

HQViT: Hybrid Quantum Vision Transformer for Image Classification

Abstract—Transformer-based architectures have revolutionized the landscape of deep learning. In computer vision domain, Vision Transformer demonstrates remarkable performance on par with or even surpassing that of convolutional neural networks. However, the quadratic computational complexity of its self-attention mechanism poses challenges for classical computing, making model training with high-dimensional input data, e.g., images, particularly expensive. To address such limitations, we propose a Hybrid Quantum Vision Transformer (HQViT), that leverages the principles of quantum computing to accelerate model training while enhancing model performance. HQViT introduces whole-image processing with amplitude encoding to better preserve global image information without additional positional encoding. By leveraging quantum computation on the most critical steps and selectively handling other components in a classical way, we lower the cost of quantum resources for HQViT. The qubit requirement is minimized to $O(\log_2 N)$ and the number of parameterized quantum gates is only $O(\log_2 d)$, making it well-suited for Noisy Intermediate-Scale Quantum devices. By offloading the computationally intensive attention coefficient matrix calculation to the quantum framework, HQViT reduces the classical computational load by $O(T^2 d)$. Extensive experiments across various computer vision datasets demonstrate that HQViT outperforms existing models, achieving a maximum improvement of up to 10.9% (on the MNIST 10-classification task) over the state of the art. This work highlights the great potential to combine quantum and classical computing to cope with complex image classification tasks.

Index Terms—Quantum computing, Quantum machine learning, Vision transformer, Self-attention, and Variational quantum algorithm.



1 INTRODUCTION

1.1 Background

IN 2017, the emergence of Transformer [1] revolutionized the landscape of deep learning. Initially, it demonstrated its strength in the field of natural language processing (NLP), quickly surpassing both RNN and LSTM architectures to become the dominant model in NLP [2], [3]. Vision Transformer (ViT) was then proposed [4], extending the transformer architecture to the computer vision (CV) domain. Once again, the Transformer showcased its remarkable performance, achieving results on par with, or even exceeding, those of convolutional neural networks (CNNs). As a result, Transformer is now applied across various CV tasks, e.g., image restoration, image generation, and image segmentation [5], [6], [7]. At the heart of the transformer architecture lies self-attention mechanism, enabling the model to weigh the importance of different parts of input data, effectively capturing long-range dependencies. This mechanism is the key to the Transformer’s outstanding learning capabilities. However, its related computational complexity grows quadratically with the sequence length, significantly increasing the computational demands. This makes the training of high-dimensional input data extremely resource-intensive for classical computing systems.

Conversely, quantum machine learning (QML) has emerged rapidly in recent years, combining the strengths of quantum computing with classical machine learning techniques [8], [9]. QML aims to leverage quantum phenomena like superposition and entanglement as new computing resources to accelerate machine learning tasks [10], [11], [12]. Motivated by this potential, many quantum counterparts to classical neural network models have been developed, such as QCNNs [13], [14], [15], [16], [17], QRNNs [18], [19], and QGANs [20], [21]. They are based on a paradigm known as a Variational Quantum Algorithm (VQA) [22], which involves replacing certain components of classical neural networks with parameterized quantum circuits and using classical optimizers to update these parameters [23].

Given the importance of image processing in machine learning and the significant impact of the transformer architecture on computer vision, a promising idea is to propose a quantum analog of Transformer. This approach aims to leverage the advantages of quantum computing to accelerate current CV tasks. Yet research on quantum Transformer models, especially for vision tasks, is in its early stage, with only a handful of studies currently in progress. Most existing work focuses on text-based tasks, and there are even fewer studies dedicated to Quantum Vision Transformers. Although some quantum Transformer models targeted for NLP can be adapted for CV tasks, they are limited in the image size they can handle or require the pre-extraction of image features. Another issue is that the current methods are not well-optimized in balancing quantum resources and classical computation acceleration. Some approaches only quantumize a specific component of the transformer [24], [25], which, while requiring fewer quantum resources, do not provide substantial acceleration

- H. Zhang, Q. Zhao (corresponding author) and L. Feng are with Faculty of Innovation Engineering, Macao University of Science and Technology, Macao 999078, China.
E-mail: h.zhang2023@hotmail.com; qlzhao@must.edu.mo; lfeng@must.edu.mo
- M. Zhou (co-corresponding author) is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA E-mail: zhou@njit.edu

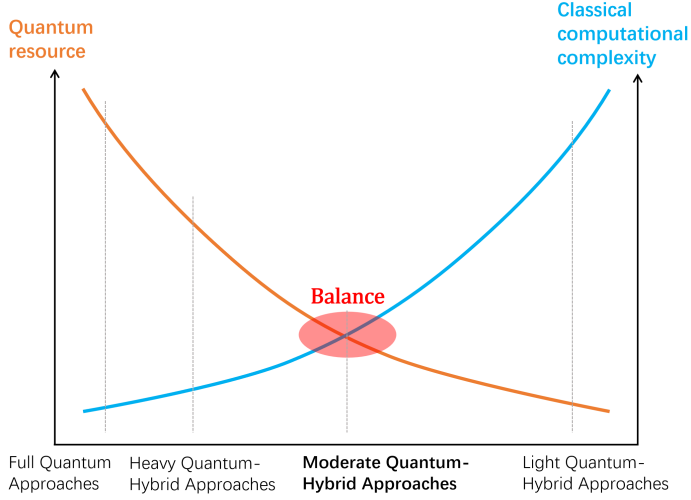


Fig. 1. The trade-off between the amount of quantum resource and classical computational complexity for existing quantum self-attention and quantum transformer models. Our motivation is to achieve a balance between the two, ensuring that the quantum resources align with the limitations of NISQ devices while maintaining high-performance and scalability. The vertical axis on the left represents the quantum resource requirements, while the vertical axis on the right represents the classical computational complexity of the model. Both axes are dimensionless, since we do not conduct a quantitative comparison between the two, merely illustrating the relative relationships of the quantum resource requirements (or classical computational complexity) across different models.

to the model training. On the other hand, some methods that adopt a fully quantum approach [26], [27], [28], can drastically reduce computational loads, but they consume substantial quantum resources, making them difficult to be scalable on Noisy Intermediate-Scale Quantum (NISQ) era, which refers to an early stage of quantum computers. NISQ devices have a limited number of qubits, and these devices are affected by noise and the limited coherence time of qubits, which prevents the quantum circuits from being designed too deeply.

1.2 Motivation

Therefore, we can conclude that there is a clear need to develop Quantum Vision Transformer model specifically tailored for image processing. As illustrated in Figure 1, excessive or insufficient quantum hybridization can lead to an imbalance between the model’s quantum resource requirements and classical complexity. Therefore, a moderate quantum-classical hybrid approach appears to be a promising pathway, striking a balance between reducing classical computational load and efficiently using quantum resources. Such a design would not only enhance the practicality of the model for NISQ systems but also ensure superior performance in CV tasks.

1.3 Contributions

To address this challenge, we propose a Hybrid Quantum Vision Transformer (HQViT) that incorporates whole-image processing and amplitude encoding. By capturing the similarity between Q and K within a quantum system,

our model computes attention coefficients in the quantum framework, significantly reducing computational complexity. Moreover, by moderately integrating quantum and classical components, our model balances the reduction of classical computational demands and the requirements of quantum resources. The design principles, advantages and contributions of our approach are shown in Fig. 2, with the specific descriptions as follows:

1) Whole-Image Processing. We process the image data as a whole rather than handling patch-by-patch separately as in conventional Transformer [1]. This enables the model to capture the relationships among tokens within the quantum framework, without needing to convert Q and K into classical data, thereby reducing classical computation complexity by $O(T^2d)$, where T is the number of patches and d is the dimension of each patch.

2) Amplitude Encoding. Our model utilizes amplitude encoding to efficiently manage high-dimensional input data without requiring a large number of qubits. This also serves as the foundation for the implementation of whole-image processing. Another beneficial byproduct of amplitude encoding combined with whole-image processing is that it simultaneously accounts for the positional information of the tokens, thereby allowing the model to be less reliant on additional positional encoding. Thus, the model architecture is further simplified.

3) Moderate quantum-classical hybrid architecture. We selectively retain some classical components because their computational complexity is relatively low, and quantizing them would require substantial quantum resources. This approach ensures the efficient use of quantum resources, which are then dedicated to addressing the most computationally intensive steps in the model. This reduces the requirement of qubits to $O(\log_2 Td)$ and the number of parameterized quantum gates to $O(\log_2 d)$, making it applicable for NISQ devices. Meanwhile, this moderate hybrid approach enhances scalability and maintains high performance, achieving a maximum improvement of up to 11% (on the MNIST 10-classification task) over the state of the art.

