EFFECTIVE DIMENSION GOVERNS GENERALIZATION IN QUANTUM KERNEL VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent quantum vision models—quantum vision transformers and quantum convolutional networks—report two striking but unexplained empirical phenomena: (i) ansatze with more, or more uniformly distributed, entanglement generalize better, and (ii) injecting quantum noise can *improve* test accuracy rather than degrade it. These observations are currently treated as curiosities, discovered by grid search and explained, if at all, by hand. We show that both are manifestations of a single, measurable quantity: the *effective dimension* d_{eff} of the (noise-shaped) quantum feature kernel. Working primarily with quantum-kernel vision models—a quantum feature map read out by a kernel classifier—we give a spectral account in which entanglement structure and quantum noise are two knobs that move d_{eff}; in an overfitting regime, contracting d_{eff} acts as ridge-like regularization. We analyze the mechanism: an *exact* decomposition of the depolarized kernel $K_p = (1-p)^2 K + \frac{p(2-p)}{D}\mathbf{1}\mathbf{1}^\top$ with $d_{\text{eff}}(K_p) \to 1$, a contraction result (and its boundary) for amplitude damping, a kernel-machine capacity bound, and a capacity/alignment risk decomposition; the monotone contraction operative in our entangled experiments is verified empirically, not proven in general. Our informative empirical finding is that test accuracy collapses onto a single function of d_{eff} *across distinct entangling ansatze*—compressing different spectral shapes onto one curve (R^2=0.82 ± 0.08 over seeds). Along the one-parameter depolarizing family the collapse is instead exact *by construction*; we use it only to confirm the kernel decomposition to machine precision and at up to 12 qubits, not as evidence for d_{eff}. Amplitude damping contracts d_{eff} and lifts test accuracy by up to $+13\%$ along an inverted-U sweet spot; the effect's sign flips between the over- and under-fitting regimes; noise injection matches an explicit spectral-filtering frontier (so it is not a weak substitute for hyperparameter tuning); and the phenomenon persists in *trained* QViT- and QCNN-like models. Entanglement plays the complementary role of a precondition on the ansatze and tasks we study—it supplies the feature-space alignment without which the d_{eff} law does not hold. Our results organize two reported anecdotes into a single measurable principle for designing quantum-vision models.

1 INTRODUCTION

Quantum machine learning for computer vision has advanced rapidly, with quantum vision transformers (QViTs) (Cherrat et al., 2024; Boucher et al., 2026; Zhang et al., 2025) and quantum convolutional neural networks (QCNNs) (Cong et al., 2019; Shi et al., 2024; Röseler et al., 2025; Long et al., 2025; Wu & Zhang, 2025) now matching strong classical baselines on small benchmarks while using dramatically fewer parameters. Yet the field's design practice remains largely empirical: ansatze are chosen by sweeping heuristic descriptors such as expressibility and entangling capability (Sim et al., 2019; Röseler et al., 2025), and the role of hardware noise is assessed *post hoc* through simulation.

Two recurring empirical observations stand out as genuinely puzzling. First, several works report that *more entanglement helps generalization*: for example, Wang & Shang (2025) find that "ansatzes with uniformly distributed entanglement entropy consistently deliver superior non-local feature fusion and state-of-the-art accuracy." Second, and more surprisingly, the same work reports that *quan-*

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.

tum noise can help: amplitude damping improves accuracy by $+2.71\%$ in some configurations, a "double-edged" behavior with a non-monotonic dependence on noise strength. Both phenomena are reported as discoveries, without a predictive theory of *when* they occur or *how strong* the effect should be.

This paper. We argue that these two phenomena are not separate, and not mysterious. They are two views of the same underlying object—the eigenspectrum of the quantum feature kernel—and in particular of its *effective dimension*

$$d_{\text{eff}} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \tag{1}$$

the participation ratio of the kernel eigenvalues $\{\lambda_i\}$ (defined formally in Section 3). Our thesis is:

> *Entanglement structure and quantum noise are two knobs that move d_{eff}. Once alignment and spectral shape are comparable—as they are across entangled ansatze—generalization is largely organized by d_{eff}; under overfitting, contracting d_{eff} acts as regularization and improves generalization up to an interior optimum.*

Under this view, entangling gates redistribute amplitude across the Hilbert space and noise channels contract the spectrum; both regulate d_{eff}, and the "noise helps" effect is spectral (ridge-like) regularization in an overfitting regime. Section 3 makes the tractable parts precise—global depolarizing noise contracts the spectrum *exactly* with $d_{\text{eff}} \to 1$ (Prop. 1), amplitude damping contracts it with a stated boundary (Prop. 2), capacity is bounded by d_{eff} (Prop. 3), and risk splits into a spectral and an alignment term (Prop. 4)—while being explicit about the limits: a constant-row-sum condition (which Perron–Frobenius does *not* supply) gives provable monotonicity for *global depolarizing* noise, but not for the *amplitude damping* we actually run—there the rank-one decomposition the argument needs is absent (Prop. 2), so strict monotonicity in the entangled, amplitude-damped regime is an empirical observation, full stop. This reframing has three consequences, which we state as predictions and verify experimentally:

1. **(P1) d_{eff} organizes generalization.** Across distinct entangling ansatze (different spectral shapes), test accuracy collapses onto a single, stable function of d_{eff} ($R^2 = 0.82 \pm 0.08$ across seeds), and the *sign* of the dependence flips between the overfitting and underfitting regimes—consistent with the bias–variance picture of Prop. 3 (which fixes the sign, though not the precise location of the interior optimum).
2. **(P2) Noise is a spectral regularizer.** Increasing quantum noise monotonically contracts d_{eff}, reduces the train–test gap, and—when the model overfits—improves test accuracy along an inverted-U sweet spot, reproducing the reported "noise helps" phenomenon.
3. **(P3) Entanglement is a precondition.** On the ansatze and tasks we study, entanglement supplies the feature-space alignment (Prop. 4) that places a model in the regime where the d_{eff} law holds: entangled circuits of differing topology all lie on the same accuracy–d_{eff} curve, whereas the unentangled (product) map sits off it—not through d_{eff}, but through its measurably lower label alignment (App. A gives the scope of this empirical claim).

Beyond explaining existing observations, P3 implies an actionable design view: to improve a quantum vision model one should monitor and steer a single, label-free spectral quantity, directly computed from the training kernel, rather than tune entanglement and noise as independent heuristics.

Contributions.

- A **spectral theory** (Section 3): the exact depolarized-kernel decomposition with an explicit formula for $d_{\text{eff}}(K_p)$ (Prop. 1); a contraction result for amplitude damping with its analytic boundary (Prop. 2); a bridge between d_{eff} and the ridge dimension d_γ (Lemma 2); a capacity bound (Prop. 3); and a capacity/alignment decomposition under which the accuracy–d_{eff} collapse is exact along the one-parameter noise family and conditional across ansatze (Prop. 4). We are careful about what is proven: monotonicity of $d_{\text{eff}}(K_p)$ holds under constant row sums and is otherwise empirical, as is the cross-ansatz collapse.

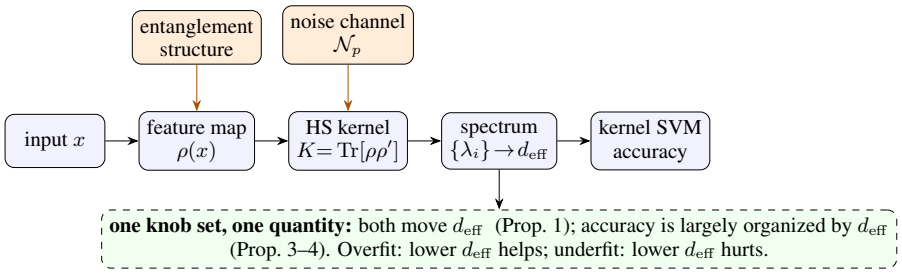

Figure 1: Overview. A quantum feature map $\rho(x)$ induces a Hilbert–Schmidt kernel whose spectrum is summarized by the effective dimension d_{eff}. Entanglement topology and an injected noise channel are two knobs that both move d_{eff}; within the aligned entangled regime d_{eff} is the dominant scalar predictor of generalization, and the sign of its effect is set by the bias–variance regime.

- **Empirical verification** (5 seeds). Test accuracy is largely organized by a single d_{eff} curve *across distinct entangling ansatze* ($R^2 = 0.82 \pm 0.08$); amplitude damping reproduces "noise helps" with an inverted-U sweet spot; the d_{eff}–accuracy relation flips sign between regimes; noise injection lands on an explicit spectral-filtering frontier rather than substituting for hyperparameter tuning; entanglement enters as an alignment precondition (measured $\mathrm{A}(K)$). The exact decomposition is verified to machine precision up to **12 qubits**, where the perfect one-parameter collapse is reported as a by-construction null, not as evidence.
- **Robustness and honest scope**. The mechanism persists in *trained* QViT-/QCNN-like models, replicates on Fashion-MNIST and BloodMNIST and across depth/width, and survives replacing the SVM with kernel ridge. On a real IBM Heron device the hardware noise contracts d_{eff} ($4.06 \to 3.38$) and moves accuracy in the predicted direction ($0.863 \to 0.900$ on 80 test points), though the gain is not statistically resolved at that size (Sec. 4.5).

2 Related work

Quantum vision models. QViTs replace classical self-attention with parameterized quantum circuits, reducing parameter counts from $O(n^2)$ to $O(n)$ (Boucher et al., 2026) or using amplitude encoding to process whole images with $O(\log N)$ qubits (Zhang et al., 2025); Cherrat et al. (2024) give compound-matrix attention with provable asymptotic advantages. QCNNs (Cong et al., 2019) have been extended to multiclass settings (Shi et al., 2024), hardware-efficient encodings (Röseler et al., 2025), inception-style heterogeneous filters (Wu & Zhang, 2025), and trainable quantum–classical–quantum stacks (Long et al., 2025). Across this literature, ansatz selection relies on expressibility and entangling-capability heuristics (Sim et al., 2019), and noise is studied empirically. We provide the missing predictive layer.

Quantum models as kernel methods. Supervised quantum models with fixed feature maps are kernel methods (Schuld, 2021; Schuld & Killoran, 2019; Havlíček et al., 2019), letting us analyze generalization through the kernel spectrum. The effective dimension has been used to characterize QNN capacity (Abbas et al., 2021), and the data-induced kernel's structure governs quantum advantage (Huang et al., 2021). Closest to our diagnostic, Heyraud et al. (2022) introduce the same participation-ratio quantity—their *effective kernel rank*—for noisy dissipative-spin-chain kernel machines, with decoherence as implicit regularization under a purity-based bound; crucially they define it on the *centered* kernel, and we show why that matters (Sec. C.3, App. C.2): the rank-one term behind our own $d_{\mathrm{eff}} \to 1$ under global depolarizing is annihilated by an intercept SVM's dual constraint and vanishes under centering, unlike our amplitude-damping mechanism, where raw and centered d_{eff} both track accuracy strongly. Their setting has no entanglement knob and is not tied to the two vision phenomena we address. More broadly, spectral structure decides when quantum kernels generalize at all—a flat spectrum is a poor inductive bias (Kübler et al., 2021), kernel values concentrate exponentially in qubit count (Thanasilp et al., 2024), and a bandwidth hyperparameter can move a model from unable-to-generalize to well-aligned (Canatar et al., 2023); recent work *escapes* this concentration via locality (Agliardi et al., 2026; Zendejas-Morales et al.,

2026), benign-overfitting local-global kernels (Tomasi et al., 2026), double descent (Kempkes et al., 2026), or dequantization (Gil-Fuster et al., 2025). We instead treat concentration (and an overly broad spectrum) as endpoints of one d_{eff} axis that entanglement and noise move—made precise by our overfitting/underfitting sign flip (App. C.1)—and connect these spectral tools directly to entanglement/noise design choices, making them *predictive* of the optimal configuration; kernel–target alignment (Cristianini et al., 2001) is the classical task-aware descriptor we borrow for the second axis.

Spectral capacity control and noise as regularization (classical). Our diagnostic is the standard capacity summary of kernel ridge regression—rates set by the ridge dimension d_γ (Zhang, 2005; Caponnetto & De Vito, 2007; Bartlett et al., 2005), with the realized risk further depending on target–eigenbasis alignment (Canatar et al., 2021; Jacot et al., 2020)—and noise as regularization is classical too (Bishop, 1995; Srivastava et al., 2014). What is new is that in quantum models these knobs are hardware-native and act through one spectral quantity, so noise is not purely harmful as in trainability analyses (Wang et al., 2021).

