

QiT: Quantum-Inspired Transformer for Visual Recognition Task

Badri N. Patro
Microsoft

Vijay S. Agneeswaran
Microsoft

Abstract

Quantum machine learning offers a compelling representational perspective: angle-encoded states inhabit Hilbert spaces in which periodic similarities and interactions can be expressed naturally. Realizing this perspective for visual recognition remains difficult, however, because present quantum neural networks are constrained by limited qubit counts, costly circuit simulation and measurement, noise, and unstable optimization on noisy intermediate-scale quantum devices. We investigate whether useful structural ideas from quantum models can instead be realized as scalable classical Transformer operations. We introduce QiT, a Quantum-inspired Transformer for vision tasks with three components: (i) angle-inspired encoding that maps image tokens to learned trigonometric Hilbert-space features analogous to quantum rotation-based state encoding; (ii) self-attention over these periodic features, inducing a classical cosine kernel approximated to quantum fidelity kernels; and (iii) gated multiplicative emulation, a trainable classical surrogate for interaction terms found in variational circuits. All components are differentiable tensor operations, so QiT claims neither quantum computation nor quantum speedup and retains the $\mathcal{O}(N^2D)$ attention complexity of a standard Vision Transformer. Across image-classification benchmarks, QiT is competitive with a matched classical Transformer while avoiding the severe runtime cost observed for a small simulated quantum Transformer. QiT-B reaches 78.3% ImageNet-1K top-1 accuracy with 45.7M parameters and 11.5 GFLOPs. These results position QiT as a scalable baseline for isolating and evaluating quantum-motivated inductive biases in visual recognition.

1. Introduction

Self-attention gives vision models a direct mechanism for relating distant image patches [10, 47]. Its query and key projections are linear, however, so nonlinear pairwise structure must emerge through depth and the surrounding feed-forward blocks. Quantum machine learning suggests a complementary design vocabulary: periodic angle encodings,

feature-space inner products, and multiplicative interactions produced by parameterized quantum circuits (PQCs) [4, 37]. Whether these ingredients are useful in a scalable classical vision model remains insufficiently separated from the question of whether a model runs on quantum hardware.

Current quantum neural networks face three obstacles to vision-scale evaluation. First, recurrent designs such as QRNNs [2] and QLSTMs [8] process tokens sequentially, limiting parallelism and making direct long-range interaction less efficient than self-attention. Second, fully quantum and hybrid Transformers remain difficult to scale on noisy intermediate-scale quantum (NISQ) hardware [32]: available qubit counts, feasible circuit depths, measurement noise, and repeated sampling constrain both model size and evaluation throughput [9, 19, 52]. Third, some parameterized circuit families develop barren plateaus, in which gradient concentration impedes optimization as the system grows [24]. Existing hybrid approaches face a dilemma: light hybrids delegate only a few layers to quantum circuits and gain little expressivity, while heavy hybrids attempt full quantum attention but cannot train at realistic sequence lengths. Consequently, quantum-attention studies are generally evaluated at substantially smaller scales than modern vision Transformers, and controlled large-scale comparisons remain scarce.

This work asks whether structural motifs associated with quantum models in Hilbert spaces can provide useful inductive biases in a Transformer that retains the scalability of classical deep learning. More precisely, can periodic angle features and multiplicative interactions be integrated into parallel visual attention without circuits, measurements, or assumptions of quantum computational advantage? We need an architecture that integrates quantum principles without quantum hardware. Figure 1 provides an overview of the resulting QiT encoder.

We propose QiT (Quantum-inspired Transformer), a classical vision architecture that translates three motifs from parameterized quantum models into differentiable tensor operations. First, *angle-inspired encoding* maps each token h to bounded angles $\theta(h) = \pi \tanh(hW_\theta + b)$ and trigonometric periodic features $\phi(h) = [\sin(\theta(h)), \cos(\theta(h))]$, analogous in form to quantum rotation-based data encoding. Second,

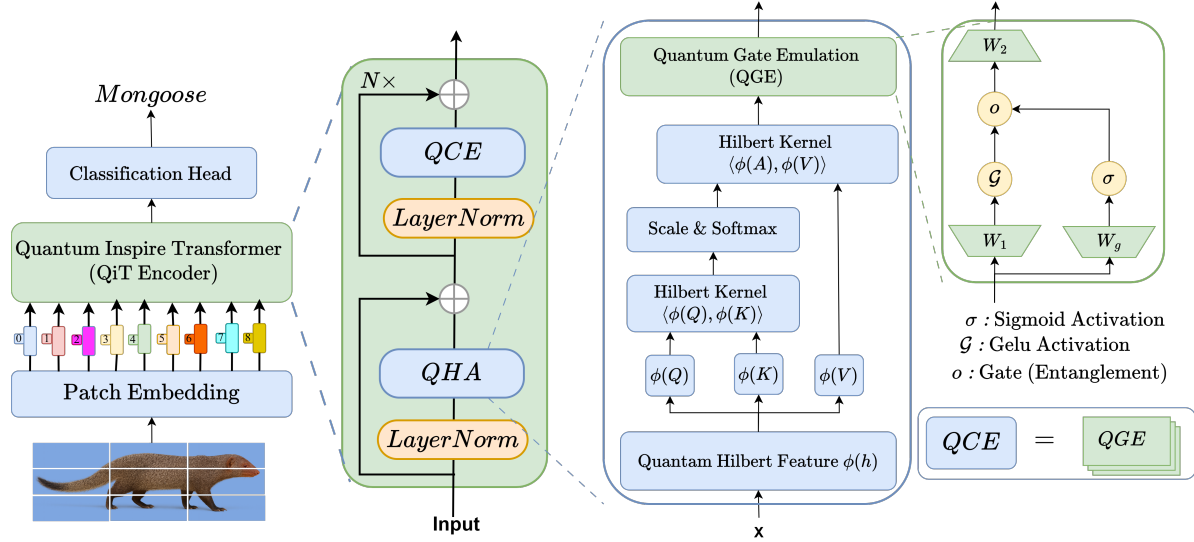


Figure 1. **QiT encoder block:** Tokens are mapped to a quantum-inspired Hilbert space via angle encoding and Fourier features. Kernel attention approximates quantum fidelity, while local quantum mixers and a circuit surrogate layer emulate bounded-depth PQCs within a Transformer architecture.

periodic-feature self-attention forms queries, keys, and values from learned projections of $\phi(h)$ in Hilbert-inspired feature space, allowing tokens to interact after a nonlinear periodic lift. Before projection, the inner product of these features is an explicit cosine kernel; after learned query and key projections, the attention score is a trainable bilinear similarity rather than a quantum fidelity kernel. Third, a *gated multiplicative mixer*, $(\text{GELU}(ZW_1) \odot \sigma(ZW_g))W_2$, introduces input-dependent channel interactions motivated by multiplicative terms in a variational quantum circuit (VQC). It does not instantiate or simulate physical entanglement.