4) Performance evaluation. Experimental results across multiple computer vision datasets of varying scales demonstrate that HQViT outperforms existing models in most classification tasks. It does so while maintaining a limited of quantum resources and relatively low classical computational complexity.

The remainder of the paper is organized as follows: Section 2 presents related work, Section 3 introduces some quantum computing concepts related to our model, Section 4 describes the methodology and model architecture, Section 5 presents the experimental implementation and results, and Section 6 concludes the study.

2 RELATED WORK

Several approaches to implementing quantum-enhanced self-attention mechanisms have emerged in recent years, each differing in the level of quantum integration. These approaches can be broadly categorized into “light quantum-hybrid”, “heavy quantum-hybrid”, and “full quantum” models.

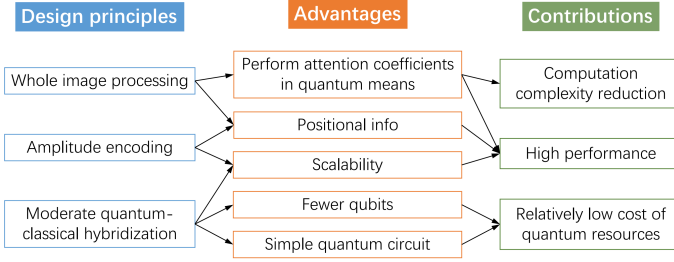


Fig. 2. Overview of HQViT’s design principles, advantages, and contributions.

Light Quantum-Hybrid Approach. In 2022, Li et al. [24] proposed a Quantum Self-Attention Neural Network (QSANN), which replaced classical linear mapping matrices (W_q , W_k , and W_v) with three parameterized quantum circuits (PQCs). These PQCs mapped the input data into Q , K , and V , followed by classical computation of attention coefficients. The motivation behind using PQCs was to project classical data into quantum-enhanced feature space, with the goal of extracting features that are difficult for classical linear mappings to capture. QSANN achieved better performance than that of its classical counterpart, demonstrating the powerful capabilities of quantum computing. Zhang et al. [25] extended this idea with an improved Quantum Self-Attention Model (QSAM). By introducing amplitude-phase decomposed measurements and more expressive PQCs, QSAM extracted quantum state information more efficiently and reduced the number of learnable parameters by one-third. However, both QSANN and QSAM require a) converting Q , K , and V from quantum states back to classical states and b) computing the attention coefficients classically, thus limiting the reduction of classical computational complexity.

Heavy Quantum-Hybrid Approach. In 2024, El Cherrat et al. introduced the Quantum Orthogonal Vision Transformer (QViT) [28], which computed the attention coefficient matrix and weighted Values entirely through quantum means. The remaining classification task was completed by a classical fully-connected layer. The model’s backbone was a quantum orthogonal neural network built by using Reconfigurable Beam Splitter (RBS) gates [29]. A dedicated data loader allowed the quantum circuit to load the entire sequence, capture relationships among tokens, and computed weighted values directly within the quantum framework. This method achieved comparable classification accuracy on the MedMNIST dataset against a classical baseline. However, the RBS gate are not a hardware-efficient gate and loading the entire image lead to an extremely deep circuit, hindering its practical deployment on NISQ devices. Another heavy quantum-hybrid quantum method is given by Zhao et al. [30], which utilizes Quantum Kernel to capture the similarity between Q and K , and then associates the similarity information to V by deferred conditional measurements. For quantum kernel methods to be effective, the measurement outcome of the output quantum state must match the initial state, which requires discarding a large number of measurement results, significantly increasing the sampling cost.

Full Quantum Approach. Shi et al. developed an end-

to-end Quantum Self-Attention Network (QSAN) in 2024 [26]. They introduced a novel quantum interpretation of self-attention mechanism called Quantum Logical Similarity (QLS), which utilizes quantum logical operations to capture the similarity among tokens, allowing uninterrupted execution of the model on quantum computers. However, QLS required a $O(T^2)$ complexity on number of qubits, making the model hard to be scalable for NISQ devices. Another “Full-Quantum” Self-Attention Neural Network (F-QSANN) was proposed by Zheng et al. [27] in 2023. This approach entangled the elements of Q , K , and V by using multi-control quantum gates, and then measured the qubits in the V register to obtain the weighted values, which carry certain relational information between Q and K . However, since the similarity relationships are captured along the depth of the circuit, the complexity of circuit depth of this method is $O(T^2)$. So it is also very challenging to make this method practical on NISQ devices.

From the above overview, it is evident that “light quantum” approaches are insufficient in reducing classical computational complexity, while “heavy quantum” or “full-quantum” models suffer from poor scalability in terms of both input size and circuit depth, Making them difficult to scale to practical image tasks. Therefore, a “moderate quantum-classical hybrid” approach appears to be a more promising solution for practical quantum computing in the NISQ era. Balancing the consumption of quantum resources and the reduction of classical computational load is key to designing effective quantum transformer models in the computer vision (CV) domain.

To address these challenges, this work proposes a novel Hybrid Quantum Vision Transformer (HQViT). It can well leverage a quantum neural network to handle image data without requirement of preprocessing. Based on the “moderate quantum-classical hybrid” concept, HQViT leverages a quantum self-attention module to handle the most computationally intensive step in Transformer, i.e., the computation of attention coefficients matrix, and selectively retains several classical modules, making it feasible for deployment on NISQ devices. Experimental results indicate that our model achieves comparable or even better performance than existing quantum self-attention and transformer models, showcasing its effectiveness and practicality in real-world applications.

3 PRELIMINARIES

Before delving into our quantum vision transformer model, it is necessary to understand a few fundamental concepts of quantum mechanics, including quantum states, ansatz, swap test and measurement.

Quantum States. In quantum computing, quantum information is usually represented by n -qubit (pure) quantum states over Hilbert space \mathbb{C}^{2^n} . A quantum state is typically represented by using Dirac notation, such as $|\psi\rangle$. For a single qubit, the state can be written as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1)$$

where $|0\rangle$ and $|1\rangle$ are the basis of Hilbert space, and α, β are amplitudes, which are complex numbers satisfying $|\alpha|^2 +$

$|\beta|^2 = 1$. The values of α and β describe the probability distribution of the qubit being in the $|0\rangle$ or $|1\rangle$ state.

Ansatz. An ansatz is a proposed form for a quantum state or a quantum circuit, often used as an initial guess in variational algorithms. It is a parameterized quantum circuit designed to approximate the desired quantum state or solve a particular problem. It contains a set of parameterized quantum gates, and the rotation angles of these gates can be adjusted during the training process. In the VQA framework, the parameters in an ansatz are typically optimized by using a classical optimizer [31]. For example, an ansatz operated on a two-qubit system can be written as:

$$|\psi(\theta_1, \theta_2)\rangle = U(\theta_1, \theta_2)|00\rangle \quad (2)$$

where $U(\theta_1, \theta_2)$ is a unitary operator parameterized by θ_1 and θ_2 .

Swap Test. A swap test is a quantum operation used to determine the similarity between two quantum states $|\psi\rangle$ and $|\phi\rangle$. It involves applying a swap operation on the two quantum states, while the swap operation is controlled by an ancilla qubit. On the ancilla qubit, a Hadamard gate is applied before and after the controlled swap operation. The ancilla qubit is then measured, and the probability of the measurement yielding 0 is:

$$Pr(0) = \frac{1 + |\langle\psi|\phi\rangle|^2}{2}. \quad (3)$$

This probability reflects the degree of overlap between the two states, serving as a measure of similarity. It provides an efficient means of comparing two quantum vectors. The swap test circuit is shown in Fig. 3.

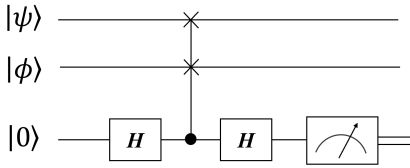


Fig. 3. The swap test circuit.

Measurement. Quantum measurements are described by a collection $\{M_m\}$ of measurement operators. These are operators acting on the state space of the system being measured. The index m refers to the measurement outcomes that may occur in the experiment. If the state of the quantum system is $|\psi\rangle$ immediately before the measurement then the probability that result m occurs is

$$Pr(m) = \langle\psi|M_m^\dagger M_m|\psi\rangle. \quad (4)$$

By selecting different projection operators, we can obtain the probability information of the quantum state projected onto different subspaces. In our model, this probability information from the different subspaces is used to obtain the attention coefficients for different token pairs.

