IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 September 2025, accepted 8 October 2025, date of publication 15 October 2025, date of current version 22 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3621607

RESEARCH ARTICLE

External Quantum Self-Attention Model

FU CHEN[1], LI FENG[1], ZHENGDONG HU[2], AND YANGBIAO REN[3]
[1]School of Computer Science and Engineering, Macau University of Science and Technology, Taipa, Macau, China
[2]College of Computer and Data Science, Putian University, Chengxiang, Putian, Fujian 351100, China
[3]Reference News, Xicheng, Beijing 100031, China

Corresponding author: Li Feng (lfeng@must.edu.mo)

This work was supported in part by the Science and Technology Development Fund, Macau, SAR, under Grant 0008/2022/AGJ, Grant 0008/2025/RIB1, Grant 0126/2025/RIA2, and Grant 0077/2025/RIA2; in part by the National Natural Science Foundation of China under Grant 61872452; and in part by Guangzhou Huangpu International Science and Technology Cooperation Project under Grant 2023GH03.

ABSTRACT Self-attention mechanisms have revolutionized machine learning domains. Quantum self-attention models have emerged as promising extensions, yet they suffer from quadratic computational complexity like their classical counterparts. To address this, we propose the External Quantum Self-Attention Model (EQSAM), which integrates external memory modules into quantum self-attention, reducing complexity from quadratic to linear. In EQSAM, two sets of fully trainable external quantum modules generate key and value memory states. The model computes similarities only between input queries and these external states, rather than pairwise among all inputs. This lowers computational demands and boosts scalability in quantum machine learning. Experiments on MNIST and Fashion MNIST classification tasks demonstrate that EQSAM achieves comparable or superior performance to pairwise quantum self-attention models with less computation. The number of external modules serves as a key hyperparameter in EQSAM, with an optimal value balancing representational capacity and generalization. Performance approaches saturation near this optimum.

INDEX TERMS Machine learning, quantum computing, external quantum self-attention mechanism.

I. INTRODUCTION

Self-attention is a mechanism that enables models to capture long-range dependencies by computing attention weights between all positions in an input sequence [1], [2], [3], [4]. This allows the model to dynamically focus on different parts of the sequence, selectively weighting the relevance of each component when generating representations. By adaptively aggregating information from across the entire sequence, self-attention provides models with enhanced contextual understanding. The mechanism's inherent flexibility and scalability have made it a foundational technology in natural language processing [5], [6], computer vision [7], [8], and other domains requiring sequential data analysis [9], [10]. Currently, with the rapid advances in quantum computing technologies [11], [12], [13], researchers have begun to explore the integration of self-attention mechanisms into quantum machine learning [14], [15], [16], proposing quantum self-attention models [17]. These models aim to leverage quantum properties such as superposition and entanglement, as well as the mathematical structure of complex Hilbert spaces, to improve representational capacity and computational efficiency.

Recent advances in quantum self-attention mechanisms have introduced several innovative computational approaches that explore the application of quantum properties to machine learning architectures. These methods can be broadly categorized into three distinct paradigms. First, fusion-based methods, exemplified by reference [18] in their QSAN model, which employ CNOT gates and parameterized quantum circuits to introduce quantum entanglement, modeling complex correlations between query and key states; attention weights are captured through entanglement mechanisms. Second, overlap-based methods calculate similarity metrics between quantum states to determine self-attention weights, extracting either real-valued scores [19], [20] or complex-valued scores [21], thereby utilizing quantum state overlaps to naturally quantify self-attention weights in accordance with quantum mechanics. Third, implicit-relation methods, such as the GQHAn model proposed by [22], use trainable

The associate editor coordinating the review of this manuscript and approving it for publication was Chin-Feng Lai.

quantum circuits to directly predict target quantum attention weights without relying on explicit pairwise similarity calculations between queries and keys. These approaches, from varied perspectives, enrich the expressive power of quantum self-attention mechanisms while demonstrating unique advantages and potential of quantum computing in machine learning.

However, like classical self-attention, quantum variants suffer from quadratic computational complexity $O(N^2)$ due to pairwise similarity calculations across N input tokens, limiting scalability for large inputs. Classical external attention (EA) addresses this by using fixed-size, learnable memory units to reduce complexity to linear $O(N)$ [23], inspiring optimizations in quantum settings.

Building on this inspiration, this paper innovatively proposes an External Quantum Self-Attention Model (EQSAM), integrating the core idea of external attention into quantum self-attention models. Specifically, we design fixed numbers of fully trainable external quantum modules for both Keys and Values to generate their respective stable external memory units. These units replace the Key and Value representations in classical quantum self-attention, thereby shifting the attention calculation from a comprehensive similarity computation between input token pairs to a similarity measurement between input tokens and shared external memory units. This design effectively reduces the computational complexity of quantum self-attention from $O(N^2)$ to a linear scale of $O(N)$, significantly enhancing the model's computational efficiency and scalability when processing large-scale quantum data. The main contributions of this paper include:

First, we propose a novel External Quantum Self-Attention mechanism. In this mechanism, the Query quantum states are generated from N classical input data blocks through quantum embedding. Meanwhile, the Key and Value quantum states are produced by a fixed number S of fully trainable external quantum modules, serving as global memory units shared across all inputs. This design reduces the self-attention computational complexity from $O(N^2)$ to $O(N)$, improving scalability while maintaining expressive power, where S is typically much smaller than N.

Second, we integrate the SWAP test, Top-Down State Preparation (TDSP), and Linear Combination of Unitaries (LCUs) to compute quantum attention weights and perform weighted superpositions. Specifically, the Swap Test measures the overlap $|\langle K_i|Q_j\rangle|$ as attention weights, which are then encoded into quantum state amplitudes using the TDSP method. These encoded weights are applied via LCU to realize the quantum-native operation $\sum_i |\langle K_i|Q_j\rangle| |V_i\rangle$, enabling scalable and efficient attention mechanisms in quantum machine learning.

Third, we validate the effectiveness of EQSAM on MNIST and Fashion MNIST datasets for binary (2C) and ternary (3C) classification tasks. In 2C, with two external quantum modules (2S) for Key and Value, it achieves test accuracies of 99.84% on MNIST and 98.05% on Fashion MNIST, comparable to the pairwise quantum self-attention model's 99.84% and 98.59%, respectively, while reducing complexity from quadratic to linear. Scaling to four modules boosts performance to 99.92% and 98.36%. In 3C, 2S yields 96.46% and 93.75%, improving to 98.28% and 95.78% with four modules, outperforming the pairwise model's 97.81% and 95.26%.

The remainder of this paper is structured as follows: Section II provides foundational preliminaries on quantum states and their evolution essential for quantum machine learning. Section III details the methods, including the general framework, quantum embedding modules, attention weight computation, encoding, linear combination of unitaries, and quantum feed-forward network. Section IV introduces the loss function used for binary classification tasks. Section V presents numerical simulations, covering datasets, experimental setup, and performance analysis on MNIST and Fashion MNIST. Finally, Section VI concludes with key findings.

II. PRELIMINARIES

This section briefly reviews key concepts of quantum states and their evolution, which underpin quantum machine learning models.

The fundamental unit in quantum computing is the qubit, which, unlike a classical bit, can be in a superposition of $|0\rangle$ and $|1\rangle$ [24]:

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,, \tag{1}$$

where $\alpha, \beta \in \mathbb{C}$ satisfy $|\alpha|^2 + |\beta|^2 = 1$. These squared amplitudes represent the probabilities of measuring the qubit in states $|0\rangle$ or $|1\rangle$.

For n qubits, the state lives in a 2^n-dimensional Hilbert space:

$$|\psi\rangle = \sum_{i=0}^{2^n-1} c_i\,|i\rangle\,, \tag{2}$$

where the $|i\rangle$ are computational basis states and $\sum |c_i|^2 = 1$. This exponential space allows simultaneous representation of many classical states.