For image tokens $X \in \mathbb{R}^{B \times N \times D}$ and labels y , QiT solves standard supervised classification by minimizing cross-entropy for $f_\Theta(X)$. The design reverses the usual hybrid construction: rather than executing classical attention through quantum circuits, it isolates selected quantum-motivated functional forms and evaluates them with conventional GPU kernels. QiT consequently preserves parallel token processing, standard backpropagation, and the $\mathcal{O}(N^2D)$ attention complexity of a Vision Transformer while avoiding circuit simulation, repeated measurement, and quantum-hardware constraints. The resulting model enables these inductive biases to be evaluated on vision benchmarks at a scale that is currently impractical for circuit-based attention.

Contributions.

1. We introduce QiT, a vision Transformer that integrates learned angle-inspired features, self-attention over their quantum-inspired Hilbert-space embeddings, and a gated multiplicative channel mixer. QiT requires no quantum hardware and retains the parallel execution and $\mathcal{O}(N^2D)$ attention complexity of a standard Vision Transformer.
2. We characterize the trigonometric feature map as an ex-

plicit positive-semidefinite cosine kernel, connecting QiT to periodic feature methods and rotation-based quantum data encoding. We state the boundary of this connection explicitly: learned query-key scores are bilinear similarities, not general quantum fidelity kernels, and gated mixing does not realize physical entanglement.

3. We evaluate QiT across image-classification benchmarks and compare it with both classical vision models and a four-qubit PennyLane Transformer. In the controlled one-epoch simulator study, QiT reduces runtime by nearly three orders of magnitude; at scale, QiT-B reaches 78.3% ImageNet-1K top-1 accuracy with 45.7M parameters and 11.5 GFLOPs.

2. Related Work

Transformers and Long-Range Sequence Modeling.

Transformers enable parallel, content-dependent token interaction across language and vision [10, 13, 47]. Vision alternatives include spectral token mixers [26, 30, 35] and state-space models [12, 23, 27–29, 49]. QiT retains softmax self-attention and changes the feature map supplied to its projections and the subsequent channel mixer.

Quantum Machine Learning. Quantum machine learning (QML) encodes classical data into quantum states, applies parameterized quantum circuits (PQCs), and measures observables to produce classical outputs [4, 14, 37]. A register of Q qubits represents states in a 2^Q -dimensional Hilbert space, but this dimensionality alone does not imply an efficiently accessible learning advantage. Quantum feature maps can induce kernels through state overlaps and generate input-dependent interactions whose utility depends on the en-

coding, circuit, measurement, and data [11, 14]. Variational quantum algorithms combine parameterized circuits with classical optimization, often favoring shallow circuits for near-term devices [6, 7, 25]. Their practical scale remains constrained by qubit availability, gate errors, decoherence, and finite-shot estimation on noisy intermediate-scale quantum hardware [32]. Moreover, barren plateaus can cause gradients to concentrate near zero for particular circuit families, objectives, and initialization regimes [24].

Quantum Deep Learning for Sequential Data. Quantum deep learning develops trainable quantum analogues of neural layers and studies their expressivity and optimization [1, 3, 17, 25, 33]. For sequential data, QRNNs and related recurrent models update quantum registers step by step [2, 20, 21, 43]; QLSTMs similarly incorporate PQCs into recurrent gating structures [8, 39]. Linear-layer variants and other hybrid modifications seek more stable or efficient training [5]. Data re-uploading and unitary or variational constructions provide additional routes to expressive circuit functions [31, 36, 38, 41]. These models demonstrate that PQCs can process ordered data, but recurrent execution remains sequential and does not provide Transformer-style all-pairs token interaction. QiT instead studies parallel interactions among image patches and executes all operations classically.

Quantum Attention and Quantum Transformers. Quantum attention methods differ in which Transformer operations are assigned to circuits. Light hybrid models such as QSANN [19] and QSAM [51] use circuits for selected projections while retaining classical attention operations. More circuit-intensive approaches, including QViT [9] and QKSAN [52], incorporate quantum subroutines or kernels into visual attention, increasing circuit-evaluation and measurement demands. QSAN [40] and F-QSANN [53] pursue fully quantum formulations, whereas HQViT [50] and quantum self-attention for molecular generation [42] investigate other hybrid allocations. Their encodings, circuit access assumptions, sampling budgets, and output readouts differ, so asymptotic quantities are not uniformly comparable and do not by themselves establish end-to-end speedups.

QiT occupies a distinct point in this landscape: it assigns no operation to quantum hardware. Instead, it translates angle-inspired periodic encoding and circuit-motivated multiplicative interactions into classical tensor operations, then applies standard softmax attention. This design enables conventional GPU training at vision scale, but its efficiency is classical and should not be interpreted as quantum computational advantage.

Positioning of QiT. QiT is neither a quantum Transformer nor a simulator of a general PQC. It isolates three design

components—learned periodic token features, attention over their projections, and gated multiplicative channel mixing—within a classical Transformer. The unprojected feature inner product is an explicit cosine kernel, whereas the learned query–key score is generally neither a positive-semidefinite kernel nor quantum-state fidelity. Likewise, multiplicative gating supplies trainable feature interactions but does not create physical entanglement. This positioning makes the quantum connection precise enough to test while permitting direct GPU evaluation. Unlike fully quantum architectures, QiT does not require quantum hardware, avoids barren plateaus and measurement noise, and scales to long sequences and large datasets. Unlike prior hybrid models that use PQCs as black-box nonlinear layers or rely on quantum kernel estimation. The small PennyLane experiment compares practical execution costs in a controlled setting; the large-scale vision results evaluate QiT as a classical architecture rather than evidence of quantum advantage.

3. Quantum-Inspired Transformer Architecture (QiT)

QiT is a classical Transformer that uses periodic token features and gated multiplicative channel mixing. These choices are motivated by angle encoding and interaction terms in parameterized quantum circuits (PQCs), but QiT neither executes nor generally simulates a PQC. We first define the implemented operations and then state the limited sense in which they relate to quantum models.

3.1. Background: Classical Multi-Head Self-Attention

Given token embeddings $H \in \mathbb{R}^{B \times L \times d}$ (batch size B , sequence length L , embedding dimension d), classical multi-head self-attention computes:

$$Q = HW_Q, \quad K = HW_K, \quad V = HW_V, \quad (1)$$

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right), \quad \text{Attn}(H) = AV, \quad (2)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_h}$ project to queries, keys, and values, and d_h is the per-head dimension. Multi-head attention parallelizes this across n_h heads with independent parameters, concatenating outputs for expressivity. This mechanism enables *direct pairwise token interactions* $i \leftrightarrow j$ without sequential processing, *parallel computation* via batched matrix operations, and *flexible multivariate integration* through embeddings—advantages over MLPs (no temporal bias without engineered features), CNNs/TCNs (indirect long-range modeling requiring depth), and RNNs/LSTMs (sequential bottlenecks, gradient degradation). However, attention incurs $\mathcal{O}(L^2)$ cost and standard linear projections lack the high-dimensional feature space expressivity promised by quantum approaches.

