. Noise-Aware Training

We compare noise-free training with noise-aware training (NAT). NAT injects the same IBM-Heron-calibrated composite noise model used in Table 3 at every training-step circuit evaluation. This repeated noisy-circuit evaluation in NAT incurs substantial computational overhead and slower convergence. In Figure 4, the noise-free Q-SAM (red-colored lines) exhibits a steep learning curve with rapid loss minimization, while the noise-induced NAT (blue-colored lines) converges slowly, with the training loss decreasing only marginally and the accuracy improving more slowly. When tested with the same composite noise used during NAT training, NAT achieves 50.52% clean and 49.42% test accuracy, while Q-SAM reaches 76.50% clean and 44.16%, respectively. Q-SAM maintains much higher clean accuracy while retaining competitive noisy accuracy. By attaining this robustness without explicit noise simulation during training, Q-SAM demonstrates how the concept of flat minima in DNNs brings a novel benefit to QNNs.

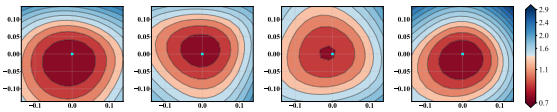


(a) Loss curve (b) Accuracy curve

Figure 4: **Comparison of noise-free and noise-aware training.** Dotted lines indicate validation results.

Table 4: **Comparison of Sharpness Metrics.** Maximum ($\lambda_{\max}$) and mean (λ_{mean}) eigenvalues of Hessian. ↓: means flatter.

Optimizer	Euclid. (H)		Rieman. ($G^{-1}H$)	
	$\lambda_{\max}$ ↓	λ_{mean} ↓	$\lambda_{\max}$ ↓	λ_{mean} ↓
Adam	18.66	0.67	81.18	2.94
SAM	**14.07**	**0.56**	71.90	2.74
QNG	16.23	0.62	78.77	2.83
Q-SAM	14.13	0.60	**60.32**	**2.65**



(a) Q-SAM (b) QNG (c) SAM (d) Adam

Figure 5: **Loss landscape visualization.** (a) Q-SAM converges to a region flat w.r.t. the Quantum Fisher Information. (b) QNG follows quantum geometry but converges to a sharper region. (c) SAM finds a minimum flat in Euclidean space but sharp in quantum geometry. (d) Adam converges to a sharper minimum.

5.4 Loss Landscape Analysis and Visualization

We analyze the loss landscape structure using both Hessian spectrum analysis and visualization. We compare **Euclidean flatness**, defined in the flat parameter space, with **Riemannian flatness**, which accounts for the local geometry of the quantum statistical manifold by QFIM, denoted as G.

Hessian Spectrum Analysis. We computed the maximum eigenvalue ($\lambda_{\max}$) of the Hessian H to quantify sharpness. For the Riemannian metric, we analyzed the generalized eigenvalues of $G^{-1}H$, which represent the curvature relative to the quantum state space. As summarized in Table 4, **Adam** converges to the sharpest minima in both metrics, confirming its lack of robustness. Notably, **SAM** achieves the lowest Euclidean sharpness ($\lambda_{\max}^{\text{Euc}} \approx 14.07$), slightly better than Q-SAM. This confirms that SAM effectively optimizes Euclidean flatness, which does not fully translate to the quantum domain. **Q-SAM** achieves the lowest Riemannian sharpness ($\lambda_{\max}^{\text{Riem}} \approx 60.32$), reducing the geometry-aware curvature by 25% compared to Adam. **QNG** ($\lambda_{\max}^{\text{Riem}} \approx 78.77$) is sharper even than SAM, confirming that the metric alone does not yield Riemannian-flatness.

Visualization. Figure 5 visualizes the loss surface projected along two random directions normalized by the Fisher Information Matrix (G). Consistent with the numerical data in Table 4, the minima found by Adam are located in steep valleys. While SAM finds a region that is relatively flat in parameter space, Q-SAM identifies a minimum that is effectively "flatter" along the directions of high curvature in the quantum manifold, validating our theoretical claim.

6 Concluding Remarks

In this work, we introduce **Q-SAM**, quantum geometry-aware training that achieves robustness in Quantum Machine Learning. By identifying the mismatch between Euclidean parameter flatness and quantum state stability, we propose an optimization strategy that minimizes the worst-case loss within a region induced by the quantum geometry. We provide a formal link between **hardware-induced state perturbations** and **quantum geometry-aware Riemannian parameter space perturbations**, providing a theoretical foundation for our approach. Our extensive experiments, ranging from diverse noise simulations to actual quantum hardware, validate that Q-SAM significantly outperforms both standard training (Adam), Euclidean sharpness minimization (SAM), and quantum-geometric minimization without flatness objective (QNG) against various quantum noise channels. Furthermore, by achieving this robustness in a strictly noise-free training regime, Q-SAM avoids the prohibitive computational costs and trainability issues associated with NAT.

Limitations and Potentials. Demonstrations are still limited to small-scale QNNs with up to 10 qubits. For this reason, we evaluate Q-SAM on a simple task rather than more complex benchmarks. MNIST and CIFAR are currently accepted as valid benchmarks for QNNs (as surveyed in Appendix C), but exploring more practical and complex benchmarks remains a clear challenge for quantum research, including our work. Also, we first discovered the potential of flat minima in QNNs, but it is nontrivial to clarify how this relates to the current witness of the QNNs' loss-surface nature (e.g., barren plateaus). In a broader impact, we emphasize that our work has the potential to bridge two distinct research areas by leveraging a concept from DL, i.e., flat minima, as a remedy for challenges in the quantum domain. On the technical side, Q-SAM can improve the utility of quantum machines in the presence of hardware noise, paving the way toward NISQ and high-qubit devices in the future.

AI USE STATEMENT

In this work, we used generative AI tools for no tasks with required disclosure. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to aid and polish writing of the manuscript. We have reviewed all AI-assisted work. We have reviewed all AI-assisted text, which was revised by the authors as needed. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or new data set releases, and we do not identify potentially harmful insights, methodologies and applications, conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance issues, or research integrity issues that require further discussion.

REPRODUCIBILITY STATEMENT

We provide the Q-SAM algorithm and experimental protocol in the main text. The assumptions, proofs, and additional theoretical derivations are provided in Appendix A. Additional robustness and scalability experiments are reported in Appendix D, including the multi-qubit settings in Appendix D.2. Detailed data preprocessing, implementation settings, optimization hyperparameters, random seeds, and computational environment are provided in Appendix E and Table 12. These sections provide the information needed to reproduce and verify the reported results.