Quantum computation proceeds via unitary operators U, which are reversible linear transformations preserving normalization:

$$U^\dagger U = UU^\dagger = I. \tag{3}$$

An initial state $|\psi\rangle$ evolves as

$$|\psi'\rangle = U\,|\psi\rangle\,. \tag{4}$$

III. METHODS

A. GENERAL FRAMEWORK

Figure 1 illustrates the framework of the External Quantum Self-Attention Model (EQSAM). The workflow starts with classical input data, divided into multiple blocks and dimensionally reduced via Principal Component Analysis

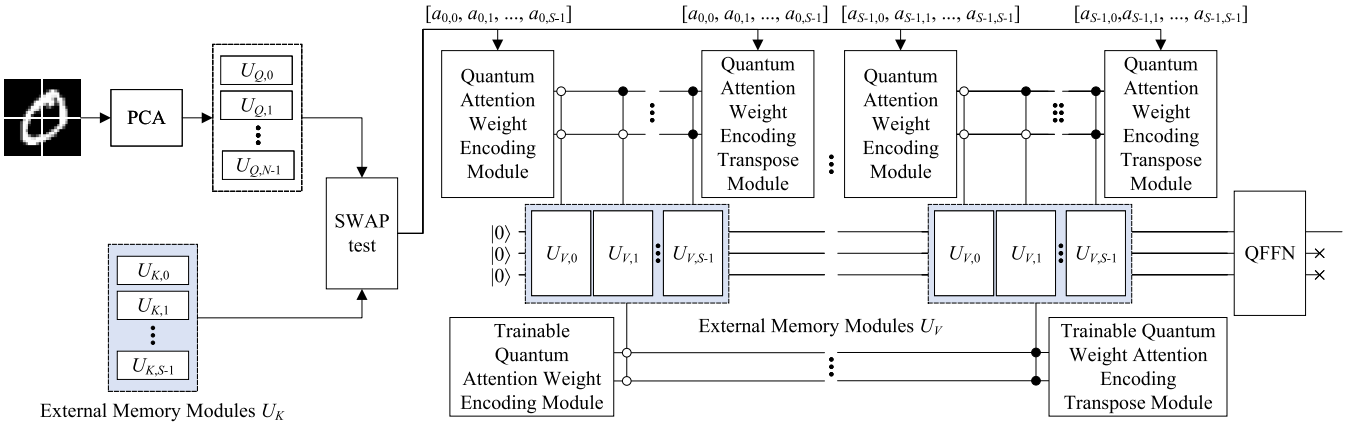


FIGURE 1. The framework of the external quantum self-attention model. $a_{j,i}$ denotes the real value obtained from the SWAP test between $|Q_j\rangle$ and $|K_i\rangle$. $U_{Q,j}$ represents $U_E(x_j, \theta_Q)$, $U_{K,i}$ represents $U_E(\theta_{K,i})$, $U_{V,i}$ represents $U_E(\theta_{V,i})$.

(PCA) [25] to match the quantum circuit's qubit count. These reduced data blocks are embedded into query quantum states using the U_Q module. Meanwhile, S fully trainable external quantum modules independently generate key quantum states. The model computes attention weights by pairing each of the N query states with the S key states through the Swap Test. These weights are encoded into amplitudes via a preparation module and integrated with S external value quantum states using Linear Combination of Unitaries (LCUs) to produce a weighted superposition $|L_j\rangle = \frac{1}{\mathcal{N}_{L_j}} \sum_{i=0}^{S-1} |\langle K_i|Q_j\rangle| |V_i\rangle$. A set of trainable scalar coefficients further combines these outputs via LCUs for global context integration, yielding $|G\rangle = \frac{1}{\mathcal{N}_G} \sum_{j=0}^{N-1} b_j |L_j\rangle$. The resulting quantum state is processed by a Parameterized Quantum Feed-Forward Network (QFFN) to boost expressiveness, followed by measurement to yield probabilities for downstream binary classification.

To enhance scalability in quantum self-attention for large inputs, EQSAM employs S fixed, fully trainable external quantum modules for key and value states, reducing complexity from $O(N^2)$ to $O(N)$. These modules create globally shared memory units, independent of specific inputs, that capture universal features and cross-sample relationships during training. This shifts attention from pairwise input similarities to query-memory overlaps, minimizing computations while enabling semantic integration across the dataset, akin to an external knowledge base.

B. QUANTUM EMBEDDING MODULE

In the proposed model, the quantum embedding module for queries plays a pivotal role in mapping classical input data into quantum states, while the key and value quantum states are generated by separate fully trainable external quantum modules independent of the classical inputs.We design two categories of quantum state embedding strategies, adapted based on the methods proposed in reference [26], to handle

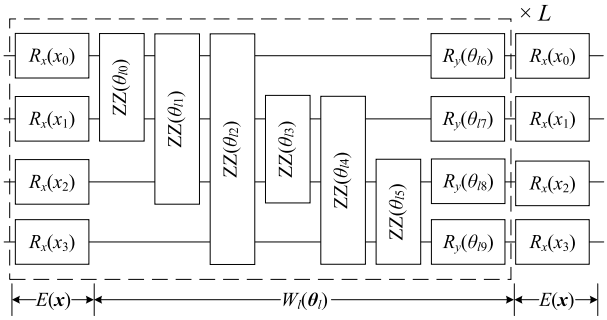


FIGURE 2. Quantum circuit architecture for the quantum embedding module.

the different functional roles within the external quantum self-attention mechanism.

Firstly, the generation of the query quantum state $|Q\rangle$ primarily relies on embeddings of the classical data. Specifically, the classical input features $\boldsymbol{x}$, after dimensionality reduction via PCA, are fed into quantum embedding module U_Q. This module consists of an initial layer of single-qubit rotation gates R_x, with rotation angles proportional to the input features, thereby achieving preliminary data encoding. Subsequently, trainable two-qubit ZZ gates together with parameterized single-qubit R_y rotation gates are incorporated to simultaneously enhance entanglement among qubits and improve the circuit's variational expressiveness. This structure, which is repeated over multiple layers where each layer sequentially executes a composition of a data encoding sublayer and a parameterized gate sublayer. Finally, to further reinforce data encoding, the module adds an additional layer of single-qubit R_x gates at the end of the last layer. The overall embedding unitary, as illustrated in Figure 2, can be expressed as:

$$U_{\mathrm{E}}(\boldsymbol{x}, \theta) = E(\boldsymbol{x}) \prod_{l=0}^{L-1} \big(W_l(\theta_l) E(\boldsymbol{x})\big), \tag{5}$$

where $E(\boldsymbol{x})$ denotes the classical data encoding layer:

$$E(\boldsymbol{x_j}) = \bigotimes_{i=0}^{n-1} R_x(x_{j,i}), \tag{6}$$

and $W_l(\theta_l)$ represents the parameterized quantum gate set of the l-th layer. This quantum embedding module effectively embeds the input features into the Hilbert space, producing the query quantum state:

$$|Q_j\rangle = U_{\mathrm{E}}(\boldsymbol{x_j}, \theta_Q)\,|0\rangle^{\otimes n}. \tag{7}$$

In contrast to the query quantum state, the key $|K\rangle$ and value $|V\rangle$ quantum states are generated independently without direct dependence on classical data inputs. Instead, they are instantiated via fully parameterized quantum circuits sharing the same circuit architecture as the query encoder.In this key and value quantum embedding module, which are external quantum modules, the classical data encoding layers are replaced by layers composing rotation angles entirely controllable by trainable parameters, as formulated by:

$$U_{\mathrm{E}}(\theta) = E(\theta_e)\prod_{l=0}^{L-1}\big(W_l(\theta_{w,l})E(\theta_e)\big), \tag{8}$$

where

$$E(\theta_{\boldsymbol{e}}) = \bigotimes_{i=0}^{n-1} R_x(\theta_{e,i}), \tag{9}$$

which $\theta_{e,i}$ and $\theta_{w,l}$ are all trainable variables fully independent from the query parameters θ_Q. The resulting key and value quantum states are then given by:

$$|K_i\rangle = U_{\mathrm{E}}(\theta_{K,i})\,|0\rangle^{\otimes n}, \quad |V_i\rangle = U_{\mathrm{E}}(\theta_{V,i})\,|0\rangle^{\otimes n}. \tag{10}$$

In our model, the quantum embedding module for the Query is designed to encode classical input data into quantum states, enabling dynamic and input-dependent representations. In contrast, the embedding modules for the Key and Value are fully parameterized and independent of the input data. They generate a fixed set of trainable quantum states that serve as stable external memory units. These external memory units store and accumulate global features and semantic information.