Table 1. Reported or asymptotic resource characteristics of quantum-attention designs. Values follow different circuit, oracle, sampling, and readout assumptions and are therefore descriptive rather than directly comparable. N denotes the number of tokens, D the embedding dimension, and Q the number of qubits. PQG denotes parameterized quantum gates.

Model	#Qubits	#PQG	#Distinct Circuits	#Total Measurements	Quantum Outputs	Reported Classical Cost	Execution Setting
CViT (Classical)	0	0	0	0	–	$O(N^2 D)$	Classical GPU
QSANN [19]	$O(D)$	$O(D)$	$O(N)$	$O(ND)$	$\{q_i\}_{i=1}^N, \{k_i\}_{i=1}^N, \{v_i\}_{i=1}^N$	$O(ND^2)$	Hybrid/simulated
QSAM [51]	$O(D)$	$O(D)$	$O(N)$	$O(ND)$	$\{q_i\}_{i=1}^N, \{k_i\}_{i=1}^N, \{v_i\}_{i=1}^N$	$O(ND^2)$	Hybrid/simulated
HQViT [50]	$O(\log ND)$	$O(\log D)$	$O(1)$	$O(N^2 \log D)^*$	$\{A_{ij}\}_{i,j=1}^N, \{v_i\}_{i=1}^N$	$O(N^2 D)$	Hybrid proposal
QViT [9]	$O(N + D)$	$O(D \log D)$	$O(N)$	$O(ND)^*$	$\{y_i\}_{i=1}^N$	$O(N^2 D)$	Hybrid/simulated
QKSAN [52]	$O(D)$	$O(D)$	$O(N^2)$	$O(N^2 D)^*$	$\{v_i\}_{i=1}^N$	$O(N^2 D)$	Hybrid/simulated
QSAN [40]	$O(N \log D + N^2)$	$O(N \log D)$	$O(1)$	$O(\log D)$	$\mathbf{y}$	$O(N^2 D)$	Fully quantum proposal
F-QSANN [53]	$O(N \log D)$	$O(N^2 \log D)$	$O(1)$	$O(1)$	Class probability	$O(N^2 D)$	Fully quantum proposal
QiT (Ours)	0	0	0	0	Classical features	None	Classical GPU

3.2. Token Embedding (Classical Preprocessing)

Given an input sequence $X \in \mathbb{R}^{B \times L \times D}$ with batch size B , sequence length L , and input dimension D , we construct token embeddings $H \in \mathbb{R}^{B \times L \times d}$ using standard embedding layers that combine value embeddings, positional encodings, and optional temporal covariates. This preprocessing step follows conventional Transformer practice and costs $\mathcal{O}(BLD)$ operations. The quantum-inspired components begin after embeddings are formed.

3.3. Quantum Angle Encoding

In quantum computing, angle encoding prepares quantum states by rotating qubits according to input features. We emulate this classically by mapping each token embedding $h \in \mathbb{R}^d$ to rotation angles:

$$\theta(h) = \pi \tanh(Wh + b), \quad \theta(h) \in \mathbb{R}^Q, \quad (3)$$

where Q is the number of learned angles (virtual qubits) and $W \in \mathbb{R}^{Q \times d}$ and $b \in \mathbb{R}^Q$ are trainable parameters. We use Q to retain the connection to a Q -qubit angle encoding, although no qubits are instantiated. The $\tanh$ nonlinearity bounds the angles to $[-\pi, \pi]$ matching the periodicity of quantum rotation gates. We then construct a trigonometric feature map in Hilbert space:

$$\phi(\mathbf{h}) = \frac{1}{\sqrt{Q}} [\sin(\theta(\mathbf{h})), \cos(\theta(\mathbf{h}))] \in \mathbb{R}^{2Q}. \quad (4)$$

This embedding corresponds to a Fourier feature representation and induces an implicit positive-semidefinite kernel:

$$k(\mathbf{h}, \mathbf{h}') = \langle \phi(\mathbf{h}), \phi(\mathbf{h}') \rangle = \frac{1}{Q} \sum_{q=1}^Q \cos(\theta_q(\mathbf{h}) - \theta_q(\mathbf{h}')). \quad (5)$$

Positive semidefiniteness follows directly because k is an inner product. The map is computed with classical trigonometric functions at cost $\mathcal{O}(BLdQ)$. For restricted product-state angle encodings, related trigonometric terms occur in state overlaps; this observation does not make k equal to the fidelity kernel of a general entangling circuit.

Connection to quantum feature maps. Angle-encoded quantum models also produce periodic functions of their inputs. QiT retains this periodic structure in a $2Q$ -dimensional real feature space, rather than representing a 2^Q -dimensional state vector. Its connection to quantum feature maps is therefore architectural and spectral, not an equivalence between the represented state spaces.

3.4. Hilbert-Space Self-Attention

Standard Transformer attention computes query-key similarities in the original embedding space. QiT instead performs attention in the quantum-inspired feature space defined by angle encoding. For each token $h_i \in \mathbb{R}^d$, we first apply the trigonometric mapping:

$$\theta_i = \pi \tanh(Wh_i + b) \in \mathbb{R}^Q, \quad (6)$$

$$\phi(h_i) = [\sin \theta_i, \cos \theta_i] \in \mathbb{R}^{2Q}. \quad (7)$$

We then form queries, keys, and values by projecting these features:

$$q_i = \phi(h_i)W_Q, \quad k_i = \phi(h_i)W_K, \quad v_i = \phi(h_i)W_V, \quad (8)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{2Q \times d_h}$ are learnable projection matrices and d_h is the per-head dimension. Attention weights follow the standard scaled dot-product formula:

$$A_{ij} = \frac{\exp(q_i^\top k_j / \sqrt{d_h})}{\sum_{j'} \exp(q_i^\top k_{j'} / \sqrt{d_h})}, \quad (9)$$

and the attention output is:

$$\text{Attn}(Q, K, V) = AV. \quad (10)$$

Before projection, $\phi(h_i)^\top \phi(h_j)$ is the cosine kernel above. After applying independently learned query and key projections, $q_i^\top k_j$ is a learned bilinear similarity in that feature space; it need not be symmetric or positive semidefinite and should not be interpreted as quantum-state fidelity. Figure 2 contrasts these two kernel constructions. For all heads, feature projection costs $\mathcal{O}(BLQd)$ and attention costs $\mathcal{O}(BL^2d)$.