The state of the system after the measurement is

$$|\psi_{\text{post}}\rangle = \frac{M_m|\psi\rangle}{\sqrt{\langle\psi|M_m^\dagger M_m|\psi\rangle}}. \quad (5)$$

This describes the change in the quantum state following measurement, where the system collapses into one of the possible measured states.

4 HYBRID QUANTUM VISION TRANSFORMER

4.1 Overall architecture

The overall framework of the proposed hybrid quantum vision transformer (HQViT) is as shown in Fig. 4, showing the integration of both quantum and classical components. Aside from the quantum self-attention module, the other classical parts largely maintain the structure of the classical Vision Transformer (ViT). First, the input image is segmented into smaller patches. Each patch is then flattened into a vector and all patch vectors are concatenated to form a single representation for the entire image, which is called whole-image embedding. This whole-image embedding is then fed into a quantum transformer block (QTB), which consists of a quantum self-attention module and a classical feed-forward neural network (FFN), both with residual connections. The quantum self-attention module, the core of the model, uses amplitude encoding and PQCs to process the self-attention mechanism and produce a set of weighted values as output. A total of L QTBs are stacked to enhance the model's feature extraction capabilities. Finally, a classical post-processing step averages the weighted values from the last QTB and feeds them into a fully-connected layer (FCL) to output the predicted category. By integrating such techniques as whole image embedding and amplitude encoding, we reduce the quantum resources required for the quantum components of the model. Additionally, we selectively retain certain classical components. Hence, the overall framework showcases our "moderate quantum-classical hybridization" design principle.

For the image segmentation, similar to ViT, an image with dimensions $H \times W \times C$ is divided into patches of size $d = h \times w \times C$, resulting in $T = \frac{H}{h} \times \frac{W}{w}$ patches. Here, we use a 4×4 grayscale image as an example. It can be divided into 4 patches of size 2×2 . By Using amplitude encoding in quantum self-attention module, the image can be represented by a 4-qubit quantum state. As shown in Fig. 4, each pixel's position is represented by each basis of the quantum state, where the first two qubits precisely encode the index of each patch. This index plays a crucial role in the subsequent quantum-based processing of attention coefficients, to be explained further.

4.2 Quantum self-attention module

The quantum self-attention module processes the entire image within the quantum system, allowing it to capture long-range dependencies among tokens and directly compute the attention coefficient matrix. Our encoding strategy encodes both the pixel values and positions, meaning the positional information of each patch is integrated into the whole-image embedding. Fig. 5 illustrates the structure of the quantum self-attention module. Below, we summarize the basic mechanism. For simplicity, assume that the input image X consists of a total of N pixels, and it is divided into T patches, each with a dimension of d .

i) We utilize three quantum registers, denoted as R_Q , R_K , and R_V , to receive the input data. Each of them is initialized with the state of $|0\rangle^{\otimes n}$, where the number of qubits $n = \log_2 N$. We are first concerned with R_Q and R_K . The whole image is fed into R_Q and R_K as a vector with amplitude encoding, obtaining $|\Psi_1\rangle$.

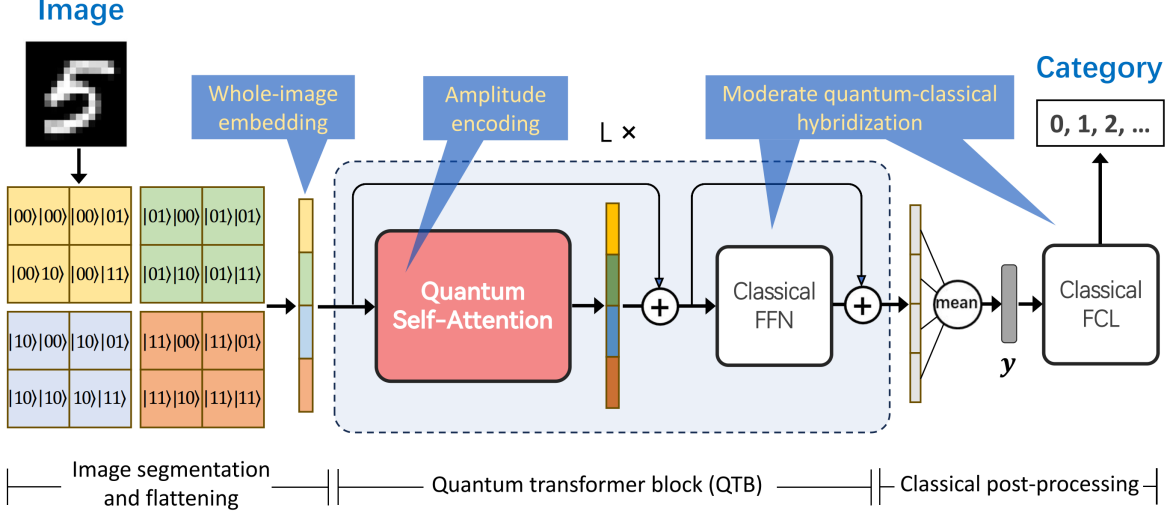


Fig. 4. Overall framework of HQViT. The image is first segmented and flattened as a whole-image embedding (we use a 4x4 grayscale image as an example), which is processed by the quantum self-attention module. The output is then passed through classical feedforward layers (FFN) and further refined by using a classical fully connected layer (FCL) for final classification. The quantized self-attention module is the core of the model, while other components are handled classically.

$$|\Psi_0\rangle = |0\rangle^{\otimes n} \otimes |0\rangle^{\otimes n} \xrightarrow{\text{Data encoding}} |\Psi_1\rangle = |\psi_X\rangle \otimes |\psi_X\rangle \quad (6)$$

ii) The encoded quantum states undergo ansatz U_Q and U_K , which act as the quantum equivalents of the classical linear mapping matrices W_q and W_k . This results in quantum states $|\psi_Q\rangle, |\psi_K\rangle$, respectively.

$$|\Psi_2\rangle = (U_Q|\psi_X\rangle) \otimes (U_K|\psi_X\rangle) = |\psi_Q\rangle \otimes |\psi_K\rangle \quad (7)$$

iii) We then combine the registers R_Q, R_K , and an ancilla qubit to perform a swap test. Thanks to our encoding strategy, the qubits in R_Q or R_K can be considered into two parts: the index subsystem encodes the index information of the tokens and the patch subsystem holds the information of each token. The swap test is applied only to the patch subsystems of R_Q and R_K to extract the similarities between each pair of tokens, q_i and k_j .

$$|\Psi_2\rangle \otimes |0\rangle \xrightarrow{\text{Swap test}} |\Psi_3\rangle \quad (8)$$

iv) We measure the index subsystems of R_Q and R_K and ancilla qubit. We keep the outcome of ancilla qubit being 0, and traverse all the outcomes of the index subsystems of R_Q and R_K . This gives us a set of probabilities $Pr(i, j, 0)_{i,j=0}^T$, where i and j are measurement outcomes of the index subsystems of R_Q and R_K , respectively. Each probability corresponds to the similarity of a particular pair of q_i and k_j . This can be expressed as:

$$|\Psi_3\rangle \xrightarrow{\text{Measurement}} \text{outcome: } \{i, j, 0\} \quad (9)$$

$$Pr(i, j, 0) \Rightarrow \langle q_i | k_j \rangle \quad (10)$$

where f is a simple functional relationship.

v) For R_V , the input data is similarly encoded and fed into ansatz U_V . Then measurements are performed to obtain

classical V . Finally, the attention coefficient matrix, after applying softmax, is combined with V to obtain the outputs of quantum self-attention, also known as the Weighted Values.

4.2.1 Data encoding

The data encoding strategy is crucial in quantum neural networks (QNNs) and can even have a decisive impact on the network's design [32]. For an input vector $\mathbf{X} = \{x_i\}_{i=0}^{N-1}$, the amplitude encoding can be written as

$$|\psi\rangle = \frac{1}{\|\mathbf{X}\|} \sum_{i=0}^{N-1} x_i |i\rangle, \quad (11)$$

where $|i\rangle$ denotes the computational basis, and $\|\mathbf{X}\|$ is normalization factor. The number of qubits required for this encoding scheme is n , which satisfies $2^{n-1} < N \leq 2^n$, with the amplitudes of unused basis states padded with zeros. Therefore, the required number of qubits is $O(\log_2 N)$.

For a whole-image embedding with T patches, each of size d , amplitude encoding is written as:

$$|\psi\rangle = \sum_{i=0}^{T-1} \sum_{l=0}^{d-1} x_{il} |i\rangle |l\rangle, \quad (12)$$

where i represents the patch index, giving each patch its positional information, and j represent the index of elements within each patch. From this perspective, the encoded quantum system can be seen as comprising two parts: the index subsystem (higher-order qubits), which carries positional information, and the patch subsystem (lower-order qubits), which encodes the patch data. These subsystems play distinct roles during the subsequent evolution.

