

Classical SU(2) Models Match or Exceed Shallow Variational Quantum Circuits on Classical Vision Benchmarks

Christopher Fulton

United States Air Force Test Pilot School

Irene Tsapara

National University

Lawrence Fulton

`fultonl@bc.edu`

Boston College

Research Article

Keywords: Quaternion neural networks, Variational quantum circuits, SU(2) geometry, Quantum machine learning

Posted Date: March 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8501568/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Quantum Machine Intelligence on August 31st, 2026. See the published version at <https://doi.org/10.1007/s42484-026-00430-4>.

Classical $SU(2)$ Models Match or Exceed Shallow Variational Quantum Circuits on Classical Vision Benchmarks

Christopher P. Fulton¹, Irene Tsapara², Lawrence V. Fulton^{3*†}

¹United States Air Force Test Pilot School, Edwards Air Force Base, California, 93524, USA .

²Department of Engineering, Data, and Computer Science; National University, San Diego, California, 92123, USA .

^{3*}Applied Analytics, Boston College, 140 Commonwealth Ave, Chestnut Hill, MA, 02467, USA .

*Corresponding author(s). E-mail(s): fultonl@bc.edu;

†These authors contributed equally to this work.

Abstract

Quaternion-valued neural networks and shallow variational quantum circuits (VQCs) both derive their local transformations from the rotation group $SU(2)$, yet their comparative performance as learning architectures on classical supervised tasks has not been systematically examined. We present a controlled comparison in which real-valued, quaternion-valued, and quantum classifiers operate on identical frozen feature representations across MNIST, FashionMNIST, and CIFAR-10, isolating performance differences to the geometric inductive biases of the classification heads. For CIFAR-10, we evaluate two feature regimes—a learned 16-dimensional bottleneck and frozen ImageNet-pretrained ResNet18 features (512-dimensional)—to assess whether observed relationships reflect fundamental architectural properties or artifacts of representation capacity. Quaternion classifiers match real-valued baselines on MNIST (93.64% vs. 93.54%) and FashionMNIST (84.47% vs. 84.60%), while shallow product-state VQCs achieve only 87.52% (MNIST) and 82.03% (FashionMNIST), despite substantially higher computational cost. On CIFAR-10, quaternion networks retain 94.3% of real-valued performance under the learned bottleneck and 95.9% under ResNet18 features, demonstrating robustness across a 32-fold increase in feature dimensionality. On CIFAR-10, product-state quantum circuits underperform quaternion classifiers by 2.8–3.5 percentage points across both feature regimes. Notably, entanglement reverses from providing modest gains (approximately +0.6 percentage points) on simple grayscale benchmarks to degrading performance by 9.3 percentage points relative to the product-state circuit when combined with high-quality pretrained features, indicating that additional quantum resources can adversely affect learning in shallow, measurement-limited regimes. These results demonstrate that classical $SU(2)$ models realized through quaternion networks provide computationally efficient classification heads that match real-valued baselines while substantially outperforming shallow variational quantum circuits on classical vision tasks.

Keywords: Quaternion neural networks, Variational quantum circuits, $SU(2)$ geometry, Quantum machine learning

Introduction

Quaternion-valued neural networks (QNNs) have emerged as compact and expressive architectures for modeling structured correlations in visual, auditory, and spatial signals [1–3]. Hypercomplex representations based on the complex numbers $\mathbb{C}$ and the quaternions $\mathbb{H}$ were originally motivated by signal-processing applications [4], where algebraic structure can encode phase relationships, spatial couplings, and directional information more efficiently than unconstrained real-valued models. Among these, unit quaternions are particularly notable: they parameterize three-dimensional rotations, form a noncommutative division algebra, and are isomorphic to the Lie group $SU(2)$ [5]. Unit quaternions under multiplication form the Lie group $S^3 \cong SU(2)$, providing a smooth four-parameter representation with constrained degrees of freedom amenable to gradient-based optimization. These properties have enabled QNNs to achieve competitive or superior performance under fixed parameter budgets in tasks involving spatial correlation, gesture recognition, and audio source separation [2, 3].

In parallel, quantum machine learning has renewed attention to models whose trainable components are intrinsically parameterized by Lie groups. Variational Quantum Circuits (VQCs) are constructed from tunable unitary operations acting on qubits, with single-qubit gates drawn from $SU(2)$ and commonly expressed as parametrized rotations $R_X(\theta)$, $R_Y(\theta)$, and $R_Z(\theta)$. These local $SU(2)$ transformations are composed via tensor products and entangling operations to form expressive quantum hypothesis classes [6, 7]. Theoretical analyses emphasize that VQC expressivity depends critically on both the richness of local rotation dynamics and the introduction of multi-qubit entanglement, which enables correlations inaccessible to purely classical systems [8]. Recent advances in hybrid optimization, adjoint differentiation, and GPU-accelerated simulators, including PennyLane Lightning and NVIDIA cuQuantum, have made systematic empirical exploration of these regimes feasible at practical scales [9, 10].

Variational quantum classifiers (VQCs) represent one of the most widely studied near-term quantum learning paradigms. In these models, classical data are embedded into quantum states via parameterized circuits, followed by measurement and classical post-processing. A central motivation for VQCs is the prospect of quantum-enhanced feature spaces, in which nonlinear embeddings induced by quantum evolution can render otherwise inseparable classical data linearly separable in Hilbert space [11]. However, end-to-end VQC pipelines conflate two distinct sources of potential advantage: quantum feature extraction and quantum classification dynamics. Moreover, unlike classical neural networks, which preserve continuous intermediate representations throughout the forward pass, VQCs must ultimately compress a high-dimensional quantum state into a small set of classical expectation values at measurement, introducing an intrinsic information bottleneck that is independent of circuit depth or parameterization.

Viewed through a geometric lens, both variational quantum circuits (VQCs) and quaternion neural networks (QNNs) derive their expressive power from trainable transformations acting on $SU(2)$. This shared structure motivates an experimental design that isolates $SU(2)$ -structured classification dynamics by fixing upstream feature representations and comparing head architectures under strictly matched conditions, ensuring that observed performance differences are attributable to the head geometry rather than to quantum or classical preprocessing. To this end, we construct a controlled evaluation based on three widely used image benchmarks—MNIST, FashionMNIST,

and CIFAR-10—chosen to span increasing levels of visual and statistical complexity. We adopt a tiered evaluation spanning increasing representational complexity: for MNIST and FashionMNIST, all models operate on a 16-dimensional learned bottleneck, while for CIFAR-10 we consider both this learned bottleneck and a frozen ImageNet-pretrained ResNet18 backbone producing 512-dimensional embeddings. This dual-regime design allows us to assess whether observed performance trends reflect fundamental architectural properties or instead arise from representation capacity, and whether local $SU(2)$ rotation structure alone can support competitive classification across feature qualities. More broadly, this framing aligns with recent work emphasizing that claims of quantum advantage must be evaluated relative to problem structure, data distributions, and strong classical baselines, rather than inferred from model class alone [12].

Despite these parallel developments, the relationship between quaternion-valued neural networks and variational quantum classifiers remains largely unexplored, even though both manipulate trainable elements drawn from isomorphic copies of $SU(2)$. Quaternion layers implement normalized Hamilton products with unit quaternions, while VQCs implement unitary evolution of quantum states using single-qubit $SU(2)$ gates. This shared local geometry does not imply equivalence of the underlying computational models, which differ fundamentally in their state spaces, measurement processes, and access to entanglement. This raises a foundational question: does classical $SU(2)$ geometry alone provide competitive performance for learning tasks addressable by product-state quantum circuits, and how does this relationship vary across feature representations of differing quality? More concretely, can classical quaternion layers match or exceed the performance of variational quantum classifiers operating on identical features, and under what conditions does entanglement introduce behavior that cannot be replicated by classical $SU(2)$ implementations?

Addressing this question clarifies the role of local $SU(2)$ geometry in learning and sharpens claims of quantum advantage that rest on rotation dynamics rather than nonclassical correlations. To this end, we adopt a controlled experimental methodology in which all models operate on identical, frozen classical representations across multiple feature regimes, ensuring that observed performance differences arise from the representational geometry and computational mechanisms of the $SU(2)$ -derived heads themselves. Using this framework, we provide the first systematic empirical comparison of quaternion-valued classifiers and variational quantum classifiers under matched $SU(2)$ constraints across MNIST, FashionMNIST, and CIFAR-10, including evaluations under both learned low-dimensional bottlenecks and high-quality pretrained feature representations.

Comparison equitability

A persistent challenge in empirical quantum machine learning is disentangling the effects of feature extraction from those of classification dynamics within end-to-end pipelines. In this study, we intentionally decouple these components by fixing shared classical feature representations and varying only the classification head. For MNIST and FashionMNIST, all models operate on a 16-dimensional learned bottleneck; for CIFAR-10, we evaluate under both this bottleneck and a frozen ImageNet-pretrained ResNet18 backbone producing 512-dimensional embeddings. This design does not aim to assess quantum feature maps or end-to-end hybrid architectures; rather, it isolates a narrower but well-defined question: whether local $SU(2)$ geometry alone, absent quantum feature extraction and deep entanglement, provides a learning advantage at the classification stage, and whether this relationship is robust across feature qualities. By holding input representations fixed within each regime while varying representation strength across regimes, we ensure that observed performance differences reflect the

inductive bias and representational geometry of the heads themselves, not unequal preprocessing capacity or co-adapted feature learning.

Unit quaternions and their correspondence to $SU(2)$

A quaternion $q \in \mathbb{H}$ is written as

$$q = w + xi + yj + zk, \quad w, x, y, z \in \mathbb{R},$$

with the imaginary units obeying Hamilton's multiplication rules

$$i^2 = j^2 = k^2 = ijk = -1.$$

From these, the cyclic products follow:

$$ij = k, \quad jk = i, \quad ki = j.$$

The reversed products acquire a minus sign, which follows directly from $ijk = -1$ and associativity:

$$ji = -k, \quad kj = -i, \quad ik = -j.$$

The general Hamilton product (quaternion multiplication) of two quaternions $q_1 = w_1 + x_1i + y_1j + z_1k$ and $q_2 = w_2 + x_2i + y_2j + z_2k$ is computed by expanding the product and applying these multiplication rules. The result is associative but generally noncommutative, $q_1 \otimes q_2 \neq q_2 \otimes q_1$, reflecting the underlying nonabelian structure inherited from three-dimensional rotations and preserved in neural network layers employing quaternion-valued weights.

The quaternion conjugate is $q^* = w - xi - yj - zk$, and the norm satisfies

$$\|q\|^2 = qq^* = w^2 + x^2 + y^2 + z^2.$$

A *unit quaternion* obeys $\|q\| = 1$, so the set of unit quaternions forms the three-sphere,

$$\{q \in \mathbb{H} \mid \|q\| = 1\} \cong S^3.$$

The same manifold appears in quantum computation through the special unitary group

$$SU(2) = \{U \in \mathbb{C}^{2 \times 2} \mid U^\dagger U = I, \det(U) = 1\},$$

whose elements act as single-qubit rotations. The Lie group of unit quaternions is isomorphic to $SU(2)$. One convenient explicit realization is given by [5]

$$\Phi : \mathbb{H} \longrightarrow SU(2), \quad w + xi + yj + zk \longmapsto \begin{pmatrix} w + iz & y + ix \\ -y + ix & w - iz \end{pmatrix}.$$

For any unit quaternion q , the image $U = \Phi(q)$ satisfies $U^\dagger U = I$ and $\det(U) = 1$, and multiplication is preserved:

$$\Phi(q_1 q_2) = \Phi(q_1) \Phi(q_2).$$

The kernel of Φ is $\{\pm 1\}$, so a unit quaternion q and its antipode $-q$ map to unitary matrices that differ only by an overall sign. In quantum mechanics, such matrices are physically indistinguishable, since global phase has no observable effect on quantum states. This establishes the identification $S^3 \cong SU(2)$, with $SU(2)$ acting as a double cover of $SO(3)$, such that each physical rotation corresponds to a pair of antipodal unit quaternions inducing the same quantum operation.

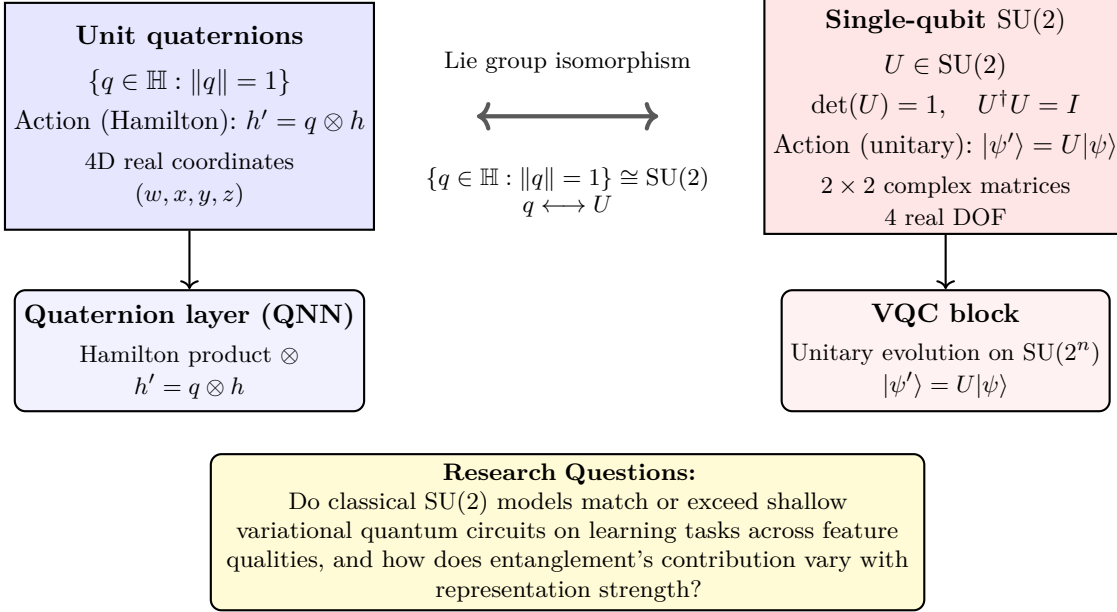


Fig. 1. Schematic correspondence between unit quaternions and single-qubit $SU(2)$ operations. The Lie group of unit quaternions $q \in \mathbb{H} : |q| = 1$ is isomorphic to $SU(2)$, and both are diffeomorphic to the 3-sphere S^3 , reflecting a shared rotation geometry. Quaternion layers act on classical feature vectors via the Hamilton product ($h' = q \otimes h$), which preserves the full four-dimensional quaternion representation, whereas variational quantum circuits apply unitary evolution ($|\psi'\rangle = U|\psi\rangle$) to quantum states followed by measurement. Although the trainable parameters inhabit the same $SU(2)$ geometry, quantum circuits uniquely permit entanglement and nonclassical state evolution; however, learning proceeds only through a low-dimensional set of measurement expectations, which can impose an information bottleneck absent in quaternion layers.

Thus, quaternion layers in neural networks and single-qubit blocks in variational quantum circuits optimize over the *same* underlying rotation group (Figure 1), implying a shared parameter geometry rather than functional equivalence. What distinguishes the quantum setting is not local $SU(2)$ structure, but the availability of measurement and multi-qubit entanglement, which introduces representational effects that no quaternion multiplication can replicate.

Contributions

This work establishes a controlled empirical connection between quaternion-valued neural networks and variational quantum classifiers under matched $SU(2)$ constraints. All models operate on identical, frozen classical feature representations across MNIST, FashionMNIST, and CIFAR-10, allowing performance differences to be attributed solely to the geometry and computational structure of the classification heads rather than to feature extraction or representation quality.

Our contributions are fourfold:

1. **Quaternion–quantum correspondence in the product-state regime:** We present the first controlled empirical comparison demonstrating that quaternion networks match or exceed the performance of product-state variational quantum classifiers on classical image tasks under shared $SU(2)$ constraints. Across both minimal 16-dimensional learned bottlenecks and rich 512-dimensional pretrained embeddings, quaternion models consistently retain 94–96% of real-valued MLP

performance, indicating a stable and robust relationship between real and quaternion architectures across a 32-fold range in feature dimensionality.