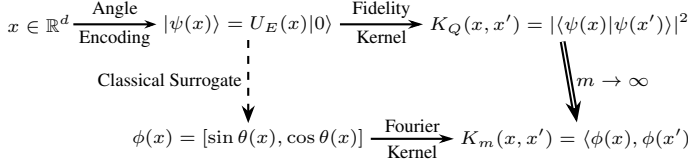


Figure 2. Two distinct routes from angle-encoded inputs. QiT uses the explicit cosine feature kernel at the bottom; it does not generally approximate the circuit-dependent fidelity kernel at the top.

Scope of the circuit interpretation. A circuit-based attention mechanism could encode a token as a normalized state

$$|\psi(h)\rangle = U_E(h)|0\rangle, \quad \langle\psi(h)|\psi(h)\rangle = 1, \quad (11)$$

and to form query and key states $|\psi_Q(h_i)\rangle = U_Q|\psi(h_i)\rangle$ and $|\psi_K(h_j)\rangle = U_K|\psi(h_j)\rangle$, where $U_Q(\theta_Q)$ and $U_K(\theta_K)$ be unitary families. Their fidelity score and normalized attention weight would be

$$s_{ij} = |\langle\psi_Q(h_i)|\psi_K(h_j)\rangle|^2 \in [0, 1]. \quad (12)$$

Then for any temperature $\tau > 0$, the attention weights

$$A_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{j'} \exp(s_{ij'}/\tau)} \quad (13)$$

form a valid probability distribution over j and induce an attention operator

$$\text{Attn}(h_i) = \sum_{j=1}^L A_{ij} v_j. \quad (14)$$

Moreover, s_{ij} defines a positive-definite kernel over token embeddings. Unitarity of U_E , U_Q , and U_K ensures all states are normalized. Unitarity guarantees normalized states, Hence $s_{ij} \in [0, 1]$ and $\exp(s_{ij}/\tau) > 0$, and softmax guarantees $A_{ij} > 0$ and so $\sum_j A_{ij} = 1$. When the same state map is used on both arguments, fidelity is a positive-semidefinite kernel because $|\langle\psi_i|\psi_j\rangle|^2 = \text{tr}(\rho_i \rho_j)$ for $\rho_i = |\psi_i\rangle\langle\psi_i|$. With distinct query and key maps, however, s_{ij} need not be symmetric and therefore is not generally a kernel.

QiT neither estimates these fidelities nor approximates them through an overlap routine such as a swap test. It uses the finite real feature map ϕ and learns task-dependent query and key projections before softmax attention. The shared periodic dependence on encoded angles motivates the design, while the learned projections allow the visual recognition objective to determine the similarity. QiT attention costs $\mathcal{O}(BL^2 d_h)$ per head, as in scaled dot-product attention. Its empirical behavior is therefore evidence about periodic classical features, not quantum-state fidelity or quantum advantage.

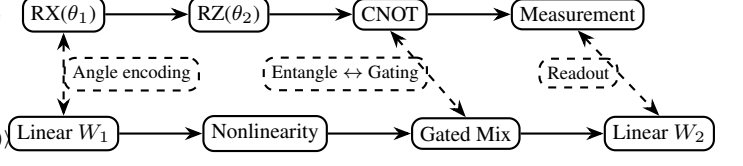


Figure 3. Design analogy between a parameterized circuit and the QiT channel mixer. The correspondence motivates the sequence of operations but does not imply equivalent states or functions.

3.5. Quantum-Inspired Gated Channel Mixer

3.5.1. Design rationale

Parameterized quantum circuits use two types of operations: single-qubit rotation gates that depend on learnable parameters, and multi-qubit entangling gates like CNOT that create non-separable correlations between qubits. A quantum circuit with parameters $\Theta = \{\theta_1, \dots, \theta_p\}$ computes:

$$f_\Theta(x) = \langle 0 | U_\Theta^\dagger(x) O U_\Theta(x) | 0 \rangle, \quad (15)$$

where x is the classical input, $U_\Theta(x)$ applies encoding and variational gates, and O is a measurement observable. Entangling gates can produce states and measurement correlations that cannot be factored into independent single-qubit contributions. Motivated by this sequence of rotation, interaction, and readout, we use a classical gated block. It introduces multiplicative feature interactions without representing a quantum state.

3.5.2. Gated multiplicative mixing

Given a token representation X , the channel mixer applies

$$H_1 = \text{GELU}(XW_1), \quad (16)$$

$$G = \sigma(XW_g), \quad (17)$$

$$\text{Mix}(X) = (H_1 \odot G)W_2, \quad (18)$$

where $\odot$ denotes elementwise multiplication, W_1 and W_g map to hidden width r , and W_2 projects back to width d . The Hadamard product creates input-dependent, non-additive feature interactions. Such products are qualitatively reminiscent of interaction terms in circuit expectation values, but no one-to-one correspondence to an entangling gate is assumed. The computational cost is $\mathcal{O}(BLdr)$, matching a gated Transformer feed-forward block of hidden width r .

3.5.3. Architectural interpretation

Figure 3 illustrates the mapping between quantum circuit elements and QiT components. Angle encoding emulates single-qubit rotations (R_X , R_Z), gated mixing emulates two-qubit gates (CNOT), and multi-head attention corresponds to parallel measurements. Residual connections act as classical analogs of quantum error mitigation by providing gradient highways that stabilize training. Angle encoding motivates

Table 2. One-epoch CIFAR simulator diagnostic with patch size 4, one block, and batch size 2. $Q=4$ denotes learned angles for QiT and simulated qubits for FQT-P. FLOPs are recorded for CViT and QiT; FQT-P is listed symbolically because its cost depends on circuit and simulator details. Accuracy values are single-run, non-converged measurements. FQT-P = Full Quantum Transformer in PennyLane.

Model	Classes	Angles/Qubits	Params	FLOPs	Energy (J)	Acc. (%)	Latency (s)	Wall-clock (s)
CViT	2	0	147,458	8,996,352	9.60	50.00	0.0010	63.72
QiT	2	4 angles	324,690	20,756,672	10.76	86.35	0.0016	102.91
FQT-P	2	4 qubits	16,394	$\mathcal{O}(N^2 2^Q)$	23,904.70	81.20	2.4987	90,745.75
CViT	4	0	147,716	8,996,608	9.40	25.00	0.0008	127.97
QiT	4	4 angles	324,948	20,756,928	10.64	71.98	0.0015	202.02
FQT-P	4	4 qubits	16,652	$\mathcal{O}(N^2 2^Q)$	13,697.42	62.42	2.3363	178,606.09
CViT	10	0	147,490	8,997,376	5.60	10.00	0.0007	253.50
QiT	10	4 angles	325,722	20,757,696	6.89	50.39	0.0013	447.23
FQT-P	10	4 qubits	17,426	$\mathcal{O}(N^2 2^Q)$	13,641.24	32.33	2.3174	445,765.80

Table 3. Unified asymptotic comparison of Transformer variants. N = tokens, D = embedding dim, Q = qubits, L = circuit depth. Runtime slowdown is relative to CViT.