REFERENCES

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

Denis Bokhan, Alena S Mastiukova, Aleksey S Boev, Dmitrii N Trubnikov, and Aleksey K Fedorov. Multiclass classification using quantum convolutional neural networks with hybrid quantum-classical learning. *Frontiers in Physics*, 10:1069985, 2022.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, 2021.

Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

Xiaodong Ding, Zhihui Song, Jinchen Xu, Yifan Hou, Tian Yang, and Zheng Shan. Scalable parameterized quantum circuits classifier. *Scientific Reports*, 14(1):15886, 2024.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *The International Conference on Learning Representations*, 2021.

Julien Gacon, Christa Zoufal, Giuseppe Carleo, and Stefan Woerner. Simultaneous perturbation stochastic approximation of the quantum fisher information. *Quantum*, 5:567, 2021.

Tak Hur, Leeseok Kim, and Daniel K Park. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 4(1):3, 2022.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Bálint Koczor and Simon C Benjamin. Quantum natural gradient generalized to noisy and nonunitary circuits. *Physical Review A*, 106(6):062416, 2022.

Michael Kolodrubetz, Dries Sels, Pankaj Mehta, and Anatoli Polkovnikov. Geometry and nonadiabatic response in quantum and classical systems. *Physics Reports*, 697:1–87, 2017.

Ioannis Kolotouros and Petros Wallden. Random natural gradient. *Quantum*, 8:1503, 2024.

Karl Kraus, Arno Böhm, John D Dollard, and WH Wootters. *States, effects, and operations fundamental notions of quantum theory: Lectures in mathematical physics at the university of Texas at Austin*. Springer, 1983.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Hyun Kyu Lee and Sung Whan Yoon. Flat reward in policy parameter space implies robust reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Taehwan Lee, Kyeongkook Seo, Jaejun Yoo, and Sung Whan Yoon. Understanding flatness in generative models: Its role and benefits. In *IEEE International Conference on Computer Vision*, 2025.

Ying Li and Simon C Benjamin. Efficient variational quantum simulator incorporating active error minimization. *Physical Review X*, 7(2):021050, 2017.

Ao Liu, Cuihong Wen, and Jieci Wang. Lean classical-quantum hybrid neural network model for image classification. *Advanced Quantum Technologies*, 8(10):2400703, 2025.

David C McKay, Ian Hincks, Emily J Pritchett, Malcolm Carroll, Luke CG Govia, and Seth T Merkel. Benchmarking quantum processor performance at scale. *arXiv preprint arXiv:2311.05933*, 2023.

John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

JP Provost and G Vallee. Riemannian structure on manifolds of quantum states. *Communications in Mathematical Physics*, 76(3):289–301, 1980.

Farina Riaz, Shahab Abdulla, Hajime Suzuki, Srinjoy Ganguly, Ravinesh C Deo, and Susan Hopkins. Accurate image multi-class classification neural network model with quantum entanglement approach. *Sensors*, 23(5):2753, 2023.

Philipp Schleich, Marta Skreta, Lasse B Kristensen, Rodrigo A Vargas-Hernández, and Alán Aspuru-Guzik. Quantum deep equilibrium models. *Advances in Neural Information Processing Systems*, 37:31940–31967, 2024.

Arsenii Senokosov, Alexandr Sedykh, Asel Sagingalieva, Basil Kyriacou, and Alexey Melnikov. Quantum machine learning for image classification. *Machine Learning: Science and Technology*, 5(1):015040, 2024.

W Forrest Stinespring. Positive functions on c*-algebras. *Proceedings of the American Mathematical Society*, 6(2):211–216, 1955.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020.

Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. Error mitigation for short-depth quantum circuits. *Physical Review Letters*, 119:180509, 2017.

Aleksei Tolstobrov, Gleb Fedorov, Shtefan Sanduleanu, Shamil Kadyrmetov, Andrei Vasenin, Aleksey Bolgar, Daria Kalacheva, Viktor Lubsanov, Aleksandr Dorogov, Julia Zotova, et al. Hybrid quantum learning with data reuploading on a small-scale superconducting quantum simulator. *Physical Review A*, 109(1):012411, 2024.

Ewout Van Den Berg, Zlatko K Minev, Abhinav Kandala, and Kristan Temme. Probabilistic error cancellation with sparse Pauli–Lindblad models on noisy quantum processors. *Nature Physics*, 19(8):1116–1121, 2023.

Joel J Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016.

Hanrui Wang, Jiaqi Gu, Yongshan Ding, Zirui Li, Frederic T Chong, David Z Pan, and Song Han. QuantumNAT: Quantum noise-aware training with noise injection, quantization and normalization. In *59th ACM/IEEE Design Automation Conference*, 2022.

Peiyong Wang, Casey R Myers, Lloyd CL Hollenberg, and Udaya Parampalli. Quantum hamiltonian embedding of images for data reuploading classifiers. *Quantum Machine Intelligence*, 7(1):35, 2025.

Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1):6961, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, Yifei Wang, and Zeming Wei. On the duality between sharpness-aware minimization and adversarial training. In *Forty-First International Conference on Machine Learning*, 2024.

Min-Gang Zhou, Zhi-Ping Liu, Hua-Lei Yin, Chen-Long Li, Tong-Kai Xu, and Zeng-Bing Chen. Quantum neural network for quantum neural computing. *Research*, 6:0134, 2023.

A Proofs and Additional Derivations

This section provides formal proofs and supporting derivations for the results in Section 4.