C. QUANTUM ATTENTION WEIGHT COMPUTATION MODULE

In the quantum self-attention mechanism, the Query quantum state $|Q\rangle$ is generated by encoding classical input data, while the Key quantum state $|K\rangle$ is produced independently through a trainable quantum circuit, serving as an external memory unit. To compute the External Quantum Attention Weight, we measure the similarity between these two states using the overlap $|\langle K|Q\rangle|$. This measure naturally reflects how much attention the model should assign to each external memory unit for a given input. We implement this similarity calculation via the Swap Test [27], [28], [29], a well-established quantum algorithm that efficiently estimates the overlap between two quantum states. This section details the Swap Test procedure.

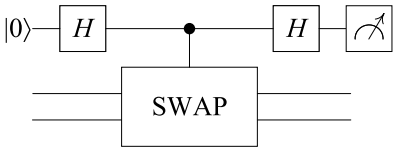


FIGURE 3. Quantum circuit architecture for the SWAP test.

The Swap Test utilizes a simple quantum circuit to estimate the similarity between two quantum states by measuring the state of an auxiliary qubit as shown in Figure 3. Its circuit involves three qubits: one auxiliary qubit (initially in $|0\rangle$) and two quantum states to be compared, $|\psi\rangle$ and $|\phi\rangle$. The step-by-step derivation of the Swap Test is as follows:

The initial state of the system is: $|0\rangle \otimes |\psi\rangle \otimes |\phi\rangle$. Apply a Hadamard gate to the auxiliary qubit, transforming $|0\rangle$ into a uniform superposition. Thus, the entire system's state becomes:

$$\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\,|\psi\rangle\,|\phi\rangle\,. \tag{11}$$

Next, apply a controlled-SWAP gate, with the auxiliary qubit as the control qubit. When the auxiliary qubit is $|0\rangle$, $|\psi\rangle$ and $|\phi\rangle$ remain unchanged; when it is $|1\rangle$, $|\psi\rangle$ and $|\phi\rangle$ swap positions. The state evolves to:

$$\frac{1}{\sqrt{2}}\left(|0\rangle\,|\psi\rangle\,|\phi\rangle + |1\rangle\,|\phi\rangle\,|\psi\rangle\right). \tag{12}$$

Apply the Hadamard gate to the auxiliary qubit again. The total state transforms into:

$$\begin{aligned}&\frac{1}{\sqrt{2}}\left[\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)\,|\psi\rangle\,|\phi\rangle + \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\,|\phi\rangle\,|\psi\rangle\right]\\ &= \frac{1}{2}\left[|0\rangle\,(|\psi\rangle\,|\phi\rangle + |\phi\rangle\,|\psi\rangle) + |1\rangle\,(|\psi\rangle\,|\phi\rangle - |\phi\rangle\,|\psi\rangle)\right].\end{aligned} \tag{13}$$

Measure the auxiliary qubit and calculate the probability of it being in the $|0\rangle$ state:

$$\begin{aligned}P(|0\rangle) &= \left\|\frac{1}{2}(|\psi\rangle\,|\phi\rangle + |\phi\rangle\,|\psi\rangle)\right\|^2.\\ &= \frac{1}{4}\left(1 + |\langle\psi|\phi\rangle|^2 + |\langle\phi|\psi\rangle|^2 + 1\right)\\ &= \frac{1}{2}\left(1 + |\langle\psi|\phi\rangle|^2\right).\end{aligned} \tag{14}$$

Therefore, the similarity is:

$$|\langle K_i|Q_j\rangle| = |\langle\psi_i|\phi_j\rangle| = \sqrt{2P(|0\rangle) - 1}. \tag{15}$$

By repeatedly running the circuit and statistically analyzing the measurement results, one can estimate $P(|0\rangle)$ and subsequently obtain the similarity between $|K_i\rangle$ and $|Q_j\rangle$.

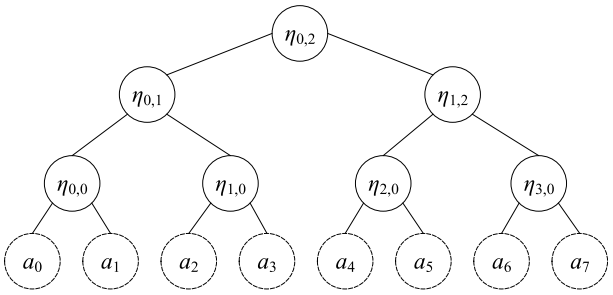


FIGURE 4. State tree representations of amplitude encoding.

D. QUANTUM ATTENTION WEIGHTS ENCODING MODULE

In the previous section, we utilized the swap test to compute the similarity $|\langle K|Q\rangle|$ between quantum states $|Q\rangle$ and $|K\rangle$. This similarity serves as the quantum attention weight, quantifying the correlation between the query and the key. These weights are classical data. To leverage them in subsequent quantum computations, particularly through the Linear Combination of Unitaries (LCUs) circuit to implement the weighted operations of the attention mechanism, we need to re-encode these classical attention weights into the amplitudes of a quantum state. This section details the encoding process, employing an efficient amplitude encoding method as described in the [30].

We first normalize the weights $|\langle K_i|Q_j\rangle|$ to obtain a set of probability values $\{\sqrt{a_{j,0}}, \sqrt{a_{j,1}}, \ldots, \sqrt{a_{j,S-1}}\}$, where $\sum_{i=0}^{S-1} a_{j,i} = 1$. Here, each $\sqrt{a_{j,i}}$ is defined by $\sqrt{a_{j,i}} = \sqrt{\frac{|\langle K_i|Q_j\rangle|}{\mathcal{N}_a}}$, with $\mathcal{N}_a$ being the normalization constant. Our objective is to construct a quantum state:

$$|\psi_j\rangle = \sum_{i=0}^{S-1} \sqrt{a_{j,i}}\,|i\rangle\,, \tag{16}$$

where $|i\rangle$ represents the computational basis states, and $\sqrt{a_{j,i}}$ are the amplitudes corresponding to these basis states. This process is known as amplitude encoding, which aims to efficiently generate the target state $|\psi\rangle$ from an initial state using a quantum circuit. The encoded quantum state $|\psi\rangle$ can then be directly used in the LCUs circuit, where the a_i serve as coefficients to control the linear combination of subsequent unitary operators, thus realizing the quantized attention weighting operation.