2. **Limits of local SU(2) dynamics without entanglement:** Under matched conditions, we observe no performance advantage attributable to local SU(2) rotations alone in product-state quantum classifiers on classical image data. Product-state quantum circuits underperform quaternion implementations by 2.4–6.1 percentage points under learned bottlenecks and by 3.5 percentage points under high-quality pretrained features, indicating that measurement-induced information compression and optimization constraints dominate in shallow quantum architectures absent entanglement.
3. **Entanglement benefit reversal under rich feature representations:** By systematically contrasting entangled and non-entangled quantum classifiers across feature regimes, we identify a clear empirical boundary: entanglement yields modest gains (0.6–0.7 percentage points) on MNIST and FashionMNIST with learned bottlenecks, provides no benefit on CIFAR-10 under the same bottleneck, and degrades performance by 9.3 percentage points on CIFAR-10 when paired with high-quality pretrained features. This reversal demonstrates that shallow entanglement can interact adversely with high-dimensional classical representations, delimiting a regime in which quantum resources hinder rather than enhance performance.
4. **Quaternions as geometry-faithful classical baselines for quantum ML:** We show that quaternion networks constitute a stronger and more appropriate classical baseline than standard real-valued MLPs for evaluating quantum machine learning proposals on classical data. The consistent quaternion–real performance relationship across datasets and feature qualities implies that claims of quantum advantage must exceed not only real-valued baselines but also parameter-efficient classical SU(2) implementations that faithfully capture the local geometry shared with shallow quantum circuits.

Together, these results delineate when quaternion layers provide faithful classical substitutes for variational quantum classifiers on classical vision tasks and identify an empirical boundary where shallow entanglement becomes detrimental under rich feature representations. The stable qualitative ordering between real-valued, quaternion, and quantum models across datasets and feature regimes indicates that these effects reflect fundamental architectural properties rather than dataset-specific artifacts. More broadly, our framework offers a systematic methodology for disentangling geometric structure from genuinely quantum contributions in near-term machine learning by enforcing shared representations across models, revealing interactions between quantum resources and classical feature quality that end-to-end pipelines typically obscure.

Materials and Methods

This section describes the datasets, preprocessing steps, model architectures, and training protocols used to evaluate the role of representational geometry in classical and quantum-inspired classifiers. All experiments are designed to enable controlled, head-to-head comparisons across model classes by holding data exposure, feature representations, optimization settings, and random seed initialization fixed wherever possible. Differences in performance can therefore be attributed to the inductive bias and parameterization of the classification heads rather than to confounding factors such as dataset imbalance, feature quality, or training protocol variation. We follow a fixed experimental pipeline (Fig. 2): datasets and preprocessing, frozen feature representations, classification heads, optimization and training protocol, and finally diagnostic evaluation.

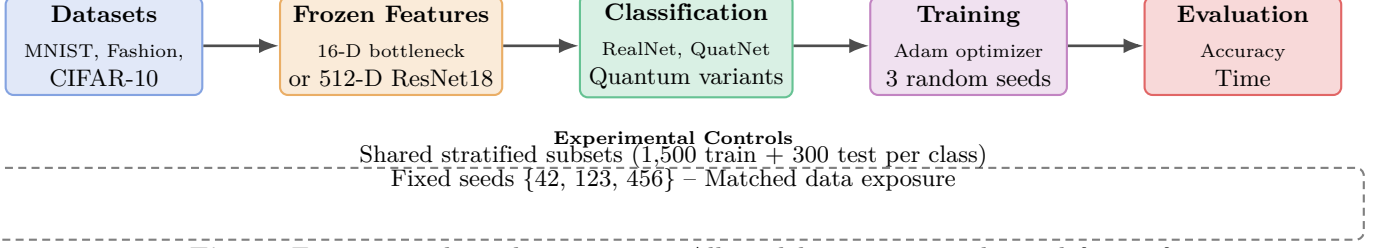


Fig. 2. Experimental pipeline overview. All models operate on identical frozen feature representations across three datasets and two feature regimes (16-dimensional learned bottleneck and 512-dimensional pretrained ResNet18). Strict controls ensure that observed performance differences arise solely from the classification head architecture rather than from data exposure, feature quality, or training protocol variations.

Datasets and preprocessing

To study how representational geometry and model architecture interact with increasing visual complexity, we conduct experiments on three standard image classification benchmarks: MNIST handwritten digits [13], FashionMNIST clothing items [14], and CIFAR-10 natural images [15]. These datasets are selected to provide a controlled progression from low-dimensional grayscale structure (MNIST), to visually richer but still grayscale imagery with greater intra-class variability (FashionMNIST), to full-color natural images with substantially higher semantic and representational complexity (CIFAR-10). This progression allows us to isolate the effects of representational geometry under increasingly demanding visual conditions while maintaining comparability with prior work.

MNIST and FashionMNIST consist of 28×28 grayscale images drawn from ten balanced classes. MNIST contains handwritten digits with relatively simple stroke-based structure, while FashionMNIST replaces digits with clothing items, introducing greater texture and shape variability while preserving identical dimensionality. CIFAR-10 consists of 32×32 RGB images from ten object categories, exhibiting substantial background variation and semantic diversity.

All datasets are normalized using mean and standard deviation statistics appropriate to the feature-extraction regime. MNIST and FashionMNIST use empirical scalar statistics computed from their respective training sets. CIFAR-10 uses per-channel normalization matched to the feature extractor: empirical per-channel statistics computed from the CIFAR-10 training subset in the learned-bottleneck regime, and standard ImageNet per-channel statistics in the pretrained ResNet18 regime. Normalization is applied as

$$\begin{aligned}
 \text{MNIST: } x &\mapsto \frac{x - 0.1307}{0.3081}, \\
 \text{FashionMNIST: } x &\mapsto \frac{x - 0.2860}{0.3530}, \\
 \text{CIFAR-10 (RGB): } x_c &\mapsto \frac{x_c - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\},
 \end{aligned}$$

where (μ_c, σ_c) denotes the per-channel mean and standard deviation used in the given regime. In the pretrained ResNet18 regime we use ImageNet statistics $(0.485, 0.456, 0.406)$ and $(0.229, 0.224, 0.225)$, while in the learned-bottleneck regime we compute (μ_c, σ_c) empirically from the CIFAR-10 training subset.

To control sample complexity and enforce strict class balance, we construct stratified subsets by randomly selecting 1,500 training and 300 test examples per class (seed 42), yielding 15,000 training and 3,000 test images per dataset. Real-valued and quaternion models are trained with batch sizes of 128 and evaluated with batch sizes of 256. Quantum models use smaller batch sizes (32 for training, 64 for evaluation) to accommodate the computational cost of circuit simulation. All methods are trained on the same fixed stratified subset, and data exposure is matched at the level of epochs over this subset rather than the number of optimizer steps. We adopt epoch-matching because it equalizes the number of times each training example is seen, which is the most direct control for sample exposure in a fixed-subset setting; step-matching would instead equalize optimizer updates while inducing unequal data passes.

For CIFAR-10, experiments are conducted under two feature-extraction regimes to disentangle architectural effects from representation-capacity effects. In the **learned bottleneck regime**, all models operate on features from a shared frozen 16-dimensional bottleneck trained as part of the real-valued baseline, matching the protocol used for MNIST and FashionMNIST. In the **pretrained CNN regime**, this bottleneck is replaced with a frozen ImageNet-pretrained ResNet18 backbone (torchvision, default weights), yielding 512-dimensional embeddings extracted from the final average pooling layer. Both regimes use identical stratified subsets, hyperparameters, and random seed control, ensuring that observed differences arise from representation strength rather than sample complexity or training protocol variations.

All architectures are trained independently on each dataset and feature regime using identical hyperparameters and comprehensive random seed control. For each experimental seed {42, 123, 456}, we initialize all random number generators (Python, NumPy, PyTorch, CUDA, and PennyLane) and enable PyTorch deterministic mode. This protocol ensures that cross-dataset and cross-regime comparisons reflect differences in visual complexity and representation quality rather than stochastic effects or experimental design choices.

Frozen feature extraction

To isolate the contribution of the classification head, we deliberately control feature extraction rather than adopting end-to-end convolutional architectures. Convolutional networks introduce strong domain-specific inductive biases, such as translation equivariance and local receptive fields, which would confound attribution by obscuring whether observed performance differences arise from the SU(2) geometry of the classification head or from architectural priors embedded in the feature extractor. By freezing feature representations, we ensure that all models operate on identical inputs within each regime, so that performance differences reflect the representational geometry and inductive bias of the classification head rather than unequal or co-adapted feature learning.

To evaluate whether conclusions drawn under this controlled setting depend on representation capacity, we consider two feature-extraction regimes. The first employs a deliberately low-capacity learned bottleneck that maximally isolates head-level effects. The second replaces this bottleneck with a substantially richer pretrained convolutional representation, providing a controlled excursion in representation strength that tests whether observed relationships persist under more expressive classical features.

In the learned-bottleneck regime, used for MNIST, FashionMNIST, and CIFAR-10, inputs are mapped through a shared feature extractor

$$f_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}^{16}, \quad f_{\theta}(x) = \tanh(Wx + b),$$

where $d = 784$ for 28×28 grayscale images and $d = 3072$ for 32×32 RGB images. The extractor consists of a single linear layer followed by a tanh nonlinearity. The bounded output of tanh ensures scale stability for both quaternion normalization and quantum data encoding, where features are interpreted directly as rotation angles; unbounded activations would otherwise induce excessively large rotations or require artificial clipping. This feature extractor is trained end-to-end as part of the real-valued baseline using seed 42 and then frozen. The resulting 16-dimensional feature vector $h = f_\theta(x)$ is treated as fixed input for all quaternion and quantum classification heads, while the three-seed protocol applies only to the head parameters.

To assess whether performance relationships observed under this restrictive bottleneck are artifacts of limited representation capacity, we conduct an auxiliary control experiment on CIFAR-10 using a frozen ImageNet-pretrained ResNet18 backbone (torchvision, default weights). Let $\phi(\cdot)$ denote the penultimate-layer embedding extracted from the final average pooling layer,

$$\phi : \mathbb{R}^{32 \times 32 \times 3} \rightarrow \mathbb{R}^{512}.$$

Only the parameters of the classification head operating on $\phi(x)$ are optimized; the ResNet18 backbone remains fixed across all seeds and model variants. This 512-dimensional representation constitutes a 32-fold increase in feature dimensionality relative to the learned bottleneck and provides a controlled test of whether quaternion–real performance relationships and quantum degradation patterns persist under substantially richer classical features.

Quantum models are retained in this pretrained-CNN regime despite increased computational cost. The same four-qubit, depth-3 variational circuit architecture used in the learned-bottleneck experiments is employed, with only the classical feature selector scaled to map 512-dimensional inputs to four encoding angles. Although training times increase from approximately four hours to eighteen hours per seed, GPU-accelerated simulation renders this regime computationally feasible and enables systematic evaluation of whether measurement-induced degradation and entanglement effects depend on representation quality.

Across both regimes, the use of frozen feature extraction ensures that observed performance differences arise from the inductive bias and representational geometry of the classification heads rather than from unequal preprocessing capacity or end-to-end co-adaptation. The dual-regime design further distinguishes head-level architectural effects, isolated by the bottleneck, from representation-quality effects revealed by the pretrained CNN control.

Real-valued baseline

The real-valued classifier (RealNet) serves as the reference architecture against which quaternion and quantum classifiers are evaluated. Its structure is adjusted across feature regimes to accommodate differing input dimensionalities while preserving comparable nonlinear depth and activation structure.

16-dimensional bottleneck regime (MNIST, FashionMNIST, CIFAR-10)

In the learned-bottleneck regime, RealNet is implemented as a two-layer multilayer perceptron operating on frozen 16-dimensional features:

$$h \in \mathbb{R}^{16} \xrightarrow{W_1, \text{ReLU}} \mathbb{R}^{64} \xrightarrow{W_2} \mathbb{R}^{10}.$$

The hidden layer width of 64 units provides sufficient nonlinear capacity for classification while avoiding additional architectural depth or inductive bias beyond that already present in the bottleneck representation. This configuration yields 1,738 trainable parameters in the classification head, excluding the frozen 12,560-parameter feature extractor.

512-dimensional CNN regime (CIFAR-10 only)

For the pretrained ResNet18 control experiment, RealNet operates on 512-dimensional frozen feature embeddings and employs a wider hidden layer to accommodate the increased input dimensionality while retaining the same two-layer structure:

$$h \in \mathbb{R}^{512} \xrightarrow{W_1, \text{ReLU}} \mathbb{R}^{128} \xrightarrow{W_2} \mathbb{R}^{10}.$$

This architecture contains 66,954 trainable parameters in the classification head, excluding the frozen 11.2M-parameter ResNet18 backbone. The increased hidden width reflects the higher-dimensional input while maintaining architectural consistency with the bottleneck regime.

Training protocol

Across all regimes, RealNet is trained using the Adam optimizer with learning rate 10^{-3} and cross-entropy loss, with early stopping based on held-out validation accuracy (maximum 100 epochs, patience 10). In the learned-bottleneck regime, the feature extractor f_θ is trained once using seed 42 and then held fixed; the three-seed protocol applies to the classification head only. In the pretrained CNN regime, the ResNet18 backbone remains frozen throughout training, and only the classification head parameters are optimized.

Together, these design choices establish RealNet as a competitive yet minimally structured baseline across feature regimes, providing a consistent reference point for evaluating quaternion and quantum classifiers without introducing additional architectural complexity that could obscure geometric comparisons. The complete architecture for all model variants, including the shared frozen preprocessor and the four classification head designs, is illustrated in Figure 4.

Quaternion classifier

The quaternion classifier (QuatNet) replaces real-valued operations with quaternion-valued layers while preserving the same frozen feature representations across regimes. Architectural variations reflect input dimensionality while maintaining identical quaternion algebra and geometric constraints.

16-dimensional bottleneck regime (MNIST, FashionMNIST, CIFAR-10)

In the primary experimental setting, QuatNet operates on the shared frozen 16-dimensional bottleneck. The real-valued feature vector is reinterpreted as four quaternions,

$$h \in \mathbb{R}^{16} \longrightarrow \mathbf{q}^{(0)} \in \mathbb{H}^4,$$

by reshaping h into quaternion components (w, x, y, z) . The classification head consists of a $4 \rightarrow 16$ quaternion linear layer followed by a $16 \rightarrow 10$ quaternion linear layer. Class logits are obtained by extracting the real component of each output quaternion. This architecture contains 1,000 trainable parameters in the quaternion head, representing a 1.7-fold reduction relative to the 1,738 parameters in the real-valued baseline.

512-dimensional CNN regime (CIFAR-10 only)

For the pretrained ResNet18 control experiment, QuatNet operates on 512-dimensional frozen feature embeddings. The input features are reshaped into 128 quaternions,

$$h \in \mathbb{R}^{512} \longrightarrow \mathbf{q}^{(0)} \in \mathbb{H}^{128},$$

and processed through a $128 \rightarrow 32$ quaternion linear layer followed by a $32 \rightarrow 10$ quaternion linear layer. This configuration yields 17,832 trainable parameters, corresponding to a 3.7-fold reduction compared to the 66,954 parameters of the real-valued baseline operating on the same features.

Quaternion operations

Each quaternion layer applies Hamilton products between learned unit-quaternion weights and input quaternions [2, 3],

$$\mathbf{q}_j^{(l+1)} = b_j + \sum_i W_{ji} \otimes \mathbf{q}_i^{(l)}, \quad W_{ji} \leftarrow \frac{W_{ji}}{\|W_{ji}\| + \varepsilon}.$$

Weights are initialized as unit quaternions and renormalized after each parameter update. Because the weighted sum and bias addition do not preserve unit norm, layer outputs are renormalized prior to application of the tanh nonlinearity.