Model	Token mixing (attn.)	Channel mixer	Total complexity	Slowdown	Feasibility (CIFAR)	Acc
CViT	$\mathcal{O}(N^2 D)$	$\mathcal{O}(ND^2)$	$\mathcal{O}(N^2 D)$	$1\times$	Fully trainable	0.50
QiT (surrogate)	$\mathcal{O}(N^2 D)$	$\mathcal{O}(\mathbf{QL})$	$\mathcal{O}(N^2 D)$	$1.1\times$	Fully trainable	0.86
FQT-P (PennyLane)	$\mathcal{O}(N^2 \cdot 2^Q)$	$\mathcal{O}(QL)$	$\mathcal{O}(N^2 \cdot 2^Q)$	$\sim 100\times$	Slow but feasible	0.81
FQT-Q (Qiskit)	$\mathcal{O}(N^2 \cdot 2^Q)$	$\mathcal{O}(QL)$	$\mathcal{O}(N^2 \cdot 2^Q)$	$\sim 1000\times$	Infeasible at CIFAR scale	–

periodic features, multiplicative gating supplies feature interactions, and a linear layer provides readout. Residual connections and normalization follow standard Transformer practice. Because QiT uses a finite classical feature map, it does not inherit the exponential state dimension or the full function class of a quantum circuit.

4. Experiments

We organize the evaluation around three questions. **(Q1) Simulator cost:** under one controlled CIFAR protocol, how does a four-qubit PennyLane implementation compare with classical execution? **(Q2) Portability:** can one QiT configuration train across datasets with different resolutions and class counts? **(Q3) ImageNet scale:** can larger QiT variants be trained on ImageNet-1K, and where do their accuracy and computational cost lie relative to published vision architectures? These experiments evaluate a classical model and do not test quantum advantage.

4.1. Experimental Setup

Datasets. The portability study comprises eight datasets. CIFAR-10, SVHN, and CIFAR-100 use 32×32 images; STL-10 uses 96×96 images; Oxford-IIIT Pets, Food-101, Flowers-102, and Caltech-256 use 224×224 images. They span 10 to 257 classes and 3,680 to 75,750 training images, as detailed in Table 5. ImageNet-1K is evaluated separately with 1.2M training images, 1,000 classes, and 224×224 inputs. The simulator diagnostic uses reduced 2-, 4-, and 10-class CIFAR tasks; these are distinct from the 300-epoch CIFAR-10 result in the portability study.

Protocols. The three studies use separate settings. For Q1, CViT, QiT, and FQT-P use patch size 4, one block, batch size 2, and one training epoch; QiT and FQT-P use $Q = 4$, where Q denotes learned angles for QiT and qubits for FQT-P. For Q2, a fixed QiT architecture uses patch size 4, depth 12, width 256, four attention heads, and $Q = 16$ learned angles. It is trained for 300 epochs with AdamW, cosine learning-rate decay, random cropping, horizontal flipping, and color jitter on four Tesla T4 GPUs; per-GPU batch sizes appear in Table 5. For Q3, the QiT-XS/S/B variants use the parameter and FLOP budgets reported in Table 4. Results must be interpreted within each protocol rather than compared across protocols.

Reporting scope. All tables report point estimates from the available runs. The current records do not include repeated seeds, uncertainty intervals, complete optimizer hyperparameters, reduced-class CIFAR split definitions, or the energy-measurement procedure. We therefore use the results to document observed accuracy, runtime, and scalability, not statistical superiority. A final reproducible submission should provide these missing details and repeated-run confidence intervals.

4.2. Comparison with Simulated Quantum Transformers

Table 2 compares a classical Vision Transformer (CViT), QiT, and a four-qubit Full Quantum Transformer simulated in PennyLane (FQT-P) under the Q1 protocol. In this table, “angles/qubits” distinguishes QiT’s four learned angles from FQT-P’s four simulated qubits.

Table 4. ImageNet-1K top-1 accuracy, GFLOPs, and Parameters are reported for representative architectures. Results are drawn from their respective papers and use different training recipes; the table provides scale and accuracy context rather than controlled pairwise comparisons. QiT rows are highlighted.

Network	Size	Params	FLOPs	Top-1
CNN				
ResNet-50 [15]	224 ²	25.5M	4.1G	78.3
ResNeXt101 [48]	224 ²	83.5M	15.6G	81.5
EfficientNet-B5 [44]	456 ²	30M	9.9G	83.6
Transformer-MLP Mixer				
gMLP-Ti [22]	224 ²	6M	1.4G	72.0
gMLP-S [22]	224 ²	20M	4.5G	79.4
gMLP-B [22]	224 ²	73M	15.8G	81.6
ResMLP-12 [46]	224 ²	15M	3.0G	76.6
ResMLP-36 [46]	224 ²	45M	8.9G	79.7
Transformer-Spectral				
Fnet-S [18]	224 ²	15M	2.9G	71.2
GFNet-Ti [34]	224 ²	7M	1.3G	74.6
GFNet-XS [34]	224 ²	16M	2.9G	78.6
GFNet-S [34]	224 ²	25M	4.5G	80.0
SpectFormer-XS [30]	224 ²	20M	4.0G	80.2
SpectFormer-S [30]	224 ²	32M	6.6G	81.7
SpectFormer-B [30]	224 ²	57M	11.5G	82.1
Transformer-State-Space				
Vim-Ti [54]	224 ²	7M	1.5G	76.0
Vim-S [54]	224 ²	26M	5.1G	80.4
LocalVMamba-T [16]	224 ²	26.0M	5.7G	82.7
LocalVMamba-S [16]	224 ²	50.0M	11.4G	83.7
SiMBA-S [27]	224 ²	15.3M	2.4G	81.7
SiMBA-B [27]	224 ²	22.8M	4.2G	83.5
VMamba-T [23]	224 ²	31.0M	4.9G	82.5
VMamba-S [23]	224 ²	50.0M	8.7G	83.6
Transformer-Attention				
ViT-B/16 [10]	384 ²	86M	55.5G	77.9
DeiT-Ti [45]	224 ²	5M	1.2G	72.2
DeiT-S [45]	224 ²	22M	4.6G	79.8
DeiT-B [45]	224 ²	86M	17.5G	81.8
Transformer-Quantum-Inspired (ours)				
QiT-XS	224 ²	18.1M	3.8G	75.0
QiT-S	224 ²	31.0M	5.6G	77.4
QiT-B	224 ²	45.7M	11.5G	78.3

Runtime and latency. Across 2, 4, and 10 classes, FQT-P requires 90,745.75, 178,606.09, and 445,765.80 seconds, respectively, versus 102.91, 202.02, and 447.23 seconds for QiT. The corresponding FQT-P/QiT wall-clock ratios are 881.8 $\times$, 884.1 $\times$, and 996.7 $\times$. Reported inference latency

shows a similarly large gap of 1557.5–1782.6 $\times$. QiT is not cost-free relative to CViT: it uses about 2.30 $\times$ as many recorded FLOPs and 2.20 $\times$ as many parameters in this setup, yielding a 1.58–1.76 $\times$ wall-clock increase. Thus the pilot demonstrates a large implementation-cost gap between classical tensor execution and this PennyLane simulation, while also quantifying QiT’s overhead over the small CViT.