4.2.2 Ansatz

The function of U_Q, U_K and U_V is to apply linear mappings on the input data to generate Q, K and V . In classical self-attention, matrices W_q, W_k and W_v map each token individually. In our quantum self-attention mechanism, the

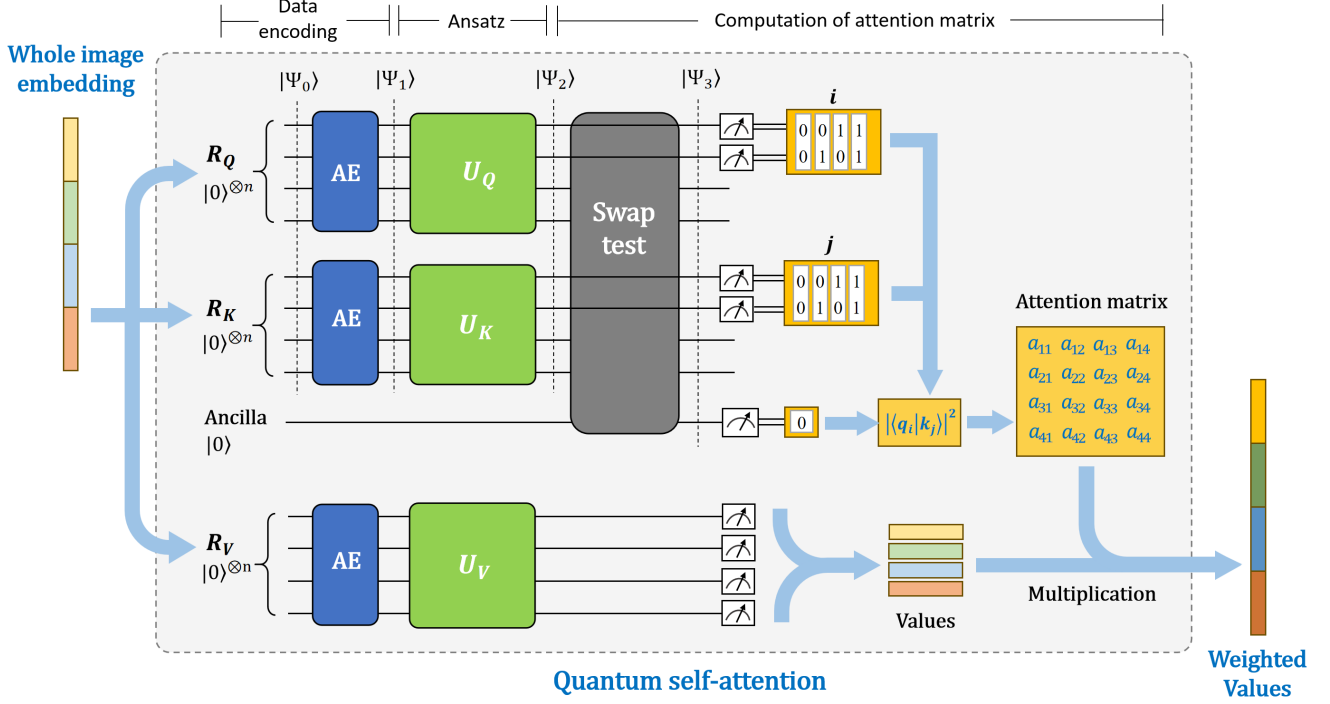


Fig. 5. Quantum self-attention mechanism for HQViT. The whole-image embedding is fed into registers R_Q , R_K , and R_V with amplitude encoding (AE), followed by unitary transformations (U_Q , U_K , U_V). A swap test is performed on the Q and K quantum states to compute the attention matrix, which is then combined with the V to produce Weighted Values.

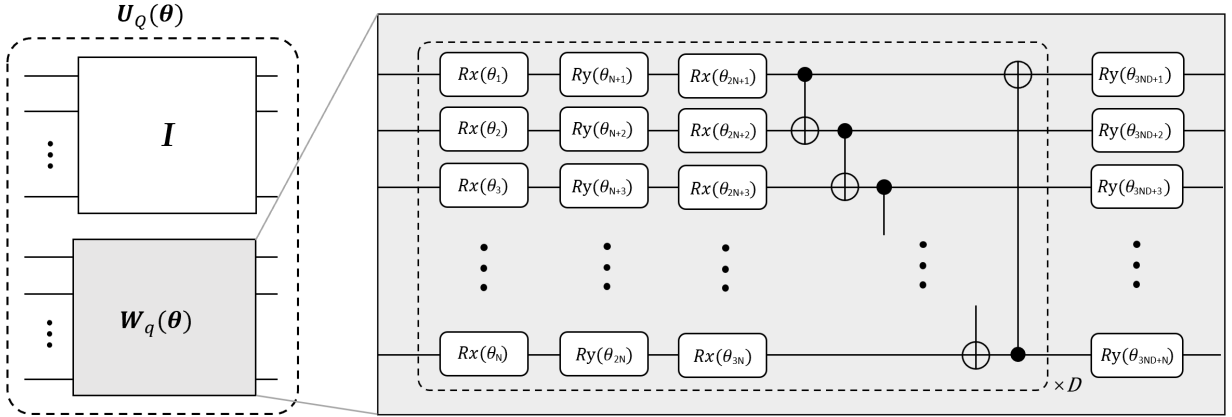


Fig. 6. The structure of U_Q , where n represents the number of qubit of R_Q and D denotes the number repeated basic units.

input is the whole image. Thus, U_Q and U_K are constructed to fulfill this linear mapping, i.e.,

$$U_Q = \begin{pmatrix} W_q & & & \\ & W_q & & \\ & & \ddots & \\ & & & W_q \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \otimes W_q. \quad (13)$$

$$U_Q |\psi_X\rangle = \begin{pmatrix} W_q & & & \\ & W_q & & \\ & & \ddots & \\ & & & W_q \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{T-1} \\ x_T \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_{T-1} \\ q_T \end{pmatrix} \quad (14)$$

Here we discuss only U_Q as an example, since the principles of U_K and U_V are completely the same as U_Q 's.

In Eq. 13, W_q is a linear mapping matrix with size of $d \times d$. So the size of U_Q is $Td \times Td$. The evaluation of Eq.(7) can be written as

The circuit implementation of U_Q is shown in Fig. 6. Here, I represents the identity transformation, and $W_q(\theta)$ is a parameterized quantum circuit, which consists of D layers of repeatable basic units. Each basic unit contains a set of single-qubit gates and a set of two-qubit gates. This structure allows the circuit to span the entire unitary group, ensuring sufficient expressibility of the model [33]. Finally,

a set of Ry gates is used for fine-tuning before outputting the quantum state.

After the linear mapping, the quantum states $|\psi_q\rangle$ and $|\psi_k\rangle$ can be written as

$$|\psi_Q\rangle = \sum_{i=0}^{T-1} \sum_{l=0}^{d-1} \alpha_{il} |i\rangle |l\rangle, \quad |\psi_K\rangle = \sum_{j=0}^{T-1} \sum_{k=0}^{d-1} \beta_{jk} |j\rangle |k\rangle. \quad (15)$$

4.2.3 Computation of attention matrix

Then, we combine registers R_Q and R_K with an ancilla qubit and perform a swap test on the combined quantum system. After the swap test, the state of the combined quantum system becomes:

$$|\Psi_3\rangle = \frac{1}{2} \left((|\Psi_2\rangle + |\Psi_{sw}\rangle) \otimes |0\rangle + (|\Psi_2\rangle - |\Psi_{sw}\rangle) \otimes |1\rangle \right), \quad (16)$$

where $|\Psi_{sw}\rangle$ is the state resulting from applying the swap operation to $|\Psi_2\rangle$. Recall that

$$|\Psi_2\rangle = |\psi_Q\rangle \otimes |\psi_K\rangle = \sum_{i=0}^{T-1} \sum_{j,k=0}^{d-1} \alpha_{il} \beta_{jk} |i\rangle |l\rangle |j\rangle |k\rangle. \quad (17)$$

Since the swap operation only involves the patch subsystems, $|\Psi_{sw}\rangle$ can be written as

$$|\Psi_{sw}\rangle = \sum_{i,l=0}^{T-1} \sum_{j,k=0}^{d-1} \alpha_{ik} \beta_{jl} |i\rangle |l\rangle |j\rangle |k\rangle. \quad (18)$$