To encode the classical attention weights into amplitudes, we use the Top-Down State Preparation (TDSP) method [30]. TDSP applies a small set of (controlled) R_y rotations on an n-qubit address register so that basis state $|i\rangle$ receives amplitude $\sqrt{a_{j,i}}$. Given the SWAP-test overlaps, normalize

$$a_{j,i} = \frac{|\langle K_i|Q_j\rangle|}{\mathcal{N}_a}, \quad \mathcal{N}_a = \sum_{i=0}^{S-1} |\langle K_i|Q_j\rangle|, \quad \sum_i a_{j,i} = 1, \tag{17}$$

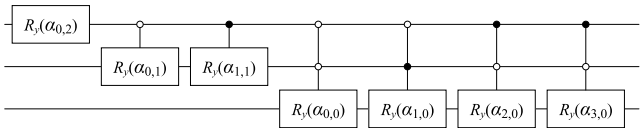


FIGURE 5. Quantum circuit architecture for top-down amplitude encoding of an eight-dimensional real vector.

and target

$$|\psi_1\rangle = \sum_{i=0}^{S-1} \sqrt{a_{j,i}}\,|i\rangle = \sum_{i=0}^{S-1} \sqrt{\frac{|\langle K_i|Q_j\rangle|}{\mathcal{N}_a}}\,|i\rangle\,. \tag{18}$$

Let $n = \lceil \log_2 S \rceil$ be the number of address qubits (if S is not a power of two, pad with $a_{j,i} = 0$ up to 2^n leaves). For readability we drop the index j below. Build the binary tree in Fig. 4 bottom-up:

$$\eta_{c,k} = \sqrt{\sum_{l=0}^{2^k-1} \left|a_{(c-1)\cdot 2^k + l}\right|^2}. \tag{19}$$

From these, compute mixing ratios and angles

$$\beta_{d,v} = \frac{\eta_{2d,v-1}}{\eta_{d,v}}, \qquad \alpha_{d,v} = 2\arcsin(\beta_{d,v}). \tag{20}$$

Apply the rotations top-down on the address register: At the root, apply $R_y(\alpha_{0,n})$ to the most significant address qubit. For each level $v = n-1, \ldots, 1$ and each node d, apply a controlled-$R_y(\alpha_{d,v})$ to the v-th qubit.

After these gates, the address register is prepared as

$$|\psi_1\rangle = \sum_{i=0}^{S-1} r_i\,|i\rangle = \sum_{i=0}^{S-1} \sqrt{a_i}\,|i\rangle\,, \tag{21}$$

which is exactly the coefficient state required by the LCUs block to realize the attention-weighted superposition over value states. An example quantum circuit illustrating the encoding of an eight-dimensional vector using the TDSP method is shown in Figure 5.

E. LINEAR COMBINATION OF UNITARIES

The Linear Combination of Unitaries (LCUs) method [31], [32] serves as the core component of our external quantum self-attention model. It enables the integration of the previously computed quantum self-attention weights with the corresponding value quantum states, allowing for an efficient weighted transformation of these states. By leveraging LCUs, the model constructs a global quantum state that encodes the self-attention weighted combination of value states, thereby facilitating effective feature representation in the quantum domain.

In quantum computing, the LCUs method is designed to implement a linear combination of a set of unitary operations. Given a collection of unitary operations $\{U_i\}_{i=0}^{S-1}$ and their associated real coefficients $\{a_i\}_{i=0}^{S-1}$, the objective of LCUs is

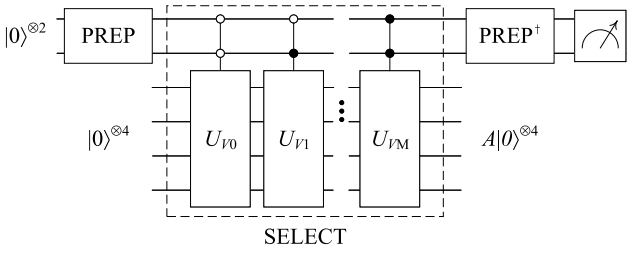


FIGURE 6. Quantum circuit architecture for the linear combination of unitaries.

to construct a quantum state transformation of the form:

$$A\,|\psi\rangle = \frac{1}{\mathcal{N}'}\sum_{i=0}^{S-1} a_i U_i\,|\psi\rangle\,, \tag{22}$$

where $|\psi\rangle$ denotes the target quantum state, a_i are the weighting coefficients for each unitary operation U_i, and $\mathcal{N}'$ is a normalization constant that ensures the probability amplitudes of the transformed quantum state satisfy the normalization condition.

To implement the weighted sum of unitaries, the LCUs method uses an ancilla register to control and apply the desired operations. The process begins with a preparation unitary U_{PREP}, which encodes the coefficients a_i into the amplitudes of the ancilla qubits, creating a superposition state from the initial $|0\rangle^{\otimes m} \otimes |\psi\rangle$, where m is the number of ancilla qubits needed to index the S unitaries.

Next, a selection unitary U_{SELECT} conditionally applies each U_i to the target register based on the ancilla state $|i\rangle$. Finally, the adjoint $U_{\text{PREP}}^{\dagger}$ is applied, followed by measurement and post-selection on the ancilla returning to $|0\rangle^{\otimes m}$, which projects the target register to the normalized linear combination $\frac{1}{\mathcal{N}'}\sum_{i=0}^{S-1} a_i U_i\,|\psi\rangle$, with orthogonal terms discarded. As shown in Figure 6, the circuit takes the input state and outputs the weighted sum after post-selection.

Within the external quantum self-attention model, the LCUs method is employed to manage attention weights and produce weighted quantum state representations. The essence of the attention mechanism lies in computing weights based on the similarity between keys ($|K_i\rangle$) and queries ($|Q_j\rangle$), and integrating these weights with value states ($|V_i\rangle$). Specifically, the coefficients $a_{j,i}$ in the LCUs framework are defined as the magnitude of the quantum inner product: $a_{j,i} = |\langle K_i|Q_j\rangle|$. This weight quantifies the similarity between $|K_i\rangle$ and $|Q_j\rangle$.

The unitary operations U_i in LCUs are implemented via parameterized quantum circuits $U_E(\theta_{V,i})$, where $\theta_{V,i}$ represents adjustable parameters that transform the initial state $|0\rangle^{\otimes n}$ into the value state $|V_i\rangle$. Combining the weights and operations, the resulting transformation is expressed as:

$$|\langle K_i|Q_j\rangle\,|U_E(\theta_{V,i})\,|0\rangle^{\otimes n} = |\langle K_i|Q_j\rangle|\,|V_i\rangle\,. \tag{23}$$

In this formulation, n denotes the number of qubits constituting the quantum state $|V_i\rangle$. $|\langle K_i|Q_j\rangle|$ governs the amplitude of the contribution from $|V_i\rangle$. This approach preserves the weighted structure of self-attention mechanisms.

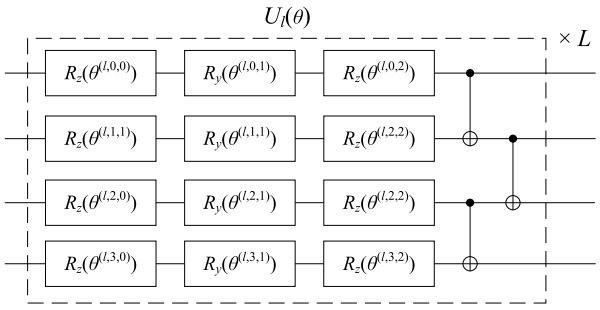


FIGURE 7. Circuit architecture for quantum feed-forward network.

Our model is to generate the following weighted state:

$$\begin{aligned}|L_j\rangle &= \frac{1}{\mathcal{N}_{L_j}}\sum_{i=0}^{S-1} |\langle K_i|Q_j\rangle\,|U_E(\theta_{V,i})\,|0\rangle^{\otimes n} \\ &= \frac{1}{\mathcal{N}_{L_j}}\sum_{i=0}^{S-1} |\langle K_i|Q_j\rangle|\,|V_i\rangle\,,\end{aligned} \tag{24}$$

where $U_E(\theta_{V,i})$ denotes the quantum circuit associated with the value state $|V_i\rangle$, and $|L_j\rangle$ represents the local weighted representation derived from the query $|Q_j\rangle$.

After generating the local weighted states $\{|L_j\rangle\}_{j=0}^{N-1}$, the model further consolidates these states through independent LCUs operations to form a global quantum state $|G\rangle$:

$$|G\rangle = \frac{1}{\mathcal{N}_G}\sum_{j=0}^{N-1} b_j\,|L_j\rangle\,, \tag{25}$$

where the input image is divided into N patches, and b_j are trainable real coefficients, with their amplitudes dynamically adjustable via parameterized controlled $R_y(\theta_j)$ gates. This design not only aggregates information from each local representation but also enhances the model's expressive capacity and adaptability to specific tasks through parameter optimization.