Accuracy and interpretation. After one epoch, QiT records 86.35%, 71.98%, and 50.39% accuracy for 2, 4, and 10 classes; FQT-P records 81.20%, 62.42%, and 32.33%. CViT remains exactly at chance (50%, 25%, and 10%), indicating that this short protocol is unsuitable for assessing its converged performance. The observed QiT–FQT-P differences are therefore optimization snapshots, not evidence of architectural superiority. Parameter counts are also unmatched: QiT has roughly 325K parameters, CViT 147K, and FQT-P 16–17K. The defensible conclusion is computational: the tested simulator does not provide a practical route to the longer and larger studies below.

Energy measurements. FQT-P’s recorded energy is three orders of magnitude above the classical runs, but the measurement procedure, simulator implementation, and device utilization are undocumented. We report these values for completeness and draw no hardware-independent energy-efficiency conclusion from them.

Asymptotic complexity analysis. Table 3 separates asymptotic complexity from measured runtime. QiT’s periodic feature projection costs $\mathcal{O}(NQD)$ per layer, attention costs $\mathcal{O}(N^2D)$, and a gated mixer of hidden width r costs $\mathcal{O}(NDr)$. Its total layer cost is therefore $\mathcal{O}(N^2D + NQD + NDr)$, retaining quadratic dependence on token count. CViT additionally incurs its channel-MLP cost $\mathcal{O}(ND^2)$. FQT-P complexity is circuit- and simulator-dependent; a dense state-vector simulator stores 2^Q amplitudes, but the measured slowdown cannot be generalized to other simulation methods or quantum hardware. We omit the proposed Qiskit row because no Qiskit experiment is recorded in the available results.

4.3. Context from Quantum Attention Models

Table 1 summarizes reported resource characteristics for QSANN [19], QSAM [51], HQViT [50], QViT [9], QK-SAN [52], QSAN [40], and F-QSANN [53]. Circuit architecture, encoding, measurements, and hardware assumptions differ across these works, so the entries provide qualitative context rather than a normalized runtime ranking. QiT has zero quantum-resource requirements by construction.

Table 5. Single-run QiT results across eight image datasets. All runs use four Tesla T4 GPUs, patch size 4, depth 12, width 256, four heads, $Q=16$ learned angles, and 300 epochs. Batch is per GPU; latency is reported per image.

Dataset	Classes	Image Size	Train	Batch/GPU	Epochs	Params (M)	GFLOPs	Acc (%)	Wall clock	Inference (ms)
CIFAR-10	10	32×32	50,000	128	300	5.48	0.355	78.84	$\sim 1.74\text{h}$	14.50
SVHN	10	32×32	73,257	128	300	5.48	0.355	95.94	$\sim 2.55\text{h}$	14.66
STL-10	10	96×96	5,000	32	300	5.61	3.386	63.50	$\sim 2.07\text{h}$	14.65
Pets	37	224×224	3,680	4	300	6.27	24.58	71.23	$\sim 20.93\text{h}$	77.18
CIFAR-100	100	32×32	50,000	128	300	5.50	0.355	51.52	$\sim 1.75\text{h}$	14.39
Food-101	101	224×224	75,750	4	300	6.29	24.58	72.00	$\sim 78.34\text{h}$	80.12
Flowers-102	102	224×224	6,149	4	300	6.29	24.65	65.00	$\sim 6.77\text{h}$	79.31
Caltech-256	257	224×224	24,000	4	300	6.33	24.58	68.23	$\sim 23.74\text{h}$	79.98

4.4. Scalability to ImageNet

Table 4 places QiT-XS/S/B alongside published ImageNet-1K results. Within the QiT family, increasing from XS to S raises top-1 accuracy from 75.0% to 77.4% while increasing compute from 3.8 to 5.6 GFLOPs. Scaling from S to B adds 14.7M parameters and 5.9 GFLOPs for a further 0.9 percentage-point gain, reaching 78.3%. This monotonic but diminishing gain indicates that the architecture remains trainable as capacity grows, although it does not establish optimal scaling.

Across published models, QiT-B matches the listed ResNet-50 accuracy and exceeds the listed ViT-B/16 result by 0.4 points with fewer reported FLOPs, but these are not controlled comparisons because input resolution, augmentation, optimization, and pretraining may differ. QiT trails DeiT-B by 3.5 points and the strongest spectral and state-space entries by 3.8–5.4 points. Accordingly, Table 4 supports ImageNet-scale feasibility and locates the current accuracy–cost trade-off; it does not support a state-of-the-art claim.

4.5. Results Across Multiple Benchmarks

Table 5 evaluates one QiT configuration without dataset-specific changes to depth, width, head count, or angle count. For 32×32 inputs, the model uses 5.48–5.50M parameters and 0.355 GFLOPs, obtaining 78.84% on CIFAR-10, 95.94% on SVHN, and 51.52% on CIFAR-100. The spread across these values reflects different tasks and class structures, so it should not be interpreted as a direct ranking of dataset difficulty. Reported latency is consistent within this resolution group at 14.39–14.66 ms per image.

Increasing resolution changes the resource profile more than the classifier size. STL-10 at 96×96 requires 3.386 GFLOPs and reaches 63.50%. At 224×224 , the four datasets require 24.58–24.65 GFLOPs and 6.27–6.33M parameters, with accuracy ranging from 65.00% on Flowers-102 to 72.00% on Food-101. Their reported latency lies between 77.18 and 80.12 ms per image. Relative to the 32×32 runs, the 224×224 configuration uses about $69 \times$ more FLOPs but approximately $5.4 \times$ the measured latency, consistent with different levels of GPU utilization across

input sizes rather than proportional latency scaling.

Wall-clock time varies with both image resolution and training-set size: Food-101 is the longest run at approximately 78.34 hours, while the smaller 32×32 datasets finish in under 2.6 hours. Taken together, the table establishes that the same QiT block can be optimized across substantial changes in resolution, sample count, and class count. Because no matched baseline or component ablation is reported under this 300-epoch protocol, the table cannot attribute accuracy gains to periodic encoding or gated mixing.