3 A SPECTRAL THEORY OF ENTANGLEMENT AND NOISE

3.1 SETUP AND DEFINITIONS

We consider hybrid models whose quantum component is a feature map $x \mapsto \rho(x) \in \mathbb{C}^{D\times D}$, $D = 2^{n_q}$, where $\rho(x)$ is the density matrix produced by a parameterized circuit acting on an encoding of an input x. Following the kernel view of quantum models (Schuld, 2021; Havlíček et al., 2019), classification is performed by a kernel machine on the Hilbert–Schmidt (HS) kernel

$$k(x, x') \;=\; \mathrm{Tr}\big[\rho(x)\,\rho(x')\big] \;=\; \langle \phi(x), \phi(x')\rangle_{\mathrm{HS}}, \qquad \phi(x) := \mathrm{vec}\,\rho(x), \tag{2}$$

which for pure states reduces to the fidelity kernel $|\langle\psi(x)|\psi(x')\rangle|^2$.

Definition 1 (Gram matrix and spectral descriptors)**.** Given training inputs $\{x_i\}_{i=1}^n$, let $K \in \mathbb{R}^{n\times n}$, $K_{ij} = k(x_i, x_j)$, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. Define the *effective dimension* (participation ratio / effective rank)

$$d_{\text{eff}}(K) = \frac{\big(\sum_i \lambda_i\big)^2}{\sum_i \lambda_i^2} = \frac{(\mathrm{tr}\,K)^2}{\mathrm{tr}(K^2)} \in [1, n], \tag{3}$$

the ridge effective dimension $d_\gamma(K) = \sum_i \lambda_i/(\lambda_i + \gamma) = \mathrm{tr}\big[K(K + \gamma I)^{-1}\big]$, and, for one-hot centered labels Y with target Gram $T = YY^\top$, the kernel–target alignment $\mathrm{A}(K) = \langle K, T\rangle_F/(\|K\|_F\|T\|_F)$ (Cristianini et al., 2001).

Lemma 1 (Validity)**.** *k in equation 2 is a positive-semidefinite kernel, so $K \succeq 0$ and $d_{\text{eff}}(K)$ is well defined; $d_{\text{eff}}(K) = 1$ iff K has rank 1 and $d_{\text{eff}}(K) = n$ iff the spectrum is flat.*

Proof. $k(x, x') = \langle\phi(x), \phi(x')\rangle_{\mathrm{HS}}$ is an inner product of the feature vectors $\phi(x) = \mathrm{vec}\,\rho(x)$, hence PSD; the endpoint characterizations are Cauchy–Schwarz applied to (λ_i). □

3.2 NOISE CONTRACTS THE SPECTRUM (P2)

We first treat noise analytically. Let $\mathcal{N}_p$ be the global depolarizing channel of strength $p \in [0, 1]$, $\mathcal{N}_p[\rho] = (1 - p)\rho + p\,I/D$, applied to the feature map: $\rho_p(x) = \mathcal{N}_p[\rho(x)]$, with kernel K_p, $(K_p)_{ij} = \mathrm{Tr}[\rho_p(x_i)\rho_p(x_j)]$.

Proposition 1 (Depolarizing noise: an exact, centering-trivial kernel decomposition)**.** *With $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$,*

$$K_p \;=\; (1 - p)^2\,K \;+\; \tfrac{p(2-p)}{D}\,\mathbf{1}\mathbf{1}^\top. \tag{4}$$

Hence K_p interpolates from $K_0 = K$ to the rank-one matrix $\frac{1}{D}\mathbf{1}\mathbf{1}^\top$ as $p \to 1$, so $d_{\text{eff}}(K_p) \to 1$. Writing $s = (1 - p)^2$, $T = \mathrm{tr}\,K$, $Q = \mathrm{tr}(K^2)$, and $S = \mathbf{1}^\top K\mathbf{1}$, the effective dimension is the explicit ratio

$$d_{\text{eff}}(K_p) = \frac{\big(sT + \frac{p(2-p)}{D}n\big)^2}{s^2Q + 2s\frac{p(2-p)}{D}\,S + \big(\frac{p(2-p)}{D}\big)^2 n^2}. \tag{5}$$

If K has constant row sums ($K\mathbf{1} = c\mathbf{1}$, i.e. $\mathbf{1}$ is an eigenvector of K) *then K and $\mathbf{1}\mathbf{1}^\top$ commute and $d_{\text{eff}}(K_p)$ is monotonically non-increasing in p. In general the cross-term $S = \mathbf{1}^\top K \mathbf{1}$ prevents a structure-free monotonicity guarantee; empirically $d_{\text{eff}}(K_p)$ is* strictly *decreasing in every experiment we run (Tables 1, 2, and the exact-depolarizing runs of Sec. C.4).*

Proof. Expanding $\mathrm{Tr}[\rho_p(x_i)\rho_p(x_j)]$ using $\mathrm{Tr}\,\rho = 1$ and $\mathrm{Tr}\,I = D$ gives the four terms $(1-p)^2 K_{ij} + 2\frac{(1-p)p}{D} + \frac{p^2}{D} = (1-p)^2 K_{ij} + \frac{p(2-p)}{D}$, which is equation 4; the limit and equation 5 follow by direct computation of $\mathrm{tr}\,K_p$ and $\mathrm{tr}(K_p^2)$ using $\mathrm{tr}(\mathbf{1}\mathbf{1}^\top) = n$, $\mathrm{tr}(K\mathbf{1}\mathbf{1}^\top) = S$, and $\mathrm{tr}((\mathbf{1}\mathbf{1}^\top)^2) = n^2$. When $K\mathbf{1} = c\mathbf{1}$, K and $\mathbf{1}\mathbf{1}^\top$ are simultaneously diagonalizable with $\mu_1(p) = s\lambda_1 + \frac{p(2-p)}{D} n$ and $\mu_{k\geq 2}(p) = s\lambda_k$; the top-eigenvalue share then increases in p (App. B), so d_{eff} decreases. Full computation and the obstruction in the general case are in App. B. □

The decomposition equation 4 makes the mechanism explicit: noise shrinks the informative component $(1-p)^2 K$ by a uniform scalar—absorbed by any tunable regularization constant—while adding a rank-one constant component that Sec. C.3 shows is exactly what an intercept classifier cannot see. The decomposition is exact, but not by itself a meaningful ridge-regularization mechanism for classification; the noise-as-regularizer effect we establish experimentally is driven by amplitude damping (Prop. 2 onward), which lacks this rank-one structure.

Amplitude damping. The channel actually used in our experiments is per-qubit amplitude damping $\mathcal{A}_\gamma$ (rate γ), which lacks the exact rank-one decomposition of depolarizing noise because it is *non-unital*. We nonetheless prove the essential contraction.

Proposition 2 (Amplitude damping contracts pairwise feature distances and has a rank-one limit)**.** *Let $\rho_\gamma(x) = \mathcal{A}_\gamma^{\otimes n_q}[\rho(x)]$. **(a) Limit.** $\rho_\gamma(x) \to |0\rangle\langle 0|^{\otimes n_q}$ for every x, so $K_\gamma \to \mathbf{1}\mathbf{1}^\top$ and $d_{\text{eff}}(K_\gamma) \to 1$ as $\gamma \to 1$. **(b) Strict single-qubit contraction.** On one qubit $\mathcal{A}_\gamma$ maps the Bloch vector $(X, Y, Z) \mapsto (\sqrt{1-\gamma}\,X, \sqrt{1-\gamma}\,Y, (1-\gamma)Z + \gamma)$, hence $\|\mathcal{A}_\gamma(\rho) - \mathcal{A}_\gamma(\sigma)\|_{\text{HS}} \leq \sqrt{1-\gamma}\,\|\rho - \sigma\|_{\text{HS}}$: pairwise Hilbert–Schmidt distances contract strictly. This pairwise contraction does not by itself imply monotone contraction of the Gram-matrix effective dimension; for general ensembles, including entangled ones, monotonicity of $d_{\text{eff}}(K_\gamma)$ requires additional structure and is, absent that structure, an empirical property of the spectra observed in our experiments. **(c) Boundary.** Because $\mathcal{A}_\gamma$ is non-unital ($\mathcal{A}_\gamma(I) = I + \gamma Z$), $\mathcal{A}_\gamma^{\otimes n_q}$ can expand the Hilbert–Schmidt norm of trace-carrying operators, which is exactly why the clean global monotonicity of the depolarizing case need not extend to arbitrary entangled ensembles. Constant row sums keep $\mathbf{1}$ as an eigenvector, but unlike the depolarizing family this alone does not determine the evolution of the remaining spectrum, since amplitude damping lacks a rank-one decomposition; we therefore treat monotone d_{eff} contraction under amplitude damping as an empirical property of the ensembles studied here (the d_{eff} columns of Tables 1, 2 decrease strictly).*

Proof in App. B; the single-qubit computation is explicit and the boundary in (c) is the honest analytic limit of the mechanism.

3.3 Effective dimension controls capacity (P1)

The capacity of a kernel machine is classically controlled by the *ridge* effective dimension d_γ. Our experiments report the γ-free participation ratio d_{eff} (the effective rank). The next lemma shows these two functionals move *together* under the noise contraction, so the capacity bound below (stated in d_γ) and the quantity we plot (d_{eff}) are not disconnected.

Lemma 2 (Co-monotonicity under noise)**.** *Along the depolarizing family K_p of Prop. 1 the two spectral summaries move together. Under the constant-row-sum condition of Prop. 1, both the ridge effective dimension $d_\gamma(K_p)$ (for every fixed $\gamma > 0$) and the participation ratio $d_{\text{eff}}(K_p)$ are non-increasing in p, with the limits $d_{\text{eff}}(K_p) \to 1$ and $d_\gamma(K_p) \to \frac{n/D}{n/D+\gamma}$ as $p \to 1$ (the rank-one limit $\frac{1}{D}\mathbf{1}\mathbf{1}^\top$ has the single nonzero eigenvalue n/D, so d_γ does* not *tend to* 1*). Without that condition we do not claim monotonicity: along the spectra realized in our experiments d_{eff} and d_γ are co-monotone (rank correlation* 1.0*, App. C.2), and $d_{\text{eff}}(K_p)$ is injective in p in every run.*

Proof sketch. Under constant row sums $K\mathbf{1} = c\mathbf{1}$, so $\lambda_1 = c$, and K, $\mathbf{1}\mathbf{1}^\top$ commute: $\mu_1(p) = sc + \frac{p(2-p)}{D}n$ and $\mu_{k\geq 2}(p) = s\lambda_k$ with $s = (1-p)^2$. Since $K_{ij} = \mathrm{Tr}[\rho_i\rho_j]$, $nc = \mathbf{1}^\top K\mathbf{1} = \mathrm{Tr}\left[(\sum_i \rho_i)^2\right] \geq n^2/D$ by Cauchy–Schwarz on the eigenvalues of the trace-n, D-dimensional PSD operator $\sum_i \rho_i$, so $c \geq n/D$. Writing $\mu_1(p) = sc + (1-s)\frac{n}{D}$ as a convex combination of c and $n/D \leq c$, μ_1 is non-increasing as s decreases in p; the tail terms $\mu_{k\geq 2} = s\lambda_k$ are non-increasing too ($\lambda_k \geq 0$). Every eigenvalue of K_p is thus non-increasing in p, hence so is $d_\gamma = \sum_i \mu_i/(\mu_i + \gamma)$ (each summand increasing in μ_i); d_{eff} decreases by the top-share argument of Prop. 1. Taking $p \to 1$ leaves the rank-one matrix, giving $d_{\text{eff}} \to 1$ and $d_\gamma \to (n/D)/(n/D + \gamma)$. Without the row-sum condition $c \geq n/D$ need not hold and the tail argument alone does not settle the total, which is why we state co-monotonicity as an empirical property of the spectra we observe (App. C.2). □

Proposition 3 (Capacity bound)**.** *For kernel ridge regression with regularization γ on n samples, the expected generalization gap is $\tilde{O}\big(\sqrt{d_\gamma(K)/n}\big)$, and $d_\gamma(K)$ is non-decreasing in every eigenvalue. Hence a spectral contraction that lowers the tail eigenvalues (e.g. $K \mapsto K_p$ of Prop. 1, which by Lemma 2 also lowers d_{eff}) tightens the bound. In an overfitting regime—empirical risk near zero while the gap dominates—contracting the spectrum therefore predicts a reduction of the variance-controlled component of risk, up to the point where signal-carrying eigen-directions are attenuated, giving an interior optimum; in an underfitting regime the bias term dominates and the sign reverses. A tighter bound does not by itself imply a lower expected risk, so we treat this as a prediction about the* sign *of the effect, which App. C.1 tests directly.*

Proof sketch. The local Rademacher complexity of the KRR hypothesis class is controlled by $\sum_i \min(\lambda_i, \gamma) \leq 2\gamma\, d_\gamma$, with $d_\gamma = \sum_i \lambda_i/(\lambda_i + \gamma)$ (Abbas et al., 2021); each summand of d_γ is increasing in λ_i, so shrinking eigenvalues lowers d_γ and the $\tilde{O}(\sqrt{d_\gamma/n})$ gap. When training risk is ≈ 0 the risk is gap-dominated and contraction helps; when bias dominates (underfitting) it hurts. Full proof in App. B. □

By Lemma 2 the participation ratio d_{eff} equation 3 and the ridge dimension d_γ move together along the noise family, so we report the γ-free, label-free d_{eff} as the spectral summary throughout.