F. QUANTUM FEED-FORWARD NETWORK

In our quantum self-attention model, we introduce the Quantum Feed-Forward Network (QFFN) to enhance the expressive power of the quantum self-attention mechanism by incorporating additional quantum circuit layers. By introducing increased entanglement and circuit complexity in the QFFN, the model gains enhanced capability to represent and manipulate quantum information

To ensure practical applicability on near-term quantum devices, we adopt a hardware-efficient implementation strategy, as detailed in [33]. The mathematical expression for its quantum circuit layer is given by:

$$U_l(\theta) = \bigotimes_{i=0}^{n-1}\left(R_z(\theta^{(l,i,0)})\,R_y(\theta^{(l,i,1)})\,R_z(\theta^{(l,i,2)})\right)U_{\text{ent}}, \tag{26}$$

where U_{ent} denotes the entanglement layer, composed of CNOT gates, which introduces entanglement between qubits.

n is the number of qubits. The parameters θ are trainable, with l indicating the layer index and i representing the qubit index. To further enhance the model's expressive power, multiple QFFN layers can be stacked, with a total of L layers. Each layer l implements an independent quantum circuit $U_l(\theta)$. This multi-layer architecture allows the model to capture intricate data patterns through deeper quantum circuits, thereby enhancing the expressive power of the quantum self-attention mechanism.

IV. LOSS FUNCTION

In this paper, our classification tasks include both binary (2 classes) and ternary (3 classes) classification. After processing the input data through the quantum circuit, we extract probabilities for each class by measuring the expectation values of specific Pauli operators on the first qubit. These measurements map the quantum state's information onto a probability distribution over the classes.

To generalize the probability extraction, we define a set of observables M_j for $j \in \{0, 1, \ldots, C-1\}$, where C is the number of classes:

$$M_j = \begin{cases} (-1)^j \sigma_z & \text{if } C = 2,\ j \in \{0, 1\} \\ \sigma_{p(j)} & \text{if } C = 3,\ j \in \{0, 1, 2\} \end{cases} \tag{27}$$

where σ_p denotes the Pauli operator acting on the first qubit. Here, $p(j)$ is a function mapping the index $j \in \{0, 1, 2\}$ to a Pauli basis: $p(0) = x$, $p(1) = y$, $p(2) = z$. Thus, $\sigma_{p(0)} = \sigma_x$, $\sigma_{p(1)} = \sigma_y$, and $\sigma_{p(2)} = \sigma_z$.

The predicted probability for class k is then given by:

$$\hat{y}_k = \frac{1 + \langle\psi|M_k|\psi\rangle}{\sum_{j=0}^{C-1}(1 + \langle\psi|M_j|\psi\rangle)}, \quad k \in \{0, 1, \ldots, C-1\}, \tag{28}$$

where $|\psi\rangle$ is the output quantum state from the circuit. This formulation ensures the probabilities are normalized.

To optimize the parameters of the external quantum self-attention model, we employ the cross-entropy loss function, which is suitable for multi-class classification tasks. Over the training dataset, the loss is defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{c=0}^{C-1} y_{i,c} \log(\hat{y}_{i,c}), \tag{29}$$

where N is the total number of samples in the training dataset, $y_{i,c}$ is the true label (one-hot encoded) for sample i and class c, and $\hat{y}_{i,c}$ is the predicted probability that sample i belongs to class c.

V. NUMERICAL SIMULATIONS

In this section, we validate the effectiveness of the external quantum self-attention model in both binary and ternary classification tasks through numerical simulations. The experiments are conducted using two classic datasets, MNIST and Fashion MNIST, from which we select two or three classes for classification depending on the task. We compare the performance of the pairwise quantum self-attention model with that of the external quantum attention model equipped with 1 to 4 external quantum modules. The following subsections provide a detailed description of the experimental setup, results, and analysis.

A. DATASETS

The experiments use two standard datasets: MNIST [34] and Fashion MNIST [35], each with 60,000 training and 10,000 test images of 28×28 grayscale pixels. For each dataset, we randomly select 512 samples per class from the training set and 128 per class from the test set.

For binary classification (2C), we select classes "0" and "1" from MNIST, and "T-shirt/top" (label 0) and "Trouser" (label 1) from Fashion MNIST. For ternary classification (3C), we add class "2" to MNIST and "Pullover" (label 2) to Fashion MNIST. This setup evaluates the model's scalability and expressiveness in quantum machine learning across different classification complexities.

B. EXPERIMENTAL SETUP

In the data preprocessing stage, each original image is first divided into four equal blocks. To align the feature dimensions with the quantum model's qubit count, we apply a fixed, non-trainable Principal Component Analysis (PCA) [36] to reduce each block's feature dimensions. This simple linear transformation excludes learnable parameters, ensuring that preprocessing does not bias the final model performance. Notably, for the Q, K, and V quantum states in the self-attention mechanism, we consistently use 4 qubits each. Quantum circuits are simulated via the TensorCircuit framework [37], integrated with TensorFlow [38] for parameter optimization. We train using the Adam optimizer [39] with a batch size of 32, over 200 epochs. The hyperparameters of our external quantum self-attention model are shown in Table 1.

TABLE 1. Hyperparameters of our external quantum self-attention model.

Setting	Model	Number of layers in QEM[1] / QFFN[2]	
		MNIST	Fashion MNIST
2C[4]	1S[3]	1/2	1/2
	2S	3/4	4/1
	3S	4/3	2/4
	4S	1/2	4/1
3C	1S	2/1	1/2
	2S	3/4	3/2
	3S	3/3	4/2
	4S	3/3	3/1

[1] QEM stands for Quantum Embedding Module.
[2] QFFN stands for Quantum Feed-Forward Network.
[3] nS: Models equipped with n external quantum embedding modules for Key and Value states.
[4] nC: n denotes the number of classes in the classification task.

C. EXPERIMENTS AND ANALYSIS

This section presents a comprehensive evaluation and detailed analysis of the proposed external quantum self-attention model on binary and ternary classification tasks

using the MNIST and Fashion MNIST datasets. We start by experimenting with the model configured to use one to four fully trainable external quantum modules each for key and value quantum state generation. The query quantum states are obtained by encoding classical input data via a quantum embedding module. In contrast, the key and value quantum states are generated by multiple fully trainable quantum embedding modules, each producing distinct sets of states. This modular design enriches the external attention mechanism with more expressive and flexible representations. Finally, we compare these results against a baseline pairwise quantum self-attention model, where query, key, and value quantum states are independently generated from the same classical input data using separate embedding modules. This comparison helps us understand the impact of different quantum embedding strategies on classification performance and training stability.

1) ABLATION STUDIES

First, we analyze the performance of the external quantum self-attention model configured with a single external module, which consists of one pair of key and value quantum embedding modules, on both datasets under binary (2C) and ternary (3C) classification settings. As shown in Table 2, in the 2C setting on the MNIST dataset, this configuration achieves a maximum average test accuracy of 53.83%, while on Fashion MNIST, it reaches 55.70%. In the 3C setting, performance drops further, with test accuracies of 34.95% on MNIST and 35.42% on Fashion MNIST. These limited results are illustrated in Figures 8 and 9, where training and testing accuracies remain around 50% or below, indicating underfitting. This indicates the model struggles to capture meaningful patterns and is prone to underfitting. These results suggest that using only one external module results in an overly simplistic external attention mechanism that lacks the representational capacity necessary to effectively model the complexities of the MNIST and Fashion MNIST datasets.