Experimental findings. The experiments answer the three evaluation questions as follows. For Q1, the tested four-qubit PennyLane implementation is $882\text{--}997 \times$ slower in wall-clock time than QiT, while QiT is $1.58\text{--}1.76 \times$ slower than the small CViT. For Q2, one QiT configuration trains across eight datasets spanning 32×32 to 224×224 inputs and 10 to 257 classes. For Q3, QiT scales to ImageNet-1K and reaches 78.3% top-1 accuracy, but remains below stronger contemporary backbones in the contextual table. The current evidence establishes execution feasibility and measured cost; it does not establish that the quantum-motivated components outperform a converged, parameter-matched Transformer. That question requires controlled ablations, repeated seeds, and a complete shared training protocol.

5. Conclusion

We presented QiT, a classical vision Transformer that forms attention projections from learned periodic token features and uses gated multiplicative channel mixing. The construction is motivated by angle-encoded quantum models but does not simulate a general quantum circuit or claim quantum advantage. QiT scales across eight image datasets and reaches 78.3% ImageNet-1K top-1 accuracy with 45.7M parameters. A controlled four-qubit pilot further shows the high cost of the tested PennyLane simulator relative to ordinary tensor operations. The current evidence establishes scalability and motivates a focused study of periodic feature maps in attention; controlled component ablations and repeated runs remain necessary to determine their advantage over matched classical alternatives.

References

- [1] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021. [3](#)
- [2] Johannes Bausch. Recurrent quantum neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1368–1379, 2020. [1](#), [3](#)
- [3] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J Osborne, Robert Salzmänn, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nature communications*, 11(1):808, 2020. [3](#)
- [4] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. [1](#), [2](#)
- [5] Yuji Cao, Xiyuan Zhou, Xiang Fei, Huan Zhao, Wenxuan Liu, and Junhua Zhao. Linear-layer-enhanced quantum long short-term memory for carbon price forecasting. *Quantum Machine Intelligence*, 5(2):26, 2023. [3](#)
- [6] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021. [3](#)
- [7] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021. [3](#)
- [8] Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L. Fang. Quantum long short-term memory. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8622–8626, 2022. [1](#), [3](#)
- [9] El Amine Cherrat, Iordanis Kerenidis, Natansh Mathur, Jonas Landman, Martin Strahm, and Yun Yvonna Li. Quantum vision transformers. *Quantum*, 8:1265, 2024. [1](#), [3](#), [4](#), [7](#)
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. [1](#), [2](#), [7](#)
- [11] Tobias Fellner, David A Kreplin, Samuel James Tovey, and Christian Holm. Quantum vs. classical: A comprehensive benchmark study for predicting time series with variational quantum machine learning. *Machine Learning: Science and Technology*, 2025. [3](#)
- [12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. [2](#)
- [13] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022. [2](#)
- [14] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram Wettroth Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567:209 – 212, 2018. [2](#), [3](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. [7](#)
- [16] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. [7](#)
- [17] Yunseok Kwak, Won Joon Yun, Soyi Jung, and Joongheon Kim. Quantum neural networks: Concepts, applications, and challenges. In *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 413–416. IEEE, 2021. [3](#)
- [18] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021. [7](#)
- [19] Guangxi Li, Xuanqiang Zhao, and Xin Wang. Quantum self-attention neural networks for text classification, 2023. [1](#), [3](#), [4](#), [7](#)
- [20] Yanan Li, Zhimin Wang, Rongbing Han, Shangshang Shi, Jiaxin Li, Ruimin Shang, Haiyong Zheng, Guoqiang Zhong, and Yongjian Gu. Quantum recurrent neural networks for sequential learning. *Neural Networks*, 166:148–161, 2023. [3](#)
- [21] Yanan Li, Zhimin Wang, Ruipeng Xing, Changheng Shao, Shangshang Shi, Jiaxin Li, Guoqiang Zhong, and Yongjian Gu. Quantum gated recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [3](#)
- [22] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlp. In *Advances in Neural Information Processing Systems*, pages 9204–9215. Curran Associates, Inc., 2021. [7](#)
- [23] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. [2](#), [7](#)
- [24] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018. [1](#), [3](#)
- [25] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018. [3](#)
- [26] Badri Patro and Vijay Agneeswaran. Scattering vision transformer: Spectral mixing matters. *Advances in Neural Information Processing Systems*, 36:54152–54166, 2023. [2](#)
- [27] Badri N Patro and Vijay S Agneeswaran. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024. [2](#), [7](#)
- [28] Badri N Patro and Vijay S Agneeswaran. Hamsa: Scanning-free vision state space models via spectralpulsenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 2408–2418, 2026.

- [29] Badri N Patro and Vijay S Agneeswaran. Nakul-med: Spectral-graph state space models with dynamics kernels for medical signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 5399–5408, 2026. 2
- [30] Badri N Patro, Vinay P Nambodiri, and Vijay S Agneeswaran. Spectformer: Frequency and attention is what you need in a vision transformer. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9543–9554. IEEE, 2025. 2, 7
- [31] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020. 3
- [32] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018. 1, 3
- [33] Yang Qian, Xinbiao Wang, Yuxuan Du, Xingyao Wu, and Dacheng Tao. The dilemma of quantum neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5603–5615, 2022. 3
- [34] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, pages 980–993, 2021. 7
- [35] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jie Zhou, and Jiwen Lu. Gfnet: Global filter networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [36] Mayra Alejandra Rivera-Ruiz, Andres Mendez-Vazquez, and José Mauricio López-Romero. Time series forecasting with quantum machine learning architectures. In *Mexican international conference on artificial intelligence*, pages 66–82. Springer, 2022. 3
- [37] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Phys. Rev. Lett.*, 122:040504, 2019. 1, 2
- [38] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021. 3
- [39] Rigui Shen, Jun Wang, Huian Jiang, Xiongjun Zhang, and Xingke Hu. Quantum implementation of the classical neural networks. *arXiv preprint arXiv:1707.03400*, 2017. 3
- [40] Jinjing Shi, Ren-Xin Zhao, Wenxuan Wang, Shichao Zhang, and Xuelong Li. Qsan: A near-term achievable quantum self-attention network. *IEEE Xplore*, 2024. 3, 4, 7
- [41] Seongwook Shin, Yong-Siah Teo, and Hyunseok Jeong. Exponential data encoding for quantum supervised learning. *Physical Review A*, 107(1):012422, 2023. 3
- [42] Anthony M. Smaldone, Yu Shee, Gregory W. Kyro, Marwa H. Farag, Zohim Chandani, Elica Kyoseva, and Victor S. Batista. A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation, 2025. 3
- [43] Yuto Takaki, Kosuke Mitarai, Makoto Negoro, Keisuke Fujii, and Masahiro Kitagawa. Learning temporal data with a variational quantum recurrent neural network. *Physical Review A*, 103(5):052414, 2021. 3
- [44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7
- [45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 7
- [46] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 2
- [48] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 7
- [49] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024. 2
- [50] Hui Zhang, Qinglin Zhao, Mengchu Zhou, and Li Feng. Hqvit: Hybrid quantum vision transformer for image classification, 2025. 3, 4, 7
- [51] Yuxuan Zhang and Qiang Ni. Quantum self-attention model (qsam), 2023. 3, 4, 7
- [52] Ren-Xin Zhao, Jinjing Shi, and Xuelong Li. Qksan: A quantum kernel self-attention network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10184–10195, 2024. 1, 3, 4, 7
- [53] Hanqi Zheng et al. Fully quantum self-attention neural network (f-qsann), 2023. 3, 4, 7
- [54] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 7

A. Supplementary Material

This supplement provides implementation details and derivations for the model in the main paper. QiT is implemented entirely with classical tensor operations. The terms “angle,” “Hilbert space,” and “circuit-inspired” describe the motivation for its parameterization; they do not imply execution on quantum hardware, simulation of a general quantum circuit, or quantum computational advantage.