We notice that if we perform projection measurement on the ancilla qubit, the probability of obtaining outcome 0 is given as:

$$\begin{aligned} Pr(0)_{anc} &= \frac{1}{2} (1 + \langle \Psi_2 | \Psi_{sw} \rangle) \\ &= \frac{1}{2} \left(1 + \sum_{i=0}^{T-1} \sum_{j=0}^{T-1} \sum_{l=0}^{d-1} |\alpha_{il}|^2 |\beta_{jl}|^2 \right) \\ &= \frac{1}{2} \left(1 + \sum_{i=0}^{T-1} \sum_{j=0}^{T-1} |\langle q_i | k_j \rangle|^2 \right). \end{aligned} \quad (19)$$

This probability gives us information about the overlaps (or similarities) between $|q_i\rangle$ and $|k_j\rangle$, $\forall i, j \in \{0, \dots, T-1\}$. Therefore, to extract the value of each pair of $\langle q_i | k_j \rangle$, we need to measure the ancilla qubit and all qubits in the index subsystems of R_Q and R_K . For a set of measurement operators corresponding to outcomes $\{i, j, 0\}_{i,j=0}^{T-1}$, where i and j represent the outcome of index subsystems of R_Q and R_K (in the form of binary string), and 0 is the outcome of ancilla qubit, we obtain the set of probability values $\{Pr(i, j, 0)\}_{i,j=0}^{T-1}$, where each $Pr(i, j, 0)$ is given:

$$Pr(i, j, 0) = \frac{1}{2} (\|q_i\|^2 \|k_j\|^2 + |\langle q_i | k_j \rangle|^2). \quad (20)$$

Since $q_i = W_q \mathbf{x}_i$, $k_j = W_k \mathbf{x}_j$, and both W_q and W_k are unitary transformations, we have $\|q_i\| = \frac{1}{\|X\|} \|\mathbf{x}_i\|$ and $\|k_j\| = \frac{1}{\|X\|} \|\mathbf{x}_j\|$. So finally we obtain:

$$Pr(i, j, 0) = \frac{1}{2} \left(\frac{1}{\|X\|^4} \|\mathbf{x}_i\|^2 \|\mathbf{x}_j\|^2 + |\langle q_i | k_j \rangle|^2 \right). \quad (21)$$

Note that $Pr(0)_{anc} = \sum_{i,j=0}^{T-1} Pr(i, j, 0)$. The detailed derivations of Eq. (19) and (20) can be found in Supplements of this paper.

In this way, by measuring the index subsystems of R_Q , R_K and ancilla qubit all at once, we can derive the similarities between q_i and k_j , i.e., the attention coefficient matrix. Take $T = 4$ and $d = 4$ as an example to demonstrate this process more clearly. Both R_Q and R_K have 4 qubits each, with their index subsystems and patch subsystems each consisting of 2 qubits. So the measurement outcomes are a bench of 5-bit binary strings, denoted as B . The first two bits of B represent i , the following two bits represent j , and the last bit corresponds to the measurement outcome of ancilla qubit, where we only care about the result of 0. When $B = 00000$, the swap test yields $Pr(0, 0, 0)$, allowing us to obtain $\langle q_0 | k_0 \rangle$; when $B = 00010$, it yields the $Pr(0, 1, 0)$ and gives $\langle q_0 | k_1 \rangle$, and so on. This process yields the entire attention coefficient matrix (ACM):

$$\begin{aligned} B &= \begin{pmatrix} 00000 & 00010 & 00100 & 00110 \\ 01000 & 01010 & 01100 & 01110 \\ 10000 & 10010 & 10100 & 10110 \\ 11000 & 11010 & 11100 & 11110 \end{pmatrix} \Rightarrow \\ ACM &= \begin{pmatrix} |\langle q_0 | k_0 \rangle|^2 & |\langle q_0 | k_1 \rangle|^2 & |\langle q_0 | k_2 \rangle|^2 & |\langle q_0 | k_3 \rangle|^2 \\ |\langle q_1 | k_0 \rangle|^2 & |\langle q_1 | k_1 \rangle|^2 & |\langle q_1 | k_2 \rangle|^2 & |\langle q_1 | k_3 \rangle|^2 \\ |\langle q_2 | k_0 \rangle|^2 & |\langle q_2 | k_1 \rangle|^2 & |\langle q_2 | k_2 \rangle|^2 & |\langle q_2 | k_3 \rangle|^2 \\ |\langle q_3 | k_0 \rangle|^2 & |\langle q_3 | k_1 \rangle|^2 & |\langle q_3 | k_2 \rangle|^2 & |\langle q_3 | k_3 \rangle|^2 \end{pmatrix}. \end{aligned} \quad (22)$$

4.3 Classical FFN and post-processing

After obtaining the weighted values from the quantum self-attention module, they are passed into a FFN to integrate the internal features of each token. This step can be implemented by using either quantum or classical methods. Since FFN is not a major computational bottleneck in the transformer architecture, a classical approach is preferred to save quantum resources. In this paper, we employ a classical Multi-Layer Perceptron (MLP), consisting of one hidden layer and one output layer. The dimensionality of the hidden layer is set based on the experimental scenario, while the output dimension matches the input dimension d .

After the last QTB, a set of outputs, \mathbf{y}_i , is produced. We take the mean of them to form a global feature vector of the whole image, \mathbf{y} . Finally, a classical fully-connected layer is applied to make the prediction: $y := \sigma(\mathbf{w}^T \cdot \mathbf{y} + b)$, where σ is the sigmoid function for binary classification, or softmax function for multi-class classification. We use cross-entropy as the loss function. Our model is based on the Variational Quantum Algorithm (VQA) paradigm, where a classical optimizer (in our case, Adam [34]) is used to train both quantum and classical parameters.

4.4 Quantum resources and complexity analysis

We analyze the quantum resource cost and reduction in classical complexity for HQViT and compare these with other quantum self-attention or quantum transformer models. From Table 1, it can be seen that HQViT maintains relatively low or the lowest levels across multiple quantum resource metrics (e.g., the number of qubits is $O(\log(Td))$,

TABLE 1

Quantum resource needs and classical complexity reduction of related models. The letters following the model names indicate the level of quantum-classical hybridization, where L stands for light, M for moderate, H for heavy, and F for fully quantum models.

Model	# Qubits	# PQG	# Distinct circuits	# Total measurement	Outputs of quantum circuits	Classical complexity reduction
QSANN [24] (L)	$O(d)$	$O(d)$	$O(T)$	$O(Td)$	$\{\mathbf{q}_i\}_{i=1}^T, \{\mathbf{k}_i\}_{i=1}^T, \{\mathbf{v}_i\}_{i=1}^T$	$O(Td^2)$
QSAM [25] (L)	$O(d)$	$O(d)$	$O(T)$	$O(Td)$	$\{\mathbf{q}_i\}_{i=1}^T, \{\mathbf{k}_i\}_{i=1}^T, \{\mathbf{v}_i\}_{i=1}^T$	$O(Td^2)$
HQViT (M)	$O(\log(Td))$	$O(\log d)$	$O(1)$	$O(T^2 \log d)^*$	$\{A_{ij}\}_{i,j=1}^T, \{\mathbf{v}_i\}_{i=1}^T$	$O(T^2 d)$
QViT ¹ [28] (H)	$O(T + d)$	$O(d \log d)$	$O(T)$	$O(Td)^*$	$\{\mathbf{y}_i\}_{i=1}^T$	$O(T^2 d)$
OKSAN ² [30] (H)	$O(d)$	$O(d)$	$O(T^2)$	$O(T^2 d)^*$	$\{\mathbf{y}_i\}_{i=1}^T$	$O(T^2 d)$
QSAN [26] (F)	$O(T \log d + T^2)$	$O(T \log d)$	$O(1)$	$O(\log d)$	\mathbf{y}	$O(T^2 d)$
F-QSANN [27] (F)	$O(T \log d)$	$O(T^2 \log d)$	$O(1)$	$O(1)$	Probability of the predicted class	$O(T^2 d)$

1: This paper proposes several quantum transformer approaches, and we select the one that is most closely aligned with the standard classical self-attention mechanism, specifically the approach described in Section 3.3.

2: This method has been implemented using both amplitude encoding and qubit encoding, but here we select the case of qubit encoding.