Next, we conduct a comprehensive comparison of the external quantum self-attention model's performance by varying the number of fully trainable external quantum modules for generating key and value quantum states, ranging from two to four modules each. On both the MNIST and Fashion MNIST datasets, increasing the number of these external quantum modules significantly boosts classification accuracy and model expressiveness in both 2C and 3C settings. In the 2C setting on MNIST, the model with two external modules achieves a test accuracy of 99.84%, rising slightly to 99.92% with three or four modules. On Fashion MNIST (2C), it reaches 98.05% with two modules and 98.36% with three or four. For the 3C setting, starting from the low baseline with one module, two modules yield 96.46% on MNIST and 93.75% on Fashion MNIST, improving to 98.23%-98.28% on MNIST and 95.36%-95.78% on Fashion MNIST with three or four modules. These improvements demonstrate that multiple external quantum modules enable richer quantum state representations, enhancing global

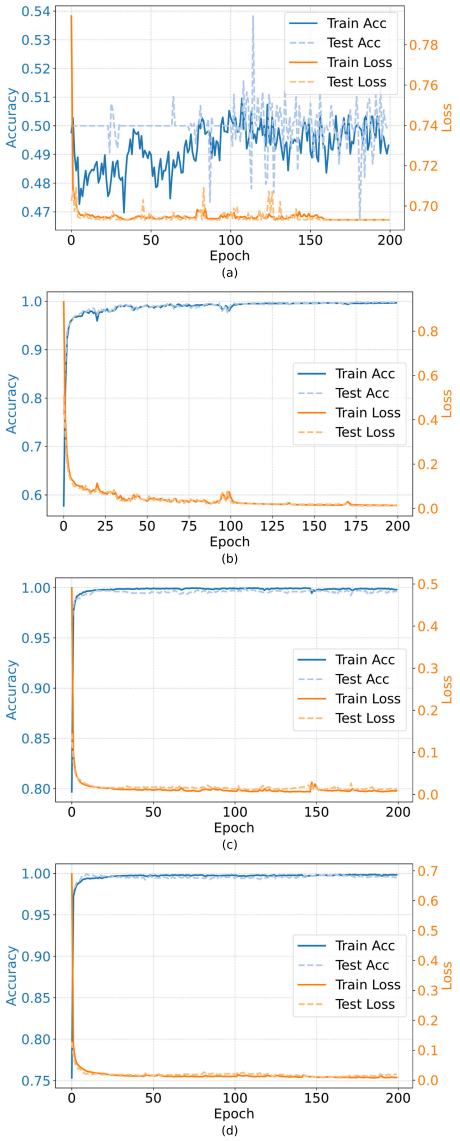


FIGURE 8. Training and testing performance of the external quantum self-attention model on MNIST with one to four External Modules. (a) One external module. (b) Two external modules. (c) Three external modules. (d) Four external modules.

feature extraction and cross-sample correlations in the external quantum self-attention mechanism. A modest number of external quantum modules effectively leverages the core benefits of external attention. Additional modules provide fine-tuned enhancements, which are particularly valuable in multi-class quantum machine learning scenarios where greater representational capacity is required.

Furthermore, as the number of external quantum modules S increases, the model's complexity rises accordingly, leading to enhanced representational power but with diminishing returns beyond a certain saturation point, where performance improvements plateau. This phenomenon aligns with the findings in [33], which demonstrate that the expected risk in quantum neural networks initially decreases with increasing

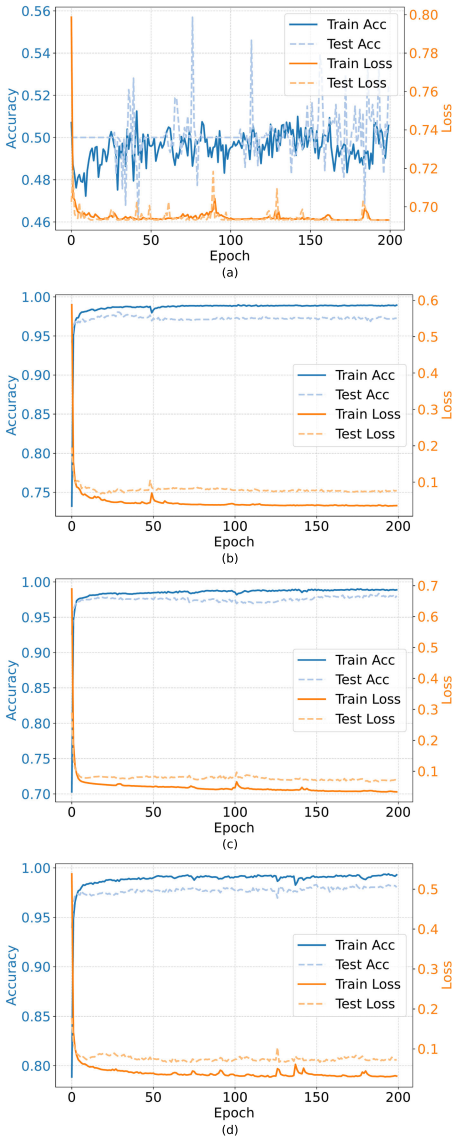


FIGURE 9. Training and testing performance of the external quantum self-attention model on fashion MNIST with one to four external modules. (a) One external module. (b) Two external modules. (c) Three external modules. (d) Four external modules.

model complexity due to improved fitting capabilities, eventually stabilizing as further complexity yields marginal gains without substantial additional benefits. Consequently, S serves as a key hyperparameter in EQSAM, whose optimal value is problem-dependent and should be empirically validated for datasets of varying complexity to balance expressiveness, generalization, and resource constraints in quantum machine learning applications.

To evaluate the contribution of the external quantum self-attention mechanism, we conducted ablation studies by comparing it against the pairwise quantum self-attention model. In these experiments, all other components, such as quantum embedding for queries, the SWAP test for overlap computation, TDSP for weight encoding, LCUs for weighted superpositions, and the quantum feed-forward network, remained identical across both models. The key difference lies in how key and value states are generated: the pairwise quantum self-attention model computes pairwise similarities between all input tokens (resulting in quadratic complexity $O(N^2)$), while our external model uses a fixed set of S fully trainable external quantum modules to produce shared key and value memory states, independent of individual inputs, reducing complexity to linear $O(N)$. Regarding the parameter count: The embedding circuits encode the $Q/K/V$ states and constitute the primary source of trainable parameters. For n qubits per $Q/K/V$ state, the total parameter count in these embeddings is $L(2S(n(n-1)/2+2n)+(n(n-1)/2+n))$ for S key/value pairs plus one query module. This count exceeds that of the pairwise baseline, which is $3L(n(n-1)/2+n)$. As a result, the external quantum self-attention mechanism leads to higher memory usage during simulation or demands more quantum gates in real quantum circuits compared to the pairwise model.

As shown in Table 2, in the 2C setting on the MNIST dataset, the ablated external model with two external modules achieves a test accuracy of 99.84%, matching the pairwise model's performance. Scaling to three or four modules yields a modest gain, surpassing the pairwise model by about 0.08%. On Fashion MNIST (2C), the two-module external variant reaches 98.05% test accuracy, approaching the pairwise model's 98.59%, with further modules closing the gap even more. In the 3C setting, the external model with two modules attains 96.46% on MNIST (versus the pairwise model's 97.81%) and 93.75% on Fashion MNIST (versus 95.26%), but with three or four modules, it improves to 98.23%–98.28% on MNIST (exceeding the pairwise) and 95.36%–95.78% on Fashion MNIST (matching or slightly surpassing). These results demonstrate that replacing pairwise token interactions with external memory-based similarities maintains or slightly improves accuracy across class settings, while leveraging global, dataset-shared representations to enhance generalization.

The ablation highlights the external mechanism's efficiency: by treating the modules as 'knowledge containers' that capture cross-sample semantics without input dependency, it fosters better feature integration across the dataset. With $S \ll N$, this approach drastically cuts computational demands, making it ideal for resource-constrained quantum settings. Even with just two modules, performance rivals the pairwise model in both binary and ternary tasks, suggesting scalability benefits for larger datasets in quantum machine learning.

2) COMPARATIVE EXPERIMENTS WITH BASELINE QUANTUM MODELS

To further assess the efficacy of our External Quantum Self-Attention Model (EQSAM), we compare its performance against several established baseline quantum models on the MNIST and Fashion MNIST binary classification tasks. Specifically, we evaluate our 3S configuration (using

TABLE 2. Ablation results: Performance of external and pairwise quantum self-attention models on MNIST and fashion-MNIST.