Training protocol

Across all regimes, the quaternion head is trained using the Adam optimizer with learning rate 10^{-3} and cross-entropy loss, with early stopping based on held-out validation accuracy (maximum 100 epochs, patience 10). The shared feature representations (f_θ in the bottleneck regime and the ResNet18 backbone in the CNN regime) remain frozen throughout training.

Implementation note: quaternion optimization

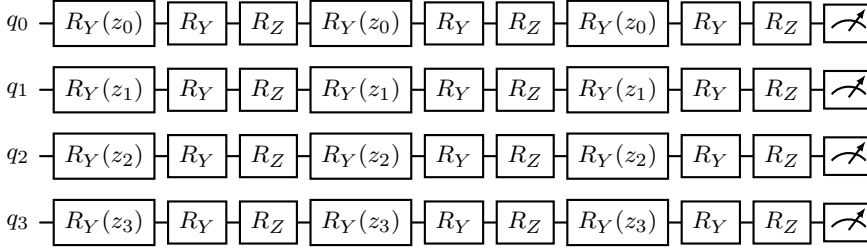
Quaternion layers are implemented using standard differentiable operations within PyTorch’s autograd framework. Although the unit-quaternion constraint defines a manifold structure (the 3-sphere S^3), we employ an extrinsic optimize-then-project strategy [16] in which parameters are updated in Euclidean space and subsequently normalized to enforce unit norm. This approach avoids explicit manifold-aware gradient computations while ensuring that all weights remain valid unit quaternions. As such, quaternion optimization here constitutes a constraint-preserving approximation to Riemannian optimization on S^3 , analogous in spirit to the Euclidean parameterizations commonly used for variational quantum circuits despite their intrinsically curved geometry.

The computational overhead of this projection is negligible: normalization incurs only $\mathcal{O}(p)$ operations per update for p parameters and requires no additional forward or backward passes. Standard Euclidean optimizers therefore suffice for quaternion networks, in contrast to quantum circuits where geometry-aware optimization typically incurs $\mathcal{O}(p^2)$ circuit evaluations per iteration.

Quantum classifiers

The quantum classifier (QuantumNet) is implemented as a variational quantum circuit (VQC) with circuit depth three, corresponding to three complete data re-uploading

(a) Quantum-NoEnt: Product-state circuit (no entanglement)



(b) Quantum-Ent: Circuit with CNOT ring entanglement

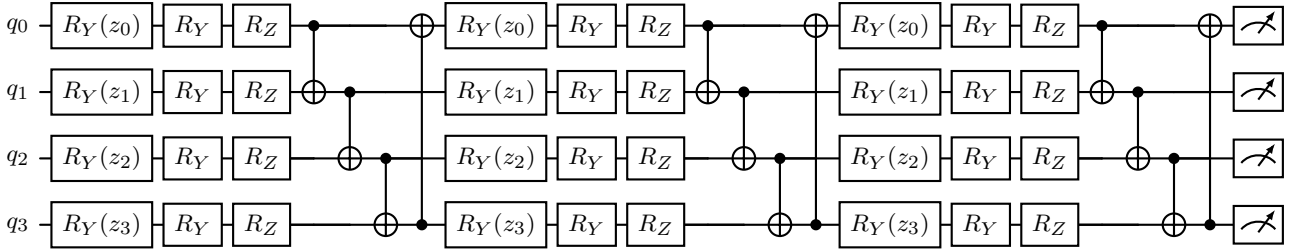


Fig. 3. Variational quantum circuit (VQC) architectures used for classification, shown here in the four-qubit configuration employed for MNIST and FashionMNIST. In each of the three data re-uploading layers, features are encoded via single-qubit rotations $R_Y(z_i)$ on qubit q_i , followed by trainable local rotations $R_Y(\theta_{i,Y}^{(\ell)})$ and $R_Z(\theta_{i,Z}^{(\ell)})$, where $\ell \in \{1, 2, 3\}$ indexes the re-uploading layer (individual parameter indices are omitted for visual clarity). Panel (a) shows the product-state variant (Quantum-NoEnt), which omits entangling operations. Panel (b) shows the entangling variant (Quantum-Ent), which inserts a CNOT ring after the local rotations in each layer, coupling $q_0 \rightarrow q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_0$. Measurements are performed only at the end of the circuit. In the four-qubit configuration, we measure Pauli-Z observables on all qubits together with two-qubit correlations Z_0Z_1 and Z_2Z_3 , yielding the six-dimensional expectation vector $(\langle Z_0 \rangle, \langle Z_1 \rangle, \langle Z_2 \rangle, \langle Z_3 \rangle, \langle Z_0Z_1 \rangle, \langle Z_2Z_3 \rangle)$ passed to a classical linear classifier. In the ResNet18 512-dimensional regime, the same depth-3 re-uploading architecture is used with eight qubits; in that case, we measure eight single-qubit Pauli-Z expectation values together with four two-qubit ZZ correlations, yielding a twelve-dimensional expectation vector prior to the final linear classifier.

layers (Figure 3). In the 16-dimensional bottleneck regime, the circuit operates on four qubits; for the 512-dimensional ResNet18 feature regime, the circuit is scaled to eight qubits to provide sufficient representational capacity for the richer input. Across regimes and entanglement variants, the circuit depth, re-uploading structure, and measurement dimensionality are held fixed, while the number of qubits and the associated classical feature selector are adjusted to match the input dimensionality. Measurement dimensionality is intentionally constrained in order to isolate the effect of SU(2) parameter geometry; expanding the measurement space would effectively introduce an additional classical post-processing layer, confounding interpretation of the variational head itself. Deeper circuits were not considered here, as increased depth introduces additional confounds (e.g., barren plateaus, noise and simulator scaling, and optimizer instability) that would obscure controlled comparisons of head geometry under fixed feature representations.

16-dimensional bottleneck regime (MNIST, FashionMNIST, CIFAR-10)

In the learned-bottleneck regime, the frozen 16-dimensional classical feature vector is mapped to four real-valued encoding angles via a trainable linear feature selector,

$$g_\phi : \mathbb{R}^{16} \rightarrow \mathbb{R}^4, \quad z = \tanh(W_\phi h + b_\phi),$$

where $W_\phi \in \mathbb{R}^{4 \times 16}$ and $b_\phi \in \mathbb{R}^4$. The parameters of g_ϕ are optimized jointly with the quantum circuit parameters, while the shared preprocessor f_θ remains frozen.

512-dimensional CNN regime (CIFAR-10 only)

For the pretrained ResNet18 control experiment, the circuit is scaled to eight qubits to accommodate richer representations, with the feature selector mapping 512-dimensional features to eight encoding angles,

$$g_\phi : \mathbb{R}^{512} \rightarrow \mathbb{R}^8, \quad z = \tanh(W_\phi h + b_\phi),$$

with $W_\phi \in \mathbb{R}^{8 \times 512}$. This increases the feature-selector parameter count from $4 \times 16 + 4 = 68$ to $8 \times 512 + 8 = 4,104$, yielding a total of 4,282 trainable parameters across the feature selector (4,104), circuit rotations (48), and final linear classifier operating on 12 measurements (130). The ResNet18 backbone remains frozen throughout training.

Circuit architecture and entanglement variants

Each circuit layer applies data re-uploading [17] through alternating feature encoding and trainable single-qubit rotations. For each layer $\ell \in \{1, 2, 3\}$ and qubit $i \in \{0, 1, 2, 3\}$, the circuit applies a feature-dependent rotation $R_Y(z_i)$ followed by trainable rotations $R_Y(\theta_{i,Y}^{(\ell)})$ and $R_Z(\theta_{i,Z}^{(\ell)})$. In the entangled variant (Quantum-Ent), each layer additionally includes a ring of CNOT gates coupling qubits sequentially as $q_0 \rightarrow q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_0$. In the product-state variant (Quantum-NoEnt), no entangling gates are applied. For the ResNet18 feature regime, the circuit is scaled to eight qubits to provide sufficient capacity for the richer 512-dimensional representations, maintaining the same depth-3 data re-uploading structure but increasing the measurement dimension from 6 to 12 observables.

Measurement and classical post-processing

Measurements are performed only at the end of the circuit. In the 16-dimensional bottleneck regime with four qubits, expectation values of Pauli-Z observables are recorded for each qubit together with two two-qubit correlations, yielding a six-dimensional measurement vector

$$(\langle Z_0 \rangle, \langle Z_1 \rangle, \langle Z_2 \rangle, \langle Z_3 \rangle, \langle Z_0 Z_1 \rangle, \langle Z_2 Z_3 \rangle)$$

passed to a classical linear classifier ($\mathbb{R}^6 \rightarrow \mathbb{R}^{10}$). In the 512-dimensional ResNet18 regime with eight qubits, twelve measurements are recorded (eight single-qubit Pauli-Z observables and four two-qubit ZZ correlations), yielding a ($\mathbb{R}^{12} \rightarrow \mathbb{R}^{10}$) linear map to produce class logits. In the product-state variant, these two-qubit observables factorize and therefore do not increase expressivity beyond the single-qubit measurements; they are retained only to keep the classical post-processing interface identical across entanglement variants and feature regimes.

Feature selector

The quantum classifiers include a trainable linear feature-selection layer g_ϕ that maps classical feature vectors to qubit rotation parameters. This component is not an auxiliary hybrid enhancement but a necessary architectural interface required to map classical inputs onto a fixed number of qubits. The same feature selector architecture and parameterization is used across all quantum variants, including both product-state and entangling circuits. As a result, performance differences between Quantum-NoEnt and Quantum-Ent cannot be attributed to feature-selection quality. Moreover, quantum underperformance persists—and in the entangling case worsens—when replacing the learned 16-dimensional bottleneck with high-quality 512-dimensional ResNet18 features. If the feature selector were the dominant limitation, increasing upstream representation quality would be expected to mitigate or offset performance loss; instead, performance degrades in the entangled setting. These patterns indicate that the primary constraints arise downstream of feature selection, from measurement-induced compression, shallow circuit expressivity, and optimization instability, rather than from the classical input interface itself.

For completeness, we note the relative contribution of the feature selector to the total number of trainable parameters in the quantum models. In the 16-dimensional bottleneck regime, the feature selector comprises 68 of 162 trainable parameters (42%), while in the 512-dimensional CNN-feature regime it accounts for 4,104 of 4,282 parameters (96%). However, parameter count alone does not determine representational bottlenecks. The feature selector performs a simple linear projection from classical feature space to rotation parameters, whereas the variational quantum circuit implements nonlinear unitary evolution followed by inherently lossy measurement. The persistence of quantum underperformance across both parameter regimes—and the reversal of entanglement effects between learned bottlenecks and rich pretrained features—indicates that circuit-level expressivity limits and measurement-induced compression dominate performance, rather than deficiencies in the classical feature-selection interface.

Further decomposing performance into feature-selection and circuit-specific contributions via additional ablations (e.g., replacing the quantum circuit with a purely classical head or freezing circuit parameters) is an interesting direction for future work. However, such decompositions are orthogonal to the conclusions of the present study, which concern end-to-end architectural behavior under a fixed and necessary quantum input interface. Accordingly, the feature selector is treated here as an integral component of the variational quantum architecture rather than as a separable source of performance limitation.

Optimization geometry and diagnostic evaluation

All quantum models are simulated using PennyLane’s `lightning.gpu` backend, which performs exact statevector evolution on $SU(2^n)$ up to floating-point precision. Circuit parameters are represented as coordinates in a Euclidean parameter space and optimized using standard gradient-based methods (Adam). Although forward quantum dynamics respect the intrinsic Lie group geometry of $SU(2^n)$, the default optimization procedure operates on a flat surrogate geometry.

To assess whether incorporating explicit geometric information materially alters short-horizon loss reduction under the present experimental design, we conducted controlled diagnostic experiments comparing Euclidean gradient descent, Adam, and Fubini–Study / quantum Fisher information (FS/QFI)–based natural-gradient methods [18]. Diagnostics were performed on fixed minibatches using identical frozen preprocessors, circuit architectures, and initialization protocols, and were deliberately restricted to short optimization horizons ($K = 10$ updates) in order to isolate local

Table 1. Optimization geometry diagnostics on the learned 16-dimensional bottleneck regime under a fixed frozen preprocessor and identical variational circuit architecture. Results are averaged over 3 random seeds and 10 minibatches per seed (30 runs per entry). Reported losses correspond to the final cross-entropy after $K = 10$ update steps on each minibatch. Wall-clock times are reported per minibatch. Cosine similarity measures alignment of the applied update direction at step 1 with the full Fubini–Study (FS/QFI) natural-gradient step. Absolute loss values remain close to the random-guessing baseline $\log(10) \approx 2.30$ by design, as the purpose of this diagnostic is to compare optimization geometry and update alignment over very short horizons rather than task-level convergence.

Optimizer	Dataset	Final loss	Loss SD	Time (s)	Cosine vs. full FS/QFI
Euclidean GD	MNIST	2.316	0.039	6.04 ± 0.36	0.940 ± 0.047
	FashionMNIST	2.301	0.052	5.95 ± 0.41	0.937 ± 0.074
	CIFAR-10	2.320	0.036	6.09 ± 0.52	0.958 ± 0.031
Adam	MNIST	2.315	0.039	5.97 ± 0.40	0.728 ± 0.073
	FashionMNIST	2.301	0.052	5.92 ± 0.47	0.725 ± 0.067
	CIFAR-10	2.320	0.036	5.97 ± 0.39	0.689 ± 0.099
FS/QFI (diag)	MNIST	2.316	0.039	7.33 ± 0.59	0.961 ± 0.030
	FashionMNIST	2.301	0.052	7.32 ± 0.68	0.959 ± 0.041
	CIFAR-10	2.320	0.036	7.48 ± 0.70	0.974 ± 0.016
FS/QFI (block-diag)	MNIST	2.316	0.039	7.44 ± 0.62	0.961 ± 0.030
	FashionMNIST	2.301	0.052	7.47 ± 0.72	0.959 ± 0.041
	CIFAR-10	2.320	0.036	7.37 ± 0.59	0.974 ± 0.016
FS/QFI (full)	MNIST	2.315	0.039	18.11 ± 1.16	1.000
	FashionMNIST	2.301	0.052	18.46 ± 1.51	1.000
	CIFAR-10	2.320	0.036	18.23 ± 1.36	1.000

geometric effects on descent direction and immediate loss sensitivity rather than global convergence behavior. For context, random-guess cross-entropy for 10-class problems is $\log(10) \approx 2.303$; the reported losses after 10 optimization steps therefore represent minimal improvement over initialization, consistent with the short diagnostic horizon. Results from these local geometric diagnostics are summarized in Table 1.

Across datasets, FS/QFI diagonal and block-diagonal approximations yield update directions that are nearly collinear with the full natural-gradient step (cosine similarity ≈ 0.96 – 0.97), whereas Adam exhibits lower directional alignment (≈ 0.69 – 0.73) due to its adaptive rescaling of gradients. In this short-horizon diagnostic setting, however, all optimizers achieve essentially identical loss reduction, with final cross-entropy means differing by less than 10^{-3} and variances remaining indistinguishable across methods (standard deviations 0.036–0.052 across all optimizers and datasets). This equivalence holds despite substantial differences in directional alignment: Adam’s updates deviate from the natural-gradient direction by cosine dissimilarity ≈ 0.31 , yet produce statistically identical outcomes in both mean performance and stability.

Training protocol and evaluation

For each architecture (Real, Quaternion, Quantum-NoEnt, Quantum-Ent), we perform independent training runs with three random seeds (42, 123, 456). Each seed initializes Python, NumPy, PyTorch, CUDA, and PennyLane random number generators, and PyTorch deterministic mode is enabled (`torch.backends.cudnn.deterministic = True`). Reported results are the mean $\pm$ standard deviation across seeds. Real and quaternion models are trained for up to 100 epochs with early stopping (patience 10).