A.1 Noise-Induced Local State Perturbation (Lemma 4.2)

A.1.1 General Bound via Stinespring Dilation

Let $\tilde{\rho} = \mathcal{E}(\rho)$ and assume a Stinespring dilation

$$\tilde{\rho} = \mathrm{Tr}_E\Big[U_{SE}\,(\rho \otimes |0\rangle\langle 0|)\,U_{SE}^\dagger\Big]. \tag{6}$$

Assume the system–environment interaction is weak so that

$$U_{SE} = e^{-i\alpha H_{SE}} = I - i\alpha H_{SE} - \frac{\alpha^2}{2}H_{SE}^2 + O(\alpha^3), \qquad \alpha \geq 0. \tag{7}$$

Define the joint input state $\Omega := \rho \otimes |0\rangle\langle 0|$ and the joint output $\Omega' := U_{SE}\Omega U_{SE}^\dagger$. Using the Taylor expansion and keeping terms up to first order yields

$$\begin{aligned}\Omega' &= \big(I - i\alpha H_{SE} + O(\alpha^2)\big)\,\Omega\,\big(I + i\alpha H_{SE} + O(\alpha^2)\big)\\ &= \Omega - i\alpha[H_{SE}, \Omega] + O(\alpha^2).\end{aligned} \tag{8}$$

Taking the partial trace over the environment gives

$$\tilde{\rho} = \mathrm{Tr}_E[\Omega'] = \rho - i\alpha\,\mathrm{Tr}_E\big([H_{SE}, \rho \otimes |0\rangle\langle 0|]\big) + O(\alpha^2), \tag{9}$$

and thus

$$\tilde{\rho} - \rho = -i\alpha\,\mathrm{Tr}_E\big([H_{SE}, \rho \otimes |0\rangle\langle 0|]\big) + O(\alpha^2). \tag{10}$$

Using $\|\mathrm{Tr}_E(X)\|_1 \leq \|X\|_1$ and $\|[A, B]\|_1 \leq 2\|A\|_\infty\|B\|_1$, we obtain

$$\begin{aligned}\|\tilde{\rho} - \rho\|_1 &\leq \alpha\,\big\|\,[H_{SE}, \rho \otimes |0\rangle\langle 0|]\,\big\|_1 + O(\alpha^2)\\ &\leq 2\alpha\,\|H_{SE}\|_\infty\,\|\rho \otimes |0\rangle\langle 0|\|_1 + O(\alpha^2).\end{aligned} \tag{11}$$

Since $\rho \otimes |0\rangle\langle 0|$ is a density operator, $\|\rho \otimes |0\rangle\langle 0|\|_1 = 1$, hence

$$\|\tilde{\rho} - \rho\|_1 \leq 2\alpha\,\|H_{SE}\|_\infty + O(\alpha^2). \tag{12}$$

Finally, to connect the above trace-distance control to the state-space distance D used in the main text, we take D as the Fubini–Study angle on the variational manifold, extended to mixed states via the fidelity angle

$$D(\rho, \sigma) \;:=\; \arccos\sqrt{F(\rho, \sigma)}, \tag{13}$$

which reduces to the standard Fubini–Study distance for pure states. Using the Fuchs–van de Graaf inequality,

$$1 - \sqrt{F(\rho, \sigma)} \;\leq\; \tfrac{1}{2}\|\rho - \sigma\|_1, \tag{14}$$

we have $\sqrt{F(\rho, \tilde{\rho})} \geq 1 - \frac{1}{2}\|\tilde{\rho} - \rho\|_1$.

We apply the elementary inequality $\arccos(1 - x) \leq \frac{\pi}{2}\sqrt{x}$ for $x \in [0, 1]$, with $x := 1 - \sqrt{F(\rho, \tilde{\rho})}$:

$$D(\rho, \tilde{\rho}) = \arccos\sqrt{F(\rho, \tilde{\rho})} \leq \frac{\pi}{2}\sqrt{1 - \sqrt{F(\rho, \tilde{\rho})}} \leq \frac{\pi}{2}\sqrt{\frac{1}{2}\|\tilde{\rho} - \rho\|_1} = \frac{\pi}{2\sqrt{2}}\sqrt{\|\tilde{\rho} - \rho\|_1}. \tag{15}$$

Combining with Equation (12) yields $D(\rho, \tilde{\rho}) = \mathcal{O}(\sqrt{\alpha})$ for the Stinespring interaction angle α, and the explicit derivations in Appendix A.1.2 confirm $D = \mathcal{O}(\sqrt{p}), \mathcal{O}(\sqrt{\gamma})$. Identifying the FS-distance scale $\lambda \sim \sqrt{p}, \sqrt{\gamma}$ as in Lemma 4.2, we conclude $D(\rho, \tilde{\rho}) \leq \epsilon = \mathcal{O}(\lambda)$ without any noise-dependent prefactor (in contrast to a Lipschitz bound that would require $c \sim 1/\sqrt{1 - F}$, divergent as $F \to 1$).

We next extend the local bound to repeated noise applications across a circuit. Let $M = M(n, L)$ denote the total number of local noise-channel applications in an n-qubit circuit with L layers, and

let $\mathcal{E}_m$ denote the CPTP noise channel applied at the m-th local noise location, $m = 1, \ldots, M$. Assume that the corresponding local system–environment interaction strengths are bounded by α and that the interaction Hamiltonians are uniformly bounded. By Equation (12), each local noise application then induces a trace-distance disturbance of order $\mathcal{O}(\alpha)$.

Let ρ_m and $\tilde{\rho}_m$ denote the ideal and noisy states, respectively, after the m-th local noise location. Using the triangle inequality, unitary invariance, and contractivity of the trace distance under CPTP maps,

$$\|\tilde{\rho}_m - \rho_m\|_1 \leq \|\tilde{\rho}_{m-1} - \rho_{m-1}\|_1 + \mathcal{O}(\alpha). \tag{16}$$

Applying this recursively over M local noise applications gives

$$\|\tilde{\rho}_M - \rho_M\|_1 = \mathcal{O}(M\alpha). \tag{17}$$

Combining Equation (17) with Equation (15), the accumulated state-space displacement satisfies

$$D(\rho_M, \tilde{\rho}_M) \leq \sqrt{\|\tilde{\rho}_M - \rho_M\|_1} = \mathcal{O}(\sqrt{M\alpha}). \tag{18}$$

Defining the corresponding single-application geometric scale as $\lambda \sim \sqrt{\alpha}$, we obtain

$$D(\rho_M, \tilde{\rho}_M) = \mathcal{O}(\sqrt{M}\lambda). \tag{19}$$

For the layer-wise local-noise setting considered in this work, $M = \mathcal{O}(nL)$, and hence

$$\epsilon_{n,L} = \mathcal{O}(\sqrt{nL}\lambda). \tag{20}$$

This establishes the accumulated-radius statement in Lemma 4.2.

A.1.2 Explicit Derivation for Physical Noise Models

To relate the FS-distance scale λ to physical noise parameters (e.g., probability p or rate γ), we directly compute the Fubini–Study distance using the Kraus representation. We focus on two representative noise models: Depolarizing and Amplitude Damping channels.