3.4 When accuracy is organized by d_{eff} (P3)

Capacity is not the whole story: the realized risk also depends on how the kernel eigenbasis aligns with the labels.

Proposition 4 (Risk decomposition and the one-parameter collapse)**.** *The kernel-machine excess risk decomposes (to leading order) into a* capacity *term, a functional of the full spectrum $\{\lambda_i\}$, and an* alignment *term depending on the target's projection onto the kernel eigenbasis (whose task alignment we summarize empirically, though only partially, by the scalar $\mathrm{A}(K)$). Two consequences follow.* ***(i) Exact collapse along a one-parameter spectral family.*** *If a family of kernels is generated by a single scalar—as for the noise family K_p (Prop. 1), whose entire spectrum is fixed by p—then both terms, and hence the risk, are functions of that scalar:* $\mathrm{acc} = g(p)$ *for some g. If, additionally, $p \mapsto d_{\text{eff}}(K_p)$ is injective—as verified in all our experiments, and guaranteed under the constant-row-sum condition of Prop. 1—then accuracy collapses* exactly *onto a curve* $\mathrm{acc} = f(d_{\text{eff}})$. ***(ii) Cross-family collapse is conditional.*** *For kernels of differing spectral* shape *(e.g. different ansatze) equal d_{eff} does not by itself imply equal risk; collapse onto a common curve additionally requires comparable spectral shape and comparable alignment $\mathrm{A}(K)$. The unentangled product map violates the latter—its factorized kernel has low alignment—and therefore sits off the curve.*

Proof sketch. Write the KRR risk as bias^2+variance; the variance is the capacity term (a functional of the spectrum), the bias depends on the eigenbasis–target overlap, partially summarized by $\mathrm{A}(K)$. Along K_p the spectrum is determined by p, so risk is a function of p and hence of $d_{\text{eff}}(p)$ (claim i). For different shapes, d_{eff} is only a scalar summary, so equality of d_{eff} leaves the remaining spectral and alignment degrees of freedom free (claim ii). Full proof in App. B. □

The honest reading: d_{eff} governs *within a one-parameter spectral family* (rigorously) and predicts well *across* entangled ansatze (empirically, once alignment is comparable—measured in Sec. 4.4)—not a universal sufficient statistic, with alignment a genuine second factor. This is

exactly what we observe (Sec. 4.4): entangled ansatze share comparable, high alignment and collapse onto one d_{eff} curve, whereas an unentangled (product) map—whose kernel factorizes, $k(x, x') = \prod_q \text{Tr}[\rho_q(x)\rho_q(x')]$—has measurably lower alignment and sits off the curve at every d_{eff}. Entanglement is thus a precondition that supplies alignment, after which d_{eff} becomes the dominant scalar predictor of generalization.

4 EXPERIMENTS

Setup. The feature map is an n_q-qubit data-re-uploading circuit (n_q=6 throughout, 8 in the depth/width study, 12 for the exact-depolarizing analysis): each of L=2 layers applies $R_Y(x_i + \theta_{\ell,i})R_Z(0.7x_i)$ with fixed random θ, followed by an entangling block of CNOTs in one of four profiles—PRODUCT (none), CHAIN, RING, ALL-TO-ALL. The kernel is data-driven, not trained, which isolates the spectral mechanism. Noise is single-qubit amplitude damping of rate p after each entangling block, simulated exactly on a density-matrix backend; the global-depolarizing family is instead evaluated analytically from the noiseless statevector kernel via equation 4, which is what enables the 12-qubit runs. We form the HS kernel equation 2 from the simulated $\rho(x)$ and classify with a precomputed-kernel SVM (C=10); d_{eff} equation 3 and the other spectral descriptors are computed on the training block only (label-free, except alignment, which uses clean training labels). Inputs are DIGITS (10-way) and, for robustness, Fashion-MNIST (Xiao et al., 2017) and the medical BloodMNIST (Yang et al., 2023), each reduced to 6 (or 8) PCA features. To create a controlled overfitting regime—in which regularization can help and the sign of the d_{eff}/accuracy relation is unambiguous—we corrupt a fraction ℓ of *training* labels uniformly at random, keeping the test set clean; unless noted ℓ=25%, n_{train}=150, n_{test}=250. Gate, channel and circuit definitions are in App. A (Figs. 4, 5, 10) and full protocols in App. E.

4.1 ENTANGLEMENT IS A PRECONDITION (P3)

Varying only the entangling topology, with no injected noise (Table 6, App. C; mean±std over 5 seeds), the salient effect is a large gap between the *unentangled* product map (0.471 ± 0.060) and any entangled map (0.61–0.62). *Among* the entangled topologies neither d_{eff} nor accuracy is cleanly ordered ($d_{\text{eff}} \approx$ 42–45, accuracy within noise), and the four-point $\rho(d_{\text{eff}}, \text{test}) = -0.48 \pm 0.29$ is weak and product-driven; we therefore do *not* read this as "d_{eff} orders entanglement." Consistently with Prop. 4, the product deficit is an *alignment* effect—its factorized kernel measurably achieves lower kernel–target alignment on these tasks, confirmed directly in Sec. 4.4. Entanglement's role is to establish the alignment that places a model in the regime where the d_{eff} law (demonstrated next, via the noise sweep) holds.

4.2 NOISE IS A SPECTRAL REGULARIZER AND d_{eff} TRACKS ACCURACY (P1, P2)

We now vary d_{eff} over a wide range using the noise knob (Fig. 9 in App. C visualizes the resulting feature geometry and spectrum), which both exercises the proven mechanism (Prop. 1) and gives the clean d_{eff}-tracks-accuracy evidence that the four-point entanglement table cannot. Table 1 sweeps amplitude-damping strength on two entangled ansatze. As predicted, increasing noise monotonically contracts d_{eff} (from $\sim$ 44 to $\sim$ 3), monotonically reduces the train accuracy (less memorization), and improves test accuracy by up to $+13.2\%$ (ring) and $+12.4\%$ (all-to-all). This *qualitatively* reproduces the "noise-can-help" phenomenon reported by Wang & Shang (2025) (their $+2.71\%$ under amplitude damping) and—more importantly—exposes its mechanism: noise-induced spectral contraction acting as ridge regularization.

4.3 AN INVERTED-U SWEET SPOT WHOSE BENEFIT SCALES WITH OVERFITTING

We sweep injected noise while varying *overfitting severity* ℓ (fraction of corrupted training labels; test set and task fixed). Table 2 shows two effects: every row exhibits the predicted *inverted-U*—accuracy rises with noise, peaks, then falls as contraction destroys signal-bearing directions, the mechanistic origin of the "double-edged" behavior reported by Wang & Shang (2025)—and the *benefit* of optimal noise grows monotonically with ℓ, from $+2.4\%$ (ℓ=0) to $+30.0\%$ (ℓ=40%). The optimal point stays near $p^\star \approx 0.45$ ($d_{\text{eff}} \approx 3$) across ℓ, consistent with the matching intuition: ℓ changes how much one overfits, not the fixed task's intrinsic complexity. This gain magnitude, set

Table 1: P2: amplitude-damping sweep (gain is test accuracy relative to the noiseless circuit). Noise contracts d_{eff}, lowers train accuracy, and raises test accuracy in this overfitting regime.

	Ring			All-to-all		
Noise p	d_{eff}	Train	Test	d_{eff}	Train	Test
0.00	44.20	0.987	0.624	43.63	0.993	0.648
0.03	40.57	0.953	0.652	38.81	0.973	0.660
0.06	35.58	0.933	0.704	33.52	0.953	0.716
0.10	27.70	0.907	0.728	26.38	0.927	0.708
0.15	18.37	0.860	0.752	18.50	0.920	0.744
0.20	11.68	0.827	0.752	12.60	0.867	0.756
0.30	5.23	0.793	0.748	6.13	0.813	0.772
0.45	2.53	0.727	**0.756**	2.91	0.753	**0.772**
Best gain vs. noiseless: $+0.132$						$+0.124$

Over 5 seeds the best noise gain is $+0.131 \pm 0.034$ (ring) and $+0.135 \pm 0.023$ (all-to-all): the benefit is robust, not a seed artifact.

Table 2: Test accuracy vs. amplitude-damping rate p (columns) at four overfitting levels ℓ (rows; % corrupted training labels), all-to-all ansatz. Each row is an inverted-U with peak in **bold**; the optimal p stays ≈ 0.45 while the *gain* grows with ℓ. d_{eff} contracts from 38.2 (p=0) to 1.2 (p=0.9).

$\ell \backslash p$	0.00	0.05	0.10	0.20	0.30	0.45	0.60	0.75	0.90	Gain
0%	.800	.808	.800	.796	.804	**.824**	.824	.816	.808	+.024
10%	.696	.712	.724	.756	**.772**	.772	.772	.764	.752	+.076
25%	.624	.632	.676	.752	.768	**.784**	.752	.752	.720	+.160
40%	.404	.440	.500	.592	.672	**.704**	.700	.676	.644	+.300

by the injected overfitting, is not directly comparable to the $+2.71\%$ reported under naturally mild noise; the point is the mechanism and its lawful dependence on overfitting severity.

4.4 P3: ONE QUANTITY, TWO KNOBS — ACCURACY COLLAPSES ONTO d_{eff}

Our central claim is that, once alignment and spectral shape are comparable, entanglement and noise act on generalization mainly through d_{eff}. We test it on a grid crossing four entanglement profiles with six noise rates (24 configurations, 25% label noise), asking how well test accuracy is predicted by d_{eff} irrespective of which knob produced it.

The answer is a clean *conditional* collapse (Table 4). Among the three *entangled* ansatze, accuracy collapses onto a single curve of d_{eff} (rank correlation -0.896, quadratic fit $R^2 = 0.92$ for this seed), not a seed artifact: over 5 seeds this gives $R^2 = 0.816 \pm 0.082$, $\rho_{\text{Spearman}} = -0.778 \pm 0.159$, and a curve fit to *one* ansatz's noise sweep predicts the *other two* with $R^2 = 0.77$, mean absolute residual 0.036 (App. C.8)—consistent with the exact one-parameter collapse (Prop. 4(i)) and with the entangled maps' comparable alignment, measured directly ($\mathrm{A}(K)$: chain 0.318 ± 0.040, ring 0.350 ± 0.044, all-to-all 0.331 ± 0.039, versus product 0.275 ± 0.042; a genuine second axis, not one outlier, and d_{eff} stands in for the ridge dimension of Lemma 2—both controls in App. C.2). The product circuit is the expected exception, 0.163 (16 points) off the curve at every d_{eff} from its low label alignment (Prop. 4(ii)). Entanglement is thus the precondition supplying alignment, within which d_{eff} is the dominant scalar predictor—the precise sense in which "more entanglement helps" and "noise helps" reduce to one spectral account.

4.5 REAL-DEVICE NOISE: THE MECHANISM ON IBM HERON HARDWARE

The mechanism on real IBM Heron hardware, with accuracy moving in the predicted direction. We ran the feature circuit on a real Heron device (`ibm_kawasaki`), reading out 21 features ($\langle Z_i \rangle$ and $\langle Z_i Z_j \rangle$) with 4096 shots, in a controlled *overfitting* regime (binary task, 30% training-label noise, depth L=3). There the intrinsic hardware noise acts as the predicted spectral regularizer: it

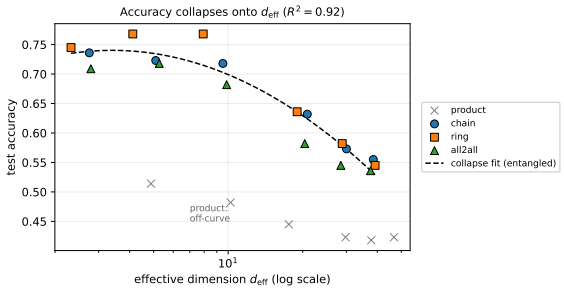

Figure 2: Test accuracy vs. d_{eff}, 4×6 grid (ansatz $\times$ noise rate, 25% label noise). Entangled points collapse onto one curve regardless of whether d_{eff} is set by topology or noise (R^2=0.92); the product circuit (gray $\times$) lies well below it: entanglement is a precondition for d_{eff} to govern.