"*" indicates that this method may require more sampling per measurement than other methods.

which is relatively low, while the number of PQCs and distinct circuits are the lowest). At the same time, it ensures a significant reduction in classical computational complexity. Although its measurement count is relatively high, overall, it achieves a well-balanced efficiency in quantum resource utilization compared to other methods. The key insights are given as follows:

qubits. HQViT uses amplitude encoding and requires $O(\log(Td))$ qubits, which is relatively low among all methods. Additionally, models with low qubit requirements include QSANN, QSAM, and QKSAN. These models use qubit encoding and process tokens one by one in the quantum circuit. As a result, the qubit complexity has a linear relationship with the token dimension, $O(d)$ (whether $O(d)$ or $O(\log(Td))$ is lower depends on the specific situation, but both are significantly lower than the qubit complexity of other models). On the other hand, other models, especially fully quantum models, require much higher qubits compared to the other models.

PQGs. In shallow parameterized quantum circuits, the number of parameterized quantum gates is generally proportional to the number of qubits for encoding a single token. However, due to the special circuit structures in QSAN and F-QSANN, their PQG requirements increase significantly. In contrast, HQViT only requires $O(\log_2 d)$ PQGs, offering a light-weight circuit structures.

Distinct circuits. This metric refers to the number of distinct circuits required to process all tokens in the sequence in parallel or to compute the attention matrix in parallel. For QSANN and QSAM, tokens are fed into the quantum circuit one by one, so mapping all tokens to QKV requires $O(T)$ distinct circuits. For QKSAN, since each circuit computes a single attention score, computing T^2 attention scores requires T^2 distinct circuits. In contrast, HQViT, QSAN, and F-QSANN use global encoding, where the attention coefficient matrix or self-attention output (or the final classification result) is computed within a holistic quantum framework, requiring only one distinct circuit.

Measurement complexity. To estimate measurement probabilities or expectation values, practical experiments typically require $O(1/\epsilon^2)$ samples (where ϵ is the target accuracy). However, in this paper, for a more intuitive comparison, we assume that applying a projection operator measurement to a single qubit counts as one measurement, ignore the detailed physical implementations. For sampling complexity, we only provide qualitative annotations (*) rather than precise calculations. Based on this standard, the measurement complexity of HQViT is $O(T^2 \log d)$, with T^2 post-selection measurements and each measuring $\log d$ qubits. QKSAN requires $O(T^2 d)$ measurements, as it runs the quantum kernel circuit T^2 times, measuring d qubits each time. QSANN and QSAM require $O(Td)$ measurements to obtain the mapped QKV vectors. In contrast, QSAN and F-QSANN only require $O(\log d)$ and $O(1)$ measurements, respectively, since the former's quantum circuit outputs the weighted features of all tokens, while the latter directly produces a predicted classification result. Although these fully quantum models have lower measurement complexity, their high resource demands in terms of qubits and circuit depth make them less scalable.

Classical complexity reduction. The computational complexity of classical transformers primarily lies in the calculation of attention coefficients, with a complexity of $O(T^2 d)$. By replacing the classical self-attention mechanism with a quantum self-attention counterpart, we transfer this computational burden to the quantum system, effectively eliminating this portion of the computational cost. In comparison to lightly quantumized models, which only reduce the classical computational overhead of the linear mapping step, $O(Td^2)$, HQViT eliminates a greater amount of classical computational overhead. Furthermore, the classical computational overhead reduced by HQViT can be comparable to that of heavy quantum-hybrid and fully quantum models, i.e., $O(T^2 d)$, while requiring fewer quantum resources.

TABLE 2
Key information about the datasets used in the experiments.

Dataset	Number of Classes	Image Size	Total samples	Description of Image Content
MNIST	10	28x28	70000	Grayscale images of handwritten digits from 0 to 9
MedMNIST2D	Diverse classes	28x28	From 100 to 100,00	A datasets collection of biomedical images, including sub-datasets such as ChestMNIST, DermaMNIST etc.
CIFAR-10	10	32x32	60000	Color images of real-life objects, including airplanes, cars, birds, etc.
CIFAR-100	100	32x32	60000	Color images of real-life objects, with more and finer categories than CIFAR-10
Mini-ImageNet	100	84x84	60000	Color images of real-life objects, commonly used for few-shot learning

5 EXPERIMENTS

5.1 Experimental settings

To comprehensively evaluate our model, we have conducted experiments on several datasets, including MNIST, MedMNIST2D, CIFAR-10, and Mini-ImageNet. These datasets vary in terms of classification categories, input image size, and feature complexity, providing a robust platform to assess the scalability of HQViT. The basic information for each dataset is provided in Table 2.

For MNIST and MedMNIST2D, we use 8 qubits to encode each image, which results in an image size of 16x16. For CIFAR-10 and Mini-ImageNet, we conduct experiments with both 8 qubits and 10 qubits to encode images with sizes 16x16 and 32x32, respectively. For the 16x16 input size, each image is segmented into 4x4 patches, yielding 16 tokens, each with a dimension of 16. For the 32x32 input size, each image is divided into 64 patches. To ensure fair comparisons, the input images for other models used in the comparison are resized to the same dimensions. We conduct these experiments by using Tencent’s tensorcircuit platform [35]. All the images used in experiments are grayscale images.

For the hyperparameters, the number of repeatable units in each PQC, denoted D , is set to 2 for most experiments, except for MedMNIST2D, where we set $D = 3$. The hidden layer of MLP is set to the same size as the dimension of each patch, and the number of QTBs L is set to 3.

5.2 Accuracy performance

Results on MNIST. On MNIST, we conduct four binary classification tasks (0 or 1, 1 or 8, odd or even, >4 or ≤ 4) and one ten-class classification task (0 to 9). To evaluate the performance of HQViT, we compare it with four other quantum image classifiers, i.e., TTN/MERA [36], two quantum CNN models (marked as QCNN-1 [16] and QCNN-2 [15]), and a quantum self-attention model (QSAM) [25].

From Table 3, we can see that HQViT achieves the highest accuracy on all classification tasks. For simpler tasks (0/1 classification and 1/8 classification), the prediction accuracy reaches 100%. And more impressively, as the task becomes more complex, our model’s leading margin becomes larger. For complex binary classification tasks such as “odd or

even” and “ >4 or ≤ 4 ”, HQViT achieves accuracy that is about 6 – 10% higher than the next best models. In the 0-9 multi-class classification task, the accuracy reaches 93.10%, surpassing the next best model, QSAM, by nearly 11%. This result demonstrates that our model’s advantage becomes even more pronounced when tackling more complex tasks. This could be attributed to HQViT’s encoding strategy, which inherently carries positional information. This allows the model to better capture contextual correlations compared to quantum self-attention models that do not incorporate positional encoding.

Results on MedMNIST2D. MedMNIST2D is a large-scale MNIST-like collection of standardized biomedical images, including 12 datasets for 2D images [37]. These datasets have image size of 28×28 , and cover diverse dataset sizes (from 100 to 100,000) and tasks (binary and multi-class). In the related work, only QOViT [28] has conducted experiments on them. Hence, we compare our results with theirs on MedMNIST2D. In Table 3, we can observe that HQViT outperforms QOViT in 3 out of 5 sub-datasets. Notably, when conducting experiments, QOViT uses classical preprocessing (e.g., fully connected layers) to extract features before encoding them into quantum states [28]. By contrast, HQViT uses only simple downsampling or upsampling without any learning-based preprocessing, highlighting the strength of HQViT’s quantum self-attention mechanism in feature extraction.

Results on CIFAR-10 and Mini-ImageNet. CIFAR-10 consists of 10 object categories, while Mini-ImageNet contains 10 main categories, each with 10 subcategories, representing finer distinctions within each category. For each dataset we conduct binary, 4-class, and 10-class prediction tasks. On these datasets, We compare HQViT with QSAM [25] and QSANN [24], and also introduce a naive classical Vision Transformer (CViT) baseline, which refers to replacing the quantum self-attention module with its classical equivalent while keeping all other components identical to those in HQViT. The classical equivalent specifically refers to a standard self-attention structure: three fully connected networks of dimension $d \times d$ that map the input data to the Q , K , and V . The attention coefficients are then computed by using Eq. (23), resulting in the weighted V . We describe it

TABLE 3

Test accuracy of HQViT compared with TTN/MERA, QCNN-1, QCNN-2 and QSAM on MNIST. The highest accuracy in each column is indicated in bold font. Note: '—' indicates that there is no relative experimental data.

Method	0 or 1	1 or 8	odd or even	>4 or ≤ 4	0 to 9
TTN/MERA [36]	0.997	—	0.848	0.791	—
QCNN-1 [16]	0.985	—	—	—	—
QCNN-2 [15]	—	0.963	—	—	0.743
QSAM [25]	0.999	0.995	0.899	0.844	0.822
HQViT	1.000	1.000	0.964	0.942	0.931

TABLE 4

Test accuracy of HQViT compared with QViT on MedMNIST2D. The highest accuracy in each column is indicated in bold font.