Depolarizing Channel. For a single-qubit depolarizing channel, the map is defined as $\mathcal{E}_p(\rho) = (1-p)\rho + \frac{p}{2}I$. Recall that the fidelity between a pure state $\rho = |\psi\rangle\langle\psi|$ and a mixed state σ simplifies to the expectation value:

$$F(\rho, \sigma) = \langle\psi|\sigma|\psi\rangle. \tag{21}$$

Substituting $\sigma = \mathcal{E}_p(\rho)$ into this definition, we derive the fidelity as follows:

$$\begin{aligned} F &= \langle\psi| \left((1-p)|\psi\rangle\langle\psi| + \frac{p}{2}I\right) |\psi\rangle \\ &= (1-p)\langle\psi|\psi\rangle\langle\psi|\psi\rangle + \frac{p}{2}\langle\psi|I|\psi\rangle \\ &= (1-p) + \frac{p}{2} \\ &= 1 - \frac{p}{2}. \end{aligned} \tag{22}$$

The Fubini–Study distance is defined as $D = \arccos\sqrt{F}$. We apply a two-step Taylor expansion. First, using the approximation $\sqrt{1-x} \approx 1 - x/2$ for small x:

$$\sqrt{F} = \sqrt{1 - \frac{p}{2}} \approx 1 - \frac{p}{4}. \tag{23}$$

Second, using $\arccos(1-\delta) \approx \sqrt{2\delta}$ for small δ (where $\delta = p/4$):

$$D \approx \sqrt{2 \cdot \frac{p}{4}} = \sqrt{\frac{p}{2}} = \frac{1}{\sqrt{2}}\sqrt{p}. \tag{24}$$

Thus, the distance scales as $D = \mathcal{O}(\sqrt{p})$, corresponding to $\lambda \sim \sqrt{p}$.

Amplitude Damping Channel. The amplitude damping channel, characterized by a decay rate γ, is defined by the Kraus operators:

$$K_0 = |0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|, \quad K_1 = \sqrt{\gamma}|0\rangle\langle 1|. \tag{25}$$

The fidelity between a pure input state $|\psi\rangle$ and the output $\mathcal{E}(\rho) = \sum K_k|\psi\rangle\langle\psi|K_k^\dagger$ is given by:

$$F = \sum_k |\langle\psi|K_k|\psi\rangle|^2. \tag{26}$$

We consider the state $|\psi\rangle = |1\rangle$ as the input, as it represents the worst-case scenario (since the ground state $|0\rangle$ is invariant under damping). Calculating the matrix elements for each Kraus operator:

- For K_0:
$$\langle 1|K_0|1\rangle = \langle 1|(|0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|)|1\rangle = \sqrt{1-\gamma}. \tag{27}$$
- For K_1:
$$\langle 1|K_1|1\rangle = \langle 1|(\sqrt{\gamma}|0\rangle\langle 1|)|1\rangle = \sqrt{\gamma}\langle 1|0\rangle = 0. \tag{28}$$

Substituting these back into the fidelity formula:

$$F = |\sqrt{1-\gamma}|^2 + |0|^2 = 1 - \gamma. \tag{29}$$

Similar to the depolarizing case, the distance $D = \arccos\sqrt{1-\gamma}$ is approximated for small γ:

$$D \approx \arccos(1-\gamma/2) \approx \sqrt{2(\gamma/2)} = \sqrt{\gamma}. \tag{30}$$

This confirms that for damping noise, $D = \mathcal{O}(\sqrt{\gamma})$, corresponding to $\lambda \sim \sqrt{\gamma}$.

A.2 Incoherent Noise Bounded by Coherent Pauli Envelope (Lemma 4.5)

We decompose the loss as $\mathcal{L}(\rho) = \ell(L_{\text{meas}}(\rho))$, where

$$L_{\text{meas}}(\rho) \;\coloneqq\; \text{Tr}(O\rho) \tag{31}$$

is the measurement-level expectation (genuinely *affine* in ρ) and ℓ is the classical post-processing (softmax + cross-entropy in our setting), locally Lipschitz with constant C_{lip} on the bounded image of L_{meas}. Lipschitz composition gives

$$\big|\mathcal{L}(\sigma_1) - \mathcal{L}(\sigma_2)\big| \;\le\; C_{\text{lip}}\,\big|L_{\text{meas}}(\sigma_1) - L_{\text{meas}}(\sigma_2)\big|, \tag{32}$$

and this separation makes the convexity argument below rigorous, since Jensen's inequality only holds for affine functionals.

Let $\mathcal{P}$ denote the n-qubit Pauli group (modulo global phase). Given a CPTP noise channel $\mathcal{E}$, define its Pauli-twirled channel as

$$\mathcal{E}_{\text{PT}}(\rho) \;\coloneqq\; \frac{1}{|\mathcal{P}|}\sum_{P\in\mathcal{P}} P^\dagger\,\mathcal{E}(P\rho P^\dagger)\,P \;=\; \sum_{P\in\mathcal{P}} q_P\,P\rho P^\dagger, \qquad q_P \ge 0, \quad \sum_P q_P = 1, \tag{33}$$

which is a stochastic Pauli channel (Wallman & Emerson, 2016) that preserves the average gate fidelity between $\mathcal{E}$ and the identity, capturing the dominant operational error strength of $\mathcal{E}$.

Since L_{meas} is affine in ρ, $L_{\text{meas}}(\mathcal{E}_{\text{PT}}(\rho)) = \sum_P q_P\,L_{\text{meas}}(P\rho P^\dagger)$, and Jensen's inequality with $\sum_P q_P = 1$ yields

$$\begin{aligned}\big|L_{\text{meas}}(\mathcal{E}_{\text{PT}}(\rho)) - L_{\text{meas}}(\rho)\big| &= \Big|\sum_P q_P\big(L_{\text{meas}}(P\rho P^\dagger) - L_{\text{meas}}(\rho)\big)\Big| \\ &\le \max_{P\in\mathcal{P}}\big|L_{\text{meas}}(P\rho P^\dagger) - L_{\text{meas}}(\rho)\big|.\end{aligned} \tag{34}$$

Composing with Equation (32) lifts this to the full loss,

$$\big|\mathcal{L}(\mathcal{E}_{\text{PT}}(\rho)) - \mathcal{L}(\rho)\big| \;\le\; C_{\text{lip}}\cdot\max_{P\in\mathcal{P}}\big|L_{\text{meas}}(P\rho P^\dagger) - L_{\text{meas}}(\rho)\big|, \tag{35}$$

which is the rigorous form with the Lipschitz constant C_{lip} explicit on the right-hand side.