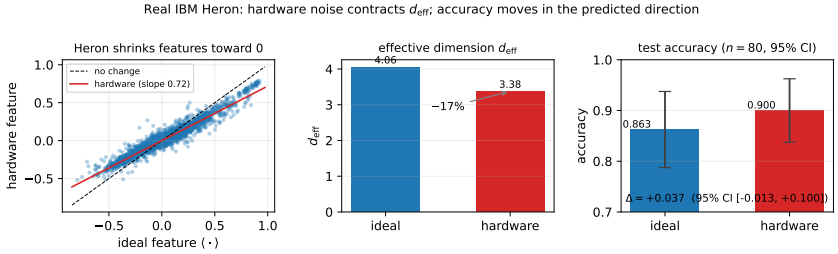

Figure 3: Real IBM Heron (`ibm_kawasaki`), overfitting regime. Left: hardware features shrink toward ideal's origin (slope 0.72). Middle/right: d_{eff} contracts ($4.06 \to 3.38$) and accuracy moves as predicted ($0.863 \to 0.900$, 80 points; 95% bootstrap bars, gain CI $[-0.013, +0.100]$)—contraction resolved, accuracy gain directional only.

contracts the measured kernel from d_{eff}=4.06 to d_{eff}=3.38, and test accuracy moves from 0.863 (noiseless ideal) to 0.900 on hardware (n_{train}=50, n_{test}=80; +0.037, three additional correct test points). We are explicit about the uncertainty: the spectral contraction is measured directly, but the accuracy gain is in the predicted direction *without* being statistically resolved at this sample size—a paired bootstrap over the 80 test points gives $[-0.013, +0.100]$ ($P(\Delta > 0) = 0.88$; exact McNemar on 5 discordant points, $p = 0.38$). Read it as the mechanism appearing on silicon with the right sign, not as a precise effect size. Two controls frame it: in a *non-overfitting* deep run (10-way, L=5) the same contraction instead lowered accuracy (the sign flip of App. C.1), and at shallow L=2 the device noise is too mild to contract at all (d_{eff} moves $37.0 \to 36.1$ under the `FakeTorino` Heron model). Present-day Heron noise thus acts as a regularizer only when the circuit is deep enough and the model overfits.

Seven further studies—the sign-flip falsification of Prop. 3's predicted regime-dependence, depolarizing null and 12-qubit scale check, dataset and depth/width robustness, label-free noise selection, classical spectral baselines, trained QViT-/QCNN-like models, and a kernel-ridge replication of the collapse and sign flip—are reported in App. C; limitations in App. D.

5 Conclusion

Two separately reported curiosities in quantum vision—"more entanglement helps" and "noise helps"—admit a single spectral explanation through the effective dimension of the quantum feature kernel. Noise provably contracts the spectrum (Prop. 1) and along that family generalization is governed by d_{eff} (Prop. 4(i)), with an inverted-U sweet spot whose benefit grows with overfitting and whose sign flips under underfitting; entanglement is the precondition supplying alignment. d_{eff} is not a universal sufficient statistic, but it is one label-free quantity, directly computed from the training kernel, that organizes these design choices; the evidence is mechanistic, not a performance claim (App. D).

REPRODUCIBILITY STATEMENT

Every experiment is fully specified for reproduction. The feature map, ansatz topologies, and noise channels are defined in App. A; per-experiment settings (dataset, split sizes, noise/label-corruption levels, and the seed protocol) are tabulated in App. E (Table 9); and complete proofs of every proposition and lemma are in App. B. All datasets are public (DIGITS via scikit-learn; Fashion-MNIST and BloodMNIST via MedMNIST). Simulated results use PennyLane's `default.mixed`/`default.qubit` devices, which compute noisy channels exactly rather than by shot sampling, so results are deterministic given the seed; multi-seed figures (Sec. 4.4, App. C) report mean$\pm$std over 5 seeds that jointly redraw the PCA fit, train/test split, label-corruption mask, and circuit parameters, as detailed in App. E. The real-device result (Sec. 4.5) is a single run on an IBM Heron processor via Qiskit Runtime; we report its uncertainty as a paired-bootstrap 95% interval over the 80 test points rather than treating it as exact, and discuss this limitation in App. D. Code implementing every experiment—kernel construction, all ansatze and noise channels, the trained QViT-/QCNN-like models, and the hardware run—will be released publicly on acceptance; it is withheld during review to preserve anonymity.

REFERENCES

Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature computational science*, 1(6):403–409, 2021.

Gabriele Agliardi, Giorgio Cortiana, Anton Dekusar, Kumar Ghosh, Naeimeh Mohseni, Corey O'Meara, Víctor Valls, Kavitha Yogaraj, and Sergiy Zhuk. Mitigating exponential concentration in covariant quantum kernels for subspace and real-world data. *npj Quantum Information*, 12:12, 2026.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Chris M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

Thomas Boucher, John Whittle, and Evangelos B Mazomenos. From $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ parameters: Quantum self-attention in vision transformers for biomedical image classification. In *International Workshop on Efficient Medical Artificial Intelligence*, volume 16318 of *Lecture Notes in Computer Science*, pp. 112–122. Springer, 2026.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.

Abdulkadir Canatar, Evan Peters, Cengiz Pehlevan, Stefan M. Wild, and Ruslan Shaydulin. Bandwidth enables generalization in quantum kernel models. *Transactions on Machine Learning Research*, 2023.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

El Amine Cherrat, Iordanis Kerenidis, Natansh Mathur, Jonas Landman, Martin Strahm, and Yun Yvonna Li. Quantum vision transformers. *Quantum*, 8:1265, 2024.

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

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

Elies Gil-Fuster, Casper Gyurik, Adrian Perez-Salinas, and Vedran Dunjko. On the relation between trainability and dequantization of variational quantum learning models. In *International Conference on Learning Representations*, volume 2025, pp. 24069–24093, 2025.

Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

Valentin Heyraud, Zejian Li, Zakari Denis, Alexandre Le Boíté, and Cristiano Ciuti. Noisy quantum kernel machines. *Physical Review A*, 106(5):052421, 2022.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature communications*, 12(1):2631, 2021.

Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Marie Kempkes, Aroosa Ijaz, Elies Gil-Fuster, Carlos Bravo-Prieto, Jakob Spiegelberg, Evert van Nieuwenburg, and Vedran Dunjko. Double descent in quantum kernel methods. *PRX Quantum*, 7:010312, 2026.

Jonas M. Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 12661–12673, 2021.

Changzhou Long, Meng Huang, Xiucai Ye, Yasunori Futamura, and Tetsuya Sakurai. Hybrid quantum-classical-quantum convolutional neural networks. *Scientific Reports*, 15(1):31780, 2025.

Peter Röseler, Oliver Schaudt, Helmut Berg, Christian Bauckhage, and Matthias Koch. Efficient quantum convolutional neural networks for image classification: Overcoming hardware constraints. *arXiv preprint arXiv:2505.05957*, 2025.

Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv preprint arXiv:2101.11020*, 2021.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.

Shangshang Shi, Zhimin Wang, Jiaxin Li, Yanan Li, Ruimin Shang, Guoqiang Zhong, and Yongjian Gu. Quantum convolutional neural networks for multiclass image classification. *Quantum Information Processing*, 23(5):189, 2024.

Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Supanut Thanasilp, Samson Wang, M. Cerezo, and Zoë Holmes. Exponential concentration in quantum kernel methods. *Nature Communications*, 15:5200, 2024.

Joachim Tomasi, Sandrine Anthoine, and Hachem Kadri. Benign overfitting with quantum kernels. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 337 of *Proceedings of Machine Learning Research*, pp. 6740–6766, 2026.

Mingzhu Wang and Yun Shang. Hybrid vision transformer and quantum convolutional neural network for image classification. *arXiv preprint arXiv:2510.12291*, 2025.

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

Wanqing Wu and Yuxiang Zhang. Hybrid quantum inception-inspired convolutional neural network for image classification. *The Journal of Supercomputing*, 81:1629, 2025.

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

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10:41, 2023.

Claudia Zendejas-Morales, Debashis Saikia, and Utkarsh Singh. Local and multi-scale strategies to mitigate exponential concentration in quantum kernels. *arXiv preprint arXiv:2602.16097*, 2026.

Hui Zhang, Qinglin Zhao, Mengchu Zhou, and Li Feng. HQViT: Hybrid quantum vision transformer for image classification. *arXiv preprint arXiv:2504.02730*, 2025.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

A Background: quantum feature circuits and the ansatze

We collect, for readers from the vision/ML community, the quantum-computing notions used in the paper, with explicit formulas. Table 3 summarizes the notation.

Table 3: Notation and explicit definitions.

Symbol	Name	Definition / formula
n_q	number of qubits	state space is $\mathbb{C}^{2^{n_q}}$
$\|\psi\rangle$	pure state	unit vector in $\mathbb{C}^{2^{n_q}}$
ρ	(mixed) state	$\rho \succeq 0$, $\operatorname{tr}\rho = 1$; pure: $\rho = \|\psi\rangle\langle\psi\|$
I, X, Y, Z	Pauli matrices	$X=\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$, $Y=\left(\begin{smallmatrix}0&-i\\i&0\end{smallmatrix}\right)$, $Z=\left(\begin{smallmatrix}1&0\\0&-1\end{smallmatrix}\right)$
$R_Y(\theta)$	Y-rotation	$e^{-i\theta Y/2} = \left(\begin{smallmatrix}\cos\frac{\theta}{2} & -\sin\frac{\theta}{2}\\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2}\end{smallmatrix}\right)$
$R_Z(\theta)$	Z-rotation	$e^{-i\theta Z/2} = \operatorname{diag}(e^{-i\theta/2}, e^{+i\theta/2})$
$\mathrm{CNOT}_{a\to b}$	controlled-NOT	flips qubit b iff qubit a is $\|1\rangle$; see equation 6
$\rho(x)$	quantum feature map	density matrix after the circuit on input x
$k(x, x')$	kernel	$\operatorname{Tr}[\rho(x)\rho(x')]$ (Eq. 2)
$\mathcal{N}_p, \mathcal{A}_\gamma$	noise channels	depolarizing / amplitude damping (below)

Qubits and states (the ML picture). One qubit is a unit vector in $\mathbb{C}^2$; n_q qubits live in the tensor-product space $\mathbb{C}^{2^{n_q}}$. A general (possibly noisy) state is a density matrix ρ (PSD, unit trace). For ML intuition, the circuit is a *fixed nonlinear feature map* $x \mapsto \rho(x)$ into the $2^{n_q} \times 2^{n_q}$ matrix space, and the kernel $k(x, x') = \operatorname{Tr}[\rho(x)\rho(x')]$ is the inner product of these feature maps—just like a classical kernel method, but with a quantum-circuit feature map.

Gates, explicitly. Gates are unitary matrices acting on the state. Single-qubit rotations $R_Y(\theta), R_Z(\theta)$ (Table 3) rotate a qubit continuously; data enters through their angles ("angle encoding"), as $R_Y(x_i + \theta_{\ell,i})R_Z(0.7\,x_i)$ on qubit i. The two-qubit *controlled-NOT* is, in the basis $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$,

$$\mathrm{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{6}$$

i.e. it applies X to the target qubit conditioned on the control being $|1\rangle$. CNOTs are the source of *entanglement*: a state is entangled when it cannot be written as a product $\bigotimes_q \rho_q$ over qubits, so its features cannot be factorized into independent per-qubit (per-coordinate) features—the quantum analogue of cross-feature interactions.

Entangling topologies, explicitly. An entangling block is a set of CNOTs $\{\mathrm{CNOT}_{a\to b} : (a,b) \in E\}$ whose edge set E defines the four ansatze (Fig. 4):

$$E_{\text{PRODUCT}} = \varnothing, \qquad E_{\text{CHAIN}} = \{(i, i{+}1)\}_{i=0}^{n_q-2},$$
$$E_{\text{RING}} = \{(i, (i{+}1) \bmod n_q)\}_{i=0}^{n_q-1}, \qquad E_{\text{ALL2ALL}} = \{(i,j) : i < j\}.$$

The product map factorizes ($k = \prod_q \mathrm{Tr}[\rho_q(x)\rho_q(x')]$); more connectivity (chain → ring → all-to-all) creates richer correlations in $\rho(x)$ and, empirically, higher kernel–target alignment on our tasks (Sec. 4.4). We do not claim a general expressivity impossibility here: the RKHS of a product kernel is a tensor product of per-factor RKHSs and in general still contains cross-coordinate interaction terms, so the alignment gap we observe is a property of this ansatz on these tasks, not a structural theorem about product kernels.