Quantum models are trained for up to 200 epochs with the same early-stopping criterion to accommodate slower convergence in variational circuit optimization; increasing the patience to 20 epochs did not materially change final accuracy. All models minimize cross-entropy loss using Adam with learning rate 10^{-3} on the fixed stratified training set. This protocol is applied consistently across datasets and feature regimes (16-dimensional learned bottleneck for MNIST, FashionMNIST, and CIFAR-10; and 512-dimensional frozen ResNet18 features for the CIFAR-10 control), with the only regime-specific change being the input dimensionality to the classification head. For each model, we report test accuracy at the epoch achieving the best validation performance, epochs to early stopping, trainable parameters in the classification head, and total wall-clock training time. Together, these metrics enable controlled comparison of representational capacity and computational efficiency across real-valued, quaternion, and quantum classifiers under matched $SU(2)$ constraints and varying feature quality.

Use of automated tools

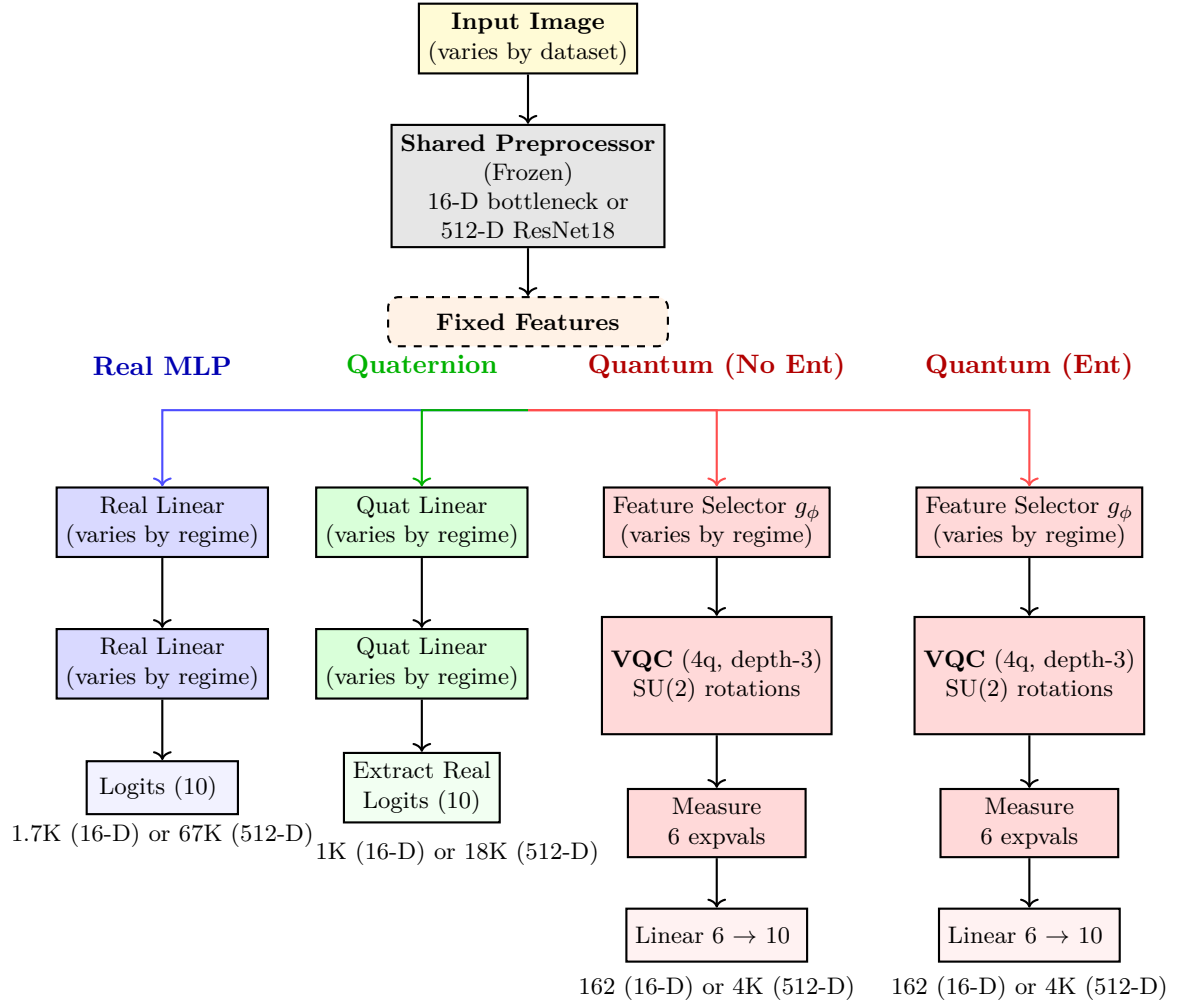
Generative artificial intelligence tools were used to assist with prose refinement, $\text{\LaTeX}$ formatting, and figure layout. All analytical decisions, experimental design, code implementation, validation, and interpretation of results were performed by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

Results

To evaluate whether classical $SU(2)$ models can match or exceed the performance of product-state quantum circuits on classical learning tasks, we compare four classification heads under strictly controlled conditions on three benchmarks: MNIST, FashionMNIST, and CIFAR-10. For each dataset (MNIST [13], FashionMNIST [14], and CIFAR-10 [15]), all models are trained and evaluated on identical stratified subsets (15,000 training and 3,000 test samples) using shared frozen feature representations, ensuring that observed performance differences arise solely from the inductive bias and representational geometry of the classification heads. The architectures compared are (i) a real-valued multilayer perceptron (RealNet), (ii) a quaternion-valued classifier (QuatNet), (iii) a non-entangling variational quantum circuit (Quantum-NoEnt), and (iv) an entangling VQC with a CNOT ring topology (Quantum-Ent). Results are reported as mean $\pm$ standard deviation across three independent random seeds $\{42, 123, 456\}$. All quantum experiments are simulated using PennyLane’s GPU-accelerated `lightning.gpu` backend.

We first present results on MNIST and FashionMNIST using a shared 16-dimensional learned bottleneck to establish baseline behavior on standard image benchmarks. We then extend to CIFAR-10 under the same bottleneck regime to assess persistence under increased visual complexity. Finally, we evaluate CIFAR-10 using frozen ImageNet-pretrained ResNet18 features (512 dimensions) to determine whether observed performance relationships reflect fundamental architectural properties or dependence on representation capacity.

Across all datasets and feature regimes, quaternion classifiers consistently match or closely approach real-valued baselines while substantially outperforming both product-state and entangled quantum circuits (Table 2, Figure 6). On MNIST and FashionMNIST, QuatNet achieves statistical equivalence with RealNet (differences < 0.2 percentage points) using comparable trainable parameters. Product-state quantum models underperform by 5–6 percentage points despite incurring roughly $700\times$ longer training times, while entanglement yields only modest gains (approximately 0.6 percentage points) at additional computational cost. On CIFAR-10 with ResNet18 features, quaternion classifiers retain 95.9% of real-valued MLP



Experimental Design:

- Two feature regimes: 16-D bottleneck (MNIST, FM, CIFAR-10) and 512-D ResNet18 (CIFAR-10)
- Preprocessor frozen after Real baseline training (16-D) or pretrained (ResNet18)
- All heads operate on identical fixed features within each regime
- Isolates the inductive bias of the head architecture, not feature extraction

Fig. 4. Architecture comparison for real-valued, quaternion, and quantum classifiers across multiple feature regimes. **Critical design feature:** All models operate on identical frozen feature representations within each regime (a 16-dimensional learned bottleneck for MNIST, FashionMNIST, and CIFAR-10, or 512-dimensional ImageNet-pretrained ResNet18 features for the CIFAR-10 control), ensuring that observed performance differences arise solely from the classification head architecture rather than feature extraction. The real-valued MLP serves as a baseline, the quaternion head applies the Hamilton product with unit quaternions, and the quantum heads implement variational circuits with four qubits and depth-3 data re-uploading, with and without CNOT entanglement gates arranged in a ring topology ($0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 0$). Layer dimensions vary by feature regime as indicated, while architectural structure remains fixed. Parameter counts for quantum heads include all trainable components of the classification head (the classical feature selector g_ϕ , variational circuit rotation parameters, and the final linear classifier) and exclude the frozen feature extractor.

performance, demonstrating robustness across representation strengths ranging from minimal learned bottlenecks to rich pretrained embeddings.

Real-valued baseline (RealNet)

RealNet combines the shared 16-dimensional bottleneck with a two-layer multilayer perceptron ($16 \rightarrow 64 \rightarrow 10$) and converges rapidly under Adam optimization (learning rate 10^{-3}) with early stopping. On MNIST, RealNet achieves $93.54\% \pm 0.52\%$ test accuracy, converging after 21.3 ± 4.5 training epochs and requiring 30.0 ± 5.9 seconds of wall-clock training time. This performance establishes a reference point for classification accuracy achievable by a simple real-valued head operating on the frozen bottleneck representation.

On FashionMNIST, RealNet attains $84.60\% \pm 0.12\%$ test accuracy, converging after 23.0 ± 2.2 epochs with a training time of 32.0 ± 3.3 seconds. The reduction in absolute accuracy relative to MNIST reflects the greater visual and semantic complexity of FashionMNIST, while the similar convergence behavior and training cost indicate that the frozen bottleneck provides a stable and comparable feature representation across both datasets.

On CIFAR-10 with the learned 16-dimensional bottleneck, RealNet achieves $40.23\% \pm 0.44\%$ test accuracy, converging after 24.0 ± 3.6 epochs and requiring 10.3 ± 1.3 seconds of training time. The pronounced drop in absolute accuracy highlights the difficulty of compressing natural images into a low-dimensional shared representation. When the same architecture operates on 512-dimensional frozen ResNet18 features (Table ??), test accuracy increases to $47.13\% \pm 0.70\%$ (+6.9 percentage points), confirming that the bottleneck regime intentionally constrains representation capacity rather than limiting the expressiveness of the real-valued classifier itself. Convergence remains stable across seeds in both regimes, with the CNN-feature model requiring 22.3 ± 4.0 epochs and 14.2 ± 2.5 seconds of training time per seed.

Quaternion classifier (QuatNet)

QuatNet replaces the real-valued MLP head with two quaternion-valued layers operating on the reshaped bottleneck features (four quaternions). Each layer applies Hamilton products with learned unit-quaternion weights, followed by normalization and a tanh nonlinearity. Class logits are obtained from the real components of the ten output quaternions.

On MNIST, QuatNet achieves $93.64\% \pm 0.15\%$ test accuracy, effectively matching the real-valued baseline (compared to 1,738 for RealNet). Convergence occurs after 31.7 ± 2.4 epochs and requires 77.4 ± 6.5 seconds of wall-clock training time, reflecting the additional computational cost of quaternion operations despite the reduced parameter count.

On FashionMNIST, QuatNet attains $84.47\% \pm 0.05\%$ test accuracy, closely tracking RealNet's $84.60\% \pm 0.12\%$ performance despite the increased visual complexity of the dataset. Training converges in 24.0 ± 2.4 epochs with a wall-clock time of 56.9 ± 6.3 seconds.

On CIFAR-10 with the learned 16-dimensional bottleneck, QuatNet achieves $37.92\% \pm 0.33\%$ test accuracy, retaining 94.3% of RealNet's performance despite a more than tenfold reduction in trainable parameters. When operating on 512-dimensional frozen ResNet18 features (Table 2), test accuracy increases to $45.21\% \pm 0.38$ (+7.3 percentage points), preserving 95.9% of the real-valued baseline. The stability of this performance ratio across feature regimes—from minimal learned bottlenecks to rich pretrained embeddings—indicates that the quaternion-real relationship is not an artifact of representation capacity. In the CNN-feature regime, QuatNet converges after

40.0 $\pm$ 0.0 epochs and requires 36.1 $\pm$ 0.2 seconds of training time per seed, using 17,832 trainable parameters compared to 66,954 for the corresponding real-valued head.

Across all datasets and feature regimes, QuatNet consistently matches or closely approaches real-valued baseline performance in the classification head. This behavior reflects the imposition of a structured inductive bias through unit-quaternion parameterization, which constrains and couples weights according to SU(2) algebra rather than expanding representational capacity. Under these constraints, quaternion-valued classifiers recover the discriminative performance of real-valued models on classical vision tasks across varying levels of dataset complexity and feature quality.

Quantum classifier without entanglement (Quantum-NoEnt)

The non-entangling quantum model employs a four-qubit, depth-3 variational circuit with data reuploading via R_Y rotations and trainable single-qubit R_Y and R_Z gates. No entangling operations are included, enforcing a strict product-state representation throughout the circuit. Final measurements yield six expectation values—four single-qubit Pauli-Z observables and two ZZ correlations—which are passed to a classical linear classifier.

On MNIST, Quantum-NoEnt achieves 87.52% $\pm$ 1.02% test accuracy, converging after 51.3 $\pm$ 16.4 epochs and requiring 24,601 $\pm$ 7,891 seconds of wall-clock training time (approximately 6.8 hours). On FashionMNIST, performance reaches 82.03% $\pm$ 0.94% after 47.3 $\pm$ 4.0 epochs with a training time of 22,285 $\pm$ 2,414 seconds ($\sim$ 6.2 hours). On CIFAR-10 under the learned 16-dimensional bottleneck, test accuracy further degrades to 35.10% $\pm$ 0.71%, with convergence after 32.3 $\pm$ 1.9 epochs and 15,504 $\pm$ 917 seconds ($\sim$ 4.3 hours) of training.

When operating on frozen 512-dimensional ResNet18 features (Table 2), Quantum-NoEnt achieves 41.71% $\pm$ 0.94% test accuracy after 48.7 $\pm$ 10.9 epochs, requiring 64,739 $\pm$ 13,400 seconds ($\sim$ 18.0 hours) per seed. Despite the substantial increase in representation quality, performance remains 5.4 and 3.5 percentage points below the corresponding real-valued and quaternion classifiers, respectively. This gap persists even as training time increases by more than an order of magnitude, indicating that the observed degradation is not alleviated by richer classical features.

Across all datasets and feature regimes, the product-state quantum classifier consistently underperforms both real-valued and quaternion baselines while exhibiting higher variance and orders-of-magnitude greater computational cost. Although Quantum-NoEnt shares the same local SU(2) rotation structure as quaternion layers, the combination of measurement-induced information loss and the absence of entanglement constrains its effective expressivity on these classical vision tasks. In this regime, classical SU(2) structure alone is insufficient to offset the representational and optimization limitations introduced by quantum measurement.

The observed reversal in the effect of entanglement—from a modest improvement on low-complexity inputs to a substantial degradation on richer learned representations—reflects optimization instability that manifests when circuit expressivity exceeds task requirements. Three patterns in the data support this interpretation:

First, variance across random seeds increases dramatically for entangling circuits when combined with high-dimensional features. On CIFAR-10 with ResNet18 features, Quantum-Ent exhibits standard deviation 3.18% compared to 0.94% for the product-state variant—a more than threefold increase in seed sensitivity. This elevated variance indicates that entangling circuits converge to qualitatively different solutions depending on initialization, consistent with a fragmented optimization landscape in which entanglement amplifies local minima sensitivity.

Second, training dynamics reveal instability: entangling circuits on ResNet18 features converge after 42.7 $\pm$ 3.9 epochs with highly variable convergence trajectories

across seeds, whereas product-state circuits reach stable solutions after 48.7 ± 10.9 epochs, despite greater dispersion in stopping time. The combination of lower epoch count but worse final performance for entangling circuits suggests premature convergence to poor local minima rather than successful optimization.

Third, the magnitude of the entanglement penalty scales with feature quality: modest gains (+0.6pp) on learned 16-dimensional bottlenecks, no effect on CIFAR-10 16-dimensional features, and severe degradation (-9.3pp) on 512-dimensional pretrained features. This monotonic relationship indicates that entanglement becomes increasingly detrimental as upstream representations grow more structured, consistent with expressivity-trainability tradeoffs documented in variational quantum circuits where increased parameterization in the absence of task-aligned structure degrades optimization.