Setting	Model	MNIST		Fashion-MNIST	
		Test (%)	Train (%)	Test (%)	Train (%)
2C	1S	53.83 ± 7.46	49.55 ± 0.64	55.70 ± 11.41	48.65 ± 1.41
	2S	99.84 ± 0.19	99.65 ± 0.27	98.05 ± 0.42	98.98 ± 0.21
	3S	99.92 ± 0.16	99.68 ± 0.17	98.36 ± 0.63	98.89 ± 0.26
	4S	99.92 ± 0.16	99.38 ± 0.29	98.36 ± 0.38	99.18 ± 0.21
	Pairwise[1]	99.84 ± 0.19	99.67 ± 0.12	98.59 ± 0.63	98.48 ± 0.69
3C	1S	34.95 ± 3.23	32.53 ± 0.75	35.42 ± 4.17	32.43 ± 1.31
	2S	96.46 ± 0.63	96.17 ± 0.65	93.75 ± 0.87	94.09 ± 0.60
	3S	98.23 ± 0.60	97.43 ± 0.39	95.36 ± 0.76	94.24 ± 0.46
	4S	98.28 ± 0.35	98.16 ± 0.19	95.78 ± 0.91	95.47 ± 0.87
	Pairwise	97.81 ± 0.54	96.55 ± 0.90	95.26 ± 1.53	94.43 ± 0.38

[1] Pairwise: Ablation baseline model using pairwise query-key quantum self-attention weights for $O(N^2)$ complexity, with all other components identical to external variants.

TABLE 3. Comparative performance of EQSAM and baseline quantum ML models on MNIST and fashion-MNIST.

Model	Qubits	MNIST		Fashion-MNIST	
		Test(%)	Train(%)	Test(%)	Train(%)
3S (Ours)	6	99.92 ± 0.16	99.68 ± 0.17	98.36 ± 0.63	98.89 ± 0.26
QCNN [40]	8	98.90 ± 0.10	/	94.30 ± 1.60	/
SQNN [41]	16	98.90 ± 0.29	/	94.30 ± 1.60	/
TRVQC [42]	4	83.73	/	/	/
QSAN [18]	8	100	100	96.8	96.77
QKSAN [19]	4	99.00	99.06	98.05	98.52

three external quantum modules for keys and values) alongside SQNN [41], TRVQC [42], QCNN [40], QSAN [18], and QKSAN [19].

In an SQNN system, several quantum devices are used as quantum feature extractors, extracting local features from an input instance in parallel. Then a separate quantum predictor then collects extracted local quantum features from quantum feature extractors via classical communication channels, learns from them, and makes prediction on them with a VQC. The TRVQC method simulates VQCs classically by representing the quantum state as a tensor ring, applying parametrized gates with low-rank via truncated SVD to maintain efficiency, and optimizing parameters for tasks like classification. QCNN employs a tree-like architecture with convolutional filters and pooling layers using identical two-qubit parameterized gates in a translationally invariant manner, reducing qubits progressively via partial traces. The final prediction is obtained by measuring the remaining single qubit in the Z-basis. Both QSAN and QKSAN exemplify traditional quantum self-attention architectures that rely on pairwise similarity computations between query and key quantum states to derive attention weights, resulting in quadratic computational complexity with respect to the input sequence length.

As shown in Table 3, compared to SQNN, TRVQC, and QCNN, our 3S EQSAM achieves superior performance while utilizing fewer qubits. SQNN, which employs multiple quantum devices as parallel feature extractors followed by a variational quantum circuit (VQC) predictor, requires 16 qubits and yields test accuracies of 98.90% on MNIST and 94.30% on Fashion MNIST. TRVQC [42], a tree-structured variational quantum classifier using reservoir computing principles, operates on 4 qubits with 83.73% on MNIST. QCNN, inspired by classical convolutional networks, uses a hierarchical structure with convolutional and pooling layers to progressively reduce qubits, operating on 8 qubits with test accuracies of 98.90% on MNIST and 94.30% on Fashion MNIST. In contrast, our 3S model operates on only 6 qubits yet attains higher test accuracies of 99.92% on MNIST and 98.36% on Fashion MNIST. This improvement underscores the advantages of quantum self-attention mechanisms over convolutional approaches or direct VQC-based feature extraction, particularly when sufficient data is available. Attention allows the model to dynamically weigh global dependencies across input patches, capturing more nuanced patterns and long-range correlations that convolution or simple VQC extraction might overlook, leading to better generalization with reduced quantum resources.

When benchmarked against other quantum self-attention models like QSAN and QKSAN, our 3S EQSAM demonstrates comparable or competitive results. QSAN, which fuses quantum states via entanglement to compute attention weights, uses 8 qubits and achieves perfect accuracies of 100% on MNIST and 96.8% on Fashion MNIST. QKSAN, employing a quantum kernel for similarity measurement in attention weights, operates on 4 qubits with accuracies of 99.00% on MNIST and 98.05% on Fashion MNIST. Our model, with its external memory-based approach, matches or slightly exceeds these on Fashion MNIST (98.36% vs. 96.8% for QSAN and 98.05% for QKSAN) and approaches

TABLE 4. Robustness of external quantum self-attention models to simulated noise on MNIST and fashion-MNIST.

Setting	Model	MNIST		Fashion-MNIST	
		Test (%)	Train (%)	Test (%)	Train (%)
2C-noise	2S	99.84 ± 0.19	99.18 ± 0.44	98.05 ± 0.78	98.14 ± 0.24
	3S	99.77 ± 0.19	99.75 ± 0.20	98.20 ± 0.53	98.67 ± 0.27
	4S	99.92 ± 0.16	99.55 ± 0.27	98.28 ± 0.72	98.77 ± 0.29
3C-noise	2S	96.88 ± 0.59	96.09 ± 0.70	93.96 ± 0.73	94.11 ± 0.65
	3S	98.23 ± 0.56	98.16 ± 0.27	95.21 ± 0.69	95.70 ± 0.50
	4S	98.28 ± 0.27	97.93 ± 0.27	95.36 ± 0.51	95.46 ± 0.59

perfection on MNIST (99.92% vs. 100% for QSAN and 99.00% for QKSAN). This parity in performance validates the effectiveness of the external quantum self-attention mechanism, which shifts from pairwise input similarities to query-memory overlaps, reducing computational complexity to linear while maintaining strong representational power through trainable, dataset-shared memory units.

3) NOISY EXPERIMENTAL RESULTS

To assess the noise robustness of our External Quantum Self-Attention Model (EQSAM), we simulated noisy conditions by adding depolarizing noise channels to each qubit at the circuit's end. Depolarizing noise, a prevalent error in NISQ devices, randomly applies Pauli X, Y, or Z operators, mixing the state toward the maximally mixed state. We used a noise level of $p = 0.1$, which is typical for current quantum hardware.

As shown in Table 4, EQSAM demonstrates strong resilience in binary (2C) and ternary (3C) classification on MNIST and Fashion MNIST. In the 2C setting with $p = 0.1$, test accuracies remain stable, with the largest drop being 0.16% (e.g., 3S on Fashion MNIST, from 98.36% to 98.20%). For 3C, the maximum test drop is 0.42% (4S on Fashion MNIST, from 95.78% to 95.36%), and training accuracies show no degradation exceeding 0.84% overall, with variations within the model's statistical uncertainty.

The observed resilience to single-qubit depolarizing noise in EQSAM can be traced back to the integration of variational quantum algorithms (VQAs) within the framework. As highlighted in [43] and [44], VQAs alleviate noise interference through the dynamic adjustment of their parameters during the optimization process. Due to this adaptive characteristic, the quantum circuit can recalibrate its variables even under noisy conditions during training, which helps to suppress error propagation effectively.