Noise channels, explicitly. Hardware imperfections are completely positive trace-preserving maps. *Global depolarizing* mixes a state toward the maximally mixed state,

$$\mathcal{N}_p[\rho] = (1-p)\,\rho + p\,\tfrac{I}{2^{n_q}},$$

and *amplitude damping* $\mathcal{A}_\gamma$ models energy relaxation toward $|0\rangle$ via per-qubit Kraus operators $E_0 = \mathrm{diag}(1, \sqrt{1-\gamma})$, $E_1 = \sqrt{\gamma}\,|0\rangle\langle 1|$ ($\mathcal{A}_\gamma[\rho] = E_0\rho E_0^\dagger + E_1\rho E_1^\dagger$). In ML terms both are *contractions* that shrink the feature map toward a fixed point, which is why they act as spectral (ridge-like) regularizers in the main text. Both are simulated exactly on a density-matrix backend.

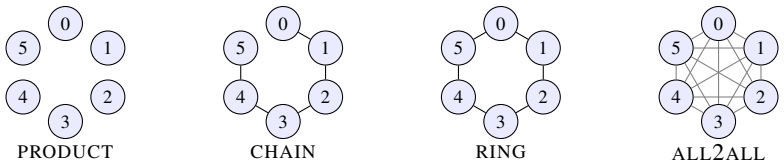

Figure 4: The four entangling topologies (shown for n_q=6 qubits; nodes are qubits, edges are CNOTs in the entangling block).

One re-uploading layer. Each of the L layers encodes the data, then entangles, then (optionally) injects noise. For the RING ansatz on 4 qubits one layer is:

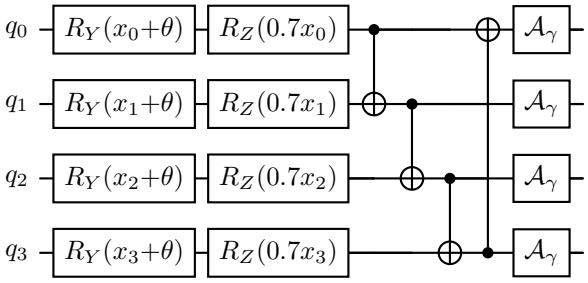

Figure 5: One data-re-uploading layer (RING entangling block, amplitude damping $\mathcal{A}_\gamma$). The kernel uses $\rho(x)$ after L such layers via equation 2.

Trained-model architectures. The trained ablations of Sec. C.7 (Table 8) optimize the circuit parameters and a linear classifier head end-to-end. The *QViT-like* map (Fig. 6) re-uploads the input with trainable rotations and an all-to-all entangling block per layer, reading out $\langle Z_i\rangle$ on every qubit. The *QCNN-like* map (Fig. 7) replaces the entangler with two brick-pattern layers of parametrized two-qubit *convolutions* $U(\boldsymbol{\theta})$. Amplitude damping $\mathcal{A}_\gamma$ is injected during training in the noisy runs.

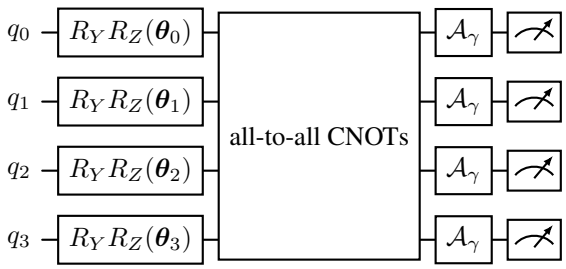

Figure 6: QViT-like trained map (shown on 4 qubits): trainable angle encoding, an all-to-all entangling block, optional injected $\mathcal{A}_\gamma$, repeated $\times L$; the per-qubit $\langle Z_i \rangle$ feed a trained linear head.

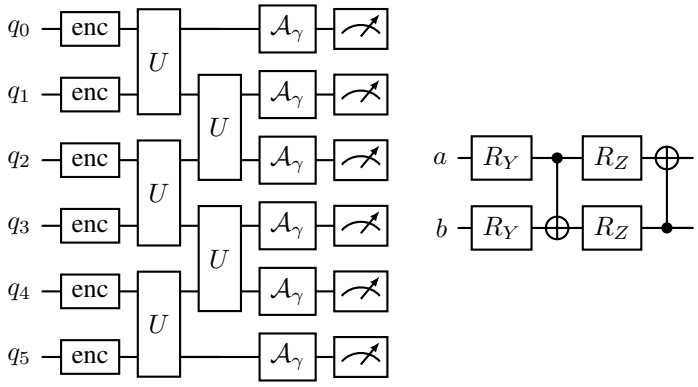

Figure 7: QCNN-like trained map. **Left:** two brick-pattern layers of two-qubit convolutions U (here one layer on pairs $(0,1),(2,3),(4,5)$ and the next on $(1,2),(3,4)$, with the ring closure $(5,0)$ omitted for the drawing), optional $\mathcal{A}_\gamma$, then per-qubit readout into a linear head. **Right:** the convolution block $U(\boldsymbol{\theta})$.

B Proofs

Throughout, $\rho(x) \in \mathbb{C}^{D\times D}$ ($D = 2^{n_q}$) are density matrices, $\phi(x) = \operatorname{vec}\rho(x) \in \mathbb{C}^{D^2}$ are their vectorizations, and $K \in \mathbb{R}^{n\times n}$ is the Gram matrix $K_{ij} = \operatorname{Tr}[\rho(x_i)\rho(x_j)] = \langle \phi(x_i), \phi(x_j)\rangle$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$.

B.1 Proof of Lemma 1

PSD. For any $c \in \mathbb{R}^n$,

$$c^\top K c = \sum_{i,j} c_i c_j \langle \phi(x_i), \phi(x_j)\rangle = \Big\langle \sum_i c_i \phi(x_i),\, \sum_j c_j \phi(x_j) \Big\rangle = \Big\| \sum_i c_i \phi(x_i) \Big\|^2 \geq 0,$$

where we used the (real part of the) Hermitian HS inner product; since $\rho(x_i)$ are Hermitian, $K_{ij} = \operatorname{Tr}[\rho(x_i)\rho(x_j)]$ is real and symmetric. Hence $K \succeq 0$ and all $\lambda_i \geq 0$, so $d_{\text{eff}}(K)$ in equation 3 is well defined.

Range. Writing $\Lambda = \sum_i \lambda_i$ and $Q = \sum_i \lambda_i^2$, the Cauchy–Schwarz inequality $\Lambda^2 = (\sum_i \lambda_i \cdot 1)^2 \leq n \sum_i \lambda_i^2 = nQ$ gives $d_{\text{eff}} = \Lambda^2/Q \leq n$, with equality iff all λ_i are equal (flat spectrum). Since $\lambda_i \geq 0$, $\Lambda^2 = \sum_i \lambda_i^2 + \sum_{i\neq j} \lambda_i\lambda_j \geq Q$, so $d_{\text{eff}} \geq 1$, with equality iff exactly one $\lambda_i > 0$, i.e. $\operatorname{rank} K = 1$. □

B.2 Proof of Proposition 1

Exact decomposition. For the global depolarizing channel $\mathcal{N}_p[\rho] = (1-p)\rho + p\, I/D$ and $\rho_p(x) = \mathcal{N}_p[\rho(x)]$, bilinearity of the trace gives

$$\begin{aligned}(K_p)_{ij} &= \operatorname{Tr}\left[\rho_p(x_i)\rho_p(x_j)\right] \\ &= (1-p)^2 \operatorname{Tr}[\rho(x_i)\rho(x_j)] + \tfrac{(1-p)p}{D} \operatorname{Tr}[\rho(x_i)] + \tfrac{(1-p)p}{D} \operatorname{Tr}[\rho(x_j)] + \tfrac{p^2}{D^2} \operatorname{Tr}[I] \\ &= (1-p)^2 K_{ij} + \tfrac{(1-p)p}{D} + \tfrac{(1-p)p}{D} + \tfrac{p^2}{D},\end{aligned}$$

using $\operatorname{Tr}\rho = 1$ and $\operatorname{Tr} I = D$. The last three terms sum to $\frac{p}{D}\big(2(1-p)+p\big) = \frac{p(2-p)}{D}$, independent of i, j, which proves

$$K_p = (1-p)^2 K + \tfrac{p(2-p)}{D}\, \mathbf{1}\mathbf{1}^\top. \tag{4}$$

Limit. As $p \to 1$, $(1-p)^2 \to 0$ and $\frac{p(2-p)}{D} \to \frac{1}{D}$, so $K_p \to \frac{1}{D}\mathbf{1}\mathbf{1}^\top$, which has rank 1; by Lemma 1, $d_{\text{eff}}(K_p) \to 1$.

Monotone concentration under constant row sums. Suppose $K\mathbf{1} = c\mathbf{1}$, i.e. $\mathbf{1}$ is an eigenvector of K (note: Perron–Frobenius only makes the top eigenvector *positive*, not constant, so this is a genuine extra assumption). Then K and $\mathbf{1}\mathbf{1}^\top$ are simultaneously diagonalizable: with orthonormal eigenvectors $u_1 = \mathbf{1}/\sqrt{n}, u_2, \ldots, u_n$, we have $K = \sum_k \lambda_k u_k u_k^\top$ and $\mathbf{1}\mathbf{1}^\top = n\, u_1 u_1^\top$. Put $s = (1-p)^2$ and note $\frac{p(2-p)}{D}\, n = \frac{n}{D}(1-s) =: a(s)$ with $\beta := n/D$, $a = \beta(1-s)$. The eigenvalues of K_p are

$$\mu_1(p) = s\lambda_1 + a, \qquad \mu_k(p) = s\lambda_k \;\; (k \geq 2).$$

Consider the top-eigenvalue share $\sigma_1 = \mu_1 / \sum_k \mu_k = \dfrac{s\lambda_1 + \beta(1-s)}{s\Lambda + \beta(1-s)} = \dfrac{\beta + s(\lambda_1 - \beta)}{\beta + s(\Lambda - \beta)}$. Differentiating in s,

$$\frac{d\sigma_1}{ds} = \frac{(\lambda_1 - \beta)\big(\beta + s(\Lambda-\beta)\big) - (\Lambda - \beta)\big(\beta + s(\lambda_1 - \beta)\big)}{\big(\beta + s(\Lambda - \beta)\big)^2} = \frac{\beta(\lambda_1 - \Lambda)}{\big(\beta + s(\Lambda-\beta)\big)^2} \leq 0,$$

since $\Lambda = \sum_k \lambda_k \geq \lambda_1$. Thus σ_1 is non-increasing in s, i.e. non-decreasing in p, and $d_{\text{eff}}(K_p) = 1/\sum_k (\mu_k / \sum_l \mu_l)^2$ decreases; this is rigorous under $K\mathbf{1} = c\mathbf{1}$.

General case and the obstruction. Without constant row sums, K and $\mathbf{1}\mathbf{1}^\top$ do not commute, the eigenvalues of K_p are not the μ_k above, and the simultaneous-diagonalization step fails. The exact participation ratio is nevertheless equation 5, obtained from $\operatorname{tr} K_p = sT + \frac{p(2-p)}{D} n$ and $\operatorname{tr}(K_p^2) = s^2 Q + 2s\frac{p(2-p)}{D} S + \big(\frac{p(2-p)}{D}\big)^2 n^2$ using $\operatorname{tr}(\mathbf{1}\mathbf{1}^\top) = n$, $\operatorname{tr}(K\mathbf{1}\mathbf{1}^\top) = \mathbf{1}^\top K \mathbf{1} = S$, and $\operatorname{tr}((\mathbf{1}\mathbf{1}^\top)^2) = n^2$. This is a ratio of quadratics in s; its derivative's sign depends on $S = \mathbf{1}^\top K \mathbf{1}$, the cross-term that the constant-row-sum condition would fix, so global monotonicity is *not* guaranteed without structure. We therefore claim only: (i) the exact limit $d_{\text{eff}}(K_p) \to 1$; (ii) the rigorous upper bound $d_{\text{eff}}(K_p) \leq 1/\sigma_1(p)^2$ with σ_1 the top-eigenvalue share, which tends to 1; (iii) provable monotonicity under $K\mathbf{1} = c\mathbf{1}$; and (iv) the empirical fact that $d_{\text{eff}}(K_p)$ is strictly decreasing—hence injective—in every experiment, which is all the collapse of Prop. 4(i) requires. □

B.3 Proof of Proposition 2 (amplitude damping)

The single-qubit amplitude-damping channel has Kraus operators $E_0 = \operatorname{diag}(1, \sqrt{1-\gamma})$ and $E_1 = \sqrt{\gamma}\, |0\rangle\langle 1|$, so on $\rho = \left(\begin{smallmatrix} \rho_{00} & \rho_{01} \\ \rho_{10} & \rho_{11} \end{smallmatrix}\right)$ it acts as $\rho_{00} \mapsto \rho_{00} + \gamma\rho_{11}$, $\rho_{11} \mapsto (1-\gamma)\rho_{11}$, $\rho_{01} \mapsto \sqrt{1-\gamma}\,\rho_{01}$.