These patterns collectively suggest that entanglement in shallow circuits introduces optimization pathologies—increased sensitivity to initialization, convergence to suboptimal basins, and amplified parameter coupling—that outweigh any representational benefits when operating on rich classical features. The elevated variance and convergence instability observed for entangling circuits are consistent with early-stage barren plateau phenomena [19], in which gradient magnitudes become exponentially sensitive to initialization as circuit depth or entanglement increases. While the depth-3 circuits employed here are too shallow to exhibit full barren plateau behavior, the reversal of entanglement’s benefit under rich features suggests that even modest entanglement can amplify optimization pathologies when feature dimensionality is high. Detailed gradient flow analysis or entanglement entropy measurements during training would provide additional mechanistic insight into these dynamics.

Quantum classifier with entanglement (Quantum-Ent)

The entangling quantum model augments the depth-3, four-qubit variational circuit by introducing CNOT gates arranged in a ring topology ($q_0 \rightarrow q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_0$) at each reuploading layer. This modification introduces multi-qubit entanglement while preserving circuit depth, parameter count, and data-encoding structure. All remaining architectural components match those of the product-state model, including feature encoding via $R_Y(z_i)$ rotations, trainable single-qubit R_Y and R_Z gates, and a linear classifier applied to six measured expectation values.

On MNIST, Quantum-Ent achieves $88.16\% \pm 0.62\%$ test accuracy, converging after 67.7 ± 16.0 epochs and requiring $36,492 \pm 8,670$ seconds of wall-clock training time (approximately 10.1 hours). On FashionMNIST, performance reaches $82.62\% \pm 0.52\%$ after 66.3 ± 27.9 epochs, with a training time of $34,446 \pm 14,570$ seconds (~ 9.6 hours). Relative to the non-entangling variant, entanglement yields modest accuracy gains of approximately 0.6–0.7 percentage points on these simpler datasets, albeit at substantially higher computational cost.

On CIFAR-10 with the learned 16-dimensional bottleneck, Quantum-Ent achieves $34.92\% \pm 0.22\%$ test accuracy after 43.3 ± 14.5 epochs and $23,117 \pm 7,498$ seconds of training (~ 6.4 hours). In this setting, entanglement provides no measurable benefit and slightly underperforms the product-state circuit, indicating that shallow entanglement does not mitigate information loss when operating on severely compressed representations.

When operating on frozen 512-dimensional ResNet18 features (Table 2), Quantum-Ent exhibits a pronounced performance collapse, achieving only $32.46\% \pm 3.18\%$ test accuracy after 42.7 ± 3.9 epochs and $66,114 \pm 5,306$ seconds (~ 18.4 hours) per seed. This represents a 9.3 percentage point degradation relative to the non-entangling variant, reversing the modest gains observed on MNIST and FashionMNIST. The result indicates that, in the presence of rich classical features, shallow entanglement interacts

adversely with optimization dynamics and measurement-induced information loss,
overwhelming any potential representational advantage.

Across all benchmarks and feature regimes, the entangling quantum classifier
remains several percentage points below both real-valued and quaternion baselines
while incurring the highest computational cost. Although entanglement introduces
nonclassical correlations unavailable to quaternion layers, these correlations do not
translate into improved performance in shallow, measurement-limited circuits. In this
regime, entanglement is neither sufficient nor reliably beneficial, and may actively
hinder learning when combined with high-quality classical representations.

Head-to-head comparison

Table 2 and Fig. 5 provide complementary views of model performance under strictly
matched conditions. While the table reports aggregated means and standard deviations,
Fig. 5 presents seed-level head-to-head comparisons across accuracy, wall-clock time,
and convergence behavior, enabling direct assessment of stability and computational
cost alongside predictive performance.

On MNIST and FashionMNIST, quaternion classifiers match the real-valued
baseline within statistical variability (differences ≤ 0.2 percentage points), despite using
fewer trainable parameters (1,000 versus 1,738). This parity is clearly visible in the
accuracy panels of Fig. 5, where RealNet and QuatNet cluster tightly across seeds with
overlapping variance. In contrast, both quantum models underperform the classical
heads by approximately 5–6 percentage points. Entanglement yields only modest gains
over the product-state circuit (approximately 0.6–0.7 percentage points), a pattern
reflected in Fig. 5 by a small upward shift accompanied by increased variance and
substantially higher computational cost. The time and epoch panels further show that
these marginal gains come at orders-of-magnitude greater training time.

On CIFAR-10 with the learned 16-dimensional bottleneck, absolute accuracies
decline for all architectures due to increased task complexity under severe
representational compression. Despite this, the relative ordering observed on simpler
benchmarks is preserved, as shown consistently across seeds in Fig. 5. RealNet achieves
 $40.2\% \pm 0.4\%$ accuracy, QuatNet retains 94.3% of this performance ($37.9\% \pm 0.3\%$), and
the product-state quantum circuit reaches only $35.1\% \pm 0.7\%$. The entangling quantum
model does not improve upon this result, indicating that shallow entanglement does not
mitigate information loss in the bottleneck regime. Seed-wise scatter in Fig. 5 confirms
that this degradation is systematic rather than driven by outliers.

When CIFAR-10 is evaluated using frozen ImageNet-pretrained ResNet18 features,
absolute accuracies increase substantially for all classical models (47.1% versus 40.2%
for RealNet), as expected from the richer representation. However, the qualitative
ordering remains unchanged, a pattern that Fig. 5 makes explicit across accuracy, time,
and epochs. QuatNet retains 95.9% of real-valued performance (45.2% versus 47.1%),
while the product-state quantum circuit reaches 41.7%, underperforming the quaternion
model by 3.5 percentage points. The entangling quantum circuit degrades further to
32.5%, underperforming even the product-state variant by 9.3 percentage points. The
persistence—and amplification—of the quantum performance gap under a 32-fold
increase in feature dimensionality demonstrates that these differences reflect
architectural and measurement constraints rather than representation capacity.

Across all datasets, random seeds, and feature regimes, the same ordering
consistently emerges in Fig. 5: RealNet $\gtrsim$ QuatNet $\gg$ Quantum-NoEnt $\gtrsim$
Quantum-Ent. The figure further shows that quantum models incur substantially higher
training times and less stable convergence, even when accuracy gains are marginal.
Taken together, these results demonstrate that classical SU(2) structure, when imposed
through quaternion parameterization, matches real-valued baselines while substantially

Table 2. Aggregated test performance (mean $\pm$ standard deviation over three seeds) under shared frozen preprocessors. For MNIST and FashionMNIST, all models use a 16-dimensional learned bottleneck preprocessor (12,560 parameters, frozen during training). For CIFAR-10, results are reported under two feature regimes: (i) the same 16-dimensional learned bottleneck, and (ii) frozen ImageNet-pretrained ResNet18 features (512-dimensional, $\sim$ 11.2M parameters). Reported times correspond to total wall-clock training time per seed. Parameter counts report *trainable parameters outside the frozen backbone*, including the classification head and any learned feature-selection layers. Note the substantially higher variance for quantum models across seeds, reflecting optimization instability under parameter-shift gradient estimation without curvature information.

Dataset	Model	Accuracy	Epochs	Time (s)	Trainable Params
MNIST	RealNet	0.9354 ± 0.0052	21.3 ± 4.5	30.0 ± 5.9	1,738
	QuatNet	0.9364 ± 0.0015	31.7 ± 2.4	77.4 ± 6.5	1,000
	Quantum-NoEnt	0.8752 ± 0.0102	51.3 ± 16.4	$24,601 \pm 7,891$	162
	Quantum-Ent	0.8816 ± 0.0062	67.7 ± 16.0	$36,492 \pm 8,670$	162
FashionMNIST	RealNet	0.8460 ± 0.0012	23.0 ± 2.2	32.0 ± 3.3	1,738
	QuatNet	0.8447 ± 0.0005	24.0 ± 2.4	56.9 ± 6.3	1,000
	Quantum-NoEnt	0.8203 ± 0.0094	47.3 ± 4.0	$22,285 \pm 2,414$	162
	Quantum-Ent	0.8262 ± 0.0052	66.3 ± 27.9	$34,446 \pm 14,570$	162
CIFAR-10 (16-D bottleneck)	RealNet	0.4023 ± 0.0044	24.0 ± 3.6	10.3 ± 1.3	1,738
	QuatNet	0.3792 ± 0.0033	36.0 ± 5.7	24.3 ± 2.7	1,000
	Quantum-NoEnt	0.3510 ± 0.0071	32.3 ± 1.9	$15,504 \pm 917$	162
	Quantum-Ent	0.3492 ± 0.0022	43.3 ± 14.5	$23,117 \pm 7,498$	162
CIFAR-10 (ResNet18 512-D)	RealNet	0.4713 ± 0.0070	22.3 ± 4.0	14.2 ± 2.5	66,954
	QuatNet	0.4521 ± 0.0038	40.0 ± 0.0	36.1 ± 0.2	17,832
	Quantum-NoEnt	0.4171 ± 0.0094	48.7 ± 10.9	$64,739 \pm 13,400$	4,282
	Quantum-Ent	0.3246 ± 0.0318	42.7 ± 3.9	$66,114 \pm 5,306$	4,282

outperforming shallow quantum models on classical vision tasks. Modest entanglement benefits observed on low-complexity grayscale benchmarks do not generalize to richer feature representations, underscoring the tight coupling between circuit depth, measurement-induced compression, and task complexity in the shallow quantum regime.

Stability across random seeds

All reported results are averaged over three independent random seeds. Across MNIST, FashionMNIST, and both CIFAR-10 feature regimes, relative performance rankings are stable across seeds, with consistently low variability for the real-valued and quaternion models and substantially higher variance for the quantum models. For all datasets and feature qualities, accuracy differences between RealNet and QuatNet remain below 2.5 percentage points (and below 0.2 percentage points on MNIST and FashionMNIST), whereas both quantum variants exhibit lower mean accuracy and markedly greater seed-to-seed variability. This behavior is illustrated in Figure 6, which shows tightly clustered accuracy ranges for the classical models and substantially wider ranges for the quantum models, particularly on the more challenging CIFAR-10 benchmark.

On CIFAR-10 with frozen ResNet18 features, quaternion classifiers exhibit especially low variance (standard deviation 0.38%) compared to both real-valued MLPs (0.70%) and quantum circuits (0.94% for Quantum-NoEnt and 3.18% for Quantum-Ent). This pattern suggests that imposing structured SU(2) constraints through unit-quaternion parameterization may stabilize optimization under high-dimensional feature representations. The exceptionally high variance observed for the entangling quantum circuit on CNN features reflects the complex interaction between entanglement, measurement, and optimization in the shallow-circuit regime.

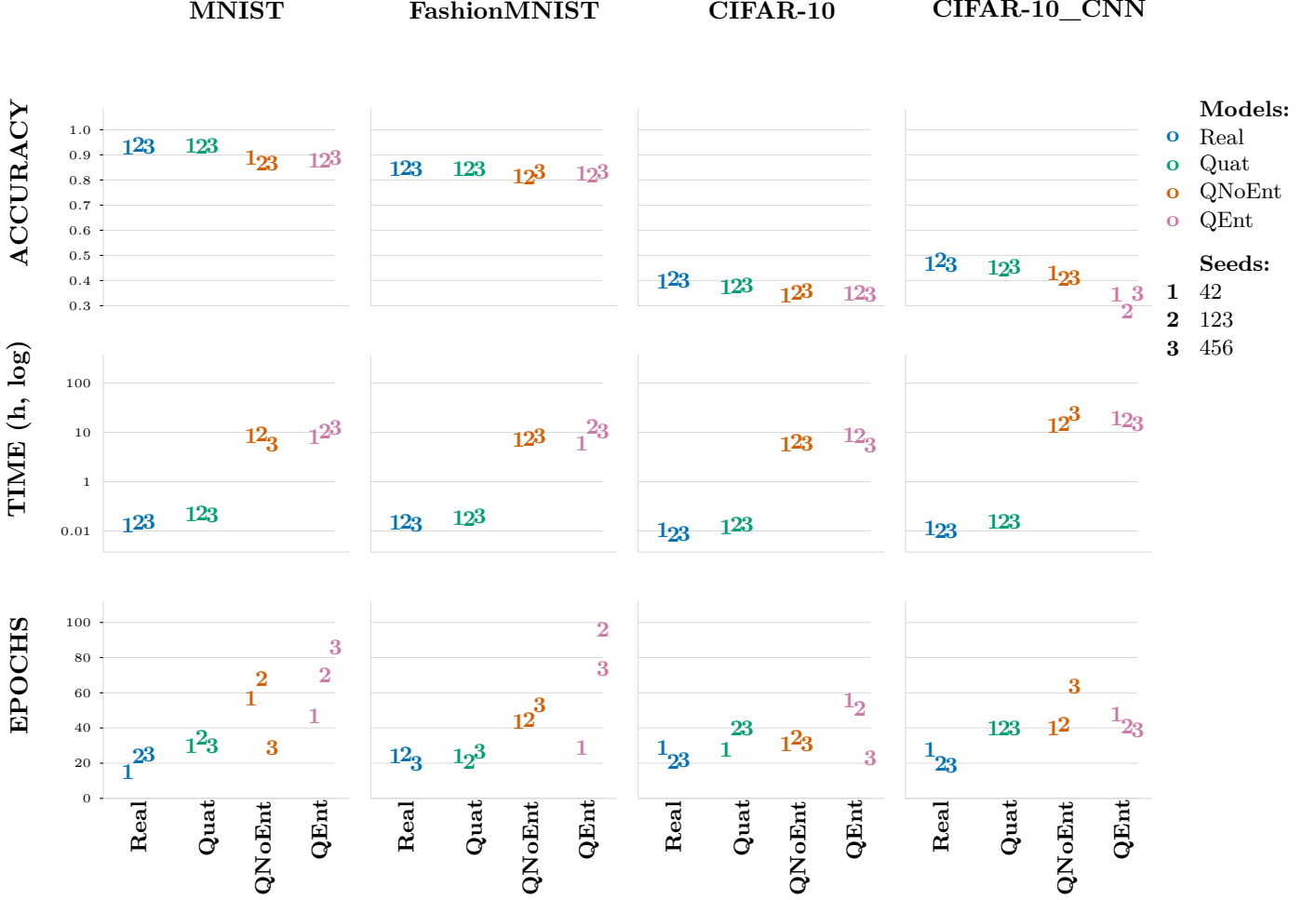


Fig. 5. Conceptual correspondence between unit quaternions and single-qubit $SU(2)$ operations. The set of unit quaternions $\{q \in \mathbb{H} : \|q\| = 1\}$ forms a Lie group isomorphic to $SU(2)$; both are diffeomorphic to the 3-sphere S^3 , reflecting the same underlying rotation geometry. In a quaternion layer, a classical feature quaternion h is transformed by quaternion multiplication (Hamilton product) as $h' = q \otimes h$, retaining a four-dimensional representation. In a variational quantum circuit, parameters specify a unitary $U \in SU(2)$ that evolves a quantum state as $|\psi'\rangle = U|\psi\rangle$, after which information is accessed only through measurement outcomes (and, for multi-qubit extensions, potentially through entanglement). Consequently, despite sharing $SU(2)$ -valued parameters, quaternion layers and quantum circuits differ in how information is represented and revealed, leading to distinct inductive biases and optimization behavior.

CIFAR-10 representation-strength control (frozen CNN features)

To assess whether the reduced CIFAR-10 accuracy observed under the shared $3072 \rightarrow 16$ bottleneck reflects representational compression rather than intrinsic differences in head architecture, we conducted an auxiliary control experiment using frozen ImageNet-pretrained ResNet18 features. The learned bottleneck was replaced by a ResNet18 backbone (512-dimensional output, 11.2M parameters, frozen), and only the classification heads were trained. All experiments used the same stratified CIFAR-10 subset (15,000 training and 3,000 test samples), identical random seeds, and the same early-stopping protocol as in the bottleneck regime.