It remains to transfer the bound from $\mathcal{E}_{\text{PT}}$ to $\mathcal{E}$. For symmetric NISQ channels (depolarizing, dephasing, bit/phase flip) the channel is already a Pauli channel, so $\mathcal{E}_{\text{PT}} = \mathcal{E}$ and Equation (35) gives the lemma directly. For non-unital channels (amplitude damping is the canonical case in our experiments), $\mathcal{E} \neq \mathcal{E}_{\text{PT}}$, but the affine decomposition above admits a rigorous fallback that bypasses the

transfer entirely. Since $L_{\text{meas}}(\mathcal{E}(\rho)) - L_{\text{meas}}(\rho) = \text{Tr}[O(\mathcal{E}(\rho) - \rho)]$, the matrix Hölder inequality $|\text{Tr}(AB)| \leq \|A\|_\infty \|B\|_1$ gives $|L_{\text{meas}}(\mathcal{E}(\rho)) - L_{\text{meas}}(\rho)| \leq \|O\|_\infty \|\mathcal{E}(\rho) - \rho\|_1$. Combined with the trace-distance bound Equation (12) from Appendix A.1 ($\|\mathcal{E}(\rho) - \rho\|_1 = \mathcal{O}(\epsilon)$) and the Lipschitz lift Equation (32),

$$\left|\mathcal{L}(\mathcal{E}(\rho)) - \mathcal{L}(\rho)\right| \;\leq\; C_{\text{lip}} \cdot \|O\|_\infty \cdot \|\mathcal{E}(\rho) - \rho\|_1 \;=\; \mathcal{O}(\epsilon), \tag{36}$$

establishing the same $\mathcal{O}(\epsilon)$ rate without requiring channel symmetry or twirling. The Pauli-envelope bound Equation (35) establishes Equation (1) directly for stochastic Pauli channels and is the form that connects to the QFIM-induced parameter ball $\mathcal{U}_G$ in Lemma 4.6. For non-unital channels such as amplitude damping, the Hölder bound Equation (36) instead provides a channel-symmetry-free $\mathcal{O}(\epsilon)$ loss bound, while the stronger geometry-aware envelope relation is evaluated empirically in Appendix A.4.

A.3 Pauli Perturbation Covered by Geometry-Aware Region (Lemma 4.6)

For the accumulated local-noise setting, recall the geometry-aware parameter region in Definition 4.4 with radius $\epsilon_{n,L}$:

$$\mathcal{U}_G(\theta, \epsilon_{n,L}) = \left\{\theta + \delta \;:\; \delta^\top G(\theta)\delta \leq \epsilon_{n,L}^2\right\}. \tag{37}$$

Fix $P \in \mathcal{P}$ and define $\rho_P := P\rho(\theta)P^\dagger$. Assume the coherent Pauli displacement is bounded by the same state-space radius $\epsilon_{n,L}$ used in Lemma 4.2, i.e.,

$$D\big(\rho(\theta), \rho_P\big) \leq \epsilon_{n,L}. \tag{38}$$

Since P is unitary, $\rho_P = P\rho(\theta)P^\dagger$ is a pure state. By Definition 4.4 together with Lemma 4.3, the QFIM-induced region $\mathcal{U}_G(\theta, \epsilon_{n,L})$ corresponds (up to second-order approximation) to the set of parameter perturbations whose realized states $\rho(\theta + \delta)$ lie within Fubini–Study distance $\epsilon_{n,L}$ of $\rho(\theta)$. Concretely, for sufficiently small δ, $D^2(\rho(\theta), \rho(\theta + \delta)) \approx \delta^\top G(\theta)\delta$, so $\theta + \delta \in \mathcal{U}_G(\theta, \epsilon_{n,L})$ so that $D(\rho(\theta), \rho(\theta + \delta)) \leq \epsilon_{n,L} + \mathcal{O}(\epsilon_{n,L}^2)$.

Since ρ_P lies within this Fubini–Study $\epsilon_{n,L}$-ball by Equation (38), and $\mathcal{U}_G(\theta, \epsilon_{n,L})$ covers the realizable pure states within the same ball, the loss deviation at ρ_P is upper-bounded by the worst-case loss deviation over $\mathcal{U}_G$:

$$\left|\mathcal{L}(P\rho(\theta)P^\dagger) - \mathcal{L}(\rho(\theta))\right| \;\leq\; \max_{\delta:\, \theta+\delta \in \mathcal{U}_G(\theta, \epsilon_{n,L})} \left|\mathcal{L}(\theta + \delta) - \mathcal{L}(\theta)\right|, \tag{39}$$

which is exactly Equation (2). The coverage step implicitly requires the parameterization to be expressive enough that the realizable states $\{\rho(\theta + \delta) : \delta \in \mathcal{U}_G\}$ span the relevant tangent direction of the observable O within the $\epsilon_{n,L}$-ball; this expressiveness is empirically verified for our ansätze in Appendix A.4, where the tightness ratio between the LHS and RHS of Equation (2) stays close to one.

A.4 Empirical Tightness and Multi-Qubit Validation of Theorem 4.7

We empirically verify that the right-hand side of Equation (3) provides a non-vacuous upper bound on the actual noise-induced loss deviation on the converged 4-qubit Q-SAM model. For each per-gate noise level p we measure: (i) LHS $= |\mathcal{L}(\tilde{\rho}) - \mathcal{L}(\rho(\theta))|$, the actual loss deviation averaged over 50 test samples; (ii) the empirical Fubini–Study distance $\varepsilon = D(\rho, \mathcal{E}_p(\rho))$, used as the realized perturbation budget so that LHS and RHS share the same ε rather than its conservative theoretical upper bound; and (iii) RHS, the parameter-space loss deviation along the Q-SAM ascent direction $\delta^* \propto G^{-1}\nabla\mathcal{L}$ scaled to the boundary of $\mathcal{U}_G(\theta, \varepsilon)$. We report two channels covering the two qualitatively distinct families of NISQ noise: *depolarizing* (representative isotropic Pauli channel, of which bit/phase flip and dephasing are subset cases) and *amplitude damping* (representative non-unital, T_1-type). The upper block of Table 5 shows that LHS $\leq$ RHS holds across both channels and all noise levels, with the ratio RHS/LHS tightening to ~ 1 in the practical NISQ regime, validating the bound and the use of the geometry-aware ascent direction $G^{-1}\nabla\mathcal{L}$ in Q-SAM. We further examine whether the bound remains informative as the circuit size increases. In the lower block of Table 5, we fix the noise strength at $p = 0.016$ for depolarizing noise and $\gamma = 0.040$ for amplitude damping and evaluate the same quantities on the 4-, 6-, and 8-qubit models. The bound remains non-vacuous across

Table 5: **Empirical validation of Equation (3).** The upper block evaluates bound tightness under increasing noise strength on the trained 4-qubit Q-SAM, while the lower block evaluates multi-qubit validity at fixed representative noise strengths. $\varepsilon = D(\rho, \mathcal{E}(\rho))$.