Method	BloodMNIST	DermaMNIST	OCTMNIST	PneumoniaMNIST	BreastMNIST
QViT ¹ [28]	0.860	0.719	0.606	0.885	0.744
HQViT	0.838	0.725	0.608	0.882	0.795

1: Here, we use the experimental results of the OrthoTransformer model proposed in that paper, which align with the model selected for complexity analysis.

TABLE 5

Test accuracy of HQViT on CIFAR-10 compared with CViT and QSAM. The highest accuracy in each column is indicated in bold font.

Method	Size = 16 × 16			Size = 32 × 32		
	2-class	4-class	10-class	2-class	4-class	10-class
CViT	0.862	0.587	0.303	0.883	0.610	0.335
QSANN [24]	0.825	0.541	0.295	0.852	0.585	0.317
QSAM [25]	0.833	0.534	0.312	0.874	0.607	0.311
HQViT	0.875	0.598	0.335	0.885	0.615	0.334

TABLE 6

Test accuracy of HQViT on Mini-ImageNet compared with CViT and QSAM. The highest accuracy in each column is indicated in bold font.

Method	Size = 16 × 16			Size = 32 × 32		
	2-class	4-class	10-class	2-class	4-class	10-class
CViT	0.875	0.582	0.321	0.880	0.603	0.343
QSANN [24]	0.845	0.509	0.284	0.850	0.485	0.317
QSAM [25]	0.875	0.534	0.290	0.867	0.530	0.341
HQViT	0.885	0.596	0.313	0.885	0.607	0.345

as ‘naive CViT’ because its structure is simplified compared to the conventional ViT architecture in [4].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (23)$$

Firstly, as seen in Tables 5 and 6, HQViT achieves superior performance over QSANN [24] and QSAM [25] in both two datasets across all tasks. The main difference between HQViT and QSANN [24]/ [25] lies in its direct calculation of the similarity between Q and K in the quantum-enhanced feature space (rather than in the classical feature space). This may allow HQViT to capture more expressive similarity features than its two peers, leading to its improved classification performance. Secondly, HQViT outperforms its classical counterpart, naive CViT, in most tasks, further demonstrating that the feature extraction capability of the quantum-enhanced space is more powerful than that of the

classical feature space of the same scale. This highlights the potential of quantum computing to enhance the performance of machine learning models.

5.3 Effect of Model Configurations

We have investigated the impact of key experimental settings, including the number of repeatable basic units in PQC and the number of QTBs. Both are contributors to the model depth, influencing model’s expressibility. As shown in Fig. 7-8, increasing the number of PQC repeatable units and stacking more QTBs both lead to better performance. This shows that HQViT scales well with depth. However, beyond three units or QTBs, the model’s training speed slows down, and its performance tends to saturate. Therefore, in our experiments, we set the number of QVBs to 3 and the number of repeated units in PQC to 2 or 3. In addition, we

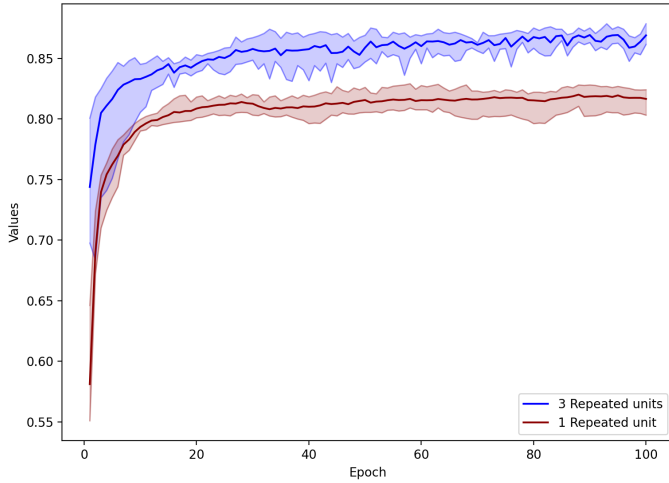


Fig. 7. The impact of the number of repeated units in the PQC on performance. The selected scenario in the figure is based on the CIFAR-10 dataset, with image size of 16x16, and 3 QTBs. The bold curves represent the averages of five experiments, with the shaded areas indicating the deviations.

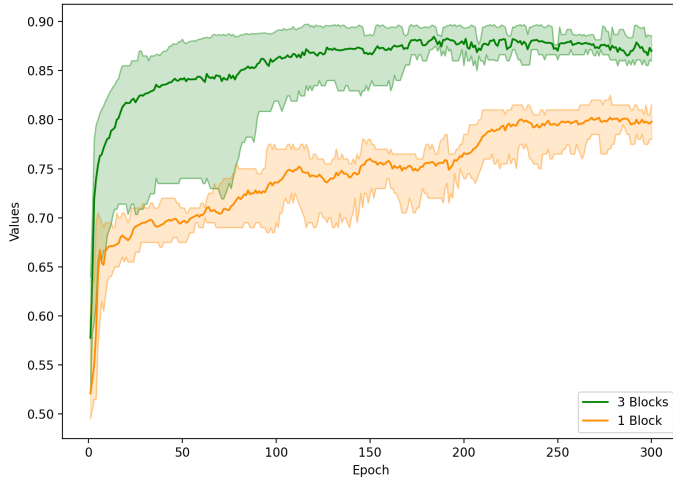


Fig. 8. The impact of the number of QTBs on performance. The selected scenario in the figure is based on the Mini-ImageNet dataset, with image size of 32x32, and 3 repeated units in the PQC. The bold curves represent the averages of five experiments, with the shaded areas indicating the deviations.

observe that the model successfully converges after several epochs, demonstrating outstanding training behavior.

6 CONCLUSION

In this paper, we have proposed an efficient hybrid quantum vision transformer (HQViT) for image classification. Our complexity analysis and experiments demonstrate that HQViT significantly reduces classical computational complexity by consuming only limited quantum resources, while delivering excellent classification accuracy. Its performance not only surpasses that of existing quantum transformer models but also outperforms other quantum image processing models, such as QCNN, particularly excelling in complex tasks involving large-size images. This highlights HQViT's ability to leverage the power of quantum comput-

ing for vision tasks, marking a new direction for quantum-enhanced models in the computer vision domain.

Three key factors contribute to our model's success: i) Selective quantum-classical hybridization. Rather than pursuing a full quantumization of a vision transformer, which would require an excessive number of ancilla qubits and deep quantum circuits, we selectively retain certain classical components. This ensures that the model remains practical for real-world tasks while maintaining high accuracy. ii) Whole-image processing. HQViT processes the entire image in one step, allowing the attention coefficient matrix to be derived through the internal evolution of a quantum circuit. This approach simplifies the model structure, better harnesses quantum advantages, and reduces classical computational complexity. iii) Efficient use of quantum resources. Amplitude encoding enables the model to process the whole image with a relatively small number of qubits, making it feasible for NISQ devices. This encoding strategy also inherently provides positional information at the input stage, improving the model's contextual understanding without additional positional encoding. However, incorporating explicit positional encoding methods, as seen in classical self-attention, remains a future direction to explore for further enhancing the model's capabilities.

Our model has been tested on various datasets, well demonstrating its scalability. HQViT's width can be easily extended by increasing the number of qubits, enabling the model to handle exponentially larger input sizes. On the other hand, the model's depth can be increased by increasing the repeated basic units and stacking more QTBs. However, this requires data conversion between quantum and classical states at the interface of two adjacent QTBs, which adds overhead to maintain the model's practicality and scalability in depth. In heavy and fully quantum models, typically only a single transformer block is used, as stacking multiple blocks is not feasible, leading to a poor scalability in depth. Therefore, developing a method to stack multiple QTBs without requiring data conversion between classical and quantum states emerges as a key direction for future research.