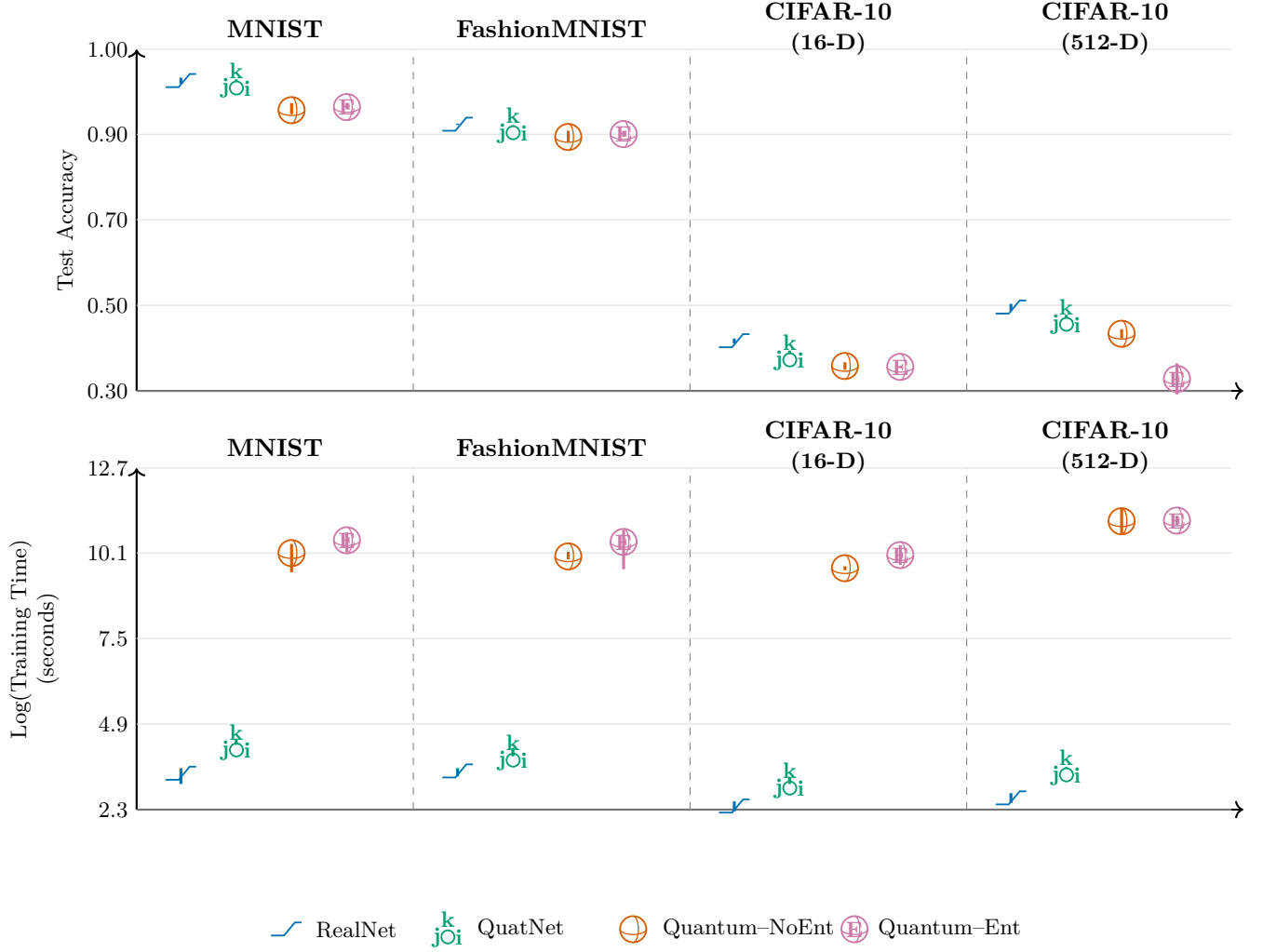


Fig. 6. Performance across three random seeds on MNIST, FashionMNIST, and CIFAR-10 under two frozen feature regimes. Top: test accuracy; bottom: log-scaled training time. Vertical bars indicate the observed range (min to max) across seeds, with glyph markers denoting the mean. Quaternion classifiers closely track real-valued baselines on MNIST and FashionMNIST and retain approximately 94% of real-valued performance under CIFAR-10 with a learned 16-dimensional bottleneck and approximately 96% under 512-dimensional ResNet18 features, indicating consistency across representation strengths. Quantum classifiers exhibit greater variance and substantially higher training time while achieving lower accuracy across all benchmarks. Entanglement provides modest gains on grayscale datasets but degrades performance under high-quality CNN features (CIFAR-10, 512-D), highlighting regime-dependent interactions between circuit structure, measurement, and feature quality.

Under this frozen-CNN regime, absolute accuracy increased for all classical architectures (Table 2). The real-valued head achieved $47.13\% \pm 0.70\%$ accuracy (+6.9 percentage points relative to the learned bottleneck), while the quaternion head reached $45.21\% \pm 0.38\%$ (+7.3 percentage points). Quaternions retained 95.9% of real-head performance under high-quality pretrained features, compared to 94.3% under the learned bottleneck, indicating that the quaternion–real performance relationship is stable across representation strengths and reflects properties of the $SU(2)$ -structured head rather than artifacts of feature capacity.

Quantum models were evaluated under the same CNN-feature regime to determine whether measurement-induced degradation persists with richer representations. The product-state circuit (Quantum-NoEnt) achieved $41.71\% \pm 0.94\%$, underperforming both real and quaternion heads by 5.4 and 3.5 percentage points, respectively. The entangling variant (Quantum-Ent) performed substantially worse, reaching only $32.46\% \pm 3.18\%$, a degradation of 9.3 percentage points relative to the product-state circuit and a reversal of the modest entanglement gains observed on MNIST and FashionMNIST. This behavior indicates that shallow entanglement interacts unfavorably with high-dimensional pretrained features, amplifying optimization difficulty and measurement-induced information loss.

All quantum experiments employ four-qubit, depth-3 circuits with approximately 4,000 trainable parameters, representing the practical upper limit for systematic multi-seed evaluation under GPU-accelerated simulation. Increasing circuit width or depth rapidly incurs prohibitive memory cost and exacerbates optimization pathologies, whereas real and quaternion models scale efficiently under standard backpropagation. Quantum circuits underperform by several percentage points, while quaternions retain nearly all real-valued performance. This asymmetry indicates that quantum degradation arises from architectural and measurement constraints rather than insufficient parameterization.

These results show that the low absolute CIFAR-10 accuracies observed under the learned bottleneck arise from intentional representational compression rather than head limitations, that quaternion–real performance ratios remain consistent across representation qualities (94–96%), and that quantum circuit limitations persist—and intensify—under high-quality pretrained features. The CNN-feature control therefore confirms that the observed performance ordering reflects fundamental architectural properties of the classification heads rather than artifacts of feature capacity. Notably, compressing high-dimensional features into a small number of circuit parameters reflects the intrinsic measurement bottleneck faced by variational quantum models; increasing classical preprocessing would shift representational burden away from the quantum head rather than alleviate it.

Discussion

The results provide empirical evidence that classical $SU(2)$ models, as realized through quaternion-valued neural networks, match or exceed the learning performance of shallow variational quantum circuits on classical supervised learning tasks. Across MNIST, FashionMNIST, and CIFAR-10, the same qualitative ordering of classification heads persists as task complexity increases from grayscale digits to natural color images, indicating that the observed differences reflect architectural inductive bias rather than dataset-specific effects. Because all models operate on identical frozen feature representations within each regime, performance differences arise solely from the representational geometry and optimization behavior of the classification heads.

To distinguish architectural effects from representational artifacts, CIFAR-10 was evaluated under two feature regimes: a shared 16-dimensional learned bottleneck and frozen ImageNet-pretrained ResNet18 features producing 512-dimensional embeddings. The quaternion–real performance relationship remains stable across this 32-fold difference in feature dimensionality, with quaternions retaining 94.3% and 95.9% of real-valued performance under the bottleneck and CNN regimes respectively. This robustness indicates that the effectiveness of quaternion-valued heads reflects intrinsic properties of $SU(2)$ -structured parameterization rather than dependence on representation quality. In contrast, quantum circuits underperform classical heads by 5–6 percentage points under

both regimes, demonstrating that the measurement bottleneck and optimization challenges persist even when high-quality classical features are provided.

These results also clarify the role of parameter count. Quantum models employ approximately 4,000 trainable parameters, representing the practical upper bound for systematic multi-seed evaluation with four-qubit circuits under GPU-accelerated simulation. Scaling beyond this regime incurs prohibitive computational cost and exacerbates known trainability limitations, including barren plateau phenomena. Real and quaternion models, by contrast, scale efficiently under standard backpropagation to 66,954 and 17,832 parameters respectively in the CNN-feature regime. Despite this disparity, quantum circuits underperform real-valued heads by 5.4 percentage points under high-quality features, while quaternions retain nearly all real-valued performance. This asymmetry indicates that quantum degradation arises from architectural and measurement constraints rather than insufficient parameterization.

A particularly informative pattern emerges from the entanglement comparisons. On MNIST and FashionMNIST, entanglement provides modest gains (approximately 0.6 percentage points) over product-state circuits at substantially higher computational cost. On CIFAR-10 with a learned bottleneck, entanglement yields no improvement. When combined with high-dimensional pretrained features, entanglement actively degrades performance, reducing accuracy by 9.3 percentage points relative to the product-state variant. This reversal suggests that shallow entanglement interacts unfavorably with rich classical representations, likely amplifying optimization instability and measurement-induced information loss rather than enhancing expressivity.

Several considerations bound the scope of these findings. The experiments are restricted to classical vision benchmarks lacking intrinsic quantum structure, the quantum circuits are shallow by design to maintain trainability, and training is conducted on moderate-sized stratified subsets to control sample complexity. These constraints define a specific but practically relevant regime in which classical and quantum architectures can be compared under matched conditions. Within this regime, the results support recent perspectives emphasizing that quantum advantage cannot be inferred from model class alone, but must be evaluated relative to task structure, data distribution, and competitive classical baselines [12].

Quaternion networks match classical baselines and exceed quantum models

Across all datasets and feature regimes, quaternion-valued classification heads achieve statistical equivalence or near-equivalence with real-valued baselines while using substantially fewer trainable parameters. On MNIST, QuatNet attains $93.64\% \pm 0.15\%$ test accuracy compared with $93.54\% \pm 0.52\%$ for RealNet, a difference well within seed-to-seed variation. A similar pattern holds on FashionMNIST, where QuatNet achieves $84.47\% \pm 0.05\%$ versus $84.60\% \pm 0.12\%$ for the real-valued head. These results demonstrate that quaternion parameterization does not sacrifice predictive performance on standard vision benchmarks despite its substantially reduced head capacity.

On CIFAR-10, absolute accuracies decline for all models under the learned 16-dimensional bottleneck, reflecting increased task complexity under severe representational compression. Nevertheless, the relative relationship between quaternion and real-valued heads remains stable: QuatNet retains 94.3% of RealNet performance (37.9% versus 40.2%). When evaluated instead on frozen ImageNet-pretrained ResNet18 features, quaternion performance increases to $45.2\% \pm 0.38\%$ while retaining 95.9% of the real-valued baseline ($47.1\% \pm 0.70\%$). The persistence of this performance ratio across a 32-fold increase in feature dimensionality indicates that the quaternion-real

relationship reflects fundamental architectural properties rather than artifacts of representation capacity.

These performance results are achieved with markedly fewer trainable parameters in the classification head. On MNIST and FashionMNIST, QuatNet uses 1,000 trainable parameters compared with 1,738 for RealNet, a 1.7-fold reduction. On CIFAR-10 with ResNet18 features, QuatNet employs 17,832 parameters compared with 66,954 for the real-valued head, a 3.7-fold reduction. All parameter counts refer exclusively to trainable head parameters; frozen feature extractors are shared across all models and therefore do not affect relative efficiency comparisons. The ability of quaternion heads to recover nearly all real-valued performance with substantially reduced capacity highlights the effectiveness of structured $SU(2)$ coupling via Hamilton products.

Quaternion models also exhibit consistently low variance across random seeds, particularly under high-quality pretrained features. On CIFAR-10 with ResNet18 embeddings, QuatNet displays lower variability (0.38%) than both real-valued MLPs (0.70%) and quantum circuits, suggesting that $SU(2)$ -constrained parameterization may stabilize optimization in high-dimensional feature spaces. This stability is consistent with the geometric coupling imposed by unit quaternions, which faithfully represent local $SU(2)$ structure [5] and constrain the hypothesis space without eliminating expressivity.

In contrast, variational quantum circuits restricted to product states offer no empirical advantage over classical quaternion layers operating on the same frozen features. Across all benchmarks and feature regimes, product-state quantum models underperform QuatNet by 6.1 percentage points on MNIST, 2.4 points on FashionMNIST, 2.8 points on CIFAR-10 with the learned bottleneck, and 3.5 points on CIFAR-10 with ResNet18 features. These results demonstrate that classical implementations of local $SU(2)$ geometry consistently match or exceed shallow product-state quantum circuits on classical learning tasks, while avoiding the measurement-induced information loss and optimization instability inherent to quantum models.

The close alignment between quaternion-valued and real-valued networks should not be interpreted as a guarantee of strict equivalence. Small but consistent performance differences remain, and on simpler benchmarks such as MNIST these may be partially masked by ceiling effects. Nevertheless, the same qualitative quaternion–real performance relationship persists across FashionMNIST, CIFAR-10 under severe bottlenecks, and CIFAR-10 with high-quality pretrained features, spanning wide ranges of visual complexity, absolute accuracy, and feature dimensionality. This robustness across regimes supports the interpretation that quaternion-valued networks are consistent with a meaningful correspondence between $SU(2)$ geometric structure and classical feature representations, rather than reflecting dataset-specific ceilings or incidental architectural coincidences.

Measurement bottlenecks, optimization geometry, and depth constraints limit shallow quantum classifiers

Across all datasets and feature regimes, the largest and most persistent separation between classical and quantum implementations of $SU(2)$ geometry appears between quaternion networks (QuatNet) and the product-state variational quantum classifier (Quantum-NoEnt). On MNIST, QuatNet achieves 93.64% test accuracy, whereas Quantum-NoEnt reaches only $87.52\% \pm 1.02\%$, a deficit exceeding six percentage points. The same pattern holds on FashionMNIST ($84.47\% \pm 0.05\%$ versus $82.03\% \pm 0.94\%$). On CIFAR-10 with the learned 16-dimensional bottleneck, QuatNet attains $37.9\% \pm 0.4\%$ accuracy compared with $35.1\% \pm 0.7\%$ for the product-state quantum model, preserving a 2.8 percentage point gap despite lower absolute performance for all

architectures. Under frozen ResNet18 features (512-dimensional), the gap widens further to 3.5 percentage points (45.2% versus 41.7%), indicating that quantum underperformance persists—and slightly intensifies—under richer classical representations. Across all settings, this accuracy deficit is accompanied by substantially higher variance and orders-of-magnitude greater training time for the quantum model.

The classical architectures employed are deliberately minimal—two-layer MLPs with a single hidden layer—to avoid confounding depth effects with geometric differences. Despite this constraint, classical heads substantially outperform depth-3 quantum circuits across all feature regimes. This framing is conservative: increasing classical depth would be expected to improve performance further, whereas deeper variational quantum circuits introduce well-documented trainability challenges, including barren plateau phenomena [19]. Accordingly, the observed performance gap should be interpreted as a conservative estimate within the shallow, measurement-limited regime studied here, rather than as an artifact of unequal architectural expressivity.

We attribute this divergence to a coupled set of structural limitations inherent to variational quantum circuits in the shallow, measurement-based regime:

(i) measurement-induced information loss, (ii) optimization challenges intrinsic to circuit parameterization, and (iii) expressivity limits imposed by shallow circuit depth. These factors should be understood jointly rather than as independent deficiencies.

At the representational level, quaternion layers preserve internal structure throughout the network. Each Hamilton-product layer maps four-dimensional real inputs to four-dimensional outputs, retaining phase-coupled information explicitly at every stage. By contrast, the product-state variational quantum circuit applies the same local $SU(2)$ rotations but must ultimately undergo a non-unitary, lossy measurement. A four-qubit quantum state $|\psi\rangle \in \mathbb{C}^{16}$ is compressed into six expectation values $(\langle Z_0 \rangle, \dots, \langle Z_2 Z_3 \rangle)$ before reaching the final classifier, discarding amplitude and phase information that remains available to quaternion networks. This measurement bottleneck limits effective expressivity independent of feature quality: whether operating on learned 16-dimensional bottlenecks or rich 512-dimensional pretrained embeddings, quantum circuits consistently trail quaternion classifiers by 2.8–3.5 percentage points.