Channel	Qubits	Strength	ε	LHS (noise)	RHS (param)	RHS/LHS
Depolarizing	4	0.004	0.165	0.148	0.339	2.29×
		0.008	0.267	0.473	0.781	1.65×
		0.012	0.329	0.850	1.137	1.34×
		0.016	0.368	1.172	1.353	1.15×
Amplitude damping	4	0.010	0.193	0.210	0.455	2.17×
		0.020	0.301	0.665	0.965	1.45×
		0.030	0.363	1.137	1.327	1.17×
		0.040	0.398	1.499	1.523	1.02×
Depolarizing	4	0.016	0.368	1.172	1.353	1.15×
	6	0.016	1.217	0.966	1.270	1.32×
	8	0.016	1.334	0.926	1.056	1.14×
Amplitude damping	4	0.040	0.398	1.499	1.523	1.02×
	6	0.040	1.092	0.974	1.245	1.28×
	8	0.040	1.216	0.937	0.997	1.06×

the evaluated scales, with RHS/LHS ranging from 1.06 to 1.32, showing no substantial loosening as the qubit count increases. The regularized QFIM also remains numerically well-conditioned, with condition numbers of 4.04 and 2.92 for the 6- and 8-qubit models, respectively. These results support that the geometry-aware parameter region remains informative and numerically stable beyond the main 4-qubit setting, although they do not establish asymptotic scalability.

B Computational Cost of Q-SAM

We analyze the cost of Q-SAM in terms of *circuit evaluations*, the standard unit for fair cross-algorithm comparison in quantum machine learning, since this is the quantum-native primitive that dominates the computational burden of any QML training pipeline.

Table 6: Per-step circuit-evaluation counts on the data-reuploading architecture (QFIM amortized over $R = 20$). Q-SAM's overall counts scale at essentially the same rate as Adam ($\times 2.23$ vs. $\times 2.25$), confirming a fixed-factor QFIM overhead rather than a scaling bottleneck.

Config	#Param. (p)	QFIM counts/step	Overall counts/step
4-qubit Adam	192	–	384
6-qubit Adam	432 ($\times 2.25$)	–	864 ($\times 2.25$)
4-qubit Q-SAM	192	144	912
6-qubit Q-SAM	432 ($\times 2.25$)	312 ($\times 2.16$)	2040 ($\times 2.23$)

Theoretical cost of the QFIM. For a circuit with p trainable parameters, computing the full QFIM (including matrix inversion) via the parameter-shift rule requires $\mathcal{O}(p^2)$ circuit evaluations. Our implementation reduces this with two standard tricks:

1. **Block-diagonal approximation** (Stokes et al., 2020): truncates the QFIM to within-block second-order terms, reducing the per-construction cost from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$. This is the same approximation adopted by the quantum natural gradient (QNG) and has been established as a faithful surrogate for the full metric in PQC training.
2. **Metric refresh**: the QFIM is recomputed only every R optimization steps and cached between updates ($R = 20$ in our experiments), reducing the amortized cost by a factor of R.

Importantly, QFIM construction therefore requires only the reduced $\mathcal{O}(p)$ circuit evaluations, keeping the same linear-in-p scaling as a vanilla gradient step. The QFIM cost does not dominate the overall procedure—it follows the same scaling behavior in p as Adam. Also, we empirically confirm that the metric refresh does not hurt the training, validating the $R = 20$ choice in our implementations (see Table 7).

Table 7: Metric-refresh ablation on the 4-qubit configuration. The first row reports the Q-SAM 5-seed mean±95% CI of composite noise evaluation from Table 3. The $R = 1$ single seed accuracies stay close to the $R = 20$ 5-seed CI throughout the sweep, indicating that the metric refresh has negligible impact on Q-SAM's performance.

Variant	**Composite noise scale** (1.0 **= composite noise of Table 3**)					
	0.0	0.2	0.4	0.6	0.8	1.0
Q-SAM ($R=20$)	76.5±1.8	75.3±2.0	73.5±3.3	69.7±4.3	58.7±4.8	44.2±4.9
Q-SAM ($R=1$)	77.3	75.8	72.0	67.3	56.6	45.6

Per-step circuit-evaluation counts. We define "per step" to mean a single optimizer update (one mini-batch gradient step). "QFIM counts/step" denotes the additional circuit evaluations required to construct the QFIM, amortized over the metric-refresh interval ($R = 20$). "Overall counts/step" is the total quantum cost per step including both gradient and QFIM contributions. Table 6 reports these counts on the data-reuploading architecture used in our main experiments, alongside an enlarged variant ($\sim 2.25\times$ more parameters) used to verify scaling.

The QFIM contribution scales linearly in p at the same rate as a vanilla gradient and accounts for roughly 15% of Q-SAM's per-step counts; the additional gradient at the perturbed point, shared with SAM, accounts for the remainder of the gap.

Wall-clock and classical-simulator overhead. The counts in Table 6 are a quantum-native cost metric: on real quantum hardware, wall-clock time can be computed by multiplying these counts and a fixed per-shot execution time, so Q-SAM's linear-in-p scaling directly carries over to hardware wall-clock. In our classical-simulator experiments, each circuit evaluation incurs an additional simulation overhead that grows as 4^n for mixed-state simulation; this overhead is optimizer-independent and applies equally to Adam, SAM, Q-SAM, and any QFIM-based optimizer (QNG (Stokes et al., 2020), QN-SPSA (Gacon et al., 2021), noise-aware QNG (Koczor & Benjamin, 2022), RNG (Kolotouros & Wallden, 2024)), which is not specific to Q-SAM.

Table 8: Representative QNN image-classification studies (2022–2025). Our configurations span multiple encoding schemes and circuit architectures from 4 to 10 qubits, within the operating range of recent NISQ-era benchmarks.