REFERENCES

- [1] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [2] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
- [3] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for nlp tasks," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 179–183.
- [4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [5] K. Han et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022.
- [6] X. Li et al., "Transformer-based visual segmentation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [7] N. Parmar et al., "Image transformer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4055–4064.
- [8] J. Biamonte et al., "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

- [9] D. Peral-García, J. Cruz-Benito, and F. J. García-Peñalvo, "Systematic literature review: Quantum machine learning and its applications," *Computer Science Review*, vol. 51, p. 100619, 2024.
- [10] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical Review Letters*, vol. 122, no. 4, p. 040504, 2019.
- [11] V. Havlíček *et al.*, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.
- [12] S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum algorithms for supervised and unsupervised machine learning," *arXiv preprint arXiv:1307.0411*, 2013.
- [13] I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks," *Nature Physics*, vol. 15, no. 12, pp. 1273–1278, 2019.
- [14] M. Henderson, S. Shakya, S. Pradhan, and T. Cook, "Quantum convolutional neural networks: powering image recognition with quantum circuits," *Quantum Machine Intelligence*, vol. 2, no. 1, p. 2, 2020.
- [15] S. Wei, Y. Chen, Z. Zhou, and G. Long, "A quantum convolutional neural network on nisy devices," *AAPPS Bulletin*, vol. 32, pp. 1–11, 2022.
- [16] T. Hur, L. Kim, and D. K. Park, "Quantum convolutional neural network for classical data classification," *Quantum Machine Intelligence*, vol. 4, no. 1, p. 3, 2022.
- [17] J. Amin *et al.*, "Detection of anomaly in surveillance videos using quantum convolutional neural networks," *Image and Vision Computing*, vol. 135, p. 104710, 2023.
- [18] J. Bausch *et al.*, "Recurrent quantum neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1368–1379, 2020.
- [19] S. Y.-C. Chen, S. Yoo, and Y.-L. L. Fang, "Quantum long short-term memory," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8622–8626.
- [20] S. Lloyd and C. Weedbrook, "Quantum generative adversarial learning," *Physical Review Letters*, vol. 121, no. 4, p. 040502, 2018.
- [21] H. Situ, Z. He, Y. Wang, L. Li, and S. Zheng, "Quantum generative adversarial network for generating discrete distribution," *Information Sciences*, vol. 538, pp. 193–208, 2020.
- [22] M. Cerezo *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021.
- [23] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Science and Technology*, vol. 4, no. 4, p. 043001, 2019.
- [24] G. Li, X. Zhao, and X. Wang, "Quantum self-attention neural networks for text classification," *Science China Information Sciences*, vol. 67, no. 4, p. 142501, 2024.
- [25] H. Zhang, Q. Zhao, and C. Chen, "A light-weight quantum self-attention model for classical data classification," *Applied Intelligence*, vol. 54, no. 4, pp. 3077–3091, 2024.
- [26] J. Shi, R.-X. Zhao, W. Wang, S. Zhang, and X. Li, "Qsan: A near-term achievable quantum self-attention network," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [27] J. Zheng, Q. Gao, and Z. Miao, "Design of a quantum self-attention neural network on quantum circuits," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2023, pp. 1058–1063.
- [28] I. Kerenidis, N. Mathur, J. Landman, M. Strahm, Y. Y. Li *et al.*, "Quantum vision transformers," *Quantum*, vol. 8, p. 1265, 2024.
- [29] I. Kerenidis, J. Landman, and N. Mathur, "Classical and quantum algorithms for orthogonal neural networks," *arXiv preprint arXiv:2106.07198*, 2021.
- [30] R.-X. Zhao, J. Shi, and X. Li, "Qksan: A quantum kernel self-attention network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [31] X. Wen and M. Zhou, "Evolution and role of optimizers in training deep learning models," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 10, pp. 2039–2042, 2024.
- [32] M. Weigold, J. Barzen, F. Leymann, and M. Salm, "Data encoding patterns for quantum computing," in *Proceedings of the 27th Conference on Pattern Languages of Programs*, 2020, pp. 1–11.
- [33] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, no. 3, p. 032308, 2020.
- [34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [35] S.-X. Zhang *et al.*, "Tensorcircuit: a quantum software framework for the nisy era," *Quantum*, vol. 7, p. 912, 2023.
- [36] E. Grant *et al.*, "Hierarchical quantum classifiers," *npj Quantum Information*, vol. 4, no. 1, p. 65, 2018.
- [37] J. Yang *et al.*, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

SUPPLEMENTARY MATERIALS A

We give the detailed derivation of Eq. (19) presented in Section 4.4. From Eq. (16), we can derive the probability of measuring outcome 00 on the ancilla qubit as:

$$\begin{aligned}
 Pr(0) &= \left| \frac{1}{2} (|\Psi_2\rangle + \Psi_{sw}) \right|^2 \\
 &= \frac{1}{4} (\langle \Psi_2 | \Psi_2 \rangle + \langle \Psi_2 | \Psi_{sw} \rangle + \langle \Psi_{sw} | \Psi_2 \rangle + \langle \Psi_{sw} | \Psi_{sw} \rangle) \\
 &= \frac{1}{2} (1 + \langle \Psi_2 | \Psi_{sw} \rangle) \tag{24}
 \end{aligned}$$

Substituting Eq. (17)-(18) into Eq. (24) yields:

$$\begin{aligned}
 Pr(0) &= \frac{1}{2} \left(1 + \sum_{i,l,j,k} \alpha_{il}^* \beta_{jk}^* \langle i | \langle l | \langle j | \langle k | \left(\sum_{i,l,j,k} \alpha_{ik} \beta_{jl} |i\rangle |l\rangle |j\rangle |k\rangle \right) \right) \\
 &= \frac{1}{2} \left(1 + \sum_{i,l,j,k} \alpha_{il}^* \beta_{jk}^* \alpha_{ik} \beta_{jl} \langle i | i \rangle \langle l | l \rangle \langle j | j \rangle \langle k | k \rangle \right) \\
 &= \frac{1}{2} \left(1 + \sum_{i,l,j,k} \alpha_{il}^* \beta_{jl} \beta_{jk}^* \alpha_{ik} \right) \\
 &= \frac{1}{2} \left(1 + \sum_i \sum_j \left(\sum_l \alpha_{il}^* \beta_{jl} \sum_k \beta_{jk}^* \alpha_{ik} \right) \right) \\
 &= \frac{1}{2} \left(1 + \sum_i \sum_j |\langle q_i | k_j \rangle|^2 \right) \tag{25}
 \end{aligned}$$

SUPPLEMENTARY MATERIALS B

We give the detailed derivation of Eq. (20) presented in Section 4.4. When measuring the index subsystems of R_Q , R_Q and the ancilla qubit, we use a set of measurement operators $\{M_{i,j}\}_{i,j=0}^{T-1}$, where each $M_{i,j}$ is given as:

$$M_{i,j} = |i\rangle\langle i| \otimes \mathbf{I} \otimes |j\rangle\langle j| \otimes \mathbf{I} \otimes |0\rangle\langle 0|, \tag{26}$$

where $\{i,j\}$ is a pair of specific index of q_i and k_j , and \mathbf{I} is an identity operator performed on the patch subsystems. According to Eq. (4), for each $M_{i,j}$ we have:

$$\begin{aligned}
 Pr(i, j, 0) &= \langle \Psi_3 | M_{i,j}^\dagger M_{i,j} | \Psi_3 \rangle \\
 &= \frac{1}{4} ((\langle \Psi_2 | + \langle \Psi_{sw} |) \otimes \langle 0 |) |i\rangle \langle i| \otimes \mathbf{I} \otimes |j\rangle \langle j| \otimes \mathbf{I} \\
 &\quad \otimes |0\rangle \langle 0| ((|\Psi_2\rangle + |\Psi_{sw}\rangle) \otimes |0\rangle)) \\
 &= \frac{1}{4} \left(\sum_{l,k} \alpha_{il}^* \beta_{jk}^* \langle l | \langle k | + \sum_{l,k} \alpha_{ik}^* \beta_{jl}^* \langle l | \langle k | \right) \\
 &\quad \left(\sum_{l,k} \alpha_{il} \beta_{jk} |l\rangle |k\rangle + \sum_{l,k} \alpha_{ik} \beta_{jl} |l\rangle |k\rangle \right) \\
 &= \frac{1}{4} \left(\sum_{l,k} \alpha_{il}^* \alpha_{il} \beta_{jk}^* \beta_{jk} + \sum_{l,k} \alpha_{il}^* \beta_{jl} \beta_{jk}^* \alpha_{ik} \right. \\
 &\quad \left. + \sum_{l,k} \alpha_{ik}^* \beta_{jk} \beta_{jl}^* \alpha_{il} + \sum_{l,k} \alpha_{ik}^* \alpha_{ik} \beta_{jl}^* \beta_{jl} \right) \\
 &= \frac{1}{4} (2 \langle q_i | q_i \rangle \langle k_j | k_j \rangle + 2 \langle q_i | k_j \rangle \langle q_i | k_j \rangle) \\
 &= \frac{1}{2} (|q_i|^2 |k_j|^2 + |\langle q_i | k_j \rangle|^2). \tag{27}
 \end{aligned}$$