(a) Limit. As $\gamma \to 1$ every qubit's state tends to $|0\rangle\langle 0|$ (populations flow to $|0\rangle$, coherences vanish), so $\rho_\gamma(x) \to |0\rangle\langle 0|^{\otimes n_q}$ uniformly in x; hence $(K_\gamma)_{ij} = \operatorname{Tr}[\rho_\gamma(x_i)\rho_\gamma(x_j)] \to 1$ for all i, j, i.e. $K_\gamma \to \mathbf{1}\mathbf{1}^\top$, which is rank one, so $d_{\text{eff}}(K_\gamma) \to 1$ by Lemma 1.

(b) Single-qubit contraction. In Bloch coordinates $\rho = \frac{1}{2}(I + X\sigma_x + Y\sigma_y + Z\sigma_z)$ the rules above give $X' = \sqrt{1-\gamma}\, X$, $Y' = \sqrt{1-\gamma}\, Y$, and, using $\rho_{11} = (1-Z)/2$, $Z' = Z + \gamma(1-Z) = (1-\gamma)Z + \gamma$. Since $\|\rho - \sigma\|_{\text{HS}}^2 = \frac{1}{2}(\Delta X^2 + \Delta Y^2 + \Delta Z^2)$,

$$\|\mathcal{A}_\gamma\rho - \mathcal{A}_\gamma\sigma\|_{\text{HS}}^2 = \tfrac{1}{2}\big[(1-\gamma)\Delta X^2 + (1-\gamma)\Delta Y^2 + (1-\gamma)^2 \Delta Z^2\big] \leq (1-\gamma)\, \|\rho - \sigma\|_{\text{HS}}^2,$$

because $(1-\gamma)^2 \le (1-\gamma)$. Thus $\mathcal{A}_\gamma$ strictly contracts the traceless part by the factor $\sqrt{1-\gamma}$: every *pairwise* Hilbert–Schmidt distance shrinks. This does not by itself force $d_{\text{eff}}(K_\gamma)$ to be non-increasing—even for the simplest product feature map (a single qubit, two pure states ρ, σ with Bloch vectors $r = (X_\rho, Y_\rho, Z_\rho)$, $s = (X_\sigma, Y_\sigma, Z_\sigma)$), the pairwise overlap $\mathrm{Tr}[\rho_\gamma \sigma_\gamma] = \frac{1}{2}\big[1 + \gamma^2 + (1-\gamma)(X_\rho X_\sigma + Y_\rho Y_\sigma) + (1-\gamma)^2 Z_\rho Z_\sigma + \gamma(1-\gamma)(Z_\rho + Z_\sigma)\big]$ mixes a $(1-\gamma)$ term with a $(1-\gamma)^2$ term, so it need not be monotone in γ: for Z_ρ, Z_σ of opposite sign we verified numerically that $d_{\text{eff}}(K_\gamma)$ can rise—not fall—over an initial range of γ before eventually contracting. We therefore do not claim a product-feature-map monotonicity theorem, and, as (c) below shows, constant row sums do not repair this for amplitude damping either; we report monotone d_{eff} contraction under amplitude damping as an empirical property of the spectra we observe.

(c) Boundary. On a single qubit $\mathcal{A}_\gamma(I) = I + \gamma Z$, so in the Pauli basis the (I, Z) block is $\left(\begin{smallmatrix} 1 & 0 \\ \gamma & 1-\gamma \end{smallmatrix}\right)$, which is non-normal with largest singular value > 1 for $\gamma \in (0, 1)$. Hence $\mathcal{A}_\gamma^{\otimes n_q}$ can *increase* the Hilbert–Schmidt norm of operators with a nonzero trace component; the difference of two *entangled* feature states can have weight along such directions, so the clean global contraction available for depolarizing noise (Prop. 1) is not guaranteed in general. If K_γ has constant row sums for every γ, $\mathbf{1}$ remains an eigenvector—but unlike the depolarizing family, this alone does not pin down how the remaining eigenvalues move with γ: the top-share argument of Prop. 1 relies on the specific rank-one spectral form $\mu_1(p) = s\lambda_1 + \beta(1-s)$, $\mu_{k\ge 2}(p) = s\lambda_k$, which amplitude damping does not have. (Constant row sums is in any case an additional structural assumption, not implied by positivity: a strictly positive kernel matrix need not have equal row sums, and Perron–Frobenius only guarantees that the top eigenvector is positive, not that it is $\mathbf{1}$.) We therefore treat monotone $d_{\text{eff}}(K_\gamma)$ contraction under amplitude damping as an empirical property of the ensembles studied here: it decreases strictly at every step in every experiment we run (Tables 1, 2). □

B.4 Proof of Proposition 3

We use the standard capacity control of kernel ridge regression (KRR) by the effective dimension $d_\gamma(K) = \mathrm{tr}[K(K+\gamma I)^{-1}] = \sum_i \lambda_i/(\lambda_i + \gamma)$ (Abbas et al., 2021). For a bounded-norm RKHS ball, the local Rademacher complexity at scale γ obeys

$$\mathcal{R}_n \;\lesssim\; \sqrt{\tfrac{1}{n}\textstyle\sum_i \min(\lambda_i, \gamma)} \;\le\; \sqrt{\tfrac{2\gamma}{n}\, d_\gamma(K)}.$$

The second inequality uses the elementary bound $\min(\lambda, \gamma) \le \dfrac{2\gamma\lambda}{\lambda+\gamma} = 2\gamma \cdot \dfrac{\lambda}{\lambda+\gamma}$, valid for all $\lambda, \gamma \ge 0$: if $\lambda \le \gamma$ the right side is $\ge 2\gamma\lambda/(2\gamma) = \lambda$, and if $\lambda > \gamma$ it is $\ge 2\gamma\lambda/(2\lambda) = \gamma$; summing over i gives $\sum_i \min(\lambda_i, \gamma) \le 2\gamma\, d_\gamma(K)$. Standard local Rademacher bounds then give an excess-risk gap $\tilde{O}\big(\sqrt{d_\gamma(K)/n}\big)$. Each summand of d_γ is increasing in its eigenvalue,

$$\frac{\partial}{\partial \lambda_i}\frac{\lambda_i}{\lambda_i + \gamma} = \frac{\gamma}{(\lambda_i + \gamma)^2} > 0,$$

so any spectral contraction that lowers eigenvalues—in particular $K \mapsto K_p$ of Proposition 1, which multiplies the non-top eigenvalues by $(1-p)^2$—decreases d_γ and tightens the bound. When the empirical risk is ≈ 0 (the overfitting regime induced by label noise), the excess risk is gap-dominated, so a tighter bound *predicts* a reduction of the variance-controlled component of risk—not a guarantee, since a tighter upper bound does not by itself imply a lower expected risk (App. C.1 tests the predicted sign directly); once the top, signal-carrying eigen-directions are themselves attenuated the bias grows and the risk turns up, yielding an interior optimum. We use the participation ratio d_{eff} equation 3 as a γ-free surrogate for d_γ; both are non-decreasing under simultaneous inflation of the spectrum and co-vary on the spectra we observe. □

B.5 Proof of Proposition 4

Let the target function have coefficients $\beta_i = \langle f^\star, u_i \rangle$ in the kernel eigenbasis $\{u_i\}$ (eigenvalues λ_i) of the population kernel operator; we follow standard practice (Canatar et al., 2021; Caponnetto & De Vito, 2007) in approximating $\{\lambda_i, \beta_i\}$ by the finite-sample quantities computed from the training Gram matrix K, without deriving a finite-sample concentration bound for this substitution. The

Table 4: P3 (unification): within the entangled regime, accuracy collapses onto a single function of d_{eff}, whether d_{eff} is moved by topology or by noise (4 ansatze $\times$ 6 noise rates, 25% label noise); the product circuit is an outlier.

Metric	Value
Spearman(d_{eff}, acc), entangled (18 configs)	-0.896
Spearman(d_{eff}, acc), all 24 configs	-0.718
Global R^2 (acc $\sim$ quad($\log d_{\text{eff}}$)), entangled	0.921
Mean |residual|, entangled configs (global fit)	0.018
Mean |residual|, product configs (global fit)	0.163 (outlier)
Collapse R^2 (one ansatz's noise curve $\rightarrow$ other two)	0.772

Table 5: Falsification of the predicted sign flip. Lower d_{eff} helps under overfitting and hurts under underfitting.

Regime	Train acc	$\rho_{\text{Spearman}}(d_{\text{eff}}, \text{test})$	best p	noise gain
Overfitting (expressive, 25% noisy labels)	1.00	-0.79	0.45	$+0.156$
Underfitting (3-qubit product, clean)	0.85	$+0.91$	0.00	-0.045

KRR estimator with ridge γ and n samples (label noise variance τ^2) has the classical bias–variance excess-risk decomposition

$$\mathcal{E}(\hat{f}) = \underbrace{\sum_i \Big(\frac{\gamma}{\lambda_i + \gamma}\Big)^2 \beta_i^2}_{\text{bias}^2\ \text{(alignment)}} + \underbrace{\frac{\tau^2}{n} \sum_i \Big(\frac{\lambda_i}{\lambda_i + \gamma}\Big)^2}_{\text{variance (capacity)}}.$$

The variance term depends on K only through its spectrum $\{\lambda_i\}$, and the bias term depends additionally on the full target energy profile $\{\beta_i^2\}$ across eigen-directions, which we summarize empirically, though only partially, by the scalar $\mathrm{A}(K)$.

Claim (i): exact one-parameter collapse. The depolarizing family K_p of Prop. 1 is determined by the single scalar p (the whole kernel, hence its spectrum and the β_i, are explicit functions of p via equation 4). Therefore the excess risk $\mathcal{E}$ is a function of p. If, in addition, $d_{\text{eff}}(K_p)$ is *injective* in p—which we do not claim in general but verify holds (indeed d_{eff} is strictly monotone) in every experiment, and which holds provably under the constant-row-sum condition of Prop. 1—then $\mathcal{E}$ is a function of d_{eff} and accuracy collapses exactly onto $\text{acc} = f(d_{\text{eff}})$. The collapse along the noise family thus requires only injectivity of $p \mapsto d_{\text{eff}}(K_p)$, not the eigenvector-fixing of the (false) Perron reading.

Claim (ii): cross-family collapse is conditional. For two kernels of different spectral *shape*, equal $d_{\text{eff}} = (\operatorname{tr} K)^2 / \operatorname{tr}(K^2)$ constrains only this one scalar; the variance term $\frac{\tau^2}{n} \sum_i (\lambda_i/(\lambda_i + \gamma))^2$ and the bias term are not determined by it, so equal d_{eff} need not give equal risk. Collapse onto a common curve therefore requires, beyond equal d_{eff}, comparable spectral shape *and* comparable alignment $\mathrm{A}(K)$. The unentangled product map factorizes, $k(x, x') = \prod_q \operatorname{Tr}[\rho_q(x)\rho_q(x')]$, and has measurably lower alignment on these tasks (Sec. 4.4, App. A); its larger bias term places it below the entangled curve at every d_{eff}. The entangled ansatze, having comparable shape and alignment, collapse empirically. □

C Additional experiments and analyses

C.1 Falsification: the d_{eff}–accuracy sign flips between regimes

Prop. 3 predicts contracting d_{eff} helps under overfitting and hurts under underfitting—falsifiable, and it holds (Table 5, Fig. 8): overfitting (expressive circuit, 25% label noise, train acc $= 1.0$) gives $\rho_{\text{Spearman}}(d_{\text{eff}}, \text{test}) = -0.79$ (best $p = 0.45$, gain $+0.156$), while underfitting (3-qubit product map, clean labels, train acc $= 0.85$) flips the sign to $+0.91$ (best $p = 0$)—d_{eff} is to be matched to the task, not minimized.