A.1. QiT Encoder Block

Algorithm 1 gives the operations in one encoder block. The input H contains patch and positional embeddings. The

Algorithm 1 One QiT encoder block

Require: Token representations $H \in \mathbb{R}^{B \times N \times d}$, number of learned angles Q

Ensure: Updated representations $H_{\text{out}} \in \mathbb{R}^{B \times N \times d}$

```
1:  $\Theta \leftarrow \pi \tanh(HW_\theta + b_\theta)$ 
2:  $\Phi \leftarrow Q^{-1/2}[\sin(\Theta), \cos(\Theta)]$ 
3: for each attention head  $a$  do
4:    $Q_a \leftarrow \Phi W_Q^{(a)}, K_a \leftarrow \Phi W_K^{(a)}, V_a \leftarrow \Phi W_V^{(a)}$ 
5:    $A_a \leftarrow \text{softmax}(Q_a K_a^\top / \sqrt{d_h})$ 
6:    $Z_a \leftarrow A_a V_a$ 
7: end for
8:  $Z \leftarrow \text{Concat}(Z_1, \dots, Z_{n_h}) W_O$ 
9:  $H' \leftarrow \text{LayerNorm}(H + Z)$ 
10:  $U \leftarrow \text{GELU}(H' W_1)$ 
11:  $G \leftarrow \sigma(H' W_g)$ 
12:  $M \leftarrow (U \odot G) W_2$ 
13:  $H_{\text{out}} \leftarrow \text{LayerNorm}(H' + M)$ 
14: return  $H_{\text{out}}$ 
```

periodic feature map is shared by the query, key, and value projections, while each attention head has separate learned projection matrices. Standard dropout operations are omitted from the notation.

The block differs from a standard Transformer encoder in two places. First, query, key, and value projections operate on Φ rather than directly on H . Second, the usual channel MLP is replaced by the gated multiplicative map $(U \odot G) W_2$. Residual connections, normalization, multi-head composition, and softmax attention remain standard Transformer operations.

A.2. Exact Property of the Periodic Feature Map

For a token representation h , let

$$\theta(h) = \pi \tanh(hW_\theta + b_\theta), \quad \phi(h) = \frac{1}{\sqrt{Q}}[\sin \theta(h), \cos \theta(h)]. \quad (19)$$

The feature vector has unit norm because

$$\|\phi(h)\|_2^2 = \frac{1}{Q} \sum_{q=1}^Q (\sin^2 \theta_q(h) + \cos^2 \theta_q(h)) = 1. \quad (20)$$

Its inner product is the finite cosine kernel

$$k(h, h') = \phi(h)^\top \phi(h') \quad (21)$$

$$= \frac{1}{Q} \sum_{q=1}^Q \cos(\theta_q(h) - \theta_q(h')). \quad (22)$$

For arbitrary inputs $\{h_i\}_{i=1}^m$ and coefficients $\{c_i\}_{i=1}^m$,

$$\sum_{i,j=1}^m c_i c_j k(h_i, h_j) = \left\| \sum_{i=1}^m c_i \phi(h_i) \right\|_2^2 \geq 0. \quad (23)$$

Thus, k is positive semidefinite. This statement applies to the unprojected feature map. QiT uses

$$q_i^\top k_j = \phi(h_i)^\top W_Q W_K^\top \phi(h_j) \quad (24)$$

inside attention. Since W_Q and W_K are learned independently, this score is generally asymmetric and is not, in general, a positive-semidefinite kernel.

For comparison, independent single-qubit R_Y encodings give the product-state fidelity

$$K_{\text{prod}}(h, h') = \prod_{q=1}^Q \cos^2 \left(\frac{\theta_q(h) - \theta_q(h')}{2} \right). \quad (25)$$

This product is different from QiT’s average cosine kernel. Entangling circuits introduce additional circuit-dependent terms. The shared periodic dependence on encoded angles motivates QiT’s feature map, but it does not establish equivalence to a quantum fidelity kernel.

A.3. Computational Cost

For batch size B , N tokens, embedding width d , Q learned angles, and gated-mixer width r , angle projection and feature construction require $\mathcal{O}(BNdQ)$ operations and store $2Q$ features per token. Projection to the attention heads requires $\mathcal{O}(BNQd)$ operations. Scaled dot-product attention requires $\mathcal{O}(BN^2d)$ operations, and the gated channel mixer requires $\mathcal{O}(BNdr)$. The resulting per-block cost is

$$\mathcal{O}(BN^2d + BNQd + BNdr). \quad (26)$$

QiT therefore retains the quadratic dependence on token count of standard softmax attention. These are classical operation counts; they are not quantum runtime bounds.

A.4. Experimental Protocol and Interpretation

The reported results come from three separate protocols and should be interpreted within their respective settings.

Simulator diagnostic. CViT, QiT, and the PennyLane model use patch size 4, one encoder block, batch size 2, and one training epoch on reduced 2-, 4-, and 10-class CIFAR tasks. QiT uses four learned angles, whereas the PennyLane model uses four simulated qubits. The models are not parameter matched, and one epoch does not measure converged accuracy. This experiment supports only the reported implementation-cost comparison for the tested configurations.

Multi-dataset study. The eight-dataset study uses one QiT configuration with patch size 4, depth 12, width 256, four attention heads, and 16 learned angles. Each model is trained for 300 epochs with AdamW, cosine learning-rate decay, random cropping, horizontal flipping, and color jitter on four Tesla T4 GPUs. Batch size varies with image resolution as reported in the main paper.

ImageNet-1K study. QiT-XS, QiT-S, and QiT-B are evaluated at 224×224 resolution with the parameter counts and FLOPs reported in the main paper. Comparisons with published architectures provide accuracy–cost context rather than controlled pairwise evidence because their training recipes can differ.

All experimental values are point estimates from the available runs. Repeated seeds, uncertainty intervals, complete optimizer settings, reduced-class split definitions, and the energy-measurement procedure are not available in the current records. Accordingly, the evidence establishes that QiT can be trained at image-classification scale and documents the measured cost of the tested implementations. It does not establish statistical superiority over a converged, parameter-matched Transformer or any quantum advantage.