VI. CONCLUSION

In this work, we introduced the External Quantum Self-Attention Model (EQSAM), which leverages fully trainable external quantum memory units to replace classical pairwise attention computations, effectively reducing the computational complexity from quadratic to linear scale. Our design not only enhances computational efficiency but also enables capturing global semantic features shared across samples via stable external key and value states. Numerical experiments on MNIST and Fashion MNIST binary classification tasks validate that EQSAM achieves competitive or improved accuracy compared to pairwise quantum self-attention models. These results highlight the potential of external attention paradigms in advancing scalable and expressive quantum machine learning architectures.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[2] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 11106–11115.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[4] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 4171–4186.

[6] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, Oct. 2021.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[8] M.-H. Guo, T.-X. Xu, J. Liu, Z.-N. Liu, P.-T. Jiang, T. Mu, S.-H. Zhang, R. R. Martin, M. Cheng, and S. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. media*, vol. 8, no. 3, pp. 331–368, 2022.

[9] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023.

[10] J. Chien and Y. Chen, "Continuous-time attention for sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 7116–7124.

[11] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018.

[12] D. P. DiVincenzo, "Quantum computation," *Sci.*, vol. 270, no. 5234, pp. 255–261, 1995.

[13] E. Knill, "Quantum computing with realistically noisy devices," *Nature*, vol. 434, no. 7029, pp. 39–44, Mar. 2005.

[14] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

[15] M. Schuld, I. Sinayskiy, and F. Petruccione, "An introduction to quantum machine learning," *Contemp. Phys.*, vol. 56, no. 2, pp. 172–185, 2014.

[16] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, "Challenges and opportunities in quantum machine learning," *Nature Comput. Sci.*, vol. 2, no. 9, pp. 567–576, 2022.

[17] R.-X. Zhao, Y. Lu, J. Shi, S. Wang, Y. Wang, and X. Li, "A review of quantum attention mechanisms: Quantum models, applications, and challenges," *Authorea Preprints*, 2025.

[18] J. Shi, R.-X. Zhao, W. Wang, S. Zhang, and X. Li, "QSAN: A near-term achievable quantum self-attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 8, pp. 13995–14008, Aug. 2025.

[19] G. Li, X. Zhao, and X. Wang, "Quantum self-attention neural networks for text classification," *Sci. China Inf. Sci.*, vol. 67, no. 4, Apr. 2024, Art. no. 142501.

[20] F. Chen, Q. Zhao, L. Feng, C. Chen, Y. Lin, and J. Lin, "Quantum mixed-state self-attention network," *Neural Netw.*, vol. 185, May 2025, Art. no. 107123.

[21] F. Chen, Q. Zhao, L. Feng, L. Tang, Y. Lin, and H. Huang, "Quantum complex-valued self-attention model," 2025, *arXiv:2503.19002*.

[22] R.-X. Zhao, J. Shi, and X. Li, "GQHAN: A grover-inspired quantum hard attention network," 2024, *arXiv:2401.14089*.

[23] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5436–5447, May 2023.

[24] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[25] M. Greenacre, P. J. F. Groenen, T. Hastie, A. Iodice d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Rev. Methods Primers*, vol. 2, nos. 1, p. 100, 2022.

[26] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, "Quantum embeddings for machine learning," 2020, *arXiv:2001.03622*.

[27] H. Buhrman, R. Cleve, J. Watrous, and R. De Wolf, "Quantum fingerprinting," *Phys. Rev. Lett.*, vol. 87, no. 16, Sep. 2001, Art. no. 167902.

[28] J. C. Garcia-Escartin and P. Chamorro-Posada, "Swap test and hong-ou-mandel effect are equivalent," *Phys. Rev. A, Gen. Phys.*, vol. 87, no. 5, May 2013, Art. no. 052330.

[29] H. Kobayashi, K. Matsumoto, and T. Yamakami, "Quantum Merlin–Arthur proof systems: Are multiple Merlins more helpful to Arthur?" in *Proc. 14th Int. Symp. Algorithms Computation*. Cham, Switzerland: Springer, Dec. 2003, pp. 189–198.

[30] I. F. Araujo, D. K. Park, T. B. Ludermir, W. R. Oliveira, F. Petruccione, and A. J. da Silva, "Configurable sublinear circuits for quantum state preparation," *Quantum Inf. Process.*, vol. 22, no. 2, p. 123, Feb. 2023.

[31] A. M. Childs and N. Wiebe, "Hamiltonian simulation using linear combinations of unitary operations," 2012, *arXiv:1202.5822*.

[32] S. Chakraborty, "Implementing any linear combination of unitaries on intermediate-term quantum computers," *Quantum*, vol. 8, p. 1496, Oct. 2024.

[33] Y. Du, Y. Yang, D. Tao, and M.-H. Hsieh, "Problem-dependent power of quantum neural networks on multiclass classification," *Phys. Rev. Lett.*, vol. 131, no. 14, Oct. 2023, Art. no. 140601.

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[35] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[36] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[37] S.-X. Zhang, J. Allcock, Z.-Q. Wan, S. Liu, J. Sun, H. Yu, X.-H. Yang, J. Qiu, Z. Ye, Y.-Q. Chen, C.-K. Lee, Y.-C. Zheng, S.-K. Jian, H. Yao, C.-Y. Hsieh, and S. Zhang, "TensorCircuit: A quantum software framework for the NISQ era," *Quantum*, vol. 7, p. 912, Feb. 2023.

[38] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[40] T. Hur, L. Kim, and D. K. Park, "Quantum convolutional neural network for classical data classification," *Quantum Mach. Intell.*, vol. 4, no. 1, p. 3, Jun. 2022.

[41] J. Wu, Z. Tao, and Q. Li, "WpScalable quantum neural networks for classification," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Sep. 2022, pp. 38–48.

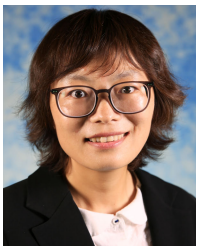
[42] D. Peddireddy, V. Bansal, and V. Aggarwal, "Classical simulation of variational quantum classifiers using tensor rings," *Appl. Soft Comput.*, vol. 141, Jul. 2023, Art. no. 110308.

[43] E. Fontana, N. Fitzpatrick, D. M. Ramo, R. Duncan, and I. Rungger, "Evaluating the noise resilience of variational quantum algorithms," *Phys. Rev. A, Gen. Phys.*, vol. 104, no. 2, 2021, Art. no. 022403.

[44] K. Sharma, S. Khatri, M. Cerezo, and P. J. Coles, "Noise resilience of variational quantum compiling," *New J. Phys.*, vol. 22, no. 4, Apr. 2020, Art. no. 043006.

FU CHEN received the B.Eng. degree in electrical engineering and automation and the M.Eng. degree in pattern recognition and intelligent systems from Fuzhou University, Fuzhou, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China.

His research interests include quantum machine learning and design and optimization of quantum circuits.



LI FENG received the M.S. degree in operation research from the Department of Mathematics, The University of Hong Kong, Hong Kong, in 2007, and the Ph.D. degree in electronic information technology from the School of Computer Science and Engineering (SCSE), Macau University of Science and Technology (MUST), Macau, China, in 2013.

She is currently a Professor with SCSE, MUST. Her research interests include wireless and mobile networks, power saving, software defined networking, and performance analysis.

ZHENGDONG HU received the B.Eng. degree in communication engineering and the M.Eng. degree in electronics and communication engineering from Southwest University of Science and Technology, Sichuan, China, in 2014 and 2018, respectively, and the Ph.D. degree in computer science from Mahasarakham University, Mahasarakham, Thailand, in 2025.

His research interests include quantum machine learning and natural language processing.

YANGBIAO REN received the Master of Engineering degree from the Communication University of China, Beijing, China, in 2018.

He is currently an Engineer at Reference News. His research directions include future networks, software-defined networks, and network applications in media convergence. His research interests include quantum machine learning and natural language processing.