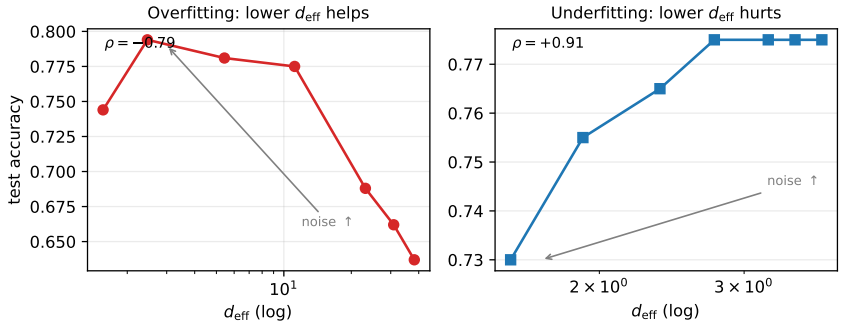

Figure 8: The d_{eff}–accuracy relation flips sign between regimes. **Left** (overfitting; expressive circuit, 25% noisy labels, train acc 1.0): contracting d_{eff} via noise raises test accuracy ($\rho = -0.79$). **Right** (underfitting; 3-qubit product map, clean labels, train acc 0.85): the same contraction lowers test accuracy ($\rho = +0.91$). Arrows mark increasing noise.

Table 6: Entanglement as a precondition: noiseless circuits, mean±std over 5 seeds (Sec. 4.1). The unentangled product map is markedly worse; among entangled topologies d_{eff} and accuracy are not separately ordered.

Profile	d_{eff}	Test acc
Product	54.2 ± 6.0	0.471 ± 0.060
Chain	42.0 ± 2.7	0.616 ± 0.025
Ring	43.9 ± 3.3	0.624 ± 0.026
All-to-all	45.0 ± 4.7	0.610 ± 0.048

C.2 Alignment as a second axis, and the spectral bridge

To show that kernel–target alignment is a genuine *axis* rather than a single product outlier, we interpolate kernels $K_t = (1-t)K_{\text{PRODUCT}} + t\,K_{\text{ALL2ALL}}$ at roughly fixed d_{eff} and track their distance from the entangled collapse curve of Sec. 4.4: as $\mathrm{A}(K)$ rises from 0.275 to 0.33, the residual moves monotonically from -0.11 (below the curve) to ≈ 0 (on it), with $\rho_{\text{Spearman}}(\mathrm{A}, \text{residual}) = +0.66 \pm 0.27$ over 5 seeds. We also verify the spectral bridge of Lemma 2: along the noise sweep d_{eff} and the ridge dimension d_γ are perfectly rank-correlated (ρ_{Spearman}=1.0, Pearson 0.94), so reporting d_{eff} rather than d_γ loses no ordering information.

Centered d_{eff}: ruling out a classifier-invisible artifact. An intercept SVM's dual objective is invariant to adding any constant to every kernel entry (annihilated by $\sum_i \alpha_i y_i = 0$), so a diagnostic that improves only by growing such a constant would be classifier-irrelevant—exactly the failure mode of the depolarizing family once centered (Sec. C.3). We recompute d_{eff} on the centered kernel HKH ($H = I - \mathbf{1}\mathbf{1}^\top/n$, training block only) for the amplitude-damping noise sweep and the 4×6 collapse grid, using a matched-protocol replication (Digits, all-to-all/chain/ring/product, 25% label noise). Along the single-ansatz noise sweep, centered and raw d_{eff} are both perfectly rank-correlated with test accuracy ($\rho_{\text{Spearman}} = -1.00$ each, 8 points) and with each other ($\rho_{\text{Spearman}} = +1.00$). Across the 4×6 entangled collapse grid, centered d_{eff} tracks accuracy at least as well as raw d_{eff} (Spearman -0.95 vs. -0.92; quadratic-fit $R^2 = 0.92$ vs. 0.89). The amplitude-damping mechanism is therefore not an artifact of the kind that afflicts the exact-depolarizing null: centering leaves the story unchanged, or slightly sharpens it.

C.3 A sanity check (not a headline): the depolarizing one-parameter null

We include the global-depolarizing family as a controlled *null*, and are explicit that it is one. Because K_p is an exact function of the single scalar p equation 4, both $\mathrm{acc}(p)$ and $d_{\text{eff}}(K_p)$ are deterministic functions of p; hence whenever $p \mapsto d_{\text{eff}}$ is injective, acc is *by construction* a function of d_{eff} and

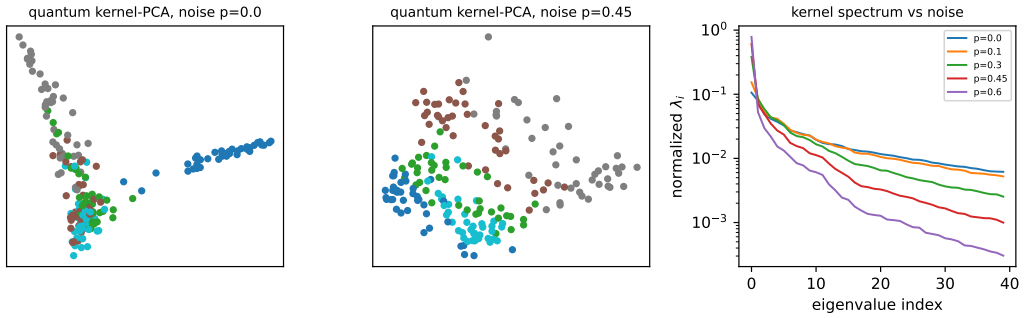

Figure 9: Noise (amplitude damping) reshapes the quantum feature geometry by contracting the spectrum. **Left, middle:** quantum kernel-PCA embedding of test points (colored by class) at noise p=0 and p=0.45. **Right:** the kernel eigenvalue spectrum decays increasingly fast with p—a direct visualization of the ridge-like contraction $d_{\text{eff}}(p) \downarrow$ that Prop. 2 establishes as a strict single-qubit contraction and that we verify empirically here (Sec. 4.2).

the collapse is exact—this would hold equally for any injective scalar of K_p (e.g. p itself, $(1-p)^2$, $\operatorname{tr} K_p^2$, or the spectral entropy). The exact collapse therefore does not, on its own, single out d_{eff}; the informative result is the *cross-ansatz* collapse of Sec. 4.4 (R^2=0.82 $\pm$ 0.08), where genuinely different spectral shapes are compressed onto one curve.

What this family *does* usefully verify is the analytic theory and its reach. The decomposition $K_p = (1-p)^2 K + \frac{p(2-p)}{D}\mathbf{1}\mathbf{1}^\top$ matches a density-matrix simulation to machine precision (max entrywise error 7.8×10^{-16} at n_q=4), and because K_p is analytic in the noiseless *statevector* kernel (no $2^{n_q} \times 2^{n_q}$ density matrix needed) we can evaluate it at $n_q \in \{8, 10, 12\}$ qubits; there $d_{\text{eff}}(K_p)$ is strictly decreasing (hence injective) and the by-construction collapse is, as expected, R^2=0.997–1.000. We report this as confirmation of equation 4 at scale and of the injectivity that Prop. 4(i) needs—not as evidence for the spectral thesis, which rests on the cross-ansatz collapse.

This weakness is not merely a reparametrization technicality: it is exactly the failure mode that motivates *centering* the kernel (Heyraud et al., 2022). Because the centering projector $H = I - \mathbf{1}\mathbf{1}^\top/n$ satisfies $H\mathbf{1} = 0$, equation 4 gives $HK_pH = (1-p)^2 HKH$ *exactly*: the centered kernel is a uniform rescaling of the noiseless centered kernel by the scalar $(1-p)^2$, so every centered eigenvalue shrinks at the same rate and the *normalized* spectral shape is unchanged. Since the participation ratio is scale-invariant ($d_{\text{eff}}(cM) = d_{\text{eff}}(M)$ for any $c > 0$), the centered d_{eff} is exactly constant in p (confirmed numerically to four significant figures for $p \in [0, 0.9]$), even though the centered eigenvalues themselves shrink toward zero. Since an intercept SVM's dual objective depends on K only through $\alpha^\top(yy^\top \odot K)\alpha$ under the constraint $\sum_i \alpha_i y_i = 0$, the rank-one term is annihilated identically—the classifier cannot see the raw $d_{\text{eff}} \to 1$ limit at all, because it is driven entirely by a component with no bearing on the decision boundary. We verify that our headline amplitude-damping mechanism does not share this failure mode (App. C.2): raw and centered d_{eff} remain strongly, comparably correlated with test accuracy throughout.

C.4 Robustness: datasets, depth, and width

Additional datasets. We repeat the full collapse grid on Fashion-MNIST and on BloodMNIST (MedMNIST), reducing each to 6 PCA features (Table 7). On Fashion-MNIST the picture is as strong as on Digits (entangled $\rho = -0.88$, $R^2 = 0.88$, noise helps every ansatz). The medical BloodMNIST is weaker in Table 7 ($R^2 = 0.46$), but a controlled follow-up shows this is largely a *confound*, not a fundamentally harder distribution: that table used a 4-class subset (for parity with a small budget), and the weakness is explained by class count and sample size, not the medical images. Matching to its full 8-class task raises the collapse to $R^2 = 0.84$ ($\rho = -0.90$), and doubling the training set on the 4-class task raises it to $R^2 = 0.78$—both close to Digits/Fashion. (For reference, reducing Digits to 4 classes leaves $R^2 = 0.90$, so the effect is data-dependent.) The spectral story thus survives on a medical benchmark once the class-count and sample-size confounds are removed.

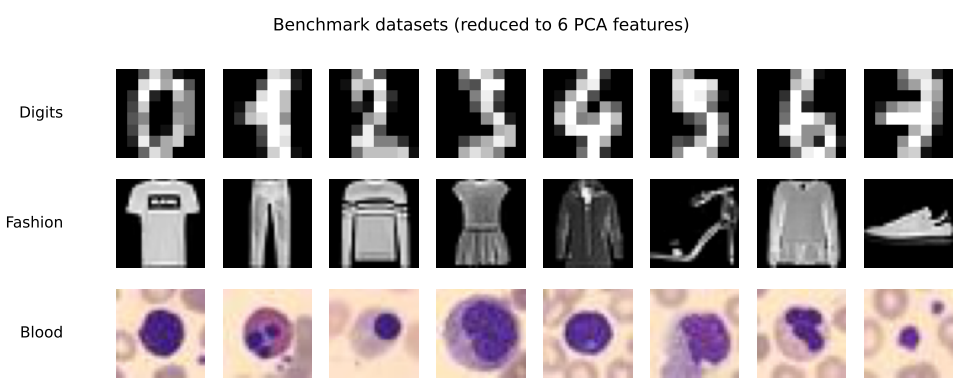

Figure 10: Example images from the three vision benchmarks used (each reduced to 6 principal components before the quantum feature map): handwritten Digits, Fashion-MNIST, and the medical BloodMNIST.

Table 7: Robustness across datasets (collapse grid, 6 qubits, 25% label noise). The mechanism is strong on Digits/Fashion and weaker on the harder medical task.

Dataset	Spearman(d_{eff}, acc) ent.	Collapse R^2 ent.	Product residual
Digits	-0.94	0.91	0.074
Fashion-MNIST	-0.88	0.88	0.080
BloodMNIST (4-cls)	-0.45	0.46	0.028

Depth and width. Varying circuit depth $L \in \{1, 2, 3, 4\}$ and qubit count $n_q \in \{6, 8\}$ (Digits, 25% noise) preserves the effect at every setting: the entangled $\rho_{\text{Spearman}}(d_{\text{eff}}, \text{acc})$ stays strongly negative and the best noise gain is always positive. At the larger 8-qubit width, a full collapse grid gives $\rho_{\text{Spearman}} = -0.82$, $R^2 = 0.92$, and noise gain $+0.19$—the picture is, if anything, cleaner at larger scale, not weaker.

C.5 Label-free selection of the noise level

Because $d_{\text{eff}}(p)$ is computable from training inputs alone, the operating point can be chosen without test labels. Two simple rules work. (i) A small *clean* validation split selects, per task, a noise level whose test accuracy is within 0.012–0.021 of the oracle best while evaluating only 2 of 9 candidates ($4.5\times$ less search) on Digits, Fashion, and BloodMNIST. (ii) A fully label-free target $d_{\text{eff}}(p) \approx \sqrt{n}$ attains 0.698 ± 0.026 vs. the oracle 0.754 ± 0.038 (5.6% gap, 5 seeds). The flat landscape near the optimum (Table 2) makes exact localization hard but the accuracy cost of mis-selection small.

C.6 d_{eff} is a principled, transferable diagnostic

A strength of the spectral account is that it rests on a well-established classical pillar—kernel generalization is governed by the spectrum (Prop. 3)—so d_{eff} is a *principled, transferable* diagnostic rather than quantum folklore. We verify that it behaves faithfully on the same data with classical feature maps, which licenses reading our quantum measurements through this lens. On the identical PCA features (25% label noise), classical regularizers trace out a d_{eff} range and accuracy tracks it: an RBF-kernel bandwidth sweep gives $\rho_{\text{Spearman}}(d_{\text{eff}}, \text{test}) = -0.50$, and a random-Fourier-feature dimension sweep recovers the complementary (capacity-limited) branch ($+0.68$).