Optimization geometry further constrains quantum performance. To assess whether geometry-aware optimization mitigates these limitations, we conducted controlled diagnostic experiments comparing Euclidean gradient descent, Adam, and several Fubini–Study / quantum Fisher information (FS/QFI) natural-gradient variants (diagonal, block-diagonal, and full) on fixed minibatches across MNIST, FashionMNIST, and CIFAR-10 (Table 1). FS/QFI-based methods produced update directions nearly collinear with the full natural-gradient step (cosine similarity ≈ 0.95 for diagonal and block-diagonal approximations), whereas Adam exhibited lower directional alignment due to its adaptive rescaling. Crucially, this difference in alignment did not translate into improved learning: final losses achieved by Adam were statistically indistinguishable from those obtained using full FS/QFI optimization, differing by less than 10^{-3} across all datasets.

Because FS/QFI-based methods incur $\mathcal{O}(p^2)$ computational cost from metric-tensor evaluation—compared with $\mathcal{O}(p)$ updates for Adam—these results indicate that explicit curvature-aware optimization is computationally inefficient in the shallow-circuit, measurement-limited regime considered here. For the architectures, datasets, and feature regimes studied, standard Euclidean optimizers capture the dominant descent structure without overcoming the fundamental representational constraints imposed by measurement.

Beyond mean performance, the product-state quantum model exhibits substantially higher variance across random seeds, reflecting intrinsic training instability. These effects are consistent with prior analyses of variational quantum circuits, which

document structured loss landscapes, sensitivity to initialization, and dataset-dependent trainability even at modest circuit depths [20]. Despite sharing identical frozen feature representations and optimization protocols, Quantum-NoEnt requires 51.3 ± 16.4 epochs and approximately 6.8 hours of GPU time per seed on MNIST, similar time on FashionMNIST, over four hours per seed on CIFAR-10 with learned bottlenecks, and approximately 18 hours per seed on CIFAR-10 with ResNet18 features. In contrast, QuatNet converges in under two minutes on MNIST and FashionMNIST, and in under one minute on CIFAR-10 under both feature regimes. Parameter-shift gradient estimation, sensitivity to initialization, and limited circuit expressivity jointly contribute to this instability despite the shared $SU(2)$ symmetry.

The restriction to shallow circuits is deliberate. Deeper variational circuits face well-documented barren plateau phenomena [19], in which gradients vanish exponentially with system size, rendering optimization impractical with current methods. Limiting circuit depth therefore ensures trainability but caps representational power. Whether reliably trainable deeper circuits can overcome measurement-induced information loss and close the observed performance gap—or whether these limitations persist at scale—remains an open question beyond the controlled regime examined here.

Quantifying the measurement bottleneck

To formalize the measurement-induced information loss, we compare the effective dimensionality of representations at the final pre-classification stage across architectures. This analysis establishes upper bounds on information-bearing capacity accessible to downstream classifiers in the shallow, measurement-limited regime studied here, independent of whether those bounds are saturated in practice.

In both the 16-dimensional bottleneck and 512-dimensional CNN-feature regimes, quaternion networks ultimately produce 10 quaternions at their penultimate layer (before real-part extraction for classification), corresponding to 40 real-valued scalar components. The architectural scaling between regimes affects intermediate layer widths but preserves the final representational dimensionality: 10 quaternions provide 40 real values from which the classifier constructs 10-dimensional logits.

Product-state quantum circuits in the 16-dimensional regime encode features into four rotation angles, apply depth-3 variational evolution to a four-qubit state (16 complex amplitudes = 32 real degrees of freedom), and then extract 6 expectation values—four single-qubit Pauli- Z observables and two two-qubit correlations. In the 512-dimensional regime, circuits scale to eight qubits (256 complex amplitudes = 512 real DOF) and extract 12 expectation values (8 single-qubit and 4 two-qubit observables). Although the quantum state space is formally high-dimensional prior to measurement, only these 6 or 12 scalar observables are accessible to the final classifier.

This constitutes a 6.7-fold compression in the 16-dimensional regime (6 measurements versus 40 quaternion components) and a 3.3-fold compression in the 512-dimensional regime (12 measurements versus 40 quaternion components). Moreover, because product-state circuits prohibit entanglement, two-qubit correlation terms factorize as $\langle Z_i Z_j \rangle = \langle Z_i \rangle \langle Z_j \rangle$ and provide no additional information beyond single-qubit measurements. The effective measurement dimensionality is thus only 4 independent observables in the 16-dimensional regime and 8 in the 512-dimensional regime, yielding 10-fold and 5-fold compressions respectively.

These formal capacity differences establish an upper bound on the information accessible to the classification stage. Although a smaller set of well-chosen measurements could in principle capture task-relevant structure more efficiently than a larger unstructured representation, the consistent underperformance of quantum models across all datasets and feature regimes (2.4–6.1 percentage points under learned bottlenecks, 3.5 percentage points under CNN features) suggests that the measurement

space is insufficient to recover the discriminative information preserved by quaternion networks’ 40-dimensional penultimate representations. The observed performance gap therefore reflects a fundamental architectural constraint within this regime: measurement-induced dimensionality reduction limits the effective expressivity of shallow variational quantum classifiers independent of the formal richness of the pre-measurement quantum state. This comparison should be interpreted as a regime-specific diagnostic rather than a universal limitation of quantum architectures; deeper circuits with richer measurement strategies may overcome these constraints.

Optimization geometry limits variational quantum circuits

Beyond measurement compression and shallow circuit depth, variational quantum circuits face intrinsic optimization challenges arising from the geometry of parameter-shift gradient estimation. Quaternion networks are trained using standard backpropagation in Euclidean parameter space, benefiting from full gradient information and adaptive second-moment estimation via Adam. In contrast, parameter-shift rules provide only local first-order directional derivatives and do not expose curvature information such as the Hessian or the Fubini–Study metric by default.

This geometric limitation manifests empirically as pronounced sensitivity to initialization and elevated variance across random seeds. On MNIST, the entangled quantum model exhibits test accuracies ranging from 87.7% to 89.0% (standard deviation 0.62%), compared with the quaternion classifier’s tightly clustered range of 93.5% to 93.8% (standard deviation 0.15%). This fourfold increase in variance reflects the optimizer’s inability to reliably distinguish favorable parameter basins from unfavorable ones in the absence of curvature guidance. The same pattern persists and intensifies under richer feature representations: on CIFAR-10 with frozen ResNet18 features, the entangled quantum circuit exhibits exceptionally high variance (standard deviation 3.18%), exceeding that of quaternion networks (0.38%) by more than a factor of eight and real-valued MLPs (0.70%) by more than a factor of four. Rather than stabilizing training, higher-quality pretrained features exacerbate the optimization fragility of shallow quantum circuits.

Controlled diagnostic experiments (Table 1) clarify the role of optimization geometry. Explicit Fubini–Study / quantum Fisher information (FS/QFI) preconditioning yields update directions nearly collinear with the full natural-gradient step (cosine similarity ≈ 0.96), yet produces final losses statistically indistinguishable from those obtained using Adam. This equivalence holds across MNIST, FashionMNIST, and CIFAR-10, and comes at approximately three times the computational cost due to the $\mathcal{O}(p^2)$ scaling of metric-tensor evaluation. These results indicate that, in the shallow measurement-limited regime studied here, optimization geometry alone does not constitute the dominant performance bottleneck; instead, expressivity constraints imposed by circuit depth and measurement compression bound attainable accuracy regardless of optimizer choice.

Nevertheless, the elevated variance across seeds reveals a persistent secondary effect. Without curvature information, gradient-based optimization cannot consistently avoid poor initialization basins. On MNIST, seed 123 converges to 87.7% accuracy while seed 42 reaches 89.0%, despite identical architectures, data, and hyperparameters. Quaternion networks, by contrast, exhibit minimal seed dependence across all datasets, with all runs converging within 0.3 percentage points on MNIST and FashionMNIST. On CIFAR-10 with ResNet18 features, quaternion classifiers display even lower variance than real-valued MLPs, suggesting that structured $SU(2)$ constraints stabilize optimization under high-dimensional feature representations rather than hinder it.

These findings show that shallow quantum classifiers are doubly disadvantaged: measurement-induced information loss and limited circuit depth cap representational capacity, while geometric fragility under first-order optimization amplifies variance and

impedes reliable convergence. Classical quaternion networks, operating in Euclidean parameter space with adaptive moment estimation, achieve both higher mean accuracy and substantially lower variance across all datasets and feature regimes. These results suggest that practical quantum advantage on classical data will require not only deeper circuits and richer entanglement, but also optimization strategies or initialization schemes that robustly exploit quantum geometric structure—neither of which is currently available at practical computational cost. In the absence of such advances, increasing feature quality or circuit complexity alone does not close the observed performance gap.

Representation-strength control via frozen CNN features

To determine whether the observed performance relationships reflect fundamental architectural properties rather than artifacts of representation capacity, we conducted a control experiment on CIFAR-10 in which the learned 16-dimensional bottleneck was replaced with a frozen ImageNet-pretrained ResNet18 backbone. This substitution yields 512-dimensional feature embeddings—a 32-fold increase in feature dimensionality—while preserving identical classification head architectures. By dramatically increasing representation quality without altering the heads, this design isolates the effect of head geometry and inductive bias from that of feature capacity.

The resulting performance patterns are remarkably consistent with those observed under the learned bottleneck. Real-valued heads achieve $47.13\% \pm 0.70\%$ accuracy, an increase of 6.9 percentage points relative to the bottleneck regime, while quaternion heads reach $45.21\% \pm 0.38\%$ (+7.3 percentage points). Under these high-quality features, quaternions retain 95.9% of real MLP performance, compared to 94.3% under the learned bottleneck. This slight increase indicates that $SU(2)$ algebraic constraints incur no additional penalty—and may even confer modest advantages—when operating on rich pretrained representations. Notably, quaternion classifiers also exhibit lower variance (0.38%) than real MLPs (0.70%) in this regime, suggesting that structured geometric constraints stabilize optimization under high-dimensional feature representations.

Quantum circuits were evaluated under the same CNN-feature regime to assess whether their underperformance under learned bottlenecks arises from weak representations or from intrinsic architectural limitations. The product-state quantum model (Quantum-NoEnt) achieves $41.71\% \pm 0.94\%$ accuracy, continuing to underperform both real and quaternion heads by 5.4 and 3.5 percentage points, respectively. This gap closely mirrors the 5–6 percentage point degradation observed on MNIST and FashionMNIST, indicating that measurement compression and optimization challenges persist independent of representation quality.

The entangling quantum model (Quantum-Ent) exhibits a striking reversal of behavior. While entanglement provides modest gains (approximately 0.6 percentage points) over product-state circuits on MNIST and FashionMNIST, and no benefit under the CIFAR-10 learned bottleneck, it substantially degrades performance under CNN features. On CIFAR-10 with ResNet18 embeddings, Quantum-Ent achieves only $32.46\% \pm 3.18\%$ accuracy, underperforming the product-state variant by 9.3 percentage points. This degradation is accompanied by exceptionally high variance—more than eight times that of quaternion classifiers—indicating that entanglement compounds optimization instability when combined with rich classical feature representations.

These results establish three central findings. First, the quaternion–real performance relationship (94–96% retention) is robust across a 32-fold increase in feature dimensionality, demonstrating that it reflects fundamental $SU(2)$ geometric structure rather than representational constraints. Second, quantum circuit degradation persists under both weak and strong feature representations, confirming that measurement compression and optimization fragility are intrinsic architectural limitations. Third, entanglement’s modest benefits on simple grayscale benchmarks reverse under

high-quality pretrained features, revealing adverse interactions between circuit depth, measurement strategy, and representation richness. These results validate the experimental design: the bottleneck regime isolates head architecture effects, while the frozen CNN control confirms that the observed performance relationships arise from fundamental architectural properties rather than feature capacity.

Entanglement: modest gains on simple features, reversal on rich features

Introducing entanglement through CNOT operations yields consistent and statistically significant improvements over product-state circuits on MNIST and FashionMNIST. The entangling classifier (Quantum-Ent) improves accuracy relative to the product-state variant (Quantum-NoEnt) by 0.64 percentage points on MNIST ($88.16\% \pm 0.62\%$ versus $87.52\% \pm 1.02\%$) and by 0.59 percentage points on FashionMNIST ($82.62\% \pm 0.52\%$ versus $82.03\% \pm 0.94\%$). These gains are consistent across all three random seeds on both datasets, indicating a reproducible representational benefit rather than a statistical fluctuation.

On CIFAR-10, however, the effect of entanglement weakens substantially. Quantum-Ent achieves $34.9\% \pm 0.3\%$ accuracy compared with $35.1\% \pm 0.7\%$ for the product-state model—a difference that is small, directionally inconsistent, and well within seed-to-seed variability. Thus, while entanglement reliably improves performance on simpler grayscale benchmarks, it fails to confer a measurable advantage under the higher visual complexity of CIFAR-10 within the shallow-circuit regime considered here.

These results demonstrate that entanglement can enable learning behavior unattainable by product-state circuits when operating on identical classical features, but that the magnitude of this benefit is both dataset-dependent and tightly constrained by circuit depth and measurement. On MNIST and FashionMNIST, the observed improvement (approximately 0.6 percentage points) corresponds to a 30–40% reduction in the error gap between product-state quantum models and quaternion baselines. Moreover, the reduced variance of Quantum-Ent relative to Quantum-NoEnt ($\sigma = 0.62\%$ versus 1.02% on MNIST and 0.52% versus 0.94% on FashionMNIST) suggests that entanglement partially stabilizes the optimization landscape in low-complexity regimes [21]. No comparable stabilization effect is observed on CIFAR-10. Although entangling gates are present, shallow depth and restricted observables limit how entanglement manifests in the loss, resulting in segmented representations that provide little advantage on higher-complexity datasets.

Implications for quantum machine learning research

These results have direct implications for near-term quantum machine learning research. First, they establish a classical performance floor for supervised learning on classical vision benchmarks. Any claim of quantum advantage must now exceed not only conventional real-valued baselines, but also parameter-efficient quaternion networks, which provide a stronger and more geometrically faithful classical comparator. This floor holds across a wide range of representation qualities—from minimal 16-dimensional learned bottlenecks to rich 512-dimensional pretrained embeddings—indicating that the quaternion baseline reflects fundamental properties of classical $SU(2)$ geometry rather than artifacts of weak feature extraction or limited representational capacity.

Second, the results substantially narrow the plausible sources of quantum advantage on classical data. Local $SU(2)$ structure alone is insufficient, and entanglement does not guarantee improved performance. Entanglement yields modest gains (approximately 0.6 percentage points) on MNIST and FashionMNIST under learned bottlenecks, provides no measurable benefit on CIFAR-10 under the same bottleneck, and actively degrades

performance by 9.3 percentage points when combined with high-quality pretrained CNN features. This progression indicates that shallow entanglement interacts adversely with rich classical representations, and that any quantum advantage on classical data—if achievable—must arise from mechanisms beyond shallow circuit depth, such as deeper trainable entanglement, alternative measurement strategies that mitigate information loss, or problem structures fundamentally misaligned with classical inductive biases.

Third, the findings suggest clear practical priorities for algorithm design. For practitioners with access to near-term quantum hardware, shallow variational circuits operating on classical features are unlikely to outperform compact classical alternatives on image classification tasks resembling MNIST, FashionMNIST, or CIFAR-10, regardless of whether features are learned or pretrained. The observed reversal of entanglement benefits under strong feature representations further cautions that naively adding quantum resources can degrade performance rather than improve it. Progress is therefore more likely to arise from (i) problems with intrinsic quantum structure, (ii) deeper circuits that remain reliably trainable while exploiting multi-qubit correlations, or (iii) hybrid classical–quantum architectures that preserve, rather than compress, information across the measurement interface.