Reference	**Qubits**	**Input dim**	**Encoding**	**Datasets**
QuantumNAT (Wang et al., 2022)	4	4×4	angle	MNIST
QCNN (Hur et al., 2022)	8	256-d (PCA)	angle / amplitude	MNIST, Fashion-MNIST
Multi-class QCNN (Bokhan et al., 2022)	8	16×16	amplitude	MNIST, Fashion-MNIST
NNQE (Riaz et al., 2023)	4	28×28	angle	MNIST, CIFAR-10
Soft Quantum NN (Zhou et al., 2023)	4	**4×4**	angle	MNIST
Tolstobrov et al. (Tolstobrov et al., 2024)	4	PCA → 4	angle (reuploading)	MNIST
HQNN-Quanv (Senokosov et al., 2024)	4–5	14×14	angle	MNIST, CIFAR-10
QDEQ (Schleich et al., 2024)	4	4×4	angle / amplitude	MNIST, Fashion-MNIST, CIFAR-10
Wang et al. (Wang et al., 2025)	3–5	32×32	angle (reuploading)	MNIST, Fashion-MNIST
SPQCC (Ding et al., 2024)	10	32×32	amplitude	MNIST
LCQHNN (Liu et al., 2025)	4	CNN-feat. → 4	angle	MNIST, Fashion-MNIST, CIFAR-10
This work (main, Appendix D.1)	4, 6, 8	4×4, 8×8, 256-d CNN feat.	angle / amplitude	MNIST, Fashion-MNIST, CIFAR-10
This work (Appendix D.2)	6, 8, 10	8×8, 16×16, 28×28	amplitude	MNIST

C Experimental Configuration Survey

Table 8 summarizes the qubit count, input dimension, encoding strategy, and image-classification datasets adopted by representative QNN/PQC image-classification studies (2022–2025). The survey

shows that MNIST remains a standard controlled dataset in NISQ-era QNN studies, often together with Fashion-MNIST, because both datasets allow evaluation under small-qubit and shallow-circuit constraints. In contrast, CIFAR-10 is typically used with additional architectural components, such as stronger classical preprocessing, quanvolutional layers, or hybrid convolutional front-ends, due to its color-channel structure and higher input dimensionality.

Each component of our experimental setup has precedent in prior QNN image-classification studies: few-qubit circuits spanning 4–10 qubits are commonly considered (Wang et al., 2022; Hur et al., 2022; Ding et al., 2024); both angle-based data reuploading (Tolstobrov et al., 2024; Wang et al., 2025) and amplitude encoding (Hur et al., 2022; Bokhan et al., 2022; Ding et al., 2024) have been used for image classification; and MNIST, Fashion-MNIST, and CIFAR-10 are all represented in prior QNN benchmarks (Schleich et al., 2024; Liu et al., 2025).

D Additional Experiments

D.1 Fashion-MNIST Robustness Evaluation

We further evaluate Q-SAM on Fashion-MNIST using the same 4-qubit data-reuploading architecture, preprocessing pipeline, optimizer settings, and noise-evaluation protocol as in the main MNIST experiments. Images are downsampled to 4×4, normalized to $[0, \pi]$, and encoded through the same angle-based data-reuploading circuit. This provides an additional benchmark under the same experimental setting, allowing us to compare robustness trends without introducing changes to the architecture, preprocessing pipeline, or noise-evaluation protocol. Unless otherwise stated, the representative noise strengths are kept identical to those in Table 2 to isolate the effect of changing the dataset while avoiding dataset-specific tuning of the noise levels.

Table 9: **Fashion-MNIST robustness comparison across five noise channels.** Test accuracy (%) of the same 4-qubit data-reuploading architecture on Fashion-MNIST-10. The representative noise strengths are matched to Table 2. Values in parentheses denote degradation relative to the nominal accuracy.

Optimizer	Train Nominal ($p = 0$)	Test (Perturbed)				
		Depolarizing ($p^* = 0.013$)	Amp. Damping ($\gamma^* = 0.024$)	Phase Damping ($\gamma^* = 0.053$)	Bit Flip ($p^* = 0.25$)	Phase Flip ($p^* = 0.026$)
Adam	70.42	30.26 (-40.16)	30.86 (-39.56)	34.42 (-36.00)	31.74 (-38.68)	29.86 (-40.56)
SAM	70.74	33.74 (-37.00)	31.42 (-39.32)	36.86 (-33.88)	33.86 (-36.88)	32.08 (-38.66)
QNG	68.08	35.26 (-32.82)	33.58 (-34.50)	39.58 (-28.50)	33.86 (-34.22)	35.08 (-33.00)
Q-SAM	**72.42**	**43.08 (-29.34)**	**42.58 (-29.84)**	**47.26 (-25.16)**	**40.58 (-31.84)**	**41.14 (-31.28)**

Table 9 shows that Q-SAM preserves the main robustness trend on Fashion-MNIST. Although Fashion-MNIST is more sensitive to the injected noise than MNIST under the same compact 4-qubit representation, Q-SAM consistently achieves the strongest noisy-inference accuracy across the evaluated channels. Compared with the strongest non-Q-SAM baseline for each channel, Q-SAM improves noisy accuracy by roughly 6–9 percentage points. This supports that the robustness benefit of QFIM-local flatness is not specific to MNIST digits, but also appears on a more challenging image-classification benchmark under the same NISQ-constrained architecture.

D.2 Scalability Experiments

We apply Q-SAM to three larger architectures with more qubits that replace data reuploading with single-pass amplitude encoding into the full 2^n-dimensional space. Table 10 lists the configurations and Table 11 shows robustness at a representative mid-range noise level, as presented in Table 2.

Beyond the main 4-qubit data-reuploading architecture, Q-SAM consistently improves robustness in the larger 6-, 8-, and 10-qubit amplitude-encoding settings. In Scaling A (6 qubits, 8×8 inputs), Q-SAM achieves the best accuracy across all five noise channels, improving over the strongest non-Q-SAM baseline by up to 6.16 points. This trend is also preserved in Scaling B (8 qubits, 16×16 inputs), and Scaling C (10 qubits, 28×28 inputs).

Table 10: Scaling architectures.