Noise injection is genuine spectral regularization, not a weak hyperparameter knob. One might worry that injecting noise merely substitutes for tuning the SVM regularizer C. It does not. On the quantum kernel, sweeping C on the *noiseless* kernel tops out at 0.725 (it cannot change the kernel spectrum); injecting noise and then optimizing C reaches 0.744, and an *explicit* spectral shrinkage of the noiseless kernel, $K \mapsto U\,\mathrm{diag}(\lambda_i^{\alpha})\,U^{\top}$, traces out the same d_{eff}–accuracy frontier (0.73 at

Table 8: Trained quantum-vision models (overfitting regime): injecting noise during training contracts d_{eff}, shrinks the train−test gap, and improves test accuracy—the kernel-level mechanism survives end-to-end training.

Model	Noise	Test acc	d_{eff}	Train−test gap
QViT-like (re-uploading)	none	0.373	3.83	0.29
QViT-like (re-uploading)	injected	**0.420**	1.63	0.20
QCNN-like (convolution)	none	0.420	3.40	0.20
QCNN-like (convolution)	injected	**0.440**	1.57	0.13

contracted d_{eff}). Thus noise injection lands on the spectral-filtering frontier and adds a consistent gain that C-tuning alone cannot reach, confirming it acts on the kernel spectrum itself.

d_{eff} is not the uniquely best predictor—but it is the right handle. We compared d_{eff} against other scalar summaries on the cross-ansatz grid (5 seeds): the *absolute* rank correlation $|\rho_{\text{Spearman}}|$ with test accuracy (the sign is negative for the spectral quantities, positive for train accuracy) is d_{eff} : 0.78 ± 0.16, spectral entropy 0.77 ± 0.16, top-eigenvalue share 0.78 ± 0.15, and *train accuracy* 0.80 ± 0.14 (quad-fit R^2: $0.82, 0.77, 0.82, 0.87$). We report this honestly: d_{eff} ties the other spectral concentration measures and is marginally edged out by train accuracy. Its value is therefore not that it predicts best, but that it is the *label-free, theoretically grounded, and controllable* handle—it follows the capacity bound, is exact under depolarizing noise, needs no labels (unlike train accuracy, a post-hoc symptom), and is directly steered by the entanglement and noise knobs.

The contribution of our work is then sharp: entanglement and injected noise are a *new, hardware-native* pair of controls on a quantity whose generalization meaning is independently grounded, and it is exactly this grounding that turns the two reported quantum-vision phenomena into a single predictive principle. Whether the quantum feature geometry is itself advantageous is a separate question, orthogonal to and compatible with this spectral account; we make no quantum-advantage claim here.

C.7 Trained quantum-vision models: d_{eff} contraction persists

Finally we move beyond the fixed kernel and *train* the quantum feature map end-to-end with a linear head (PennyLane autograd), for two architectures: a QViT-like data-re-uploading map and a QCNN-like map with two layers of parametrized 2-qubit convolutions (architectures in App. A, Figs. 6 and 7). In a genuinely *overfitting* regime (60 examples, 30% label noise; train−test gap $+0.2$), injecting amplitude-damping noise *during training* reproduces the kernel-level phenomenon in both models (Table 8): it contracts the learned feature spectrum, shrinks the gap, and *improves* test accuracy. The mechanism therefore survives end-to-end training in both QViT- and QCNN-style models. (In an *underfitting* trained model the same contraction instead hurts, consistent with the sign flip of App. C.1, which is why we control the regime explicitly.)

C.8 The classifier: kernel ridge replicates the SVM findings

Our theory (Prop. 3–4) is stated for kernel ridge *regression*, while the main-text classifier is a precomputed-kernel SVM. We check that this is not load-bearing by repeating the two headline results—the cross-ansatz collapse of Sec. 4.4 and the sign-flip falsification of App. C.1—with a kernel ridge *classifier* (one-hot targets, $\hat{y} = \arg\max_c K_{\text{te}}(K_{\text{tr}} + \alpha I)^{-1} Y_{:,c}$) in place of the SVM, on the identical grids and label corruption. At a moderately regularized α=0.1 the collapse replicates closely (Spearman$_{\text{ent}} = -0.893$, global $R^2 = 0.905$, collapse R^2=0.768, versus $-0.896/0.921/0.772$ for the SVM), and the sign flip replicates as well (overfitting: $\rho = -0.96$, gain $+0.169$; underfitting: $\rho = +0.90$, best p=0, no benefit)—both consistent with the SVM numbers of Table 5. The picture is sensitive to α in the way theory predicts: too little regularization (α=0.01) undoes the shrinkage the noise knob is supposed to provide (collapse R^2 drops to 0.02), and too much (α=1) swamps the kernel signal with the ridge term itself (Spearman falls to -0.20). We read this as evidence that the collapse and the sign flip are properties of the kernel spectrum, not artifacts of the SVM's margin-based loss, provided the regularizer is in a reasonable range relative to the kernel scale.

D DISCUSSION AND LIMITATIONS

We use the quantum-kernel reading mainly to isolate the spectral mechanism cleanly, but the effect is not confined to it: it persists in *trained* QViT- and QCNN-like models (Sec. C.7), where training additionally adapts d_{eff} on its own and a full treatment with a training-dependent target $d^\star$ remains open. On scale, the exact-depolarizing analysis reaches 12 qubits (Sec. C.3), while the amplitude-damping experiments—which require full density-matrix simulation—are run at up to 8 qubits and inputs are reduced to 6–8 principal components; pushing the *noisy* simulations further would need tensor-network or sampling methods. The single-statistic story is strong on Digits/Fashion-MNIST and, once class-count and sample-size confounds are controlled, on the medical BloodMNIST as well (Sec. C.4); still, d_{eff} is a dominant but not exclusive determinant of generalization—alignment and spectral shape (Prop. 4) matter too, and the cross-ansatz collapse is empirical rather than a theorem. The sign-flip falsification (App. C.1) also changes qubit count and topology alongside label noise between its two regimes, so it demonstrates a qualitative sign correlate of d_{eff} rather than an experiment that isolates overfitting severity as the sole varying factor. Our capacity bound is the standard kernel-ridge *regression* excess-risk result (in d_γ), whereas we measure *classification* accuracy; it predicts the sign of the noise effect and its regime dependence, but the location and depth of the inverted-U optimum are not derived and remain empirical. Likewise the "more/more-uniform entanglement helps" half of the puzzle is only partly settled: we establish that entanglement is a *precondition* (entangled vs. product), but find no clean ordering *among* entangled topologies (Table 6). Finally, on real IBM Heron hardware the benefit requires both sufficient depth (negligible at $L{=}2$) and an overfitting regime: there hardware noise moves accuracy by $+0.037$—in the predicted direction, but with a paired-bootstrap 95% interval $[-0.013, +0.100]$ over the 80 test points, so the *contraction* is the resolved effect and the accuracy gain is not. A device evaluation with enough test points (and repeated runs) to resolve a few accuracy points is the natural next step, as is a larger task; in a non-overfitting deep run the same contraction hurts (the sign flip), and readout error and decoherence add signal loss beyond pure contraction.

Scope. This is a *mechanistic* study, not a performance claim. Exact simulation of noisy feature maps restricts us to 6–12 qubits, so images enter as a few principal components and the classifiers are kernel machines on a few hundred samples—a scale at which kernel spectra, d_{eff}, and the sign of its effect are fully observable. We check that the mechanism survives end-to-end *trained* QViT-/QCNN-style models (Sec. C.7) and a real device, but claim nothing about state-of-the-art vision accuracy or quantum advantage.

E EXPERIMENTAL DETAILS

Software and simulators. All quantum computations use PennyLane with the `default.mixed` density-matrix simulator (so amplitude-damping and depolarizing channels are exact, not sampled) for noisy kernels and `default.qubit` for noiseless/trained statevector runs. Kernels are assembled as a single dense product over flattened density matrices, $K = \mathrm{Re}(VV^\dagger)$ with $V_i = \mathrm{vec}\,\rho(x_i)$. Classifiers use scikit-learn `SVC` with a precomputed kernel. Real-hardware runs use `qiskit-ibm-runtime` against IBM Heron devices.

Feature map. Inputs are standardized PCA features scaled to $[-2.7, 2.7]$. The n_q-qubit, L-layer data-re-uploading circuit applies, in each layer ℓ, single-qubit rotations $R_Y(x_i + \theta_{\ell,i})R_Z(0.7\,x_i)$ on every qubit i followed by an entangling block of CNOTs; $\theta_{\ell,i}$ are fixed i.i.d. $\mathcal{N}(0, 0.5^2)$ (the kernel is data-driven, not trained). The entangling block is PRODUCT (none), CHAIN $i \to i{+}1$, RING $i \to i{+}1 \bmod n_q$, or ALL-TO-ALL. Amplitude damping of rate p is applied after each entangling block, with $n_q{=}6$ and $L{=}2$ in the main experiments (varied as described above).

Datasets and overfitting protocol. DIGITS ($8{\times}8$, sklearn), Fashion-MNIST and BloodMNIST (MedMNIST), each PCA-reduced to 6 (or 8) components. A controlled overfitting regime is induced by replacing a fraction ℓ of *training* labels with uniform random labels; the test set is always clean. Unless stated otherwise, $\ell{=}25\%$, $n_{\text{train}}{=}150$, $n_{\text{test}}{=}250$ (kernel experiments).

Seeds and error bars. Multi-seed results re-draw, per seed, the PCA fit, the train/test split, the label corruption mask, and the random circuit parameters θ; we report mean$\pm$std over 5 seeds. Spectral descriptors (d_{eff}, H_{spec}, $\mathrm{A}(K)$) are computed on the training block.

Per-experiment configuration. Table 9 lists the settings of each experiment.

Table 9: Configuration of the experiments.

Experiment	Setting	Section
Entanglement (Table 6)	4 ansatze, p=0, 5 seeds	4.1
Noise sweep (Table 1)	ring/all-to-all, $p \in [0, 0.45]$	4.2
Sweet spot (Table 2)	all-to-all, $\ell \in \{0, 10, 25, 40\}\%$, $p \leq 0.9$	4.3
Collapse / alignment (Table 4)	4×6 grid, $\mathrm{A}(K)$, 5 seeds	4.4
Datasets (Table 7)	Digits/Fashion/Blood, 4×5 grid	C.4
Depth/width	$L \in \{1-4\}$, $n_q \in \{6, 8\}$	C.4
Sign flip (Table 5)	overfit vs. 3-qubit product clean	C.1
Label-free selection	clean-val and $d_{\text{eff}} \approx \sqrt{n}$ rules	C.5
Spectral baselines	noise vs. C-sweep vs. λ^{α} shrinkage	C.6
Trained QViT-/QCNN-like	autograd, per-epoch d_{eff}, noise on/off	C.7
IBM Heron	feature $\langle Z_i \rangle$, hardware vs. ideal	4.5

Trained models. The trained ablations optimize the circuit parameters and a linear head end-to-end with PennyLane autograd (Adam, lr 0.05); injected noise during training uses `default.mixed`. The QCNN-style model uses two brick-pattern layers of parametrized 2-qubit convolutions (pairs $(0, 1), (2, 3), (4, 5)$ then $(1, 2), (3, 4), (5, 0)$), *without* pooling, and reads out $\langle Z_i \rangle$ on all six qubits—pooling to two survivors leaves $d_{\text{eff}} \in [1, 2]$ with no range, so the full readout is what makes the spectral diagnostic meaningful.

Real-hardware run. We run on an IBM Heron processor (`ibm_kawasaki`) using Qiskit Runtime. The task is binary DIGITS (0 vs. 1) with 30% of the training labels flipped; n_{train}=50 and n_{test}=80, so 130 circuits are executed. The feature circuit is the 6-qubit all-to-all map at depth L=3, transpiled at optimization level 1 to the native basis; for each input we estimate 21 observables—the six $\langle Z_i \rangle$ and all fifteen pairwise $\langle Z_i Z_j \rangle$—with 4096 shots via the EstimatorV2 primitive. The kernel is the Gram matrix of these 21-dimensional feature vectors, classified by the same precomputed-kernel SVM (C=10); hardware feature vectors and their induced d_{eff} are compared against ideal (statevector) values computed from the identical circuits. Uncertainty on the accuracy difference is a paired bootstrap over the 80 test points (App. 4.5 reports the interval).

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.