Scope, limitations, and future directions

This study provides a controlled and reproducible demonstration that quaternion-valued architectures, as classical implementations of $SU(2)$ geometry, match or exceed the learning performance of shallow variational quantum classifiers on classical supervised learning tasks lacking intrinsic quantum structure. Across MNIST and FashionMNIST, quaternion heads match real-valued baselines in accuracy while substantially outperforming both product-state and entangled quantum circuits. Extending the analysis to CIFAR-10 under two representation regimes—a learned 16-dimensional bottleneck and frozen ImageNet-pretrained ResNet18 features—shows that this qualitative ordering persists under increased visual complexity and across a 32-fold increase in feature dimensionality (from 16 to 512 dimensions). Quaternion networks retain 94–96% of real-valued performance independent of feature quality, whereas quantum classifiers degrade by 5–6 percentage points under both learned and pretrained representations. Because all models operate on identical frozen feature extractors, these differences isolate the inductive bias and representational geometry of the classification heads rather than confounding effects of feature learning.

The present findings also connect naturally to the growing literature on quantum kernel methods. Prior work has shown that, on classical data, quantum kernels induced by parametrized circuits often fail to outperform well-chosen classical kernels, with performance primarily determined by data structure rather than quantum-specific effects [22]. From this perspective, the frozen-feature design employed here can be interpreted as a controlled comparison of kernel-induced representations: classical heads (real-valued and quaternion-valued) and quantum circuits act on identical fixed features, differing only in the geometry and parameterization of the induced transformation. The consistent underperformance of shallow variational quantum classifiers relative to quaternion and real-valued counterparts is therefore consistent with kernel-based analyses, which predict limited advantage for quantum feature maps on classical data lacking intrinsic quantum structure. Our results extend these observations beyond kernel alignment metrics to explicit supervised learning performance across multiple datasets and feature qualities. Whereas prior analyses focus primarily on kernel expressivity bounds and generalization theory, we provide direct empirical evidence that classical $SU(2)$ implementations outperform quantum circuits by 2.4–6.1 percentage points on standard classification benchmarks under matched conditions, with the performance gap persisting across both minimal learned bottlenecks and rich pretrained

features. Moreover, we establish that quaternion-valued neural networks—which share the local $SU(2)$ geometric structure of single-qubit gates—provide a stronger and more appropriate classical baseline than standard real-valued models when evaluating quantum advantage claims on classical vision tasks.

The conclusions are necessarily bounded by several design choices that define the scope of validity. First, all benchmarks considered (MNIST, FashionMNIST, and CIFAR-10) are classical vision datasets without inherent quantum structure, such as phase-sensitive features, interference effects, or many-body correlations. The benchmarks employed here are deliberately chosen as standard, well-characterized datasets that enable controlled comparison and reproducibility. While these are relatively low-resolution tasks with balanced classes and full supervision, they span a meaningful range of visual complexity, from simple grayscale digits to textured clothing items to natural color images with substantial background variation. The dual-regime experimental design on CIFAR-10, comparing learned 16-dimensional bottlenecks against rich 512-dimensional pretrained features, further distinguishes representation-quality effects from architectural properties. More challenging vision tasks—including fine-grained recognition, few-shot learning, imbalanced classification, or higher-resolution natural images—would provide additional empirical coverage, and the experimental methodology developed here (frozen features, matched seeds, systematic entanglement comparisons) transfers directly to such settings. However, the consistent qualitative ordering observed across datasets and feature regimes ($\text{RealNet} \gtrsim \text{QuatNet} \gg \text{Quantum-NoEnt} \gtrsim \text{Quantum-Ent}$) suggests that the core architectural relationships reflect intrinsic properties of shallow $SU(2)$ model parameterizations rather than artifacts of dataset-specific performance ceilings. Extension to more complex vision tasks remains valuable future work, particularly for assessing whether deeper quantum circuits or alternative measurement strategies can reverse the observed performance gaps. Problems with genuine quantum structure may therefore exhibit qualitatively different behavior, including regimes in which classical $SU(2)$ implementations are no longer sufficient.

Parameter counts are reported throughout for transparency but should not be interpreted as direct measures of model capacity, as parameters in different architectures (real-valued weights, unit quaternions, rotation angles) occupy different geometric spaces and contribute differently to expressivity. The frozen-feature experimental design controls for representational capacity by ensuring all models operate on identical inputs, rendering such comparisons secondary to performance under matched conditions.

Second, the quantum models are intentionally shallow. Depth-3 circuits with limited parameter counts are chosen to ensure trainability under current optimization methods rather than to maximize expressivity. The consistent underperformance of quantum circuits across all datasets and feature regimes—by 2.4–6.1 percentage points under learned bottlenecks and by 3.5 percentage points under high-quality pretrained features—suggests that shallow circuits are particularly susceptible to measurement-induced information loss and optimization difficulties, independent of representation strength. Deeper circuits, more aggressive data reuploading, or alternative parameterizations could in principle reduce this gap, provided that advances in training methods (such as quantum natural gradients [18], layer-wise training, or improved initialization strategies) can mitigate barren plateau phenomena [19]. Whether such improvements scale reliably remains an open question.

Third, the entanglement structure explored here is deliberately simple, relying on a fixed CNOT ring topology. While entanglement yields consistent but modest improvements on MNIST and FashionMNIST (approximately 0.6 percentage points), its benefit diminishes on CIFAR-10 under the learned bottleneck and reverses dramatically under ResNet18 features, where performance degrades by 9.3 percentage points relative to product-state circuits. This pattern is consistent with shallow entanglement

interacting adversely with high-dimensional classical representations, compounding measurement compression and optimization instability rather than compensating for limited circuit depth. Alternative entanglement patterns—such as all-to-all, hierarchical, or problem-adaptive connectivity—may alter this tradeoff, but at the cost of increased circuit depth, simulation expense, and susceptibility to noise. Whether more structured entanglement can reverse this degradation on rich features, or whether the adverse interaction persists at scale, remains an open empirical and theoretical question.

Fourth, the use of frozen classical feature extractors is both a strength and a limitation. It enables a clean comparison of head architectures while restricting the ability of quantum models to perform quantum feature extraction. The robustness of the observed ordering across both learned and pretrained feature regimes indicates that the performance gap reflects fundamental architectural properties rather than artifacts of representation quality or dataset simplicity. Nevertheless, end-to-end hybrid training may expose complementary roles for quantum preprocessing, although existing evidence suggests that classical components often dominate hybrid architectures in practice [23].

Finally, all quantum experiments are conducted using GPU-based simulation. While this allows precise control and reproducibility, it does not capture the noise, gate errors, and coherence limitations of near-term quantum hardware [24]. Such effects may further degrade variational quantum classifier performance or favor alternative algorithmic strategies not examined here.

Future work should extend this analysis to datasets with intrinsic quantum structure, such as quantum chemistry simulations encoded as classification problems, phase estimation tasks with superposition-based data, or problems exhibiting global symmetries that entanglement can exploit efficiently but classical product structures cannot. Additional directions include exploring deeper and more expressive circuit architectures under improved training schemes, systematically characterizing the interaction between entanglement and feature dimensionality, comparing quaternion networks with other hypercomplex models (including complex- and octonion-valued networks), and validating findings on real quantum hardware. Analytical results identifying when classical $SU(2)$ implementations are provably sufficient—and when they must fail—would provide a valuable complement to empirical studies.

These findings delineate a practically relevant regime in which classical quaternion models achieve near-equivalence with real-valued networks (retaining 94–96% of performance across feature qualities) while substantially outperforming shallow quantum classifiers (by 2.4–6.1 percentage points) as data complexity and representation strength increase. The unexpected reversal of entanglement’s benefit under high-quality pretrained features—from modest gains on simple datasets to substantial degradation on rich representations—establishes a novel empirical boundary that warrants theoretical explanation. Within this regime, quaternion architectures emerge as efficient classical alternatives to product-state quantum circuits on classical data, while leaving open the central question of when genuinely quantum resources can deliver unambiguous advantages beyond the reach of any classical $SU(2)$ -based model.

Conclusion

This work identifies a well-defined empirical regime in which classical $SU(2)$ models, implemented through quaternion-valued neural networks, match or exceed the performance of shallow variational quantum circuits on classical supervised learning tasks. Under shared frozen feature representations—ranging from minimal 16-dimensional learned bottlenecks to rich 512-dimensional ImageNet-pretrained embeddings—quaternion classifiers consistently match or closely approach real-valued baselines. On MNIST and FashionMNIST, quaternions achieve statistical equivalence

with real MLPs (differences below 0.2 percentage points). On CIFAR-10, quaternions retain 94–96% of real MLP performance across both learned and pretrained feature regimes, demonstrating robustness independent of representation quality.

Across all benchmarks and feature regimes, shallow variational quantum classifiers underperform quaternion networks by 2.4–6.1 percentage points despite dramatically higher computational cost and substantially greater variance across random seeds. Introducing entanglement yields modest gains on simple datasets with learned bottlenecks (approximately 0.6 percentage points on MNIST and FashionMNIST), provides no benefit on CIFAR-10 with learned bottleneck, and actively degrades performance by 9.3 percentage points on CIFAR-10 with high-quality pretrained features. This reversal shows that adding quantum resources does not guarantee improved performance and can instead impair learning when combined with rich classical feature representations. Local $SU(2)$ rotation geometry alone is therefore insufficient to confer a practical learning advantage on classical image data under current near-term training paradigms, and shallow entanglement can interact adversely with high-dimensional embeddings.

The performance gap between classical and quantum models arises from structural differences rather than parameter count. Quaternion networks preserve continuous intermediate representations throughout the forward pass, whereas quantum models must ultimately compress a high-dimensional quantum state into a small number of classical expectation values at measurement. This measurement-induced compression, together with optimization challenges associated with parameter-shift gradient estimation and architectural limits imposed by shallow circuit depth, constrains the effectiveness of variational quantum classifiers independent of feature quality. The persistence of quantum degradation across both minimal learned bottlenecks and rich pretrained representations confirms that these limitations are intrinsic to the shallow quantum architecture rather than artifacts of weak feature extraction.

The scope of these conclusions is necessarily bounded. Our experiments focus on classical vision benchmarks, shallow circuits, frozen feature extractors, and simulated quantum execution. Tasks with intrinsic quantum structure, deeper or more expressive circuits, end-to-end hybrid training, or improved quantum optimization methods may exhibit qualitatively different behavior. In particular, the entanglement reversal observed under high-quality features warrants systematic investigation to determine whether this adverse interaction persists at greater circuit depths or reflects a more fundamental boundary between shallow quantum processing and rich classical representations. These results do not preclude advantages from deeper circuits or richer measurement strategies; rather, they delimit the regime in which local $SU(2)$ structure alone governs performance.

Within the regime examined, quaternion networks emerge as efficient, stable, and expressive classical alternatives to shallow variational quantum circuits, retaining 94–96% of real MLP performance across feature qualities while substantially outperforming both product-state and entangled quantum models. More broadly, they provide a practically relevant reference point for disentangling geometric effects from genuinely quantum sources of expressivity, reframing the boundary between classical and quantum advantage as an empirical question rather than an assumption. The observed reversal of entanglement’s benefit under rich feature representations—from modest gains on simple datasets to substantial degradation—establishes a new empirical constraint that challenges prevailing intuitions about the universal utility of quantum resources and underscores the importance of controlled, geometry-aware comparisons in quantum machine learning research.

Disclosures

Competing Interests.

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions.

All authors contributed equally to the conceptualization, methodology, software development, analysis, visualization, and writing of this manuscript.

Use of Artificial Intelligence Tools.

Generative artificial intelligence tools were used to assist with prose refinement, $\text{\LaTeX}$ formatting, and figure layout. All analytical decisions, experimental design, code implementation, validation, and interpretation of results were performed by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

Ethics Approval and Consent to Participate.

This study does not involve human participants, human data, or animal subjects and therefore did not require ethical approval or informed consent.

Data and Code Availability.

All datasets used in this study are publicly available benchmark datasets. Code used to generate the results is available from the authors upon reasonable request.

References

1. Parcollet, T., Morchid, M., Linarès, G.: A survey of quaternion neural networks. *Artificial Intelligence Review* **53**, 2957–2982 (2020) <https://doi.org/10.1007/s10462-019-09752-1>
2. Parcollet, T., Ravanelli, M., Morchid, M., Linarès, G., Trabelsi, C., De Mori, R., Bengio, Y.: Quaternion Recurrent Neural Networks. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2019)
3. Gaudet, C.J., Maida, A.S.: Deep Quaternion Networks. In: *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2018). IEEE
4. Mandic, D., Goh, V.S.L.: *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. John Wiley & Sons, Chichester, UK (2009)
5. Kuipers, J.B.: *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace, and Virtual Reality*. Princeton University Press, Princeton, NJ (1999)
6. Schuld, M., Bocharov, A., Svore, K.M., Wiebe, N.: Circuit-Centric Quantum Classifiers. *Physical Review A* **101**(3), 032308 (2020)
7. Beer, K., Bondarenko, D., Farrelly, T., Osborne, T.J., Salzmann, R., Scheiermann, D., Wolf, R.: Training Deep Quantum Neural Networks. *Nature Communications* **11**(1), 808 (2020)

8. Sim, S., Johnson, P.D., Aspuru-Guzik, A.: Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies* **2**(12), 1900070 (2019) <https://doi.org/10.1002/qute.201900070>
9. Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Ahmed, S., Ajith, V., Alam, M.S., Alonso-Linaje, G., Narayanan, B.A., Asadi, A., et al.: PennyLane: Automatic differentiation of hybrid quantum-classical computations. arXiv preprint arXiv:1811.04968 (2018)
10. NVIDIA Corporation: cuQuantum SDK: A High-Performance Library for Quantum Circuit Simulation. <https://developer.nvidia.com/cuquantum-sdk> (2021)
11. Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., Gambetta, J.M.: Supervised learning with quantum-enhanced feature spaces. *Nature* **567**(7747), 209–212 (2019) <https://doi.org/10.1038/s41586-019-0980-2>
12. Schuld, M., Killoran, N.: Is quantum advantage the right goal for quantum machine learning? *PRX Quantum* **3**(3), 030101 (2022) <https://doi.org/10.1103/PRXQuantum.3.030101>
13. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)
14. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
15. Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical Report CSRG-TR-2009-1, University of Toronto, Toronto, Canada (2009)
16. Absil, P.-A., Mahony, R., Sepulchre, R.: *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ (2008)
17. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., Latorre, J.I.: Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020) <https://doi.org/10.22331/q-2020-02-06-226>
18. Stokes, J., Izaac, J., Killoran, N., Carleo, G.: Quantum natural gradient. *Quantum* **4**, 269 (2020) <https://doi.org/10.22331/q-2020-05-25-269>
19. McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. *Nature Communications* **9**(1), 4812 (2018) <https://doi.org/10.1038/s41467-018-07090-4>
20. Huembeli, P., Dauphin, A.: Characterizing the loss landscape of variational quantum circuits. *Quantum Science and Technology* **6**(2), 025011 (2021) <https://doi.org/10.1088/2058-9565/abbf6b> arXiv:2008.02785
21. Holmes, Z., Sharma, K., Cerezo, M., Coles, P.J.: Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum* **3**(1), 010313 (2022) <https://doi.org/10.1103/PRXQuantum.3.010313>
22. Huang, H.-Y., Kueng, R., Preskill, J.: Power of data in quantum machine

- learning. *Nature Communications* **12**(1), 2631 (2021) <https://doi.org/10.1038/s41467-021-22539-9>
23. Kölle, M., Maurer, J., Altmann, P., Sünkel, L., Stein, J., Linnhoff-Popien, C.: Disentangling quantum and classical contributions in hybrid quantum machine learning architectures. *arXiv preprint arXiv:2311.05559* (2023) [arXiv:2311.05559](https://arxiv.org/abs/2311.05559) [quant-ph]
 24. Preskill, J.: Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018) <https://doi.org/10.22331/q-2018-08-06-79>