Setting	qubits n	input dim d	encoding	#blocks	#param
Main (Section 5)	4	16 (4×4)	reuploading	16	192
Scaling A	6	64 (8×8)	amplitude	12	216
Scaling B	8	256 (16×16)	amplitude	12	288
Scaling C	10	784 (28×28)	amplitude	12	360

Table 11: **Robustness comparison across five noise channels of 6, 8, 10 qubits.** Test accuracy (%) at a representative noise level per noise channel. Values in parentheses denote the performance degradation relative to the nominal accuracy.

Scale	**Optimizer**	**Train Nominal** ($p = 0$)	**Test (Perturbed)**				
			Depolarizing ($p^* = 0.011$)	**Amp. Damping** ($\gamma^* = 0.027$)	**Phase Damping** ($\gamma^* = 0.50$)	**Bit Flip** ($p^* = 0.17$)	**Phase Flip** ($p^* = 0.50$)
Scaling A (6q, 8×8)	Adam	82.14	32.86 (-49.28)	40.42 (-41.72)	55.08 (-27.06)	31.24 (-50.90)	54.42 (-27.72)
	SAM	81.42	43.26 (-38.16)	44.58 (-36.84)	69.14 (-12.28)	38.86 (-42.56)	66.74 (-14.68)
	QNG	81.08	43.74 (-37.34)	44.26 (-36.82)	68.74 (-12.34)	39.86 (-41.22)	67.42 (-13.66)
	Q-SAM	79.08	**49.86 (-29.22)**	**48.14 (-30.94)**	**71.42 (-7.66)**	**44.08 (-34.82)**	**69.74 (-9.26)**
Scaling B (8q, 16×16)	Adam	82.74	34.74 (-48.00)	37.42 (-45.32)	65.42 (-17.32)	30.08 (-52.66)	62.08 (-20.66)
	SAM	82.58	34.58 (-48.00)	34.74 (-47.84)	69.42 (-13.16)	26.74 (-55.84)	70.58 (-12.00)
	QNG	84.08	50.08 (-34.00)	46.14 (-37.94)	82.14 (-1.94)	42.74 (-41.34)	79.58 (-4.50)
	Q-SAM	83.42	**54.58 (-28.84)**	**51.42 (-32.00)**	**82.74 (-0.68)**	**48.08 (-35.34)**	**82.74 (-0.68)**
Scaling C (10q, 28×28)	Adam	90.92	47.88 (-43.04)	49.56 (-41.36)	88.83 (-2.09)	41.38 (-49.54)	88.34 (-2.58)
	SAM	91.12	52.34 (-38.78)	52.82 (-38.30)	88.31 (-2.81)	46.50 (-44.62)	88.03 (-3.09)
	QNG	90.32	52.06 (-38.26)	53.14 (-37.18)	88.27 (-2.05)	52.02 (-38.30)	88.83 (-1.49)
	Q-SAM	90.98	**63.74 (-27.24)**	**64.48 (-26.50)**	**89.84 (-1.14)**	**56.68 (-34.30)**	**89.54 (-1.44)**

In Table 11, we observe that phase damping and phase flip produce relatively mild degradation in the larger amplitude-encoding settings, especially for Scaling B and C. This suggests that, under the present ansatz and measurement configuration, the learned decision function is less sensitive to pure phase perturbations than to population-changing or isotropic errors. For example, bit flip and amplitude damping directly alter computational-basis populations or relaxation behavior, which can more strongly affect the measured logits, whereas phase damping/phase flip primarily perturb coherence phases. Accordingly, the phase-noise results should be interpreted as architecture- and measurement-dependent rather than as universal evidence that these channels are intrinsically harmless.

Overall, these results support that Q-SAM is not restricted to the 4-qubit data-reuploading circuit used in the main experiments. Its advantage is most pronounced for noise channels that strongly perturb the measured decision statistics, such as depolarizing noise, amplitude damping, and bit flip, whereas phase-type noise induces comparatively mild degradation in the present amplitude-encoding configuration.

E Implementation Details

E.1 Dataset and Preprocessing

For MNIST dataset of our experiments, we accommodate the limited qubit capacity ($n = 4$), the original 28×28 images were downsampled to 4×4 resolution ($d = 16$) via average pooling. For computational efficiency during the extensive simulation of quantum noise channels, we constructed a fixed subset of the dataset comprising 5,000 training samples and 1,000 test samples, balanced across the 10 classes. The pixel values were normalized to the range $[0, 1]$ and multiplied by π before angle encoding, so that each pixel maps to a rotation angle in $[0, \pi]$.

For the main Fashion-MNIST experiment, images are downsampled to 8×8 and encoded using amplitude encoding in the 6-qubit QNN. We additionally evaluate Fashion-MNIST in Appendix D.1 using the same 4-qubit, 4×4 data-reuploading configuration as the MNIST experiment, isolating the effect of changing the dataset while keeping the circuit architecture and encoding scheme fixed.

For CIFAR-10, the 32×32 RGB images are processed by a frozen CNN feature extractor to obtain a 256-dimensional feature vector, which is normalized before 8-qubit amplitude encoding.

E.2 HYPERPARAMETERS

Table 12 summarizes the architecture, data, and optimization settings for the three main experimental configurations reported in Table 2. All experiments were conducted using the PyTorch framework integrated with PennyLane for quantum circuit simulations. To ensure reproducibility, five different random seeds were applied to all stochastic processes, including weight initialization and data shuffling. All experiments were run on Intel Xeon Gold CPUs.

Table 12: Architecture, data, and optimization settings for the three experimental configurations in Table 2.

Category	**Parameter**	**MNIST**	**Fashion-MNIST**	**CIFAR-10**
Architecture	Number of Qubits (n)	4	6	8
	Input Dimension (d)	16 (4×4 pixels)	64 (8×8 pixels)	256 (frozen CNN features)
	Encoding	Angle (data re-uploading)	Amplitude	Amplitude
	Ansatz	16 Strongly Entangling blocks	4 Strongly Entangling layers	QCNN (conv. + pooling)
	Variational Parameters	192	72	74
	Measurement Basis	Pauli-Z	X (Hadamard layer)	X (Hadamard layer)
Data	Training Samples	5,000	50,000	20,000
	Test Samples	1,000	1,000	1,000
	Random Seeds	5	5	5
Training	Optimizer	Adam / SAM / QNG / Q-SAM		
	Learning Rate	0.005	0.005	0.01
	Batch Size	64	32	32
	Epochs	150	80	80
	Training Noise	None (0.0)		
	Perturbation Radius (η)	0.008	0.3	0.6